University of Nottingham
UK | CHINA | MALAYSIA

Exploring Non-Verbal Methods in Voice Interaction Systems for Autonomous Driving Applications

By Chenwen LIN

Supervisors Prof. Xu Sun

Assoc. Prof. Qingfeng Wang

A master thesis submitted for the degree of Master of Mechanical Engineering of Nottingham University

Department of Mechanical, Materials, and Manufacturing Engineering

May 2025

Abstract

As autonomous driving technology continues to evolve, in-vehicle voice interaction has become more natural and personalized. However, these systems often face limitations when managing tasks that require rapid responses or precise control. Traditional voice input may not be suitable for all scenarios due to challenges with processing speed and response latency. To address these constraints this study explored the combination of non-verbal sounds and voice input in autonomous driving, with a focus on system activation and continuous non-driving-related tasks. In Experiment 1, participants used non-verbal sounds to wake-up the system and compared this method with traditional wake-up words and wake-up free approach. Results showed that many users still preferred traditional wake-up methods, although snapping fingers did not show a significant disadvantage in terms of interaction duration. In Experiment 2, non-verbal sound input was further developed for continuous task control and was tested alongside multiple voice commands and the Stop input for continuous non-driving-related tasks. While the combination of non-verbal input methods was innovative, the Stop command was highly favored by participants, likely due to its higher accuracy and lower subjective workload, which may have been influenced by task design. Overall, this study introduces a novel approach to non-verbal sound input, offering new insights into voice input design and future interactions in autonomous vehicles.

Content

List of Abbreviations

The following abbreviations are used in this study:

- **AV**: Autonomous Vehicle
- **HMI**: Human-Machine Interaction
- **NDRTs**: Non-Driving Related Tasks
- **WUWs**: Wake-up Words
- **ASR**: Automatic Speech Recognition
- **TLX**: Task Load Index

List of Figures

List of Tables

1. Introduction

1.1 Background

As autonomous driving gains popularity globally, leading automakers and technology giants such as Tesla, Waymo, and General Motors are actively investing in the development of autonomous driving technologies, aiming to make vehicles a comfortable, safe, and enjoyable companion for passengers. However, the current research focus extends beyond automation in driving, concentrating on enhancing the overall driving experience and user satisfaction [1]. The motivation behind this shift lies in the vision of transforming vehicles from traditional transportation tools into mobile living spaces. In such an intelligent environment, users can seamlessly engage in activities such as work, entertainment, or social interactions while in transit[2][3][4]. By prioritizing user-centered design, automakers aim to redefine vehicles as multifunctional spaces that integrate comfort, connectivity, and convenience.

With the development of smart cockpits, innovative human-vehicle interaction (HVI) modes have rapidly emerging, transforming the way drivers interact with their vehicles. As in-car electronic devices continue to advance, human-machine interaction (HMI) technologies are becoming more diverse, providing consumers with novel and intuitive user experiences [5]. For instance, XPeng Automobiles has adopted a more diversified approach in HMI technology. In addition to the traditional touch screen and voice control, XPeng has introduced gesture recognition technology, enabling drivers to control specific vehicle functions through simple, clear gestures, such as changing music or adjusting air conditioning temperature[6]. This introduction of gesture recognition not only enhances the interactivity of the interactions but also promotes driving safety. Similarly, Mercedes-Benz has introduced the MBUX virtual assistant, which realizes the visual interaction of voice commands through generative AI technology and active intelligence technology[6]. In addition, BMW's i3 model is equipped with an Augmented Reality Horizontal View Display (AR-HUD), which can display crucial information directly in the driver's field of view in real-time, further optimizing the driving experience[7]. Together, these advancements illustrate how cutting-edge HMI technologies are reshaping HVI, focusing on enhancing usability, interactivity, and safety.

Voice, as a natural method of HMI similar to human communicate, has been widely used in various machine operations and plays an important role in HMI[8]. It reduces the need for manual input, enhancing both driving safety and convenience. Current in-vehicle voice systems primarily rely on predefined wake-up words (WUWs) to activate voice assistants, which are crucial for enhancing the user experience. An ideal wake-up mechanism should activate promptly when needed while avoiding unintentional triggers. This requires precise threshold settings and dynamic adjustment strategies[11]. However, traditional wake-up methods present limitations. Fixed WUWs may fail to accommodate users with strong accents or dialects, while cognitive overload during complex tasks can hinder users from promptly recalling and stating these words[12][13]. In response, wake-

up free technology has emerged, allowing speech recognition to continuously monitor input without specific activation words. While this approach improves efficiency and user experience, it raises privacy concerns as continuous listening could inadvertently capture private conversations or sensitive information[13][14].

Compared to both traditional WUWs and wake-up free systems, non-verbal sounds offer a promising alternative. They bypass pronunciation and vocabulary challenges, require minimal learning effort, and avoid many privacy risks associated with continuous listening[12]. By addressing some of the core limitations of existing voice interaction methods, non-verbal sounds hold significant potential for advancing in-vehicle HMI.

Executing user's voice commands is another critical function of speech recognition systems, but current methods are restricted to recognizing predefined word-based commands [9]. This limitation reduces the system's ability to handle unexpected inputs or complex, continuous tasks [10]. For example, adjusting a car window's height may require multiple sequential commands like "a little up" or "a little down". These tasks require continuous input in the form of sequential motions, leading to interaction discontinuities that undermine the overall user experience.

This study explores the potential of non-verbal sounds in autonomous driving by focusing on two primary applications: replacing traditional wake-up words (WUWs) and enhancing natural language for executing continuous non-driving-related tasks (NDRTs). By reviewing current advancements and challenges in Human-Vehicle Interaction (HVI), this research identifies key limitations in existing voice interaction systems, particularly in the context of autonomous vehicles.

1.2 Research Objectives and Questions

While speech input has become the main method for interacting with In-Vehicle Information Systems (IVIS), most studies on non-verbal sounds have focused on their role in estimating factors such as emotion[83], gender, and language[84], demonstrating that accurate recognition can be achieved without relying on text. Non-verbal sounds have also been used as output signals to provide feedback to users[85][86]. However, in the context of autonomous vehicles, there is limited research exploring the use of non-verbal sounds as input signals. To address this gap, this study aims to investigate how non-verbal sounds can improve interaction efficiency, enhance user experience, and address privacy concerns.

The research consists of two experiments: the first evaluates the effectiveness of non-verbal sounds in activating the voice interaction system, while the second assesses their feasibility as input signals for continuous tasks.

Experiment 1: Comparing the effectiveness of non-verbal sound input with two other methods in waking up the voice interaction system and evaluating their impact on user experience.

RQ1: How do WUWs, non-verbal voice, and wake-up-free modes affect interaction efficiency in non-driving-related tasks (NDRTs)?

RQ2: Can the use of snapping fingers to reduce false negatives and thus improve

interaction efficiency?

RQ3: What are users' preferences and subjective experiences with different wake-up methods for voice interaction systems?

RQ4: Are users willing to sacrifice some privacy for more convenient interaction?

Experiment 2: Exploring the feasibility of using non-verbal sounds as input signals in continuous tasks.

RQ1: How do different forms of hybrid voice input (Multiple Voice, speech +stop or snapping fingers and continuous voice) affect task performance in autonomous vehicles, with performance being evaluated based on task completion time and accuracy?

RQ2: Do different forms of speech input influence on the user's cognitive load?

RQ3: What are user experiences regarding different sound input methods?

RQ4: What are users' preferences regarding different sound input methods?

2. Literature review

2.1 HMI in the Automatic Driving Scenarios

Autonomous driving technology is changing the role of the driver in the car and redefining how they interact with the vehicle[17]. The Society of Automotive Engineers (SAE) classifies driving automation into six levels, from manual driving (Level 0) to full autonomy (Level 5), each level indicating the degree of automation and the corresponding driver responsibilities [18]. At Levels 0-2, drivers are fully responsible for vehicle control, while at Level 3 (conditionally autonomous vehicles), the system assumes control of both lateral and longitudinal motion, as well as object detection and response. However, the driver must remain alert and prepared to take over in case of system failure or emergencies that the system cannot handle effectively [19]. At Level 4 and 5 (Fully Autonomous Vehicles), the vehicle takes over nearly all driving tasks, transforming the driver's role from active operator to passive passenger, with their attention shifting to non-driving-related tasks (NDRTs) such as entertainment, communication, or productivity. Despite this shift, drivers may still be required to interact with the vehicle, adjusting the infotainment system or intervening in driving tasks for non-emergency purposes, such as directing the vehicle to pick up a passenger [20][21][22]. The changes in autonomous driving levels not only affect the driver's behavior patterns, but also put forward higher requirements for the design of human-machine interfaces (HMIs). Therefore, HMI design must prioritize both enhancing the user experience through precise command execution and supporting the growing need for NDRTs. As the ability to perform these tasks increases, ensuring that HMI systems are designed to handle them efficiently has become a critical factor in autonomous vehicle (AV) design [23].

To meet these evolving demands, HMI systems must not only support complex task execution but also provide intuitive interfaces for diverse use cases [24]. The rise of automation has significantly expanded opportunities for NDRTs, such as reading, watching videos, or messaging[25][26][27]. The extent and nature of driver involvement in these activities significantly impact the driver's ability to take over of the vehicle[28], which has become one of the critical area of research[24]. . Advanced HMI designs enable seamless transitions between NDRTs and driving tasks, reducing cognitive load during system takeovers[100].

In general, the design of human-machine interfaces (HMI) in autonomous driving contexts poses a significant challenge, requiring designers to not only understand the potential shifts in user behavior within this new environment but also to anticipate and accommodate the varied needs of future travelers. Achieving a truly seamless and user-centric interactive experience is the ultimate goal. This focus has driven significant innovation in HMI technology, positioning it as a key driver in the continued evolution and advancement of the autonomous vehicle industry.

2.2 Prevalent Interaction Methods

The common ways of HMI in vehicles include touch screen interaction, voice interaction and gesture interaction, etc. (Fig. 1).

Touch screen interaction is one of the most intuitive and natural forms of interaction[31]. Drivers interact with the system by tapping, swiping, or long-pressing icons on the touchscreen. However, one key drawback of this tactile interaction is that it requires drivers to divert their gaze from the road, which can affect driving safety[35]. With the advancement of autonomous driving technology, the reassignment of driving functions has prompted a transformation in drivers' core responsibilities, shifting from traditional driving tasks to handling NDRTs [17]. This change has led to innovations in driving control interfaces, such as steering wheels and cockpit designs, thereby affecting transformations in HMI modes. The limitations of traditional tactile interaction methods are becoming increasingly apparent, especially as drivers transition to handling more NDRTs. In contrast, touchless interaction methods are gaining widespread favor due to their convenience and efficiency. Pierstefano Bellani et al. have found that touchless interfaces are more user-friendly and appealing[36]. In the control of autonomous vehicles based on maneuver, touchless interactions bring about a more positive emotional perception of the interaction [37].

Gesture interaction, as a form of touchless interaction, manifests in various forms, including mid-air gestures, steering wheel gestures, and finger-pointing. Steering wheel gestures allow drivers to control the system with thumb movements above the steering wheel[40][41]. This approach reduces visual demands and enhances driving safety compared to traditional systems[40][42]. However, some studies also suggest that this method may have some negative impacts on driving performance and perceived workload [41]. Similarly, pointing gestures convey directional information to the system through the precise pointing of the finger. Robert Tscharn et al. proposed the combination of voice and indicating gestures, which can convey spatial instructions more naturally and intuitively [43]. Another form of gesture interaction is mid-air gestures, where drivers interact with the system by performing gestures toward the center console or windshield without touching any surface, often supplemented by ultrasound feedback [44][46]. Although mid-air gestures enhance the user experience, but both indicative gestures and pointing gestures face the same challenge: they often provide only limited information, and may not be enough to meet all needs in complex situations.

