Interrupting Drivers for Interactions: Predicting Opportune Moments for In-vehicle Proactive Auditory-verbal Tasks

AUK KIM, KAIST, South Korea
WOOHYEOK CHOI, KAIST, South Korea
JUNGMI PARK, Samsung Research, Samsung Electronics, South Korea
KYEYOON KIM, Hyundai Motor Company, South Korea
UICHIN LEE*, KAIST, South Korea

Auditory-verbal interactions with in-vehicle information systems have become increasingly popular for improving driver safety because they obviate the need for distractive visual-manual operations. This opens up new possibilities for enabling proactive auditory-verbal services where intelligent agents proactively provide contextualized recommendations and interactive decision-making. However, prior studies have warned that such interactions may consume considerable attentional resources, thus negatively affecting driving performance. This work aims to develop a machine learning model that can find opportune moments for the driver to engage in proactive auditory-verbal tasks by using the vehicle and environment sensor data. Given that there is a lack of definition about what constitutes interruptibility for auditory-verbal tasks, we first define interruptible moments by considering multiple dimensions and then iteratively develop the experimental framework through an extensive literature review and four pilot studies. We integrate our framework into OsmAnd, an open-source navigation service, and perform a real-road field study with 29 drivers to collect sensor data and user responses. Our machine learning analysis shows that opportune moments for interruption can be conservatively inferred with an accuracy of 0.74. We discuss how our experimental framework and machine learning models can be used to design intelligent auditory-verbal services in practical deployment contexts.

CCS Concepts: • **Human-centered computing → User interface management systems**; **Ubiquitous and mobile computing**;

Additional Key Words and Phrases: Interruptibility; In-vehicle information system; Human-vehicle interaction; Auditory-verbal interface, Speech-based interaction

ACM Reference Format:
Auk Kim, Woohyeok Choi, Jungmi Park, Kyeyoon Kim, and Uichin Lee. 2018. Interrupting Drivers for Interactions: Predicting Opportune Moments for In-vehicle Proactive Auditory-verbal Tasks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 175 (December 2018), 28 pages. https://doi.org/10.1145/3287053

*This is the corresponding author

Authors' addresses: Auk Kim, KAIST, Daejeon, South Korea, kimauk@kaist.ac.kr; Woohyeok Choi, KAIST, Daejeon, South Korea, woohyeok.choi@kaist.ac.kr; Jungmi Park, Samsung Research, Samsung Electronics, Seoul, South Korea, jungmi.park@samsung.com; Kyeyoon Kim, Hyundai Motor Company, Uiwang, South Korea, kyekim@hyundai.com; Uichin Lee, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, South Korea, uclee@kaist.ac.kr.

2474-9567/2018/12-ART175 $15.00
https://doi.org/10.1145/3287053

1 INTRODUCTION

Auditory-verbal interfaces are becoming increasingly popular for in-vehicle information systems. For example, recent smart vehicles allow drivers to read, reply to, and send text messages via such interfaces [42, 46]. There will be soon be many *proactive services* that understand drivers and their situations. This would facilitate proactive auditory-verbal interactions to deliver personalized services, ranging from information delivery to decision-making [51]. For example, on the way to the office, a vehicle might give an early reminder of an upcoming meeting or help to re-schedule a meeting because of traffic jams [77]. As another example, before arriving at a destination, the vehicle might inform the driver of nearby parking spaces and their expected costs and then ask the driver to make a selection [21]. Although these services have not yet been fully deployed in commercial vehicles, similar concepts have been demonstrated by some of the major manufacturers (Chrysler, Nissan, and Honda) [46]. It is expected that this concept will be re-introduced to drivers in the next few years in the form of more intelligent, advanced, and human-like proactive services as in Google Duplex [85].

The reason for the popularity of auditory-verbal interfaces is that compared to visual-manual interfaces (e.g., visual buttons/knobs), for a given secondary task, they help to balance driving and secondary task performance. This is mainly because interactions with auditory-verbal interfaces do not require drivers to take their eyes off the road, which is the major departure from traditional visual-manual interfaces (e.g., tuning radios) [6]. Even though auditory-verbal interfaces reduce visual-manual distractions, prior studies warn that they can cause potential cognitive distraction. For example, Faure et al. [22] showed that the use of auditory-verbal interfaces can negatively affect driving performance, particularly when a driver is overloaded with driving. Our intuition is that just as human companions know roughly when to interrupt a driver for conversation, we can analyze driving context data to judge whether the driver can safely engage in proactive auditory-verbal interactions.

Finding such an opportune moment has been one of the active research areas in the domains of ubiquitous computing and human-computer interaction. Traditionally, prior studies have considered mostly computing environments with desktop computers or mobile devices, where task-switching is feasible [56, 79]. For example, Bailey and Konstan found that opportune moments for information delivery occur naturally at the end of ongoing tasks in desktop work contexts [5]. However, the body of work cannot be directly applicable to the vehicular contexts, because task-switching is not feasible while driving. In other words, drivers should concurrently execute ongoing driving tasks (e.g., steering, observing, and operating the controls) as well as any interruption task (e.g., comparing parking prices of nearby parking spaces), which represents *dual-tasking* situations.

As a first step toward interruptibility research in driving contexts, Kim et al. assumed that the moments when drivers naturally engage in *user-initiated secondary tasks* of visual-manual operations such as drinking and radio tuning while driving can be considered as opportune moments for interruption [37]. However, this definition of interruptibility has several limitations. Typical auditory-verbal tasks require varying levels of cognitive demands by nature [1, 2, 18], but visual-manual operations may not be primarily cognitive. For example, drinking and eating are largely visual-manual operations, requiring a minimal cognitive load. Furthermore, visual-manual operations while driving are one of the major causes of road accidents [3]. Thus, it is difficult to say that such moments are appropriate for drivers to safely engage in proactive auditory-verbal tasks with various cognitive demands. To the best of our knowledge, none of the prior studies investigated *interruptible moments of proactive auditory-verbal tasks* or system-initiated secondary tasks in a real-world driving setting. Our work is based on real-world driving, because prior studies showed that drivers' behaviors and attitudes (e.g., risk perception) in simulated driving could be very different from their behaviors and attitudes in real-world driving [65].

Despite the popularity of auditory-verbal interfaces and proactive intelligent agents [2, 51], there are still insufficient systematic studies of what constitutes opportune moments in driving contexts and of how they can be measured and predicted for delivering proactive in-vehicle auditory-verbal tasks. In this study, we first defined opportune moments for auditory-verbal interactions, and then iteratively developed an experimental framework through extensive literature reviews and a series of simulation/real-road pilot trials. The key distinction is that our framework considers multiple dimensions, namely driving safety, auditory-verbal performance, and overall perceived difficulty. We then implemented our experimental framework using OsmAnd [70], an open-source navigation app, and deployed the data collection system in a real-world field trial with 29 drivers (IRB Approval No. KH2016-49). Finally, we built machine learning models for predicting opportune moments. The major contributions of this work can be summarized as follows:

- We carefully defined drivers' interruptibility in various interruption contexts, such as multiple levels of secondary task complexity and various levels of driving complexity and proposed the key interruptibility metrics such as driving safety, auditory-verbal performance, and overall perceived difficulty.
- We iteratively developed the experimental interruptibility framework, by performing two simulator studies (n = 8, n = 17) for developing the procedure and deployment method of the auditory-verbal secondary task, and two real-road pilot studies (n = 4, n = 6) for evaluating task triggering strategies.
- Our experimental interruptibility framework was integrated into an open-source navigation app, and we built a comprehensive sensor data collection tool whose data collection ranges from in-vehicle OBD-II/CAN data to mobility and environmental data, including dashcam videos. In our main study, we then collected a total of 3,480 hours of a real-world driving dataset with 29 drivers who engaged in proactive (or system-initiated) auditory-verbal tasks.
- Our data analyses showed the importance of interruptibility metrics on defining interruptible moments in vehicular contexts. We used diverse combinations of interruptibility definition and evaluated various machine learning models by considering varying window sizes, feature selection, and personalization. Our model was able to achieve an accuracy of 0.74 even under the conservative definition of interruptibility (i.e., simultaneously considering all three metrics).

The remainder of this paper is organized as follows. We begin by reviewing the related works in Section 2. We then define what constitutes the interruptibility in Section 3, and present the details of our experimental framework in Section 4. Our data collection in real-road contexts is illustrated in Section 5. Using this dataset, we investigate interruptibility prediction in Section 6. Finally, we discuss our main findings and their limitations in Section 7 and conclude the work in Section 8.

2 BACKGROUND AND RELATED WORK

2.1 Interruptibility and Driving Contexts

In driving environments, drivers must concurrently engage into both ongoing driving task (i.e., primary task) and interrupting task (i.e., secondary task) [66]. Unlike driving environments, the majority of prior studies in the domain of interruption management have typically considered the environments in which people suspend an ongoing task and then switch to an interrupting task (e.g., interruptions in the workplace or daily routine) [4, 11]. In these environments, various measures have been used to estimate interruptibility. Turner et al. [79] suggested three broad categories: 1) the performance of an ongoing (primary) task (e.g., the elapsed time to complete and resume interrupted tasks), 2) a user's perceived receptiveness of interruptions (e.g., a self-reported interruptibility), and 3) physiological reaction reflecting the cognitive workload (e.g., pupil size) [57]. In addition, several studies utilized objective measures to assess whether a user is capable of engaging in interruption tasks, such as the reaction time to interruptions [56]. Advances in pervasive computing technologies facilitate the use of a variety of data from sensors and computing devices for estimating opportune moments, such as physical activities [53], physiological responses [37], and environmental factors [56].