Moreover, the use of gestures in HMI systems may lead to issues such as accuracy concerns and unconscious manipulation [50]. These limitations highlight the need for further research to determine the optimal input modes for in-vehicle infotainment systems (IVIS) and to address the challenges of gesture interaction in real-world driving scenarios.

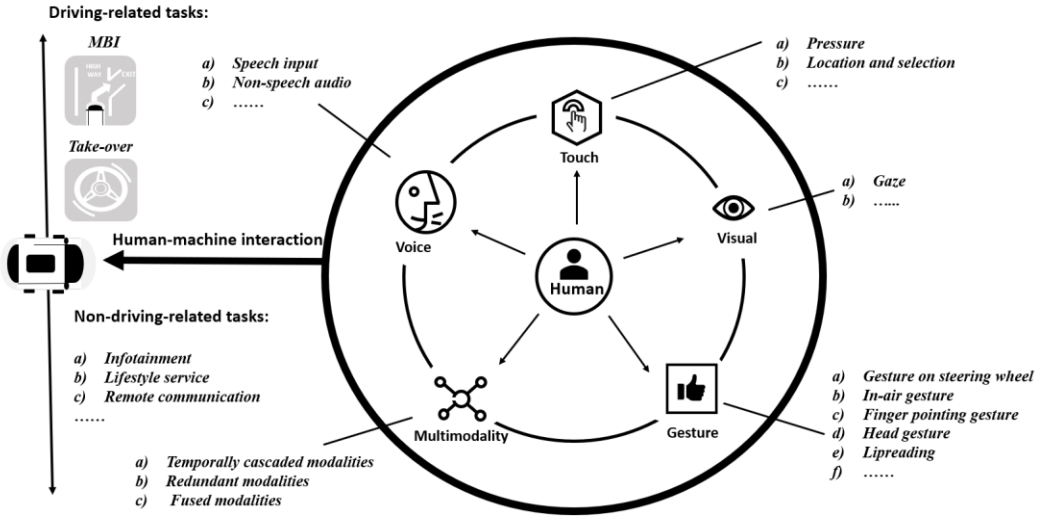

Fig. 1. In-vehicle HVI.

2.3 Voice and Non-verbal Sound Interaction

Voice input is one of the most common modes of interaction in both traditional and autonomous vehicles. As a touchless interaction method, Voice interaction can address the limitations of touch screen interaction, particularly in situations that require remote and hands-free operation[32]. Compared to gesture-based input, it can reduce the driver's visual distraction on the road and provide easier access to and control of IVIs. This contributes to both [33][34][35].

Typical speech-based interfaces rely on speech recognition systems, which convert spoken language into text commands that trigger appropriate actions[50]. However, speech is a complex sound that includes verbal and non-verbal cues like pitch and rhythm, which convey contextual and emotional information[51]. In addition to conveying speech, the sound channel also includes non-verbal auditory signals such as laughter, coughing, and other sounds that can serve a communicative function[53]. Previous studies typically classified non-verbal sounds based on how they are pronounced or their intended purpose [54], Yilmazyildiz et al proposed a definition-based classification. This approach provides a more accurate understanding of non-verbal sounds, referred to as non-semantic speech, which is divided into four categories: babble (“meaningless phonetic strings”), paralinguistic speech (“independent sound events”), musical speech (sounds based on musical theory), and non-verbal speech (other non-verbal sounds)[55]. This refined classification helps to deepen the understanding of the role and impact of non-verbal sound in interaction design, and reveals its great potential for improving user experience and interaction effectiveness. Studies have shown that non-verbal sounds can be leveraged to control interactive applications by using subtle variations in sound, such as pitch and volume [50]. For example, Seo J H et al. used the sound of clapping as a trigger signal to start gesture tracking, significantly improving the clarity and operability of the

interaction process [56].

In-vehicle voice recognition systems, for discrete tasks, such as activating car air conditioning or music, voice input is quite practical. However, it becomes inefficient when dealing with finer-grain and real-time control of continuous actions, like scrolling through pages or zooming in on maps. This is due to the limitations of voice input, which is not ideally suited for continuous and incremental operations [58]. In contrast, non-verbal sound interaction showcases its strengths in real-time and continuous operation control, independence from specific languages, and convenience in use. Moreover, non-verbal sound input can be processed reliably, even in noisy environments, providing stable and consistent system responses [50]. However, this interaction method does have limitations in terms of its information-carrying capacity, which makes it less suitable for conveying complex or multifaceted commands [8].To effectively leverage the benefits of both speech and non-verbal sound interaction, integrating these modes presents a promising solution. Research by Kaur has shown that combining voice and non-verbal sounds enables smooth and continuous control of a computer's mouse pointer. For example, users can produce different vowel sounds to indicate the desired direction of movement [54]. Similarly, window operation can be managed using a command like "roll the window uuuuuup," where the duration of the sound "up" determines how far the window opens. The window stops when the user stops making the sound. This method enhances the interaction's naturalness and fluidity, and significantly expands the use of non-verbal sounds in the complex environment of HVI.

2.4 Applications and Technologies in Voice Input

Voice interaction technology allows users to interact with computer systems through natural language [54]. Central to this technology is speech recognition, which converts the user's speech input into text or executable commands. The development of speech recognition technology includes the analysis, recognition and understanding of speech patterns, as well as the improvement of robustness in different environments [55] [56]. Voice interaction technology is widely used, including but not limited to computer-telephone integration, voice portals, virtual personal assistants, such as Apple’s Siri. [57]. These systems offer users the convenience of hands-free control, making them an integral part of modern smart environments.

In contrast to speech recognition, non-verbal sound recognition, such as detecting the sound of clicking fingers, relies on different detection methods and is applied in distinct scenarios. Non-verbal sound recognition primarily focuses on identifying specific sound events, such as a snap or a clap, rather than converting sound into text information. The technology behind non-verbal sound recognition detects specific features in the audio signal, such as frequency patterns or temporal characteristics, to recognize and trigger corresponding actions [59]. This distinction highlights the versatility of voice interaction technologies, with speech recognition handling more complex language inputs, while non-verbal sound recognition is better suited for discrete, event-based tasks.

2.4.1 Wake-up Technology

In voice interaction, wake-up word technology acts as a critical interface between the user and the system, evolving from simple voice recognition to involving

complex signal processing and pattern recognition methods. Traditionally, wake-up functionality primarily relies on keyword detection algorithms, which activate the system by matching a predefined voice template. However, this method has limitations, including high model complexity and the need for extensive training data that covers a wide vocabulary range [65][66]. With the advancement of deep learning technologies, particularly speech recognition models based on Deep Neural Networks (DNNs), such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), Christin Jose et al. have demonstrated the use of CNNs to accurately detect the start and end positions of WUWs, offering more efficient and precise localization of trigger words [66][67]. These advancements in wake-up word detection have significantly enhanced the overall performance of voice interaction systems. Despite these technological advances, the naturalness and flexibility of WUWs are still limited. In contrast to real-life conversations, where a gesture or smile can signal the intent to start a dialogue, WUWs typically require explicit spoken words or phrases [68]. Common WUWs like "Ok Google" or "Hey Alexa" are often rigid and limited by pronunciation, vocabulary knowledge, and cultural differences. [69]. Additionally, noisy environments can affect the performance of WUWs, including accidental activations (also known as false positives) and failures to correctly capture the trigger word. These errors, particularly false positives, significantly impact the user experience and reduce user expectations and frequency of use of the technology [70]. To address these challenges, many researchers are currently exploring alternatives to WUWs, such as gaze [71] or gesture signals [72][73].

2.4.2 Wake-up Free Technology

As intelligent interaction technologies evolve, users increasingly expect to initiate conversations naturally without relying on explicit commands. This demand has led to the development of wake-up free technology, which enables voice recognition systems to automatically detect user commands without the need for specific WUWs. This technology employs advanced Speech Activity Detection (SAD) and Speaker Verification technologies, which continuously monitor environmental sounds and initiate interactions when specific vocal characteristics of a user are recognized [74][75]. Additionally, Scholars have proposed integrating multimodal inputs, combining voice signals with face detection, to better understand the user's intent to interact, thereby creating a more natural and seamless voice interaction experience [76][77].

However, despite its potential, wake-up free technology faces several challenges. In automotive environments, for example, the interaction subject might not be within the current field of view. This issue arises when passengers in the back seat are out of the system's detection range, or when facial information is obstructed by car seats. Additionally, the "always listening" environment may pose a potential threat to personal privacy [11]. Moreover, during routine interactions, users might not provide complete information in a single attempt, as a result, they may prefer using WUWs to interrupt or readjust the conversation, providing them with a greater sense of control over the system [79]. Therefore, while removing WUWs can improve

system usability, it requires careful consideration to maintain the user experience without compromising privacy or usability. It is essential to explore new, more flexible ways to activate voice interaction systems.

WUWs are designed to mimic the rich "call-and-answer" interaction patterns found in natural conversation, but they essentially serve as tools for activating the system, much like pressing a button [80]. According to binary input theory, research has shown that a single statement of any other Non-verbal Auditory Input (NVAI) mode can be used to trigger binary input [52]. Compared to traditional voice commands, non-verbal sounds tend to be easier for users to master since they do not rely on complex language skills or knowledge of pronunciation and cultural nuances. Furthermore, when the user knows that the interaction is with the machine, the user is actually “repelled” by excessive politeness and repetition [81]. Given these factors, exploring the potential of non-verbal sounds as WUWs represents a promising direction for future research. Non-verbal sounds such as clapping or snapping fingers are intuitive, less prone to linguistic barriers, and provide a more accessible method for initiating interactions. As such, they offer an opportunity to replace traditional verbal WUWs, resulting in a more flexible and user-friendly voice interaction experience.

NO.	Author, Publication year, Country	The Way to Activate Voice Interaction System	Main Findings
1	Albert S, & Hamann M. (2021), Spain	Wake word	Although prosody cues offer potential for enhancing voice interfaces, we should explore more flexible ways to initiate interactions with virtual agents[80].
2	Jung H, & Kim H. (2019), Ireland	Wake word	First, wake words give users control over the VUI. Second, they seem to emotionally project onto sound agents [79].
3	Combs M, Hazelwood C, & Joyce R. (2022), USA	Wake word	(1) the number of false positives is related to wake word; (2) number of false positives is related to Amazon Echo hardware; (3) false positives decrease over time[82].
4	Bleakleya A, Wua Y, Pandeyb A, et al. (2021), Ireland	Wake word	The limited choice of wake phrases may exclude users who speak different languages or interact with IPAs in a non-native language[69].
5	Pomykalski P, Woźniak M P, Woźniak P W,	Gesture	We conducted gesture elicitation to identify five candidate gestures. Initial results indicate that the snap gesture shows the most potential[72].

	et al. (2020), Poland		
6	Zhao S, Westing B, Scully S, et al. (2019), USA	Voice & gesture	A novel approach to activating Intelligent Virtual Assistants (IVAs) on smartwatches: raise your hand and speak naturally for accurate and energy-efficient detection [73].
7	McMillan D, Brown B, Kawaguchi I, et al. (2019), Sweden	Gaze	Gaze can be used to augment, or even replace, the wake-work in initiating interaction with speech agents[71].
8	Zhang H, Wang J, Yang S, et al. (2022), China	Wake-free (voice & video)	Make full use of both voice and video mode information to solve challenging multi-mode activation task[76].
9	Dong X. (2019), China	Wake-free (voice & video)	Building on the original TVM ticket purchase process, voice recognition adds wake-up-free input, Chinese phonetic alphabet input, and fuzzy location inquiry functions[77].
10	Vertegaal R, Slagter R, Van der Veer G, et al. (2001)	/	The user's eye gaze can form a reliable source of input for conversational systems that need to establish whom the user is speaking or listening to [78].