To the best of our knowledge, few studies have investigated interruptibility in a driving context. In their seminal work, Kim et al. investigated user-initiated secondary tasks (e.g., drinking, eating food, smoking, audio tuning, or using portable devices) as candidates for interruptible moments [37]. This approach, however, is limited for the following reasons. First of all, user-initiated secondary tasks are different from typical auditory-verbal tasks; user-initiated secondary tasks are not usually primarily cognitive [2, 18]. For example, eating mainly requires primarily visual-manual operations but it involves a small cognitive load. In addition, user-initiated secondary tasks are often *self-paced*, whereas the pace of user interaction is led by the system in system-initiated secondary tasks (i.e., self-paced vs. system-led interaction) [81]. Second, there is a lack of systematic consideration of driving safety. Unfortunately, many accidents happen due to *user-initiated secondary tasks* while driving. For example, on US roads in 2015, estimated 885,000 vehicle crashes were mainly due to the drivers' user-initiated secondary tasks (e.g., using portable devices or drinking coffee) [3]. Third, their model building did not consider compensatory driving behaviors that happen when drivers engage in the secondary tasks while driving (e.g., reducing speed for coffee drinking or phone calling) [27]. In this case, a data window *prior* to the moment of engaging in the secondary task should be used for model building and testing. Our work complements this prior work in that we explore multiple dimensions for driver interruptibility with a specific emphasis on driving safety for system-initiated (or proactive) auditory-verbal tasks, design an experimental framework with auditory-verbal tasks, and collect a real-road driving dataset where drivers actually perform auditory-verbal tasks for interruptibility classification.

When considering the prior studies that specifically focused on cognitive workload due to auditory-verbal interactions, in their simulator-based study (ConTRE [47]), Rajan et al. [57] introduced a driver workload inference model based on psycho-physiological signals. This showed that pupil dilation is an important measure for mediating in-vehicle auditory-verbal binary notification tasks (e.g., responding in binary answers). Our study complements this work in four areas. First, we used in-vehicle sensor and contextual data instead of physiological signals. Second, we proposed an experimental framework that considers multiple dimensions to systematically define interruptible moments (i.e., driving safety, auditory-verbal performance, and overall perceived difficulty). Third, we implemented our experimental framework into an open-source navigation app and collected a *real-world driving* dataset for interruptibility classification. Our work can be extended by considering physiological signals and other auditory-verbal tasks.

2.2 Driver Interruption, Dual-Tasking, and Driver Distraction

Although insufficient studies have been carried out on interruptibility management in driving environments, abundant studies have examined the consequences of concurrent execution of driving and secondary tasks in the domain of driver distraction. Driver distraction has been defined differently across the studies, but commonly it is indicated as diversion of attention away from driving caused by an attention-induced competing activity [64]. Further, an interruption has been defined as "an externally generated randomly occurring, discrete event that breaks continuity of cognitive focus on a primary task" [12]. From these definitions, we can combine these two concepts into one: interruption tasks would cause driver distraction. According to this combination, a variety of factors on driver distraction should be interpreted in terms of interruption management, as follows.

Dual-tasking and Compensatory Behavior: Driving environments necessarily lead to concurrent execution of primary task (driving) and secondary task (auditory-verbal interactions) [33, 66]. The dual-tasking nature with a limited amount of cognitive resources results in a variety of *compensatory behaviors* such as speed reduction [67]. A domain of driver distraction revealed compensatory behaviors while engaging in the secondary task: (1) changes in driving performance: e.g., reducing speed [58], increasing in inter-vehicular distance [76], and variation in lateral control of vehicle position [54]), and/or (2) stopping engaging in the secondary task [29]. Such a trade-off between performances of primary and secondary tasks implies that interruptibility evaluation for driving contexts should carefully consider both primary and secondary tasks.

Auditory-verbal Interfaces: Regarding sensory channels of the secondary task, sources of driver distraction are separated into visual, manual, and cognitive [61]. Among these sources, prior studies reported that the auditory-verbal interface has lower disruptive effects on driving performance, decreased self-reported workload, and reduced visual demands when compared to traditional visual-manual interfaces [49]. Despite the fact that auditory-verbal interfaces reduce visual-manual distraction, prior studies also warn about the potential cognitive distraction of such interfaces [22].

Measuring the Effects of Driver Distraction: As previously mentioned, interruptibility can be measured via the performance level of the primary task. Prior studies on driver distraction also have utilized several performance measurements that reflect the effects of distraction, including mean speed, lateral position, brake jerks, and steering wheel reversal rate (SRR) [54]. Likewise, it is important to consider what are the impacts of interrupting tasks in the contexts of interruptibility prediction as well. Unlike existing distraction studies [54], interruptibility studies [37, 57] did not systematically examined how interrupting tasks affect driving performance by quantifying the change in deriving performance.

Driving and Secondary Task Demands: Prior studies on interruptibility prediction have considered the environmental context at the moment of interruption [37, 57]. However, these studies have not systematically varied the cognitive demand of interrupting tasks. Traditionally, existing driver distraction studies have considered not only the demand of primary driving (or driving contexts), but also the demand of a secondary task (or an interrupting task). This is due to the fact that a driver's (finite) attentional resources are shared for both tasks. Thus, the effect of the secondary task on driving performance is highly dependent on the aggregated workload of dual-tasking [22, 65]. These prior studies highlight that for interruptibility judgment in driving contexts, the secondary tasks should be able to consider varying levels of cognitive demand, and the tasks should be performed in various driving contexts. In our experimental framework, we incorporated a well-known auditory-verbal task with varying cognitive demands and considered various real-road driving contexts.

Labeling Interruptible Moments: Prior studies on interruptibility prediction have often utilized users' perceived interruptibility to label highly-interruptible moments [79]. Likewise, we can consider how users think about dual-tasking when judging interruptibility. However, we should take this approach with care. This subjective measure may not truly reflect interruptibility in driving contexts because drivers tend to overestimates their driving capability [83]. As discussed earlier, it is important to guarantee that dual-tasking does not cause any safety problems, such that driving performance can be considered as one important measure for predicting interruptibility.

3 DEFINING DRIVER INTERRUPTIBILITY FOR PROACTIVE AUDITORY-VERBAL TASKS

Our goal is to develop an automatic driver interruptibility classifier that infers interruptible moments for a driver to engage in proactive in-vehicle services via auditory-verbal interfaces. To start with, we define what constitutes interruptible moments for in-vehicle auditory-verbal interactions. In driving, we consider the unique constraint of dual-tasking, which is very different from traditional work environments in which task-switching is feasible [5]. In other words, drivers must concurrently execute both driving and the auditory-verbal tasks [33]. Drivers are required to share their limited capacity of cognitive resources for both tasks [22, 65], leading to a performance trade-off between the primary task of driving, and the secondary task of auditory-verbal interaction [67]. This contextual uniqueness with respect to prior interruptibility studies implies that driving safety must be considered first.

In this work, we propose using following dimensions to define interruptible moments: (1) *driving safety*: auditory-verbal engagement should not negatively affect driving performance for safety reasons, (2) *auditory-verbal task performance*: the driver should be able to successfully perform auditory-verbal tasks, and (3) *overall perceived difficulty*: the driver should not feel a considerable burden when performing a dual-task.

Driving safety is critically important for in-vehicle information services, where unfavorable effects on driving performance can negatively affect the safety of drivers, passengers, other road drivers, and pedestrians. Owing to the capacity limit of cognitive resources, driving performance may decrease (e.g., paying less attention to tracking the vehicle's position) to allocate more of their cognitive resources to secondary tasks [27, 58]. Prior studies on driver distraction examined the negative effects of the concurrent execution on driving performance [73]. Beyond dual-task interference, another critical factor is that people tend to overestimate their driving capabilities, which encourages them to engage in complex secondary tasks while driving [83]. Therefore, for safety reasons, the concurrent execution of driving and auditory-verbal tasks must not negatively affect driving performance.

Auditory-verbal task performance is important from a practical point of view, as the driver should be able to successfully finish an auditory-verbal task. Prior interruptibility studies did not systematically consider the performance of *interrupting* tasks. When evaluating auditory-verbal task performance, we can consider various metrics, such as response delay and engagement time of the task, as proxy measures of task performance [57]. Traditionally, earlier studies in the domain of driver distraction have used *auditory-verbal pseudo tasks*, such as a delayed digit recall task [50] and an auditory operation span task [73].

Overall perceived difficulty is particularly important, because it indicates the overall perceived difficulty in performing a dual-task, which may indicate the willingness to engage in interrupting tasks. Prior literature states that the perceived level of difficulty of a technology is negatively correlated with its perceived level of usefulness [23]. Furthermore, the perceived level of difficulty of a technology is negatively correlated with how often the technology is used [23]. These findings imply that a higher level of overall perceived difficulty during proactive in-vehicle services will reinforce unfavorable attitudes toward those services.