Table 1. Summary of Previous Research on the Way to Activate Voice Interaction System

3. Exploring Non-verbal Wake up method for autonomous driving applications

3.1 Methodology

3.1.1 Interactive Tasks and NDRTs

To investigate the potential of non-verbal sounds as WUWs. The study designed two interactive tasks for participants in controlled study environment. These tasks were selected to represent common Non-Driving-Related Tasks (NDRTs), where the driver interacts with the system using voice commands. The study chose two tasks with varying levels of complexity: a music-playing task and a social media browsing task. The complexity of each task was measured based on the number of steps required and the type of actions involved [87]. These tasks were selected because they reflect common in-car activities where voice interaction could improve convenience and reduce distractions for drivers. The specific operation process is as follows:

Task 1: Navigation task

(1) Open the navigation software
(2) Input navigation address
(3) Navigate to the detail address
(4) Cancel the command, change the address
(5) Re-enter your address
(6) Navigate to the detail address
(7) Confirmed
(8) Close the navigation software

Task2: Social media browsing tasks

(1) Open social media software
(2) Open the first news
(3) Browse
(4) Close the social media software

3.1.2 The Way and Use Strategy of Activating Voice Interaction System

There are several strategies for activating the voice interaction system. Once the system is awakened, if no further voice input is received within 5 seconds, the system will return to sleep mode. To issue a command, users must reactivate the system. Three kinds of awakening voice interaction system way:

Traditional WUWs:

Traditional wake-up words are designed to be short enough to ensure easy pronunciation while being long enough to avoid accidental activations. The study chose "Alexa" as the wake-up word (WUWs) due to its balance between simplicity and effectiveness.

Nonverbal sounds (Snapping Fingers):

Snapping Fingers serve as binary input and discrete input, which is often

preferred in NVAI modalities [52], In this experiment, snapping fingers are used to trigger the voice interaction system, providing a non-verbal alternative to traditional WUWs.

Wake-up free:

In this mode, users can directly issue voice commands without needing to activate the system with a WUW or gesture. This provides a more seamless and natural interaction, where the system continuously listens for commands without the need for explicit activation.

3.1.3 Setups

The experiment was conducted in a quiet indoor environment. A display simulating the exterior environment of an autonomous vehicle was placed in front of the participants, with a tablet was used to provide visual feedback. The participants' voices were captured using a Logitech microphone, which transmitted the audio to a laptop on the right. The system uploaded the audio to the cloud for voice recognition (Fig. 2).

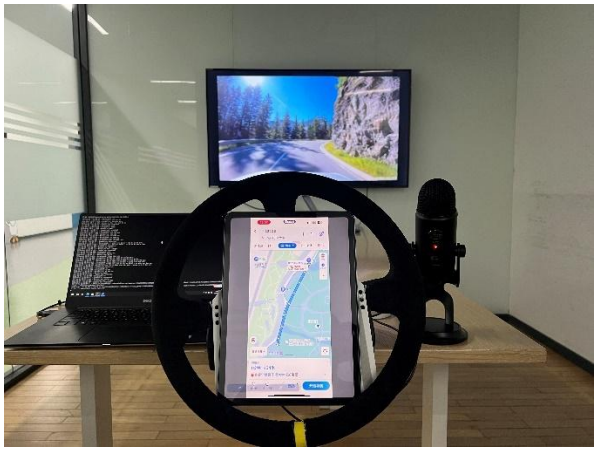

Fig. 2. Experiment Setups (Experiment 1).

3.1.4 Participants

The study invited a total of 20 participants (Male = 9, Female = 11), with ages ranging from 18 to 60 years (M = 31.75, SD = 10.27). All participants had normal or corrected vision and hearing and came from diverse backgrounds. Additionally, the study surveyed participants on their experience with voice interaction systems, with the results as follows:

Frequency of Use	Number of Participants
Never Used	1
Occasionally Used	13
Frequently Used	6

Table 2. Frequencies of Voice Interaction System Use Among Participants

Before conducting the experiments, ethical approval was obtained from the University of Nottingham Ningbo China (UNNC) Ethics Committee. All participants were provided with detailed information about the study and gave their informed consent before participating.

3.1.5 Experiment Design

To assess different methods of activating the voice interaction system, the study used the Wizard of Oz method [88].This method ensures that the system only triggers actions after a participant has completed a full command (either a voice command or a combination of a wake-up word and voice command). The system has accurately recognized the speech and converted it into text. This approach minimizes biases caused by system performance and standardizes system behavior. The study employed a within-subjects design with three independent variables: three activation methods (wake-up word vs. snapping fingers vs. wake-up free) and two interaction tasks (navigation and social media browsing), both classified as NDRTs.

3.1.6 Evaluation Index

The NASA-TLX score was used to estimate subjective workload, asking participants to evaluate six demand dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration level [89]. The User Experience Questionnaire (UEQ-S) was used to assess both the hedonic quality (user experience) and pragmatic quality (usability) of the three activation methods, as well as participants' preferences [90]. This information helps us gain a more comprehensive understanding of user needs and expectations.

The study also recorded the total interaction time and the number of false negatives in wake-up word detection. A false negative occurs when a wake-up word is spoken but not recognized by the system, preventing task activation. This quantitative data not only reflects the efficiency and accuracy of task completion but also provides clear direction for subsequent performance optimization.

3.1.7 Procedure

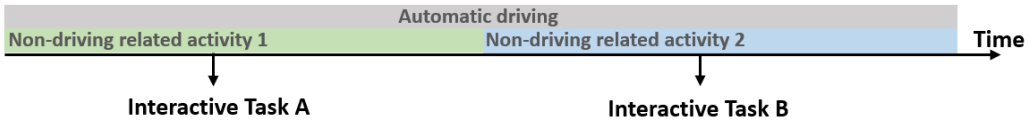

Fig. 3. Experimental Process.

Upon arrival, the researcher briefed the participants on the purpose and procedure of the study, emphasizing that all data would be used anonymously for scientific purposes. After the briefing, participants signed an informed consent form. Before starting the experiment, participants completed a pre-questionnaire to collect demographic information such as age and prior experience with voice interaction systems. Next, participants were trained on how to use the three activation methods and perform the two interaction tasks (navigation and social media browsing). The researcher confirmed that participants understood the tasks and activation patterns, providing further explanations if needed.

As depicted in Fig. 3, at the start of the experiment, the simulator's screen was set to autopilot mode. Once participants were acclimated to this setup, they were assigned to perform the tasks. Each participant was required to employ the three activation methods in a randomized sequence. The experiment's duration for each participant was approximately 20 minutes.

In the experiment, the researchers recorded the interactive duration and the failure rate within the task. after each activation methods, the participants filled out questionnaires related to usability, privacy, and subjective load, and the study ended with an activation method preference questionnaire and participant reports.

3.2 Result

This study employed multiple data analysis methods to evaluate the performance of different hybrid input methods. The Shapiro - Wilk test was used to determine the normality of the data, and the Levene test was used to determine the homogeneity of variances. If the data did not meet the conditions for parametric tests, the Friedman test was used for inter - group comparisons, followed by the Wilcoxon signed - rank test for post - hoc analysis. In the assessment of subjective workload, the NASA - TLX was used to calculate the total and average scores. For the evaluation of pragmatic and hedonic quality, the UEQ - S was used, and the Cronbach's alpha coefficient was calculated, along with the calculation of average scores.

3.2.1 Qualitative Measures

3.2.1.1 Subjective Workload

The NASA-TLX scores were used to estimate subjective workload. Participants were asked to rate their workload on a 20-item scale, which assessed six demand dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration level [89].The results indicated that participants in the Snapping Fingers group reported a higher average subjective workload (M=38.1, SE=18.7) compared to those in the Alex group (M=26.3, SE=16.5) and the Wake-up Free group (M=24.5, SE=16.1).

Statistical tests confirmed the normal distribution of data across all groups (Alex: P=0.123; Snapping Fingers: P=0.234; Wake-up Free: P=0.064) and homogeneous variances (P=0.663). A one-way ANOVA revealed a significant impact of wake-up method on workload (F=3.580, P=0.034). Post-hoc analysis (Tukey's method)indicated a significant difference between Snapping Fingers and Wake-up Free groups (P=0.043).

3.2.1.2 Pragmatic Quality and Hedonic Quality

The User Experience Questionnaire (UEQ-S) was used to assess the pragmatic quality (usability) and hedonic quality (user experience) of three wake-up methods [90]. The UEQ-S scale ranges from -3 to 3. To assess the consistency of the scale, we used Cronbach's α, where a value above 0.7 is considered sufficiently consistent. The results for Cronbach's α were as follows: Pragmatic quality—α_Alex=0.90, α_Snapping Fingers=0.81, α_Wake-up Free=0.95; Hedonic quality—α_Alex=0.95, α_Snapping Fingers=0.94, α_Wake-up Free=0.88.

The Wake-up Free group reported higher pragmatic quality mean scores (M=2.063, SD=1.076) compared to the other three groups. However, the Snapping Fingers group exhibited higher scores in hedonic quality (M=1.363, SD=1.182) (Table 3).

In terms of total UEQ-S score, S-W tests confirmed the data's normal distribution across all groups (Alex: D=0.953, P=0.417; Snapping Fingers: D=0.933, P=0.177; Wake-up Free: D=0.971, P=0.781). Levene's Test showed no significant difference in variance across groups (P = 0.322).A one-way ANOVA comparing the total UEQ-S scores across wake-up modes revealed no significant differences (F = 0.719, p = 0.492).In terms of pragmatic quality, the results show that not all of them conform to the normal distribution （Alex: D=0.872, P=0.013; Snapping Fingers: D=0.963, P=0.598; Wake-up Free: D=0.838, P=0.003）. The result of Levene's Test (p=0.774) shows that there is no statistically significant difference in the variability (variance) of the data in the groups examined, Friedman's test was used to compare pragmatic quality across the three wake-up methods. The results indicated significant differences (χ^2 = 7.719, df = 2, p = 0.021). Further post hoc test (Wilcoxon signed rank test) showed that there was a significant difference between wake-up free and Snapping Fingers (p=0.014), while no significant difference was observed between other pairs.

In terms of hedonic quality, the data for all three wake-up modes conformed to normal distribution (Alex: D = 0.961, P = 0.568; Snapping Fingers: D = 0.946, P = 0.308; Wake-up Free: D = 0.952, P = 0.339). Levene's Test indicated no significant difference in variance (P = 0.301). An ANOVA comparing hedonic quality across wake-up modes found no significant differences (F = 2.584, p = 0.084).

UEQ-S		**Input Form**		
		Alex	**Snapping Fingers**	**Wake-up Free**
Pragmatic quality	Mean	1.838	1.100	2.063
	SD	1.182	1.165	1.076
Hedonic quality	Mean	0.413	1.363	0.925
	SD	1.594	1.182	1.144
Overall	Mean	1.125	1.231	1.494
	SD	1.187	0.906	0.902

Table 3. UEQ-S Scores for Three Kinds of Wake Modes.

3.2.1.3 Usage Preferences

The study asked each participant to rank their preferred wake-up methods from most favored to least favored (Fig. 4). A clear preference divide was observed between the snapping fingers and wake-up free methods. 75% of participants preferred 'Alex' as their wake-up method, followed by wake-up free and then snapping fingers. Notably, 50% of participants ranked snapping fingers as their least preferred wake-up method.