It is possible to use a combination of these measures both disjunctively and conjunctively depending on the purpose of the auditory-verbal interruptions. Driving safety must be included as this is always the top priority for any in-vehicle system. Driving safety alone could be appropriate for cases of delivering simple-context information, such as navigation information and weather forecasts. The conjunctive form of driving safety and auditory-verbal performance could be appropriate for cases of delivering critical information that requires a driver response, such as informing about a low fuel level and asking for permission to reroute to a nearby gas station. Although it is very conservative, the conjunctive form of the three measures (i.e., driving safety, auditory-verbal performance, and overall perceived difficulty) could be used for auditory-verbal tasks that involve interactive decision-making (e.g., comparing trade-offs between cost and convenience for choosing parking spots).

A summary of the dimensions for interruptible moments are given below. The detailed metrics for measuring each dimension are illustrated in the following section.

- ***Driving safety*** measures how safely a user drives a vehicle (e.g., steering wheel control performance).
- ***Auditory-verbal performance*** measures how well a user performs an auditory-verbal task (e.g., auditory perception and cognitive processing involving working memory).
- ***Overall perceived difficulty*** measures how difficult it is to perform a dual task (e.g., Likert scale rating of difficulty level).

4 EXPERIMENTAL FRAMEWORK FOR DRIVER INTERRUPTIBILITY STUDIES

In this section, we describe our experimental framework to systematically collect an interruptibility dataset. Our framework consists of three parts: 1) *secondary-task procedure* for emulating proactive in-vehicle auditory-verbal interactions with variable levels of cognitive demand, 2) *task-triggering method* for triggering secondary tasks in diverse driving contexts, and 3) *key metrics* to evaluate a driver's interruptibility in terms of driving safety, auditory-verbal performance, and overall perceived difficulty. We iteratively develop our experimental framework by performing two simulator studies (n = 8, n = 17) and two real-road pilot studies (n = 4, n = 6) for developing the procedure and implementation of the auditory-verbal secondary task, and for evaluating various task triggering strategies. In this section, we first describe each part of the framework, including notable findings in its development, followed by a brief summary of the pilot studies used for the framework development.

4.1 Secondary Task Selection and Procedure

4.1.1 Secondary Task Selection. When selecting secondary tasks for interruptibility research, we consider the following aspects of proactive auditory-verbal tasks, namely variable cognitive demand, measurable performance, and variable task duration. In dual-tasking contexts, since a driver's attentional resource is limited, concurrent execution of primary and secondary tasks causes interferences, thereby negatively influencing task performances. As in prior distraction research [27, 58, 67], the experimental framework should consider variable cognitive (or attentional) demands in both primary/secondary tasks, and the performances of both driving and secondary tasks should be measurable to understand the impact of dual-task interferences. Furthermore, since auditory-verbal operations tended to take longer time than traditional visual-manual operations [1, 34], the secondary tasks should be flexible enough to support a reasonable length of interaction.

In this work, we consider a secondary task that can systematically induce varying levels of cognitive demand and that support performance measurement and variable task length. As shown in Figure 1, we chose *n*-back tests [50], in which users are asked to verbally recall *n*-back digits repeatedly (that the system speaks). The levels of cognitive demand are varied by adjusting the number of delayed digits, and the length of interactions. In practice, *n*-back tests have been widely utilized as an auditory-verbal secondary task since 1993 [88] in numerous real-road driving studies [50], including ISO-associated standards research [8], and prior works carried out by the U.S. National Highway and Transportation Safety Administration [60]. We chose *n*-back tests as an auditory-verbal secondary task for the following reasons:

- *Similarity of cognitive engagement*: This task has a similar cognitive engagement (e.g., auditory attention and memory components) as when a driver engages in an externally paced verbal task such as interacting with the auditory-verbal interface of in-vehicle devices or answering a cell phone [50].
- *Consistently structured demand*: This task induces consistently structured levels of cognitive demand, because we can vary the number of delayed digits and the length of interactions [50].
- *Reference task of various auditory-verbal interactions*: This task has been widely used as a reference task for studying auditory-verbal interfaces in commercial vehicles [48, 50]. We can design an *n*-back task that matches the cognitive demand of a real-world auditory-verbal task. Likewise, we can use this task as a reference task for investigating the interruptibility of proactive auditory-verbal interactions with various cognitive demands.

There are other widely used pseudo cognitive tasks called operation span (OSPAN) tasks [73–75, 82] in which drivers are asked, for example, to solve several math problems and to recall unrelated words appeared in

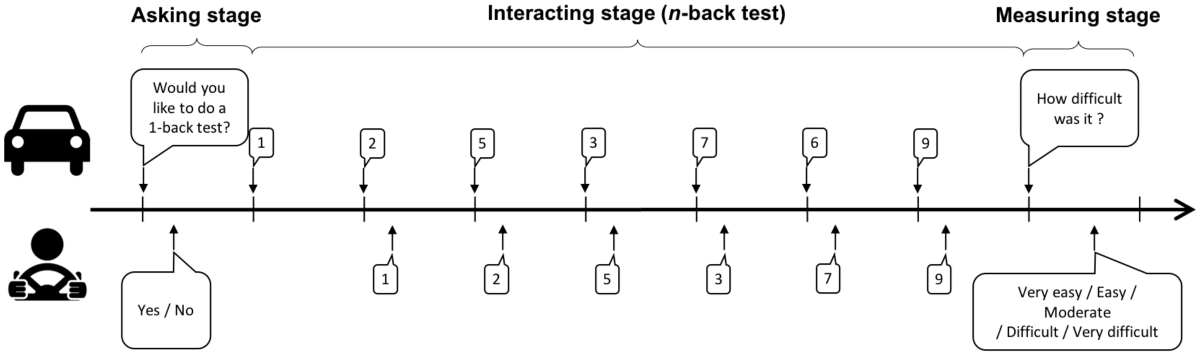
Asking stage
Interacting stage (n-back test)
Measuring stage
Would you like to do a 1-back test?
1
2
5
3
7
6
9
How difficult was it ?
Yes / No
1
2
5
3
7
9
Very easy / Easy / Moderate / Difficult / Very difficult

Fig. 1. Procedure of a secondary task (*N*-back test type = 1-back test).

between math problems. OSPAN tasks are somewhat similar to n-back tasks, but controlling cognitive demands is cumbersome with OSPAN tasks. In voice interface research, various voice interaction scenarios were used for evaluation such as phone usage, music selection, navigation control, and email manipulation [48, 49, 73]. In addition, prior research used simple factoid Q&A tasks (e.g., true/false answering of factual questions) [57] or even topic-based free conversations [41]. Although these kinds of scenario-based tasks are more realistic than n-back tasks, designing representative task sets for proactive auditory-verbal interaction is difficult due to its diverse service domains, and developing *standardized scenarios with variable cognitive demands* is very challenging as that requires considerable time and resources for assuring consistency and reliability, as well as safety validation in real-world driving. As a first step towards interruptibility research, we leverage a body of prior studies on driver distraction where n-back tasks have been widely used to investigate dual-tasking scenarios.

4.1.2 Secondary Task Procedure. The entire procedure of our task consisted of three stages: 1) asking stage, 2) interacting stage (n-back test), and 3) measuring stage.

Asking stage: The asking stage is when a driver is asked whether he or she is willing to engage in the remaining stages after hearing which n-back type will be given in the following stage. The remaining stages are only presented if the driver answers "Yes" within 2.5 seconds; otherwise, the current task is immediately stopped. We introduced the asking step to 1) check whether a user fails to notice a secondary task, possibly due to vehicle background noise, and 2) ensure that any interactions are performed under safe conditions.

Interacting stage: Next, the driver performs the n-back test. We adopted the n-back protocol previously [50]. As shown in Figure 1, the n-back test sequentially presents seven randomly-selected numbers from 0 to 9 to the driver at 2.25 second intervals. The driver is required to repeat back the single digits by following one of three tasks: 0-back (a very mild task demand), 1-back (a moderate level), or 2-back (a high level of task demand).

Measuring stage: The procedure finishes with the measuring stage, in which the driver is asked to verbally indicate the overall difficulty during the previous stage (interacting stage). The driver is asked to indicate the difficulty level using natural language: "very easy, easy, moderate, difficult, very difficult." Our approach of asking the overall task demand immediately after completing the task does not notably interfere with the primary task performance [15].

Note that the statistical analysis in our simulator pilot study 2 showed that 1) this verbal rating measured the overall perceived workload reasonably well (correlation with NASA-RTLX [30], $r = 0.72, p < 0.01$) and 2) not surprisingly, our secondary task induced varying cognitive demand across the tasks with different n-back types, $F(1.425, 17.103) = 9.162, p < 0.001$. We initially used a numerical rating for difficulty rating, but the participants expressed considerable difficulties in making such rating, by often commenting that rating was cognitively demanding. After several iterations, we found that our participants felt more comfortable with verbally expressing a difficulty level, as shown earlier.