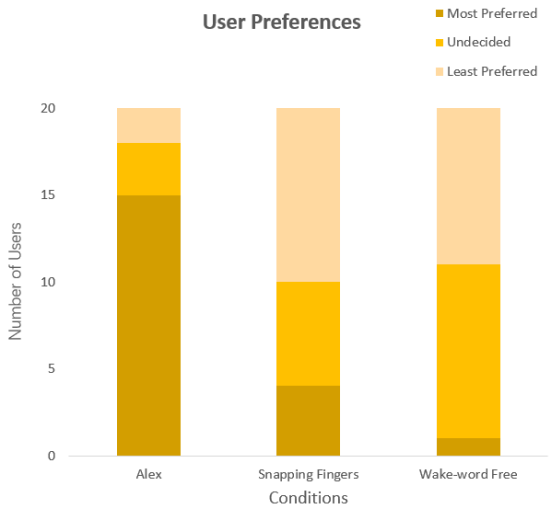

Fig. 4. User Preferences for Three Kinds of Wake Modes.

3.2.1.4 Privacy Concerns

The study designed a questionnaire to assess participants' privacy concerns regarding voice interaction systems. Four questions were included to capture their attitudes and sensitivities. As shown in Fig. 5, the results clearly highlight participants' privacy concerns when using voice interactive systems.

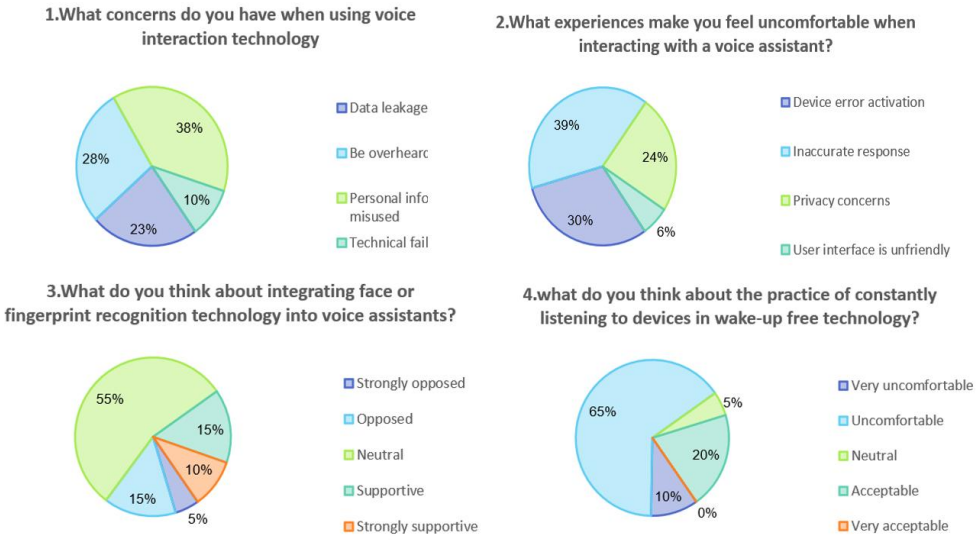

Fig. 5. Participants' Privacy Concerns.

3.2.2 Quantitative Measures

3.2.2.1 Interactive Duration

By measuring the interaction duration in different wake modes, we find that the average interaction duration of the wakeup free group is significantly lower than the other two groups (Alex and Snapping Fingers) (Table 4).

In the condition of task 1, we analyzed interaction durations across different wake-up modes, which were normally distributed (Table 4). Levene's Test showed homogeneous variances across groups (F=2.802, P=0.69), validating the use of ANOVA, which revealed significant differences in interaction durations among the modes (F=10.221, P < 0.001). Tukey's post-hoc test confirmed significant shorter durations for Wake-up Free compared to both Snapping Fingers and Alex (P<0.001 each), with no significant difference between Alex and Snapping Fingers.

In the condition of task 2, the interactive duration is normally distributed in different wake modes (Table 4), and Levene's Test showed no significant variance differences across groups (F=0.551, P=0.580). And ANOVA analysis demonstrated significant differences in interaction durations between modes (F=13.179, P < 0.001). Tukey's post-hoc test highlighted significant shorter durations for the Wake-up Free method compared to both Alex and Snapping Fingers (P<0.001 for each), with no notable difference between Alex and Snapping Fingers.

The average interactive duration under different task types was normal (Task 1: D=0.970, P=0.141; Task 2: D=0.984, P=0.618), the result of Levene's Test, which showed that the variance of the data in the examined group was statistically significant (F=21.502, P<0.01). Therefore, we used Wilcoxon Signed-Rank Test to compare the subjective load under two task types, and the result showed that the duration of task 2 was significantly lower than that of task 1 (N=60, MR = 30.50, SR = 1830.00). There was a significant difference between the two (Z=-6.736, p < 0.001). The results of the Wilcoxon signed rank test strongly support this difference and rule out that it is due to randomness.

Interactive Duration		**Alex**	**Snapping Fingers**	**Wake-up Free**
Task 1	Mean	43.436	42.000	32.338
	SE	10.5503	7.0397	7.2804
	S-W Test(P)	0.318	0.805	0.154
Task 2	Mean	19.680	20.985	14.930
	SE	4.6000	3.5432	7.2804
	S-W Test(P)	0.096	0.069	0.735

Table 4. Interactive Duration for Three Kinds of Wake Modes.

3.2.2.2 False Negatives

The false negative rates varied significantly across the three wake-up modes. The Snapping Fingers method exhibited the highest false negative rate (M=14.00%, SD=0.18890), surpassing both Alex and Wake-up Free modes (Table 3). The data were not normally distributed (Snapping Fingers: D= 0.727, P <0.001 wake-up Free: D=0 P=0). Levene's Test confirmed significant variance differences among the groups (P<0.001). Therefore, we used Friedman's test to compare the false negative indicators under the three wake-up modes, showing that there are significant differences in the false negative indicators for at least two of these wake-up modes. (χ^2 = 12.667, df=2, p = 0.002). Post-hoc analysis (Wilcoxon

signed-rank test) identified significant differences between the Snapping Fingers and Wake-up Free modes (P=0.011) and between the Snapping Fingers and Alex modes (P=0.018). However, no significant difference was observed between the Alex and Wake-up Free groups.

False Negatives	Input Form Alex	Snapping Fingers	Wake-up Free
Mean	1.25%	14.00%	0
SD	0.05590	0.18890	0

Table 5. False Negatives for Three Kinds of Wake Modes.

3.3 Discussion

This study explored the feasibility use of non-verbal sounds, specifically snapping fingers, as a wake-up method for voice interaction systems in NDRTs. The findings revealed that snapping fingers did not significantly improve interaction time, reduce subjective workload, or lower false negative rates compared to traditional WUWs or the wake-up free method. These findings prompt us to reconsider the methods for replacing traditional WUWs. This research will discuss the potential reasons behind these conclusions with the aim of enhancing the efficiency and user satisfaction of future technologies. Regarding interaction time and false negatives (Q1 & Q2), the results indicated that the wake-up free method performs best in improving interaction efficiency. This advantage is likely due to the elimination of the WUW steps and allows users to enter commands directly, thereby saving time. Additionally, Task 2 exhibited significantly shorter interaction duration than Task 1, suggesting that the complexity or type of task is a key factor influencing interaction time [91]. Higher task complexity tends to increase the proportion of incorrect responses [92], which can negatively affect the overall user experience. However, the sensitivity of different tasks to the wake-up method is limited, and further research may be needed in the future, considering a wider variety of NDRTs and more complex interaction scenarios. Regarding false negatives, the wake-up free method also achieved the lowest rate, enhancing responsiveness and reliability of the system. In contrast, snapping fingers method is often affected by individual differences among users and the system's limited recognition capacity, resulting in a higher rate of false negatives and lowering user expectations of the technology [70].

Regarding users' concerns (Q3), this study surveyed participants about their concerns when using voice interaction technology, and most participants expressed significant privacy concerns, including fears of eavesdropping, misuse of personal information, and inappropriate disclosure of data, followed by concerns about technical failures. These concerns mirror findings in related studies, highlighting the growing mistrust in voice interaction systems [88]. Participants also expressed discomfort with the constant monitoring inherent in wake-up free methods, which may explain their preference for traditional WUWs despite the higher efficiency of wake-up free. We asked participants about the most uncomfortable experiences with voice assistants. The results indicated that the majority of participants reported problems with device mis-activation or

inaccurate system responses. These experiences often lead to user frustration and may also lead to a decrease in trust in the system [89]; In addition, we explored participants' perceptions of constant detection in wake-up free technology. Many participants said they felt uncomfortable with this constant monitoring mechanism; When evaluating whether face recognition technology should be integrated into voice assistants, we found that most participants held a neutral attitude, some users had reservations, and a small number of participants chose to support it, considering the trade-off between privacy and data security. In conclusion, future research needs to take measures to reduce users' privacy concerns.

Regarding subjective workload (Q4), The wake-up free method achieved the best scores, reflecting its simplicity and reduced cognitive demand compared to snapping fingers and traditional WUWs. Additionally, snapping Fingers, as a wake-up method, required more physical effort, making it more demanding for users, which may affect the user's preference for this form of wake-up. Therefore, this research proposes that designers should take the subjective burden of users into consideration during the design process, so that users can maintain their physical and mental health while experiencing the convenience of technology in order to achieve the best interaction.

Regarding pragmatic quality and hedonic quality (Q4), all wake-up modes showed high overall scores, suggesting that wake-up methods have limited impact on overall ratings of user experience and usability. However, in terms of utility quality, the Wake-up Free group presents the best pragmatic quality indicator, which may indicate that users may feel freer and more direct in the absence of explicit instruction constraints [92]. Future research could further explore the causes of these differences, such as how the user's personal preferences, cultural background, or use environment affect the evaluation of different wake-up method.

Regarding users' specific preferences (Q4), this study found that participants preferred traditional WUWs (Alex) over the wake-up free and snapping fingers methods. Despite the efficiency of the Wake-up Free method in reducing interaction time, many users expressed concerns about continuous monitoring, which may lead to privacy breaches. It may cause users to be unwilling to use or even refuse to accept the services offered by smart car providers [93]. Similarly, Snapping Fingers was also met with skepticism due to uncertainty regarding its sound threshold and concerns about physical strain. Many participants were also uneasy about the potential safety risks of snapping fingers while driving, even in an autonomous vehicle scenario.

To be specific, they feared that this action may interfere with the driver taking control of the autonomous vehicle, creating a safety hazard. This potential uncertainty, combined with concerns about driving stability, left participants uneasy, as they realized that such small movements could have unintended consequences during emergencies. These concerns show that the choice of how users interact is not simply a matter of technology adoption, but also a combination of personal privacy, health, and convenience and security.

4. Exploring non-verbal sounds as input signals in continuous tasks for autonomous driving application

4.1 Methodology

4.1.1 Ways to Hybrid Sound Input

We designed four hybrid voice input methods to explore different ways of combining verbal and nonverbal sounds to control tasks:

(1) Multiple voice input:

Voice commands can be input multiple times, allowing users to repeat or modify tasks as needed.

(2) Voice + nonverbal sounds or say "stop":

Once a voice command is issued, the task will continue to execute until another command or a stop command is given. Users can end the task by snapping their fingers or saying "stop." The study utilizes standard automatic speech recognition (ASR) for voice commands, while finger snapping is detected based on specific audio features that identify this sound event.

(3) Continuous voice:

This method allows users to control the task by prolonging the final vowel of a word. The task will continue as long as vowel sound persists and it stops when the sound volume decreases to zero. ASR is used for detecting the initial voice command, while continuous control relies on volume-based detection to monitor task duration.

4.1.2 Setups

The empirical study was conducted in a quiet indoor environment. A display simulating the exterior environment of an autonomous vehicle was placed in front of the participants, with a tablet providing visual feedback. Audio was captured using a Logitech microphone and transmitted to a laptop, where the data was processed and uploaded to the cloud for voice recognition (Fig. 6).

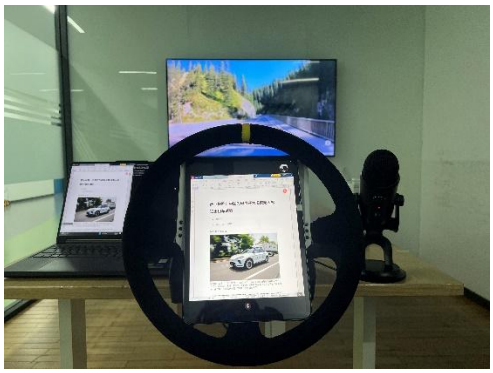

Fig. 6. Experimental Setups (Experiment 2).

4.1.3 Participants

The experiment invited 37 participants (16 males and 21 females), ranging in

age from 18 to 60 years (M=33.00, SD=9.829). All participants had normal or corrected-to-normal vision and hearing, and came from diverse backgrounds. During the study, we asked about the participants' familiarity with speech interaction technology: 2 participants said they were not familiar with the technology, 14 participants described their knowledge as moderate, 18 participants said they were very familiar, and 3 participants said they were very familiar and understood the underlying principles of the technology.

Before conducting the experiments, ethical approval was obtained from the University of Nottingham Ningbo China (UNNC) Ethics Committee. All participants were provided with detailed information about the study and gave their informed consent before participating.

4.1.4 Experiment Design

In Experiment 2, the tasks themselves were refined to better assess the real-time interaction between voice commands and non-verbal sounds. For example, continuous sound inputs, such as humming or snapping fingers, were incorporated into tasks that required fluid, ongoing control, such as adjusting volume or scrolling through content. This adds a layer of complexity compared to the discrete control tasks used in Experiment 1.

To reliably evaluate several hybrid voice input modes, the study used an in-subject design with four independent variables: four modes (Multiple Voice input vs Voice + say “top” vs Voice +Snapping Fingers vs Continuous Voice) and two interactive tasks (music volume adjustment task and social media browsing task) (Fig.7), we designed the continuity of the driving task related tasks include volume and social media browsing tasks, specific as follows:

Task1: music volume task

(1) Input the name of the songs to

(2) Adjust the volume to the target level (indicated by a red arrow on the display).

Task2: Social media browsing task

(1) open social media software

(2) browse to the target line on the screen.

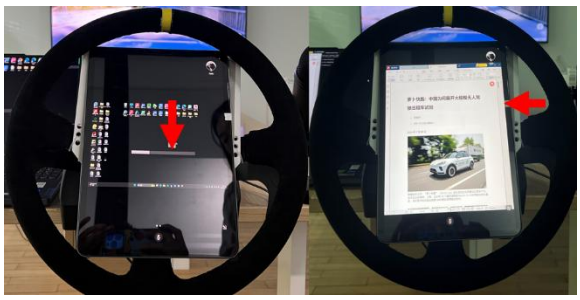

Fig. 7: Experiment Target (The left image shows Task 1, and the right image shows Task 2).

4.1.5 Evaluation Index

Subjective workload was assessed using the NASA TLX scores, which measure six dimensions: mental need, physical need, time need, performance, effort, and frustration [90]. The U-S questionnaire was used to assess the hedonic quality (user

experience) and pragmatic quality (usability) of the four hybrid voice inputs[90], as well as their specific preferences. This information will help us to better understand the needs and expectations of our users.

The study recorded the interactive duration of task execution, and the calculation results with the preset target deviation rate. For example, in task 1, we set the target value when the volume is adjusted to a red arrow position, and in Task 2, we set the progress bar to move to the red arrow position as the target value. Such quantitative data can not only reflect the efficiency and accuracy of task completion, but also provide a clear direction for the subsequent performance optimization.

4.1.6 Procedure

Upon arrival, participants were briefed on the study's purpose, and informed consent was obtained. A pre-questionnaire was administered to collect demographic data and assess familiarity with voice interaction technology. Before the experiment, researchers trained the participants on four hybrid voice input methods and the two interaction tasks: adjusting music volume and browsing social media. The lead researcher explicitly asked if participants understood the input methods and provided further explanations as needed.

The experiment officially began with the simulator screen set to autonomous driving mode, allowing participants time to acclimatize. The researcher then informed the participants of the start time for the interaction tasks. Each participant was required to complete the two tasks using the four hybrid voice input methods. During the experiment, researchers recorded the duration and accuracy of task completion. After each experiment session, participants completed questionnaires related to subjective workload and user experience. The study concluded with a specific preference survey and participant debriefing.

4.2 Result

This study employed multiple data analysis methods to evaluate the performance of different hybrid input methods. The Shapiro - Wilk test was used to determine the normality of the data, and the Levene test was used to determine the homogeneity of variances. If the data did not meet the conditions for parametric tests, the Friedman test was used for inter - group comparisons, followed by the Wilcoxon signed - rank test for post - hoc analysis. In the assessment of subjective workload, the NASA - TLX was used to calculate the total and average scores. For the evaluation of pragmatic and hedonic quality, the UEQ - S was used, and the Cronbach's alpha coefficient was calculated, along with the calculation of average scores.

4.2.1 Qualitative Measures

4.2.1.1 Subjective Workload

Use NASA-TLX scores to estimate subjective workload. Participants were asked to respond to a 20-point scale covering six dimensions of needs, including psychological needs, physical needs, time needs, performance, effort, and frustration[89]. We observed that the mean subjective load index of the continuous voice group was significantly higher than that of the other three groups,

indicating a higher subjective workload (Table 6).

The study assessed the subjective load total score of different input methods, and confirmed they did not show obvious normal distribution characteristics through S-W test (Table 6). Additionally, Levene's Test (F = 4.778, P = 0.03) suggested that at least one group had unequal variances. Therefore, we used Friedman's Test which showed there was a significant difference in the subjective load among the groups (χ^2 = 44.622, DF=3, p < 0.01). Subsequent post-hoc tests (Wilcoxon signed-rank test) show that there are significant differences between Continuous Voice and the other three input modes respectively（P <0.001 for each). There are no significant differences between the other input modes.

Subjective Workload		Input Form			
		Multiple Voice	Stop	Snapping Fingers	Continuous voice
Mean		29.20	27.40	27.80	54.89
SD		20.251	20.333	18.852	28.471
S-W	D	0.850	0.847	0.882	0.934
Test	P	<0.01	<0.01	0.01	0.037

Table 6. Subjective Workload for 4 Kinds of Input Forms.

4.2.1.2 Pragmatic Quality and Hedonic Quality

The User Experience Questionnaire (UEQ-S) was used to determine the quality of pragmatic (usability) and hedonic quality (user experience) of the three wake-up modalities [90]. The value of UEQ-S can range from -3 to 3. We measure the consistency of the scale α Cronbach's indicator, and the alpha value should be greater than 0.7 to be considered consistent enough. (Pragmatic quality:α _multiple voice=0.89 ,α_stop=0.86 ,α_snapping=0.90 ,α_continuous voice=0.93 ; Hedonic quality: α_multiple voice=0.80 , α_stop=0.84 , α_snapping=0.92, α _continuous voice=0.89).

Compared with the other three groups, the Stop group showed a higher pragmatic quality index (M=2.129, SD=0.748), but the Snapping Fingers group showed a higher index (M=1.563, SD=1.225) in the hedonic quality, and the specific parameters are as follows. A normality test (S-W test) was performed on the total score of the UEQ-S to confirm the distribution of each data set. The results showed that the total UEQ-S scores of all input methods did not show obvious normal distribution characteristics (Multiple Voice: D=0.953, P=0.0138; Stop: D=0.970, P=0.443; Snapping Fingers: D=0.931, P=0.031; Continuous voice: D=0.953, P=0.140), and the Levene's test indicates that at least one set of variances is unequal (F=3.157, P=0.027). We compared the UEQ-S total scores using the Friedman's test and found a significant difference in the UEQ-S total scores between groups (χ^2 = 44.622, DF = 3, p <0.01). Further post-hoc tests (Wilcoxon signed-rank test) showed that there was a significant difference between Continuous Voice and the other inputs (P_ Multiple Voice=0.003, P_ Stop=0.01, P_ Snapping Fingers=0.002), and there was no significant difference

between the other inputs.

In terms of pragmatic quality, the S-W test was performed on the data of different input methods (Multiple Voice: D=0.883 P=0.001, Stop: D=0.970 P=0.443; Snapping Fingers: D=0.931 P=0.031; Continuous Voice: D=0.953 P=0.140), the results showed that it did not conform to the normal distribution. The results of Levene's Test ($p<0.001$) illustrate statistically significant differences in the variability (variance) of the data within the groups. Therefore, we applied the Friedman test, identifying significant differences in pragmatic quality across the methods ($\chi^2=41.671$, DF=3, $P<0.001$). Further post-hoc tests (Wilcoxon signed-rank test) showed that there were significant differences between continuous voice and the other three input modes respectively (P_ Multiple Voice=0.001, P_ Stop=0.001, P_ Snapping Fingers=0.001), with no significant differences between multiple voice, stop and snapping fingers.

In terms of hedonic quality, the S-W test was applied, and the results showed that not all of them were normally distributed (Multiple Voice: D=0.951 P=0.118, Stop: D=0.941 P=0.060; Snapping Fingers: D=0.881 P=0.001; Continuous Voice: D=0.907 P=0.006). The results of Levene's Test ($p<0.855$) illustrate that there was no statistically significant difference in the variance of the data in the four groups. Since the data are not normally distributed, we used the Friedman's test to compare the indicators of hedonic quality, and the results showed that there was no significant difference between the four input methods ($\chi^2 = 5.250$, DF=3, $p=0.154$).

4.2.1.3 Usage Preferences

Participants ranked their preferred input methods, with 71.4% favoring the Stop input the most, followed by Multiple Voice and Snapping Fingers. Conversely, 68.6% of participants found the Continuous Voice to be their least preferred option (Fig. 8).

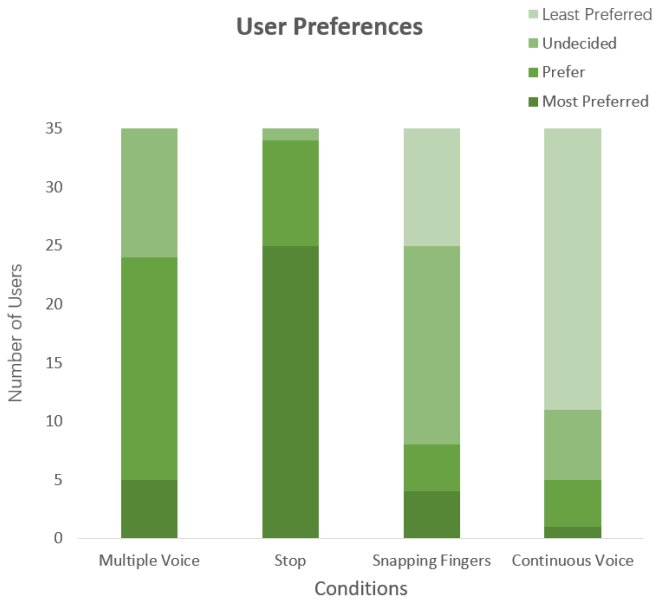

Fig. 8: User Preferences for 4 kinds of input forms.