In our framework, we decided to use the Wizard of Oz approach to collect and code the driver responses, owing to the several practical issues that we found in our simulation study 1. We originally considered two other approaches: 1) automatic speech recognition, and 2) Bluetooth button attachment at a driver-preferred place on a steering wheel (see Figure 2). We found that the speech recognition provided by the Android operating system was quite erroneous, because there was considerable vehicle background noise, and the drivers' response words (e.g., Yes, No) were fairly short [9]. In the case of the Bluetooth button use, the drivers reported considerable inconvenience because certain maneuvers require frequent steering movements. This inconvenience also resulted in a response delay, because users often pressed the button after a full maneuver finished.

4.2 Task Triggering Method

To trigger the secondary tasks in various driving contexts, we designed our method that presents the tasks based on a hybrid approach of two triggering methods, i.e., random-interval triggering and location-based triggering.

This was due to the weaknesses found in the iterative development when using a single method. In the following, we present the details of two triggering methods and their weaknesses, as well as the hybrid approach.

Random-interval triggering method presents tasks at random intervals between 30 and 90 seconds. The interval was defined after several trials in our iterative development stages (simulator pilot study 1). Random-interval triggering should be used with special care. Our pilot study results (simulator pilot study 2 and real-road pilot study 1) showed that the number of tasks triggered when the vehicle was running on straight roads was greater than that when the vehicle was cornering—in practice, most roads have more straight parts than curves.

Location-based triggering method presents tasks at specific locations associated with various driving environmental factors (e.g., various types of driving maneuvers). These factors have been highlighted as important factors for determining the driving environment complexity and its effects on driving task demands [22, 65]. Note that when employing only location-based triggering, we found that in our iterative development stage (real-road pilot study 2), the distribution of *n*-back tasks did not reflect realistic road situations, such as traffic jams. In this case, if the distance between two consecutive locations is too great, tasks are not triggered in between. *Hybrid method* was developed to complement the weaknesses of above methods as follows. For a given round-trip driving course, we carefully choose a number of predetermined locations for location-based triggering (e.g., 40 locations with 20 per direction), and then, random-interval triggering is used to present additional tasks if the distance to the next predetermined location for location-based triggering is sufficiently long. For example, assuming a vehicle speed of 40 km/h and a baseline random-interval of 30 seconds, the minimum distance is set as 334 m. In our iterative development stage (real-road pilot study 2), we confirmed that our hybrid triggering approach presented an approximately equal number of cases for three *n*-back tasks in various driving environments. In our main field trial, our data collection status showed a similar result (see Table 2).

4.3 Metrics for Measuring Drivers' Interruptibility

For auditory-verbal engagement, we consider three metrics to evaluate driver interruptibility: driving safety, auditory-verbal performance, and overall perceived difficulty.

Driving safety was measured based on the difference between driving performances when concurrently executing driving and non-driving auditory-verbal tasks, and when only executing the driving task. As a measure for driving performance, we use steering wheel angles that can quantify effects of secondary tasks on a driver's lateral control behavior [54]. In particular, we use steering wheel reversal rate (SRR) that counts the frequency of a steering wheel reversal event (or wheel rotation) whose angle difference between start and end points exceeds a certain minimum angle (or the gap size). In general, driving performance is inverse proportional to SRR, and conducting secondary tasks will increase SRR, thereby lowering driving performance. Interestingly, prior studies showed that reversal angles differ according to the secondary task types: cognitive tasks mainly induced very small reversals (less than 1 degree), while most of the reversals in the visual task condition were in the range of 2-6 degrees [39, 54]. This means that we can use a small gap size to optimize SRR's sensitivity to cognitive auditory-verbal tasks. Another important characteristic of SRR is that its value tends to vary according to the road shapes. Given that road topologies are readily available, we can segment roads based on road shapes (e.g., turn vs. straight) and separately calculate SRR values for each road segment. Having said that, we define a moment as interruptible if the SRR during the *n*-back test period at a given road segment is less than or equal to the baseline SRR of its corresponding road shape. The baseline SRR value can be measured by introducing a baseline driving session in which a driver does not perform any secondary tasks.

Auditory-verbal performance was measured based on the accuracy of the *n*-back test, which represents the performance achieved for information processing and presentation for the auditory-verbal interaction. The accuracy was given by the ratio of the number of correctly answered to the total number of items that the driver is required to answer. In this study, we adopted a conservative approach: auditory-verbal interaction was deemed interruptible if a driver correctly answered all the items in the *n*-back test (100% accuracy).

Overall perceived difficulty was used to measure the overall demand for conducting a dual-task of driving and auditory-verbal interaction. We used a five-point verbal Likert scale (very easy: 1 – very difficult: 5). This rating task was administered after the *n*-back test. We determined that auditory-verbal interactions were interruptible if the normalized value of the overall perceived difficulty was less than or equal to 0.5. We normalized the value of the perceived difficulty within a range of values for each driver, as each driver had a different range of values. For example, consider that a user makes several recall mistakes in a 2-back test; one driver may rate this instance as "very difficult," whereas others may rate it as "difficult." For feature scaling, we used the min-max normalization [69] that rescales the range of the values to [0, 1].

4.4 Simulator and Pilot Studies for Framework Development

We iteratively developed our experimental framework via two simulator studies and two real-road pilot studies, with a total of 35 participants (average age = 33.8, SD = 12.4, 8 female). For all studies, the *n*-back tests were consistently employed as an auditory-verbal task, whereas other settings for the experimental framework, such as the secondary task procedure and task triggering method, varied across the studies.

The simulator-based studies (study 1: n = 8, study 2: n = 17) were mainly conducted to iteratively develop the procedure and deployment approach of the secondary task. As shown in Figure 2, we used the Euro Truck Simulator 2 [72] in a custom mode to simulate sedan driving (instead of truck driving) [20], with a vehicle cockpit module comprising a steering wheel, shifting lever for an automatic transmission, brake, and accelerator. We customized the vehicle parameters to provide a realistic driving experience, similar to popular midsize sedans in Korea. For collecting a driver's responses, we initially used Android's automatic speech recognition (study 1) and then used a Bluetooth button attached at the steering wheel (study 2). After two trials, we faced several critical problems (e.g., high errors in speech recognition and inconvenience of button usage), and thus, the Wizard of Oz approach was finally adopted. In study 2, we asked drivers to measure subjective cognitive load using NASA-RTLX [30] across each case of their dual-tasks (driving and secondary tasks), while they review the recorded scenes of their driving. In addition, we discovered that drivers felt more comfortable with verbally expressing difficulty levels in corresponding words than that in numerical rating.

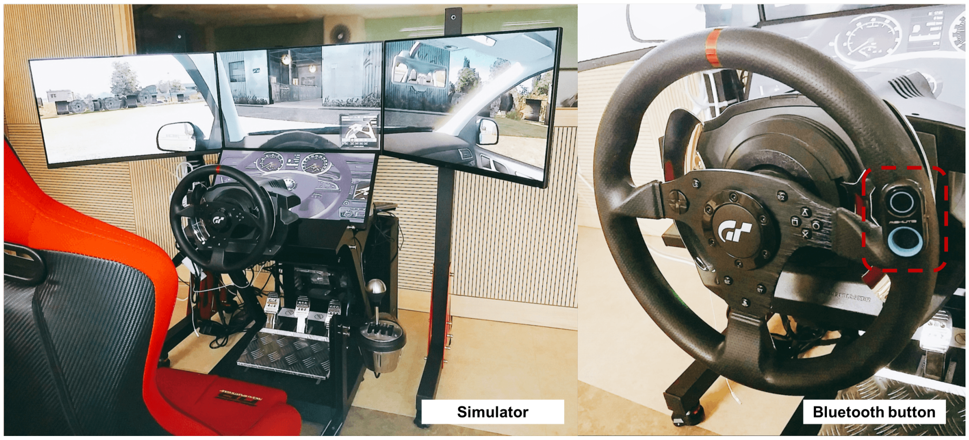

Fig. 2. Euro Truck Simulator 2 [72] with a custom cockpit simulating sedan driving, and a Bluetooth button for driver input.

The real-road pilot studies (study 1: n = 4, study 2: n = 6) were primarily conducted to iteratively develop the task triggering method. We initially studied how to configure random-intervals in the simulator pilot study 1. In real-road pilot study 1, we tested a random-interval triggering method by recruiting four participants and asking them to drive their own vehicles. We discovered that random-interval triggering tended to result in an imbalanced dataset with the events happened at the straight roads. For real-road pilot study 2, we tested location-based triggering and hybrid triggering methods by recruiting six drivers and asking them to drive our test vehicle on one of two chosen round-trip driving routes, including the driving course used in our main field study. We found that location-based triggering may fail to consider realistic road situations, such as a traffic jam, if the distance between two consecutive locations is too far. We addressed these issues by devising a hybrid approach that considers both random-interval and location-based triggering method.

5 COLLECTING INTERRUPTIBILITY DATA IN REAL-ROAD DRIVING ENVIRONMENTS

We used our experimental framework to collect the real-road driving dataset. In this section, we describe our participants, data collection procedure, driving routes, data types, and data collection results.