4.2.2 Quantitative Measures

4.2.2.1 Interactive Duration

We analyzed interactive duration under four input methods across two tasks. In task 1, Multiple Voice group was recorded the shortest average interaction durations among all groups (Table 7). The interaction durations for all input methods approached normal distribution (Table 7). Levene's Test revealed significant variance differences among the groups (F=5.347, P=0.002). We used the Friedman's test to compare the interaction duration of the four input modes, and the results showed that there was a significant difference (χ^2 = 16.131, df = 3, p = 0.01). Further analysis with Wilcoxon post-hoc tests (Wilcoxon signed-rank test) highlighted significant differences between the Stop and Multiple Voice (P=0.02), and Continuous Voice when compared to other methods (Multiple Voice: P<0.001; Stop: P=0.014; Snapping Fingers: P<0.001), while no other significant differences were observed.

In task 2, the Stop group showed the shortest average interaction duration. Normal distribution tests (S-W test) for the data showed were close to normal distribution (Table 7). and Levene's Test indicated unequal variances among groups (F=9.270, p<0.01). We used the Friedman's test to compare the interaction duration of the four input modes, and the results showed that there was a significant difference in interaction durations (χ^2 = 84.840, DF=3, p <0.01). Further post-hoc tests (Wilcoxon signed-rank test) showed that there was a significant difference in the interaction duration between Stop and voice (p<0.01), some significant differences between Continuous Voice and Multiple voice, Stop or Snapping Fingers (P<0.01 for each), and there was no significant difference between the other inputs.

The analysis of interaction durations across different task types indicated that the durations were not normally distributed (Task 1: P=0.740; Task 2: P<0.01). Levene's Test confirmed significant differences in variance (F=65.961, P<0.01). The Wilcoxon Signed-Rank Test revealed that interaction durations in task 2 were significantly shorter than those in task 1 (Z = -10.265, P < 0.001), showing a notable distinction in performance between the two tasks.

Interactive Duration			Input Form			
			Multiple Voice	Stop	Snapping Fingers	Continuous Voice
Task 1	Mean		9.8277	10.7026	10.3420	11.3591
	SD		1.83375	1.31899	1.01160	1.10576
	S-W Test	D	0.980	0.954	0.981	0.972
		P	0.750	0.145	0.793	0.512
Task 2	Mean		21.4846	13.3583	13.65660	15.1543
	SD		2.38374	1.09085	1.204180	1.51558
	S-W Test	D	0.9	0.979	0.969	0.956
		p	0.181	0.725	0.405	0.176

Table 7. Interactive Duration for 4 Kinds of Input Forms.

4.2.2.2 Deviation Rate

In this study, we statistically analyzed the effects of four different input modes (Multiple Voice, Stop, Snapping Fingers and Continuous Voice) on deviation rates. Because the results of the tasks entered by voice are 100% correct, the deviation rate is 0. The average deviation rate of Stop is lower than that of Snapping Fingers and Continuous Voice of the two types of tasks (Table 8).

In the condition of task 1, through normal test (S-W Test) to confirm the distribution of the data set. The results show that the deviation rate data in all input modes is not in obvious non-normal distribution (p_ Stop=0.058, p_ Snapping Fingers =0.185, p_ Continuous Voice=0.200). Levene's test showed equal variance among the groups (F=1.405, P=0.205). ANOVA analysis showed that there was no significant difference in the performance of all input methods (F=38.674, $P < 0.01$). In multiple comparison, we adopted the Tukey HSD test to evaluate the significant difference between the input modes. The Stop input had a significantly lower deviation rate compared to both Snapping Fingers and Continuous Voice ($P < 0.001$). Although the difference in deviation rates between Snapping Fingers and Continuous Voice was smaller, it remained statistically significant ($P = 0.023$), with Snapping Fingers showing a lower rate.

In the condition of task 2, the S-W test indicated that deviation rates for all input methods were approximately normally distributed (Multiple Voice: D = 0.957, P = 0.181 Stop: D = 0.986, P = 0.929; Snapping Fingers: D=0.969 P=0.405; Continuous Voice: D=0.956 P=0.176). However, the results of Levene's Test (F=136, $p < 0.01$) indicate that at least one variance of the deviation rate is not equal between the groups. We used the Friedman's test to compare the deviation rates of the four input methods, and the results showed significant differences in the deviation rate between the groups ($\chi^2 = 84.840$, df=3, $p < 0.01$). Further post hoc tests (Wilcoxon signed rank test) showed that There is a significant difference

between Stop and the other input modes (P<0.01 for each). There is no significant difference between the other two modes.

Deviation Rate		Input Form			
		Multiple Voice	Stop	Snapping Fingers	Continuous voice
Task 1	Mean	0	3.3810%	-8.1905%	-4.5714%
	SD	/	0.0449193	0.0673217	0.0544345
Task 2	Mean	0	1.5063%	1.5899%	1.8192%
	SD	/	0.0117481	0.0143020	0.0106637

Table 8. Deviation Rate for 4 Kinds of Input Forms.

4.3 Discussion

In this study, we selected a variety of hybrid input methods as research objects, aiming to explore their application performance in continuous NDRTs. The conventional voice input mode was used as a baseline to compare the effectiveness of the various input methods in real-world usage. Our findings revealed key insights into user preferences and the performance of these methods across multiple metrics. One of the most notable findings was the superior performance of the "stop" command compared to snapping fingers as a means to terminate a task. The "stop" command demonstrated better results in both interaction duration and deviation rate, and was the most preferred method among participants. Conversely, the prolonged sound (continuous voice) method performed the worst across most metrics. This method resulted in higher subjective workload, longer interaction times, and greater deviation rates. Therefore, this section will focus on the causes of these, so as to provide useful references and improvement methods in the future design of HVI systems.

Regarding interaction time (RQ1), our results revealed mixed performance across tasks. In Task 1 (volume adjustment), the multiple voice input method resulted in significantly shorter interaction times compared to the "stop" control method. However, in Task 2 (social media browsing), the results showed a completely opposite trend, with the "stop" control method has better performance on time efficiency than multiple voice input. These contrasting results may be attributed to the inherent differences in task complexity. For simpler tasks like volume adjustment, multiple voice inputs may allow for more direct and quicker control. Additionally, the continuous voice method shown poor interaction efficiency in both tasks. The participants suggested that in a shorter task environment, the continuous voice method may provide a smoother experience and lower physical burden. This reminds us that future designers should take into account the characteristics and expectations of different types of tasks when designing interactive interfaces. Designers can simulate various scenarios to test different input methods and find the most suitable strategy for a specific task. This will not only improve user

experience but also ensure that the system remains efficient and stable under complex usage conditions.

Regarding deviation rate (RQ1), the 0% deviation rate set by multiple voice inputs was due to system settings. Participants noted that when the gradient of a single voice command was too large, command accuracy tended to decrease. Conversely, if the gradient was small but the target value was distant, participants were forced to make frequent readjustments, adding to their operational load. Additionally, without the ability to monitor their input in real-time, participants found it challenging to achieve optimal results through repeated adjustments, which negatively impacted their overall experience. The “stop” input method performed better, allowing for real-time control of volume adjustments and page scrolling. However, the continuous voice method exhibited a much higher deviation rate than other input methods, likely due to users’ limited stamina and skill in maintaining continuous vocalization. Moreover, there was an approximate 1-second delay between the end of vocalization and the system stopping actions, such as volume increase or page scrolling. While previous research has shown that non-verbal vocalizations, like humming, can improve input accuracy compared to speech in gaming contexts[101], our findings differ. This discrepancy may be attributed to the fact that their studies leveraged the pitch of non-verbal sounds for more precise control. To improve user experience, developers must optimize algorithms to minimize unnecessary delays and explore additional information from non-verbal vocalizations, allowing users to achieve more precise control through real-time feedback, thereby creating a smoother and more natural interaction process.

Regarding subjective workload (RQ2), the continuous voice input method had the greatest workload among the input methods. This high workload stems from the need for continuous vocalization, which naturally increases physical pressure on the participants. As one participant mentioned, the method could be better suited for tasks requiring fine-tuned adjustments, where brief but precise vocalization might be useful. Several participants also suggested that traditional voice input could be simplified by converting repetitive commands into a more concise form (e.g., “volume up, up, up” instead of saying “volume up” multiple times) This adjustment could reduce the subjective effort required and enhance the overall user experience. Regarding the lack of significant differences among the other input methods, future research could focus on exploring various task types and the intensity required to complete them. A more detailed understanding of task characteristics could facilitate the development of input tools better suited for specific tasks, allowing users to perform tasks with minimal physical workload. Regarding pragmatic quality and hedonic quality (RQ3),In terms of pragmatic quality, the continuous voice method performed the worst, likely due to the increased physical burden it imposed on users. The need for sustained vocalization contributed to a less favorable user experience. On the other hand, the other input methods, including multiple voice input and Stop input, showed similar pragmatic quality scores, In terms of hedonic quality, there was no obvious gap between the four input methods, which may be because the system provided similar feedback mechanisms [99]. Future designs should take into account the user's physiological burden and system feedback mechanisms to improve the user experience.

In terms of user preferences (QR4), this research indicated that Stop input method was the most favored, likely due to its combination of low physical effort and high effectiveness across both tasks. It was simple to learn and intuitively understood by users. Traditional voice input also received support from some participants. An interesting suggestion from participants was to use the loudness of the user's voice to directly control the music volume. This offers valuable insights for future voice interaction systems, specifically in how to better integrate user actions with system functionality to improve the overall user experience and satisfaction.

5. Conclusion

This research provided a comprehensive investigation into the application of non-verbal sound technology in autonomous driving, with a particular focus on its use in voice activation and the management of continuous NDRTs. The study was divided into two parts: the first part examined the effectiveness and user experience of finger snapping as a novel wake-up method., while the second part evaluated the role of non-verbal sounds in managing continuous NDRTs within AVs

In the first experiment, we investigated the potential of using non-verbal sounds to wake up voice interaction systems. The results indicated that, although snapping fingers to wake up the system did not significantly improve interaction efficiency, the wake-up-free method performed well in reducing false negatives and shortening interaction duration. However, the most participants still preferred traditional wake-up methods, likely due to their familiarity and trust in these interactions. Therefore, while non-verbal sound technology offers unique advantages, further improvements are needed in terms of user experience, task efficiency, and wake-up reliability.

In the second experiment, we further explored the performance of non-verbal sounds in continuous tasks by comparing the effectiveness of Stop input methods with Continuous Voice (vowel sustain) for managing NDRTs. The results showed that the Stop input method performed better at maintaining the text deviation rate and was preferred by the majority of participants. However, continuous voice did not exhibit a significant advantage in terms of interaction efficiency when compared to other sound inputs. In addition, the workload required to complete certain tasks appears to be large for continuous voice input, suggesting potential flaws in task design. To address this problem, we suggest that continuous voice may be more suitable for tasks requiring fine-tuning in the future.

In conclusion, while this research highlights the potential of non-verbal sound interaction technology, it also underscores the challenges and areas for improvement. Our initial findings show promise under specific conditions, but further research is necessary to confirm the broader applicability of these technologies. Future research can start from user experience, refining system design, and optimizing algorithms to enable more efficient HVI. Through ongoing technological iteration and the incorporation of user feedback, non-verbal sound technology has the potential to become a key driver of AI development, offering richer and more seamless interactive experiences.

6. Reference

[1] Pilataxi J, Vinan W, Chavez D. Design and implementation of a driving assistance system in a car-like robot when fatigue in the user is detected[J]. IEEE Latin America Transactions, 2016, 14(2): 457-462.

[2] Kim H S, Yoon S H, Kim M J, et al. Deriving future user experiences in autonomous vehicle[C]//Adjunct proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications. 2015: 112-117.