5.1 Participants and Procedure

We recruited 29 drivers who have a valid driving license, at least one year of driving experience, and who drive at least 30 minutes per day. All of our drivers drove a similar sized car to the one that we used for our study, which was a Kia K3, a popular midsize sedan in Korea. The average age of our participants was 43.2 (SD = 11.4); their ages ranged from 19 to 64 (20s = 3, 30s = 9, 40s = 6, 50s = 7, 60s = 4). Fourteen drivers were male and 15 drivers were female. The intention of the broad age range was to improve the repeatability of our dataset [59]. Drivers received a compensation of approximately 50 USD. The experiment took approximately four and a half hours to complete, of which 50 minutes were for breaks (10 minutes x 2 + 30 minutes) and two hours were for driving (30 minutes x 4). Each of the drivers was assigned either one of two chosen time slots (9:00 – 13:30 or 13:30 – 18:00).

Prior to our study, drivers were asked to read and sign a consent form, as specified in the IRB of our institute (No. KH2016-49). Each driver performed two driving sessions (baseline session and secondary-task session), followed by an interview session. The order of driving-session conditions was counter-balanced. The baseline session involved no secondary tasks and was used as a baseline to compare driving performances using SRR rates. The secondary-task session involved drivers performing secondary tasks while driving. Before each session, a campus drive was provided to familiarize drivers with the vehicle. Afterwards, the drivers drove a round-trip course (the two routes shown in Figure 3).

For the secondary-task session, we provided a training session of *n*-back tasks that involved an instruction and trials for each *n*-back test, and a situation when the test type was randomly triggered. We used in-car audio systems as the main speakers. The open-source navigation app (OsmAnd v2.6 [70]) was modified to guide the next directions at least 10 seconds before or after secondary tasks to avoid interruption of the tasks. During two driving sessions, one researcher was seated in the rear seat of the vehicle (see 4) for Wizard of Oz, participant-observation, and ensuring safe vehicle operation and experimental instrument checking. After two driving sessions, drivers conducted a survey, followed by an exit interview with the authors.

5.2 Driving Routes

Figure 3 shows Route A of the round-trip driving course employed in our main study. The return route (Route B) is slightly different from Route A due to traffic rules. The driving course was designed to include various driving conditions, various types of driving maneuvers, differing levels of traffic density, and other potential factors such as pedestrian traffic and school zones. These conditions have been highlighted as important factors that determine driving complexity and affect driving task demand [22, 65]. The light-orange flags show the 20

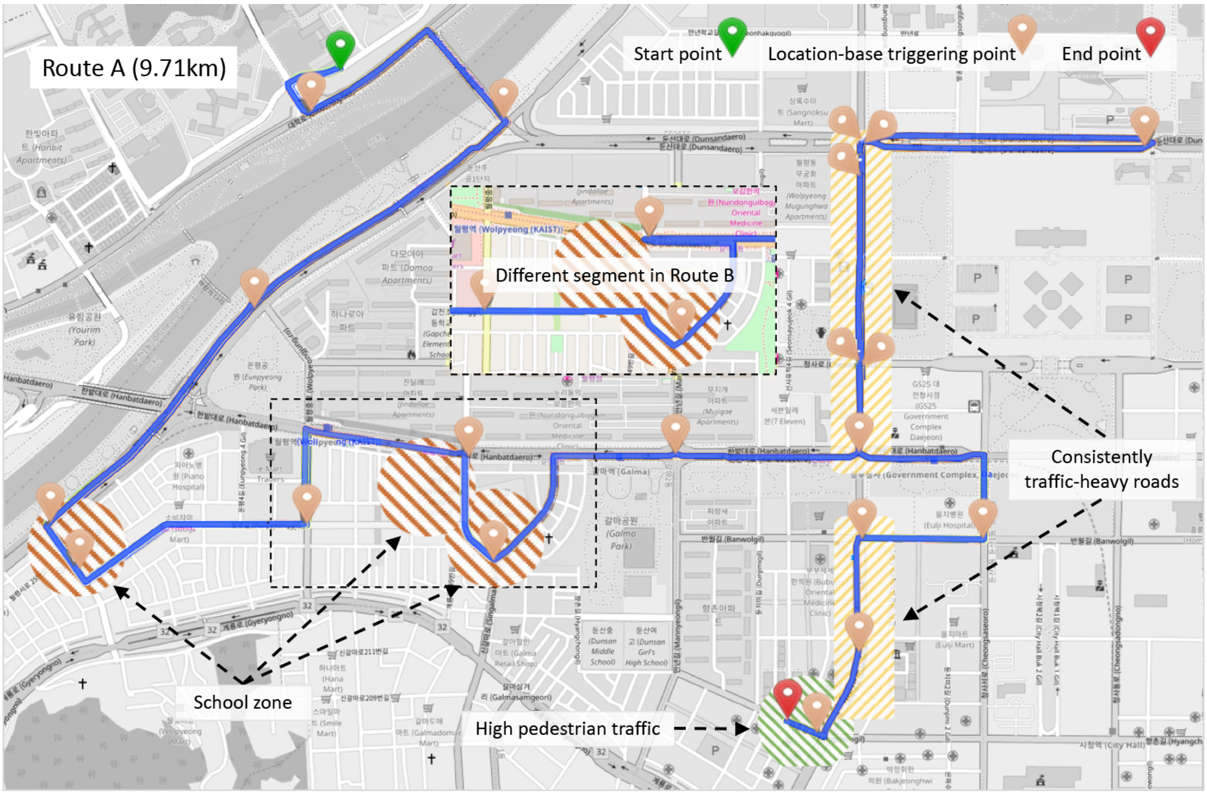
Route A (9.71km)
Start point
Location-base triggering point
End point
Different segment in Route B
Consistently traffic-heavy roads
School zone
High pedestrian traffic

Fig. 3. Route A in the round-trip driving course. Route B (returning route) is slightly different from Route A (see the dotted box for a different route).

predetermined locations where our secondary tasks were presented via a location-based task triggering method, to covers the different conditions; the total number of location-based tasks was 40 with round-trip driving.

The green flag signifies the starting point of each route, and the red flag indicates the end points. Route A runs from the campus to a department store on the main street of downtown and Route B returns is the return journey. Note that Route B is slightly different to Route A, due to due to differences in traffic rules (see the dotted box in the figure). Therefore, the length of Route A (9.71 km) is slightly longer than Route B (9.21 km); however, both routes require approximately 30 minutes of driving time. To reflect various types of maneuvering, both routes have at least four left turns, right turns, and lane changes, while Route A has one U-turn and Route B has two U-turns.

In addition, the number of lanes on the roads varied from one to four. To include various traffic density levels, drivers drove consistently *traffic-heavy roads* (3.82 km, yellow zone). For potential considerable factors, we included roads with *high pedestrian traffic* (0.5 km, green zone) and *three school zones* (1 km, red zone). The result of our exit-survey showed that approximately 60% of the routes were well-known roads, and the remaining roads were not familiar to the drivers.

5.3 Data Collection Equipment and Collected Data Types

As shown in Figure 4, we equipped a test vehicle (Kia K3) with a CAN-bus logger, a smartphone, and four dashcams (in-vehicle, front, and left and right sides) (Transcend's DrivePro 200 [78]). A CAN-bus logger was

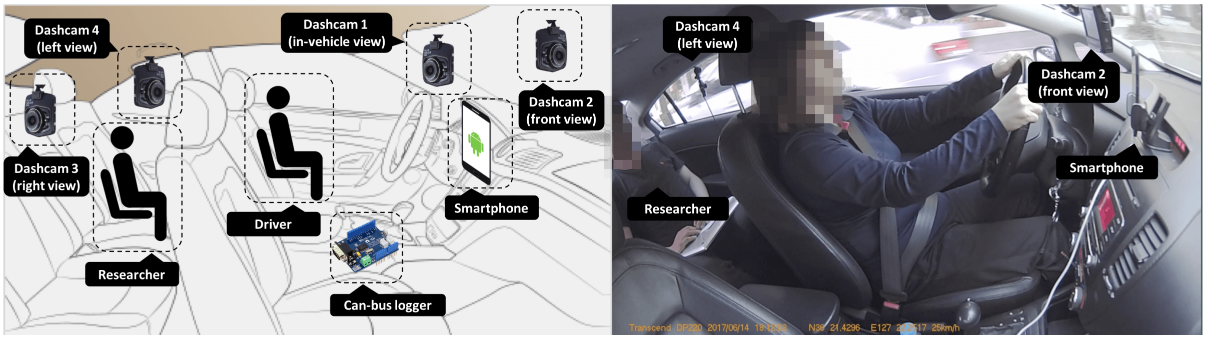

Fig. 4. Equipment setting and driving scene.

Table 1. List of collected data. Lt = left turn, Rt = right turn, HZ = hazard warning (situation where both lights are on), Off = none, Ut = U-turn, Str. = straight, L.Ln = lane change to left, R.Ln = lane change to right, Stp. = stop; Overall per. diff. = Overall perceived difficulty; DBQ = Driving Behavior Questionnaire; DSI = Driving Skill Inventory.

Category	**Source**	**Name**	**Type**	**Category**	**Source**	**Name**	**Type**
Vehicle	OBD	Rel. throttle position	Percentage	Driver	Survey	Age	Continuous
		Abs. throttle position	Percentage			Gender	{Male, Female}
		Engine road	Percentage			Driving experi.	Continuous
		Accel. pedal position	Percentage			DBQ	5-Likert scale
		RPM	Continuous			DBI	5-Likert scale
		Speed	Continuous				
	CAN	Steering wheel angle	-450 ~ 450	Task	APP	*n*-back type	{0, 1, 2}-back
		Turn indicator	Lt / Rt / Hz. / Off			Engagement in test	Binary
		Brake pressure	Percentage			Overall per. diff.	5-Likert scale
Driving Env.	Dashcam	Type of maneuver	Lt / Rt / Ut / Str.			Test accuracy	Percentage
			L.Ln / R.Ln / Stp			Time of each stage	millisecond
		Number of vehicles	Continuous			Location	Lat / Lon
		Dist. to adjacent vehicle	Continuous		Dashcam	In-vehicle view	Video

custom-built by using Arduino and MCP2515/MCP2551 (CAN transceiver/controller) (see [68] for the details). As shown in Table 1, we collected four categories of data: 1) vehicle-status information, 2) driving-environment information, 3) driver-related information, and 4) task-related information. For vehicle-status information, we collected two types of data that varied by source: 1) standard OBD-II data, and 2) reverse engineered controller area network (CAN) data [14, 32].

Driving-environment information included the type of maneuvering, the distance to adjacent cars, and the number of cars in the front, left lane, and right lane relative to the vehicle. These data were manually labeled by reviewing the recorded videos of three dashcams (front, left/right side). The type of maneuvering was labeled based on vehicle movement and road type by reviewing the recorded videos. We considered the major type of maneuvering: straight, left/right turn, lane change, and stop. Next, similar to a previous study [37], the distance to adjacent cars and the number of cars in the front and either side of the vehicle was manually labeled to capture the levels of traffic density. For reasonably accurate distance estimation, using our test vehicle, we conducted a real-world calibration measurement with dashcam video recording, by increasing the distance to a target vehicle in the front, left lane, and right lane by one meter up to 40 meters. This calibration allows us to place meter-level markers on the recorded videos. These markers were then used to label the distance to adjacent cars in the front and either side of the vehicle.

For driver information, we collected demographic information and self-reported driving skills and habits, using Driving Behavior Questionnaire [63] and Driving Skill Inventory [40]. Finally, for task-related information, we collected the following data for each task: *n*-back type, engagement in test, self-reported overall difficulty level, accuracy of the driver answer for the given test, and time and GPS location for each stage in the task.

5.4 Data Collection Status

The drivers drove an average of 119.99 minutes (SD = 15.07). For the natural-driving session, journey time was an average of 59.91 minutes (SD = 7.70). Meanwhile, for the secondary-task driving, journey time was an average of 60.08 minutes (SD = 10.20). Among the 29 drivers' data (n = 1,413 cases), we excluded one driver's data (n = 25 cases) for Route A data due to a recording problem. Thus, we used 28 sets of data for Route A, and 29 sets of data for Route B. With these data, we successfully collected 1,388 task cases from the 29 drivers without any further loss. There were only two cases in which drivers denied the secondary tasks by saying 'no' in the asking stage (which happened to belong to the excluded data). In contrast, for all 1,388 cases, the drivers performed the secondary tasks by indicating 'yes' in the asking stage of the task. Note that we discuss the implications of such biased responses in Section 7.2.

To analyze the distribution of collected cases, we arranged the task cases across types of *n*-back test and maneuvers. Our results show that our drivers experienced approximately an equal number of cases for each type of *n*-back test under the various driving environments. As shown in Table 2, the average number of cases for each driver was 47.86 (SD = 6.83, range = 25–63). After excluding the 25 cases for the driver with the recording issues, the minimum number of cases was 41. Our drivers performed at least one case for different *n*-back types across each type of maneuver, and similar number of cases for each type of *n*-back test.

For the maneuver type, the majority of cases involved *straight* type. While examining the distribution of data across maneuver types, we found that each case often involved more than one type of maneuvers. The most frequently appearing types were *straight* and *stop*. Among the total cases, straight and stop accounted for 93% and 80%, respectively. *Turn* and *lane_change* type accounted for 10% and 24%, respectively. When considering the average number of type for each driver, 44.86 and 38.34 cases were for *straight* and *stop*, respectively, while 11.34 and 4.66 cases were for *turn* and *lane_change*, respectively.

Table 2. Average frequency of task cases across *n*-back type and maneuver types for a driver.

Mean (SD)		**Frequency**			
		Overall	***n*-back type**		
			0-back	**1-back**	**2-back**
Overall		47.86 (6.83)	15.52 (4.38)	16.10 (3.80)	16.24 (3.51)
Maneuver type	**Straight**	44.48 (7.00)	14.14 (4.30)	15.28 (3.92)	15.07 (3.63)
	Turn	11.34 (3.10)	4.03 (1.96)	3.55 (1.83)	3.76 (2.18)
	Lane_change	4.66 (2.31)	1.76 (1.48)	1.52 (1.10)	1.38 (1.22)
	Stop	38.34 (6.29)	12.45 (4.17)	12.97 (3.75)	12.93 (3.11)

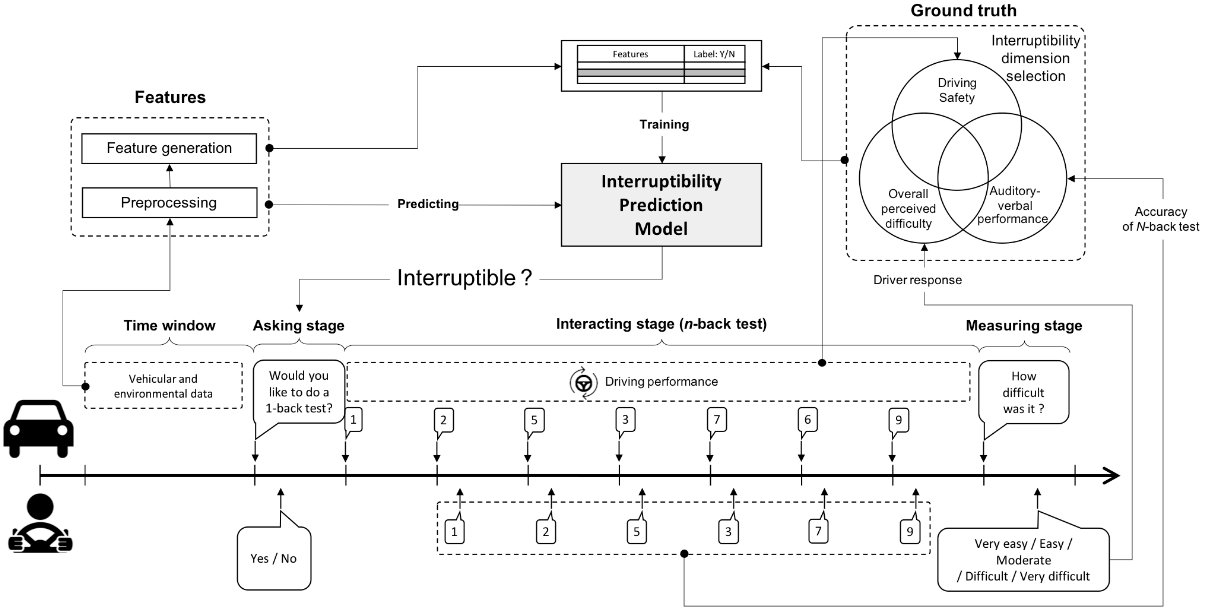
Ground truth
Interruptibility dimension selection
Driving Safety
Overall perceived difficulty
Auditory-verbal performance
Accuracy of N-back test
Driver response
Features
Feature generation
Preprocessing
Features
Label: Y/N
Training
Predicting
Interruptibility Prediction Model
Interruptible ?
Time window
Asking stage
Interacting stage (n-back test)
Measuring stage
Vehicular and environmental data
Would you like to do a 1-back test?
Driving performance
How difficult was it ?
1
2
5
3
7
6
9
Yes / No
Very easy / Easy / Moderate / Difficult / Very difficult

Fig. 5. The visual description of training an interruptibility prediction model for proactive in-vehicle auditory-verbal services. Vehicular and driving environmental data prior to a driver's interaction are used to generate features. Interruptibility labeling is made by considering driving safety (SRR), auditory-verbal performance, and overall perceived difficulty.

6 INTERRUPTIBILITY ANALYSIS AND PREDICTION

We now analyze our measured data to study how different criteria affect interruptibility labeling, and how interruptible moments are related to task difficulty and maneuver type. As illustrated in Figure 5, we then build a machine learning model to predict opportune moments for proactive in-vehicle auditory-verbal services.