[3] Janssen C P, Kun A L, Brewster S, et al. Exploring the concept of the (future) mobile office[C]//Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications: Adjunct proceedings. 2019: 465-467.

[4] Krome S, Batty J, Greuter S, et al. Autojam: Exploring interactive music experiences in stop-and-go traffic[C]//Proceedings of the 2017 conference on designing interactive systems. 2017: 441-450.

[5] Tan Z, Dai N, Su Y, et al. Human–machine interaction in intelligent and connected vehicles: a review of status quo, issues, and opportunities[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(9): 13954-13975.

[6] Zhang Y, Angell L. Pointing towards future automotive HMIs: The potential for gesture interaction[C]//Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. 2014: 1-6.

[7] Jianan L, Abas A. Development of human-computer interactive interface for intelligent automotive[J]. International Journal of Artificial Intelligence, 2020, 7(2): 13-21.

[8] Yoshimura N, Yoshida H, Matulic F, et al. Extending discrete verbal commands with continuous speech for flexible robot control[C]//Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. 2019: 1-6.

[9] Kaur S. Mouse movement using speech and non-speech characteristics of human voice[J]. Int. J. Eng. Adv. Technol, 2012, 1(5): 368-374.

[10] Hone K S, Baber C. Modelling the effects of constraint upon speech-based human–computer interaction[J]. International Journal of Human-Computer Studies, 1999, 50(1): 85-107.

[11] Kim Y, Gruber T R, Bridle J. Dynamic thresholds for always listening speech trigger: U.S. Patent 10,789,041[P]. 2020-9-29.

[12] Bleakleya A, Wua Y, Pandeyb A, et al. “Hey Guguru”: Exploring Non-English Linguistic Barriers for Wake Word Use[J]. 2021.

[13] Brédart S. Strategies to improve name learning[J]. European Psychologist, 2019.

[14] Këpuska V Z, Klein T B. A novel wake-up-word speech recognition system, wake-up-word recognition task, technology and evaluation[J]. Nonlinear Analysis: Theory, Methods & Applications, 2009, 71(12): e2772-e2789.

[15] Schönherr L, Golla M, Eisenhofer T, et al. Unacceptable, where is my privacy? exploring accidental triggers of smart speakers[J]. arXiv preprint arXiv:2008.00508, 2020.

[16] Banks V A, Stanton N A. Analysis of driver roles: Modelling the changing role of the driver in automated driving systems using EAST[J]. Theoretical issues in ergonomics science,

2019, 20(3): 284-300.
[17] Rosenfeld A, Bareket Z, Goldman C V, et al. Towards adapting cars to their drivers[J]. AI Magazine, 2012, 33(4): 46-46.
[18]SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," Apr. 2021. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/
[19] Jose C, Mishchenko Y, Senechal T, et al. Accurate detection of wake word start and end using a CNN[J]. arXiv preprint arXiv:2008.03790, 2020.
[20] Ataya A, Kim W, Elsharkawy A, et al. How to interact with a fully autonomous vehicle: Naturalistic ways for drivers to intervene in the vehicle system while performing non-driving related tasks[J]. Sensors, 2021, 21(6): 2206.
[21] H. Detjen, S. Geisler and S. Schneegass, "Maneuver-based Control Interventions During Automated Driving: Comparing Touch, Voice, and Mid-Air Gestures as Input Modalities," 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 2020, pp. 3268-3274, doi: 10.1109/SMC42975.2020.9283431.
[22] Tscharn R, Latoschik M E, Löffler D, et al. “Stop over there”: Natural gesture and speech interaction for non-critical spontaneous intervention in autonomous driving[C]//Proceedings of the 19th acm international conference on multimodal interaction. 2017: 91-100.
[23] Sun X, Cao S, Tang P. Shaping driver-vehicle interaction in autonomous vehicles: How the new in-vehicle systems match the human needs[J]. Applied ergonomics, 2021, 90: 103238.
[24] Kun A L. Human-machine interaction for vehicles: Review and outlook[J]. Foundations and Trends® in Human–Computer Interaction, 2018, 11(4): 201-293.
[25] Carsten, O., Lai, F. C. H., Barnard, Y., Jamson, A. H., & Merat, N. (2012). Control Task Substitution in Semiautomated Driving: Does It Matter What Aspects Are Automated? Human Factors, 54(5), 747-761. https://doi.org/10.1177/0018720812460246
[26] Lin Q F, Lyu Y, Zhang K F, et al. Effects of non-driving related tasks on readiness to take over control in conditionally automated driving[J]. Traffic injury prevention, 2021, 22(8): 629-633.
[27]R. Neßelrath, M. M. Moniri and M. Feld, "Combining Speech, Gaze, and Micro-gestures for the Multimodal Control of In-Car Functions," 2016 12th International Conference on Intelligent Environments (IE), London, UK, 2016, pp. 190-193, doi: 10.1109/IE.2016.42.
[28] Wandtner B, Schömig N, Schmidt G. Effects of non-driving related task modalities on takeover performance in highly automated driving[J]. Human factors, 2018, 60(6): 870-881.
[29] Cui C, Ma Y, Cao X, et al. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 902-909.
[30] JIANG Q, ZHUANG X, MA G. Evaluation of external HMI in autonomous vehicles based on pedestrian road crossing decision-making model[J]. Advances in Psychological Science, 2021, 29(11): 1979.
[31] Wang Y, Xu Q. A filed study of external HMI for autonomous vehicles when interacting with pedestrians[C]//HCI in Mobility, Transport, and Automotive Systems. Automated Driving and In-Vehicle Experience Design: Second International Conference, MobiTAS

2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22. Springer International Publishing, 2020: 181-196.
[32]Murali P K, Kaboli M, Dahiya R. Intelligent In-Vehicle Interaction Technologies[J]. Advanced Intelligent Systems, 2022, 4(2): 2100122.
[33]Shanmugarajah S, Tharmaseelan J, Sivagnanam L. AI Approach In Monitoring The Physical And Psychological State Of Car Drivers And Remedial Action For Safe Driving[C]//2020 2nd International Conference on Advancements in Computing (ICAC). IEEE, 2020, 1: 186-191.
[34]Bajaj J S, Kumar N, Kaushal R K, et al. System and method for driver drowsiness detection using behavioral and sensor-based physiological measures[J]. Sensors, 2023, 23(3): 1292.
[35]Stecher M, Michel B, Zimmermann A. The benefit of touchless gesture control: An empirical evaluation of commercial vehicle-related use cases[C]//Advances in Human Aspects of Transportation: Proceedings of the AHFE 2017 International Conference on Human Factors in Transportation, July 17− 21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8. Springer International Publishing, 2018: 383-394.
[36]Bellani P, Picardi A, Caruso F, et al. Enhancing User Engagement in Shared Autonomous Vehicles: An Innovative Gesture-Based Windshield Interaction System[J]. Applied Sciences, 2023, 13(17): 9901.
[37]Ataya A, Kim W, Elsharkawy A, et al. How to interact with a fully autonomous vehicle: Naturalistic ways for drivers to intervene in the vehicle system while performing non-driving related tasks[J]. Sensors, 2021, 21(6): 2206.
[38]Turk M. Multimodal interaction: A review[J]. Pattern recognition letters, 2014, 36: 189-195.
[39]H. Detjen, S. Geisler and S. Schneegass, "Maneuver-based Control Interventions During Automated Driving: Comparing Touch, Voice, and Mid-Air Gestures as Input Modalities," 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 2020, pp. 3268-3274, doi: 10.1109/SMC42975.2020.9283431.
[40]Döring T, Kern D, Marshall P, et al. Gestural interaction on the steering wheel: reducing the visual demand[C]//Proceedings of the sigchi conference on human factors in computing systems. 2011: 483-492.
[41]Angelini L, Baumgartner J, Carrino F, et al. Comparing gesture, speech and touch interaction modalities for in-vehicle infotainment systems[C]//Actes de la 28e conférence francophone sur l'Interaction Homme-Machine on-IHM'16, 25-28 octobre 2016, Fribourg, Suisse. 25-28 Octobre 2016, 2016.
[42]Zhao D, Wang C, Liu Y, et al. Implementation and evaluation of touch and gesture interaction modalities for in-vehicle infotainment systems[C]//Image and Graphics: 10th International Conference, ICIG 2019, Beijing, China, August 23–25, 2019, Proceedings, Part III 10. Springer International Publishing, 2019: 384-394.
[43]Nickel K, Stiefelhagen R. Visual recognition of pointing gestures for human – robot interaction[J]. Image and vision computing, 2007, 25(12): 1875-1884.
[44]Bellani P, Picardi A, Caruso F, et al. Enhancing User Engagement in Shared Autonomous Vehicles: An Innovative Gesture-Based Windshield Interaction System[J]. Applied Sciences, 2023, 13(17): 9901.
[45]Zheng P, McDonald M, Pickering C. Effects of intuitive voice interfaces on driving and in-vehicle task performance[C]//2008 11th International IEEE Conference on Intelligent

Transportation Systems. IEEE, 2008: 610-615.
[46]Harrington K, Large D R, Burnett G, et al. Exploring the use of mid-air ultrasonic feedback to enhance automotive user interfaces[C]//Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications. 2018: 11-20.
[47]May K R, Gable T M, Walker B N. A multimodal air gesture interface for in vehicle menu navigation[C]//Adjunct proceedings of the 6th international conference on automotive user interfaces and interactive vehicular applications. 2014: 1-6.
[48]Wu H, Wang Y, Liu J, et al. User-defined gesture interaction for in-vehicle information systems[J]. Multimedia Tools and Applications, 2020, 79: 263-288.
[49]Large D R, Harrington K, Burnett G, et al. Feel the noise: Mid-air ultrasound haptics as a novel human-vehicle interaction paradigm[J]. Applied ergonomics, 2019, 81: 102909.
[50] Igarashi T, Hughes J F. Voice as sound: using non-verbal voice input for interactive control[C]//Proceedings of the 14th annual ACM symposium on User interface software and technology. 2001: 155-156.
[51]Nygaard L C, Tzeng C Y. Perceptual integration of linguistic and non‐linguistic properties of speech[J]. The handbook of speech perception, 2021: 398-427.
[52]Funk M, Tobisch V, Emfield A. Non-verbal auditory input for controlling binary, discrete, and continuous input in automotive user interfaces[C]//Proceedings of the 2020 CHI conference on human factors in computing systems. 2020: 1-13.
[53]Trouvain J, Truong K P. Comparing non-verbal vocalisations in conversational speech corpora[C]//4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2012). European Language Resources Association (ELRA), 2012: 36-39.
[54]Zhang B J, Fitter N T. Nonverbal sound in human-robot interaction: a systematic review[J]. ACM Transactions on Human-Robot Interaction, 2023, 12(4): 1-46.
[55]Yilmazyildiz S, Read R, Belpeame T, et al. Review of semantic-free utterances in social human–robot interaction[J]. International Journal of Human-Computer Interaction, 2016, 32(1): 63-85.
[56]Seo J H, Yang J Y, Kim J, et al. Autonomous humanoid robot dance generation system based on real-time music input[C]//2013 IEEE RO-MAN. IEEE, 2013: 204-209.
[57]Kim H S, Yoon S H, Kim M J, et al. Deriving future user experiences in autonomous vehicle[C]//Adjunct proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications. 2015: 112-117.
[58]M ü ller C, Weinberg G. Multimodal input in the car, today and tomorrow[J]. IEEE MultiMedia, 2011, 18(1): 98-103.
[59]Kaur S. Mouse movement using speech and non-speech characteristics of human voice[J]. Int. J. Eng. Adv. Technol, 2012, 1(5): 368-374.
[60]McTear M F. Spoken dialogue technology: enabling the conversational user interface[J]. ACM Computing Surveys (CSUR), 2002, 34(1): 90-169.
[61]Schafer R W. Scientific bases of human-machine communication by voice[J]. Proceedings of the National Academy of Sciences, 1995, 92(22): 9914-9920.
[62]Nakatsu R, Suzuki Y. What does voice-processing technology support today?[J]. Proceedings of the National Academy of Sciences, 1995, 92(22): 10023-10030.