6.1 Interruptibility Labeling and Analyses

6.1.1 Interruptibility Labeling Methods. As illustrated in Section 3, our experimental framework of interruptibility considers multiple dimensions, namely driving safety, auditory-verbal performance, and overall difficulty. We labeled the interruptibility of each secondary task as a binary outcome (i.e., interruptible or uninterruptible). If multiple dimensions are used as interruptibility criteria, we labelled a secondary task as interruptible if all the considered dimensions are interruptible. We summarized each dimension and its metric as follows (see Section 4.3 for the details):

- ***Driving safety***: Interruptible if the SRR value during the *n*-back test period at a given road segment was less than or equal to the baseline SRR of its corresponding road shape. The baseline SRR of each driver was separately calculated for each road shape (e.g., turn, straight). The gap size was set to 0.1 degree [54] and was identical for all drivers.
- ***Auditory-verbal performance***: Interruptible if a driver correctly answered all the items in a given *n*-back test.
- ***Overall perceived difficulty***: Interruptible if the normalized value of the overall perceived difficulty was less than or equal to 0.5. Here, the min-max normalization was used [69].

Table 3. Frequency of interruptible and uninterruptible moments.

n = 1388	**Interruptible moments**	**Uninterruptible moments**
Driving safety (safety)	1241	147
Auditory-verbal performance (secondary)	1190	198
Overall perceived difficulty (difficulty)	1119	269
Safety + secondary + difficulty	939	449

Table 4. Percentage of interruptible moments across *n*-back type and maneuver types for a driver.

Mean (SD)		**Percentage of interruptible moments**			
		Overall	***n*-back type**		
			0-back	**1-back**	**2-back**
Overall		0.68 (0.18)	0.88 (0.10)	0.67 (0.25)	0.51 (0.28)
Maneuver type	**Straight**	0.68 (0.17)	0.88 (0.10)	0.67 (0.24)	0.51 (0.28)
	Turn	0.62 (0.22)	0.82 (0.20)	0.61 (0.36)	0.41 (0.36)
	Lane_change	0.60 (0.28)	0.76 (0.36)	0.67 (0.43)	0.25 (0.40)
	Stop	0.68 (0.19)	0.88 (0.11)	0.65 (0.25)	0.52 (0.32)

6.1.2 Interruptible Moment Analyses. We arranged 1,388 cases after judging interruptibility in each dimension. Table 3 shows that considering only driving safety yielded 1,241 interruptible cases (89%), considering only auditory-verbal performance yielded 1,190 interruptible cases (86%), and considering only overall perceived difficulty yielded 1,119 interruptible cases (81%). If all the dimensions are simultaneously considered, we have 939 interruptible cases (68%).

We analyzed the interruptible moments across the types of *n*-back tests and driving maneuvers using the most conservative interruptibility definition of considering all three dimensions (see Table 4). In this case, we showed earlier that only 68% (939 of out of 1,388) were interruptible. The results showed that as the difficulty of *n*-back tests increases, the fraction of interruptible moments considerably decreases (from 88% in 0-back tests to 51% in 2-back tests). When comparing interruptibility across different maneuver types, the cases with a maneuver type of *straight* or *stop* were slightly more interruptible than those with a maneuver type of *turn* or *lane_change*. In particular, 2-back tests during these maneuvers had the lower levels of interruptibility when compared with other maneuvers.

6.2 Prediction Models and Evaluation

For interruptibility prediction, we considered both general and user-specific models. We first illustrate how we preprocess data and generate features. We then selected the best machine learning algorithm and window size for our general model, and then investigated the important features to predict interruptibility. Next, we evaluated the model performance across types of interruptibility measurements, with difference definitions of successful interruptible moments. Finally, we evaluated the user-specific models.

For interruptibility definition, we used a combination of *driving_safety & auditory-verbal_performance & overall_perceived_difficulty*. Our data set is unbalanced because the fraction of interruptible moments is greater than that of uninterruptible moments. Even with this most conservative definition, the ratio of interruptible moments to uninterruptible moments is about 7:3. To handle unbalanced classes in model training, we used a technique of balancing training samples by oversampling called the Synthetic Minority Over-sampling Technique (SMOTE) [31]. Over-sampling was only applied to the training dataset according to the guideline.

For the evaluation of our model, we used the leave-one-subject-out (LOSO) method. Each model was trained with data of all drivers except one driver, and this driver's data was used to test the trained model. This method was chosen due to the concern of a neighborhood bias in n-fold cross validation [28], and the model would be more generalizable if a driver's data is not included in model building.

For the performance metric in the evaluation, we used the mean f-measure that is the average geometric mean of precision and recall for each class [28]. The f-measure is known to be less sensitive to changes in data distribution when comparing to accuracy [31]. The equation for the mean f-measure is as follows:

$$\mathrm{f}_1 = \frac{2}{c}\sum_{i=1}^{\mathrm{c}} \frac{\mathrm{precision}_i \times \mathrm{recall}_i}{\mathrm{precision}_i + \mathrm{recall}_i}$$

where c is the number of classes in the data-set (c = 2 in our case).

As for the confusion matrix, we used an aggregated confusion matrix [44]. The aggregated matrix was computed by summing the entries in all the confusion confusion matrices each of which was produced after a leave-one-subject-out evaluation. For example, the true-positive value in the aggregated confusion matrix is the sum of all the true-positive values in each of the confusion matrices.

We performed the statistical tests for model comparisons [16]. The performance results of tested models were normally distributed according to the Shapiro-Wilk tests ($p > 0.05$)—the assumption of normality is met if the p-value of the Shapiro-Wilk test is greater than 0.05 [62]. For multiple comparison tests (i.e., comparing more than two models), we used the repeated-measures (or within-subjects) ANOVA. For the pairwise comparison (i.e., comparing two models), we carried out the paired t-test for comparing two general models, or the unpaired t-test (or independent t-test) for comparing general models with user-specific models [16]. The unpaired t-test was used because different datasets were used to build the general and user-specific models. For repeated tests in the post-hoc analysis, we adjusted the p-values based on Bonferroni correction [24].

6.2.1 Preprocessing and Feature Generation. To generate features, we used the vehicle and environmental data that were collected over a specific time window (1–5 seconds) before the start of a secondary task execution. The time difference between the starting point of a secondary task and the starting point of the n-back test in the task also varied, depending on how quickly the drivers responded in the asking stage of the task (in seconds, M = 1.8, SD = 0.6, range = 0.6–6.8, distance = 20.34 m for 40 km/h). The time difference between each secondary task (i.e., window) varied but was at least 30-second intervals.

In the preprocessing stage, for the numeric vehicle data, we generated vectors and absolute vectors representing the directed quantities and the magnitude of data, respectively. The vector was calculated by dividing the difference in value by the difference in time between two consecutive values. The numeric vehicle data include relative absolute throttle position, absolute throttle position, engine road, accelerator pedal position, RPM, speed, steering wheel angle, and brake pressure (for the details of each data type, see Table 1).

Next, we generated features for each time window. For each type of numeric data, including vectors and absolute vectors, we calculated a set of features by taking two or more of the mathematical operators in Table 5. Similarly, we also calculated a set of features for positive-valued and negative-valued vectors, representing acceleration and deceleration of data, respectively, by removing either non-negative or non-positive values for the given window.

For the non-numeric data, which were exclusively turn indicator data, we generated three features by taking the most frequently appearing value in the whole window, and the first and second half of each window. In total,

Table 5. Feature generation for numeric vehicle data.

Dataset	Mean	SD	Min	Max	Mdn	Skew.
Raw data	yes	yes	yes	yes	yes	yes
Raw vectors	yes	yes	yes	yes		
Absolute vectors	yes	yes				
Negative-valued vectors	yes	yes	yes	yes		
Positive-valued vectors	yes	yes	yes	yes		

we extracted 163 features from the vehicle data. In addition to the vehicle data, we included an *n*-back type of an incoming *n*-back test as a feature. Since it is system-initiated user interaction, we assume that the cognitive demand (*n*-back type) of an incoming secondary task can be measured a priori. Furthermore, we included driving environment information prior to the secondary task as features. The environment information includes the type of maneuvering (straight, left/right turn, left/right lane change, and stop), the distance to adjacent cars, and the number of nearby cars in the front, left lane, and right lane of the vehicle. In total, we had 171 features for model building.

6.2.2 Selection of Best-performing Machine Learning Algorithm and Window Size. We first examined the general models that used an aggregated dataset of all drivers. As for interruptibility labeling, we took a conservative approach where all three metrics were conjunctively considered (see Section 6.1.1); i.e., a task is labelled as interruptible if each of all metrics is interruptible. To determine the best machine learning (ML) algorithm and window size for our general models, we considered four well-known ML algorithms and five window sizes. For the algorithms, we used the decision tree (DT), naïve bayes (NB), support vector machine (SVM), and random forest (RF). For DT, we used the C4.5 algorithm. For the SVM, we used the C-SVC algorithm with a radial-kernel function and gamma in the function set as 1/k (k = number of features) [10]. Detailed information about these algorithms can be found in Witten et al. [84]. We trained each ML model using windows with a size varying from 1 to 5 seconds. As shown in Table 6, the performance of each model varied with the ML algorithms, regardless of its window size. Namely, the RF model achieved the best performance greater than 0.70 among the four ML algorithms. The other algorithms achieved lower performance than the RF model. When considering the performance of the RF model across different window sizes, a 2-second window resulted in the highest performance of 0.74, as confirmed by the ANOVA test with the Greenhouse-Geisser correction (window-size effect: $F_{(2.944, 82.431)} = 16.287, p < 0.001$, ML-algorithm effect: $F_{(2.046, 57.284)} = 677.866, p < 0.001$, window-size × ML-algorithm effect: $F_{(7.510, 210.275)} = 8.373, p < 0.001$) and post-hoc comparison tests ($p < 0.05$). The aggregated

Table 6. Performance (F-measure) of general models against machine learning (ML) algorithm and window sizes.