[63]McTear M F. Spoken dialogue technology: enabling the conversational user interface[J]. ACM Computing Surveys (CSUR), 2002, 34(1): 90-169.
[64]Porcheron M, Fischer J E, Reeves S, et al. Voice interfaces in everyday life[C]//proceedings of the 2018 CHI conference on human factors in computing systems. 2018: 1-12.
[65]Francisco Carlos, C Alinne et al. "An Analysis of Visual Speech Features for Recognition of Non-articulatory Sounds using Machine Learning." (2019). 1-9.. Francisco Carlos; C Alinne; Y. Carolina; Patricia Pupin; Alessandra Alaniz.
[66]Kumar R, Rodehorst M, Wang J, et al. Building a robust word-level wakeword verification network[J]. 2020.
[67]Jose C, Mishchenko Y, Senechal T, et al. Accurate detection of wake word start and end using a CNN[J]. arXiv preprint arXiv:2008.03790, 2020.
[68]Këpuska V Z, Klein T B. A novel wake-up-word speech recognition system, wake-up-word recognition task, technology and evaluation[J]. Nonlinear Analysis: Theory, Methods & Applications, 2009, 71(12): e2772-e2789.
[69]Bleakleya A, Wua Y, Pandeyb A, et al. "Hey Guguru": Exploring Non-English Linguistic Barriers for Wake Word Use[J]. 2021.
[70]Cámbara G, López F, Bonet D, et al. Tase: Task-aware speech enhancement for wake-up word detection in voice assistants[J]. Applied Sciences, 2022, 12(4): 1974.
[71]McMillan D, Brown B, Kawaguchi I, et al. Designing with gaze: Tama--a gaze activated smart-speaker[J]. Proceedings of the ACM on Human-Computer Interaction, 2019, 3(CSCW): 1-26.
[72]Pomykalski P, Woźniak M P, Woźniak P W, et al. Considering wake gestures for smart assistant use[C]//Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020: 1-8.
[73]Zhao S, Westing B, Scully S, et al. Raise to speak: An accurate, low-power detector for activating voice assistants on smartwatches[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 2736-2744.
[74]Yadav S, Legaspi P A D, Alink M S O, et al. Hardware implementations for voice activity detection: Trends, challenges and outlook[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2022, 70(3): 1083-1096.
[75]Atal B S. Automatic recognition of speakers from their voices[J]. Proceedings of the IEEE, 1976, 64(4): 460-475.
[76]Zhang H, Wang J, Yang S, et al. A Multimodal Activation Detection Model for Wake-Free Robots[C]//Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. Singapore: Springer Nature Singapore, 2022: 97-109.
[77]Dong X. Innovation and Application of Speech Recognition Technology in Automatic Fare Collection of Rail Transit[C]//2019 2nd International Conference on Mathematics, Modeling and Simulation Technologies and Applications (MMSTA 2019). Atlantis Press, 2019: 195-198.
[78]Vertegaal R, Slagter R, Van der Veer G, et al. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes[C]//Proceedings of the SIGCHI conference on Human factors in computing systems. 2001: 301-308.
[79]Jung H, Kim H. Finding contextual meaning of the wake word[C]//Proceedings of the 1st

International Conference on Conversational User Interfaces. 2019: 1-3.
[80]Albert S, Hamann M. Putting wake words to bed: We speak wake words with systematically varied prosody, but CUIs don't listen[C]//Proceedings of the 3rd Conference on Conversational User Interfaces. 2021: 1-5.
[81]Luger E, Sellen A. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents[C]//Proceedings of the 2016 CHI conference on human factors in computing systems. 2016: 5286-5297.
[82]Combs M, Hazelwood C, Joyce R. Are you listening?–an observational wake word privacy study[J]. Organizational Cybersecurity Journal: Practice, Process and People, 2022, 2(2): 113-123
[83]Pierre-Yves O. The production and recognition of emotions in speech: features and algorithms[J]. International Journal of Human-Computer Studies, 2003, 59(1-2): 157-183.
[84]Lamel L, Gauvain J L. A phone-based approach to non-linguistic speech feature identification[J]. Computer Speech & Language, 1995, 9(1).
[85]Read R, Belpaeme T. How to use non-linguistic utterances to convey emotion in child-robot interaction[C]//Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction. 2012: 219-220.
[86]Read R, Belpaeme T. Non-linguistic utterances should be used alongside language, rather than on their own or as a replacement[C]//Proceedings of the 2014 ACM/IEEE International Conference on Human-robot interaction. 2014: 276-277.
[87]Sakamoto D, Komatsu T, Igarashi T. Voice augmented manipulation: using paralinguistic information to manipulate mobile devices[C]//Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services. 2013: 69-78.
[88]Dahlbäck N, Jönsson A, Ahrenberg L. Wizard of Oz studies: why and how[C]//Proceedings of the 1st international conference on Intelligent user interfaces. 1993: 193-200.
[89]Byers J C, Bittner A C, Hill S G. Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary[J]. Advances in industrial ergonomics and safety, 1989, 1: 481-485.
[90]Schrepp M, Hinderks A, Thomaschewski J. Design and evaluation of a short version of the user experience questionnaire (UEQ-S)[J]. International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108., 2017.
[91]Darbutas T, Juodžbalienė V, Skurvydas A, et al. Dependence of reaction time and movement speed on task complexity and age[J]. Medicina, 2013, 49(1): 4.
[92]Aykin N, Aykin T. Complex Task Performance under Speed-Accuracy Tradeoff: Single Task versus Dual Task[C]//Proceedings of the Human Factors Society Annual Meeting. Sage CA: Los Angeles, CA: SAGE Publications, 1987, 31(2): 161-165.
[93]Cheng P, Roedig U. Personal voice assistant security and privacy—a survey[J]. Proceedings of the IEEE, 2022, 110(4): 476-507.
[94]Turner C W, Safar J A, Hardzinski M, et al. The effects of service availability and recognition errors on trust in voice user interfaces[C]//Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Sage CA: Los Angeles, CA: SAGE Publications, 2005, 49(5): 656-660.
[95]Pfoh E R, Hong S, Baranek L, et al. Reduced cognitive burden and increased focus: a mixed-

methods study exploring how implementing scribes impacted physicians[J]. Medical care, 2022, 60(4): 316-320.

[96]Anderson A J, Vingrys A J. Small samples: does size matter?[J]. Investigative Ophthalmology & Visual Science, 2001, 42(7): 1411-1413.

[97]Deci E L, Ryan R M. The support of autonomy and the control of behavior[J]. Journal of personality and social psychology, 1987, 53(6): 1024.

[98]Ostern N, Eßer A, Buxmann P. Capturing Users' Privacy Expectations To Design Better Smart Car Applications[C]//PACIS. 2018: 97.

[99]Silva B, Costelha H, Bento L C, et al. User-experience with haptic feedback technologies and text input in interactive multimedia devices[J]. Sensors, 2020, 20(18): 5316.

[100] Naujoks, Frederik, et al. "Improving usefulness of automated driving by lowering primary task interference through HMI design." Journal of Advanced Transportation 2017.1 (2017): 6105087.

[101] Sporka A J, Kurniawan S H, Mahmud M, et al. Non-speech input and speech recognition for real-time control of computer games[C]//Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility. 2006: 213-220.

Appendix A: NASA-TLX Questionnaire

NASA Task Load Index (TLX)

Name: **Task:** **No.:**

1. **Mental Demand**: How much mental and perceptual activity was required? (e.g., thinking, deciding, calculating, remembering, looking, searching) /心理需求：需要多少心理和知觉活动？（例如，思考、决定、计算、记忆、观察、搜索）

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low (1) Very High (20)

非常低（1） 非常高（20）

2. **Physical Demand**: How much physical activity was required? (e.g., pushing, pulling, turning, controlling, activating) /体力需求：需要多少体力活动？（如推、拉、转、控制、激活）

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low (1) Very High (20)

非常低（1） 非常高（20）

3. **Temporal Demand**: How much time pressure did you feel due to the pace of the task or your inability to complete it in time? /时间需求：由于任务的节奏或你无法及时完成它，你感到有多大的时间压力？

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low (1) Very High (20)

非常低（1） 非常高（20）

4. **Performance**: How successful do you think you were in accomplishing the goals of the task? /**表现**：您认为自己在完成任务目标方面有多成功？

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low (1) Very High (20)

非常低（1） 非常高（20）

5. **Effort**: How hard did you have to work to accomplish your level of performance? **/努力：**完成任务时，您需要付出多少努力才能达到自己的表现水平？

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low (1) Very High (20)

非常低（1） 非常高（20）

6. **Frustration Level**: How insecure, discouraged, irritated, stressed, and annoyed were you? **挫折感：**您在任务过程中感到有多不安、沮丧、恼火、压力大或烦恼？

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low (1) Very High (20)

非常低（1） 非常高（20）

Appendix B: User Experience Questionnaire (UEQ-S)

Name: **No.:**

The questionnaire is designed to evaluate your user experience with different input methods. Please answer each question based on how you really feel. /该问卷旨在评估您对不同的输入方式的用户体验。请根据您的真实感受回答每个问题。

For each question, please select the rating that best matches your actual experience with the system.

For example/例如,

For the first question/第一题:

If you found the system to be very obtrusive, choose a score of 1/如果您觉得系统非常碍手碍脚，选择 1 分；

If you found the system to be very supportive, choose a score of 7 如果您认为系统非常能提供辅助，选择 7 分；

In all other cases, choose the appropriate score (2-6) based on your experience/其他情况根据您的感受选择适当的分值（2-6 分）。

1. Obstructive/碍手碍脚的 Supportive/能提供辅助的

2. Complicated/复杂的 Easy/简单的

3. Inefficient/低效的 Efficient/高效的

4. Confusing/令人眼花缭乱的 Clear/一目了然的

5. Boring /乏味的 Exciting /带劲的

6. Not interesting /无趣的 Interesting /有趣的

7. Conventional /常规的 Inventive /独创的

8. Usual /传统的 Leading Edge /新颖的

1 2 3 4 5 6 7

Appendix C: Preference Questionnaire (Experiment 1)

Name: **No.:**

Please rank the following wake-up methods based on your personal preference, with 1 indicating the most preferred and 3 indicating the least preferred. Rank the wake-up methods according to your experience/请根据您的个人体验对不同唤醒方式进行排名，其中 1 表示最偏好，3 表示最不偏好。请根据您的使用体验给以下唤醒方式排序：

- Wake-up word activation/唤醒词唤醒:_____
- Snapping finger activation/打响指唤醒:_____
- Wake-up-free activation/免唤醒:_____

Appendix D: Preference Questionnaire (Experiment 2)

Name: **No.:**

Please rank the following input methods based on your personal preference, with 1 indicating the most preferred and 4 indicating the least preferred. Rank the wake-up methods according to your experience/请根据您的个人体验对不同输入方式进行排名，其中 1 表示最偏好，4 表示最不偏好。请根据您的使用体验给以下唤醒方式排序：

- Multiple voice input/多次语音输入:_____
- Voice + say “stop” /语音+停止的输入方式:_____
- Voice + snapping finger/语音+打响指的输入方式:_____
- Continuous voice/连续性语音输入:_____