		Window size (in seconds)				
		1	2	3	4	5
ML algorithm	Decision Tree	0.57	0.59	0.59	0.56	0.58
	SVM	0.18	0.15	0.14	0.13	0.13
	Naïve Bayes	0.73	0.70	0.65	0.61	0.61
	Random Forest	0.73	0.74	0.70	0.71	0.69

Table 7. Performance of models that predict each of interruptibility dimensions, and an ensemble model that classify interruptibility by aggregating the result of the models.

Interruptibility Definition	**ML algorithm**	**Avg. F-score**	**Confusion matrix**				**Interruptibility Freq.**	
	(Win. size = 2s)		**Positive (inter.)**		**Negative (uniter.**		**Inter.**	**Uninter.**
			True	**False**	**True**	**False**	**moments**	**moments**
Driving safety (safety)	RF	0.87	1196	18	129	45	1241	147
Aud.-verbal performance (secondary)	RF	0.82	1118	42	156	72	1190	198
Overall perceived difficulty (difficulty)	SVM	0.81	1057	65	200	84	1119	269
Safety + secondary + difficulty	RF	0.74	786	140	309	153	939	449
	Ensemble	0.73	742	117	332	197		

confusion matrix for the RF model with a 2-second window shows that the number of true positives was 786 among 939 interruptible cases whereas that of true negatives was 309 among 449 uninterruptible cases. Our results indicate that future interruptible moments for a driver can be reasonably predicted using a ML algorithm. For the following model training and evaluation, we used the RF model with a 2-second window.

6.2.3 Alternative Model Building Strategies. So far we built a model based on the dataset where interruptibility of a task was labelled by considering all the dimensions simultaneously (driving_safety + auditory-verbal_performance + overall_perceived_difficulty). As an alternative, we can use an ensemble method [17] as follows. We built three separated models that predict each of interruptibility dimensions, and the overall interruptibility was classified by aggregating the results of the three models. In other words, a task is classified as interruptible if all three models classified it as interruptible. We first examined the performance of the three separated models. We trained each of four ML models (Random forest, SVM, Decision Tree, Naïve Bayes) using a 2-second window for each interruptibility dimension. As shown in Table 7, we found that each of the separated models (driving safety = 0.87, auditory-verbal performance = 0.82, overall perceived difficulty = 0.81) had a better predictive performance than the model with conservative labeling (safety + secondary + difficulty = 0.74). Next, we built an ensemble model by aggregating the results of the three models. The new model resulted in a slightly lower performance as opposed to the original model, with a decrease of 0.01. However, we did not find a statistically significant performance difference according to our pairwise comparison result with the Paired t-test ($t_{(28)} = 0.194, p = 0.86$).

6.2.4 Important Features in Prediction of Driver Interruptibility. We investigate the important features to predict interruptibility. Interruptibility labeling differs depending on how we define interruptibility (e.g., individually, conjunctively, or disjunctively). In the case of conjunctive labeling, we consider the case of considering all three dimensions (driving_safety + auditory_verbal_performance + overall_perceived_difficulty). To find important features, we removed irrelevant or redundant features via the correlating feature selection (CFS) algorithm [87]. CFS determines a subset of relevant features with the highest predictive ability, and a low inter-correlation among them [87]. We performed feature selection in each training model in the leave one-subject out evaluation (LOSO) and counted the frequency of each feature. We performed 29 CFS processes since we had 29 iterations (i.e., training models) in LOSO.

Regardless of interruptibility definitions, the number of features that were selected by the CFS at least once was about one fourth of 171 features (overall interruptibility: n = 62, driving safety: n = 49, auditory-verbal performance: n = 42, overall perceived difficulty: n = 37). When considering the frequency of the selection across feature type in the results, majority of features were selected less than a half (i.e., 15 times) of total 29 CFS processes. For the overall interruptibility class, the mean of the frequency of the selection was 14.3 (SD = 11.4). For the driving safety, the mean of the frequency of the selection was 14.0 times (SD 11.8). For the auditory-verbal

Table 8. Common features in the correlating feature selection (CFS). The CFS was applied to each training model in leave one-subject our evaluation (LOSO). Common features are the features that were selected more than or equal to half times (15 times) in 29 iterations of LOSO evaluation. rel. = relative, wind. = window; abs. = absolute, pos. = positive, neg. = negative, vec. = vector

Common feature type	**Frequency of the selection**			
	Type of Interruptibility dimension			**Overall inter.**
	Drive safety	**Aud.-verbal perform.**	**Overall. perc. diff.**	
N-BACK TYPE of the incoming secondary task	29	29	29	29
Current MANEUVER_TYPE	29	15	15	29
DISTANCE_TO_ADJACENT_CAR in the left lane	29			27
NUMBER_OF_CARS in the front	29			29
NUMBER_OF_CARS in the left lane	29			28
ACCELERATOR_PEDAL_POSITION median				26
ACCELERATOR_PEDAL_POSITION pos. vec. SD	29			29
ACCELERATOR_PEDAL_POSITION max.				29
ACCELERATOR_PEDAL_POSITION median		17		20
ACCELERATOR_PEDAL_POSITION pos. vec. min.				29
ACCELERATOR_PEDAL_POSITION pos. vec. SD			25	20
BRAKE_PRESSURE_MAX	23			
BRAKE_PRESSURE_MEDIAN	27			
BRAKE_PRESSURE_MIN	29	29		29
BRAKE_PRESSURE neg. vec. mac.			27	26
BRAKE_PRESSURE skewness				23
ENGINE_LOAD mean				21
ENGINE_LOAD median	15			
ENGINE_LOAD neg. vec. SD	15			29
REL._THROTTLE_POSITION abs. vec. mean				26
REL._THROTTLE_POSITION min.				21
REL._THROTTLE_POSITION neg. vec. SD	29			29
RPM max.	29			15
RPM median	22			24
RPM vec. max.	25			27
SPEED mean				28
SPEED pos. vec. SD				18
SPEED vec. mean	20			
STEERING_WHEEL_ANGLE abs. vec. SD	29			29
STEERING_WHEEL_ANGLE max.				29
STEERING_WHEEL_ANGLE median	29			
STEERING_WHEEL_ANGLE pos. vec. SD	29			29
STEERING_WHEEL_ANGLE vec. mean.	29			29
TURN_INDICATOR most freq. value in 2nd half wind.	29		15	29
TURN_INDICATOR most freq. value in whole wind.	29			29

performance, the mean was 5.30 (SD = 5.7). For the overall perceived difficulty, the mean was 6.98 (SD = 6.24). The selection of a wide range of features may be due to the change of a training set for each iteration in LOSO. In the iteration, a driver's data is removed from a training dataset and a new driver's data is added to the dataset.

Among the selected features in CFS, we reported the commonly-appeared features (common feature) that selected more than or equal to half times (15 times) in 29 iterations in the LOSO. Table 8 shows the common features across the type of interruptibility definition. Regardless of types of the interruptibility definition, *n*-back type of the incoming secondary task was commonly selected by the CFS for each iteration in LOSO, since the *n*-back type indicates the amount of cognitive distraction for the incoming secondary tasks and interruptibility. In

Table 9. Performance of models across types of interruptibility definition

Interruptibility Definition	**Avg. F-score**	**Confusion matrix**				**Interruptibility Freq.**	
		Positive (inter.)		**Negative (uniter.**		**Inter.**	**Uninter.**
		True	**False**	**True**	**False**	**moments**	**moments**
Driving safety (safety)	0.87	1196	18	129	45	1241	147
Safety + Auditory-verbal performance (secondary)	0.79	954	60	257	117	1071	317
Safety + secondary + overall perc. difficulty	0.74	786	140	309	153	939	449

addition, the maneuver type before engaging in a secondary task was also commonly selected. These observations are consistent with our earlier descriptive findings in Table 4. In the case of driving safety, detailed vehicle operations (e.g., braking, steering wheel angle) were often selected as features. In addition, the features that describe traffic levels on the road, such as distance to the adjacent vehicle in the left lane, and the number of vehicles in the front and left-hand lane of the vehicle, were also selected.

6.2.5 Variance in Definition of Interruptible Moments. We were interested in the performance of the prediction models across types of interruptibility definition (see Section 3). As shown in Table 9, we basically considered interruptibility definitions of *driving_safety & auditory-verbal_performance & perceived_overall_difficulty* and *driving_safety*. In addition, we trained a model with interruptibility definition of *driving_safety & auditory-verbal_performance*, and measured its performance. Overall, the performance decreased as we consider more dimensions in interruptibility definition. The model with interruptibility definition of *driving_safety*, which may be appropriate for the scenarios of simple factual conversations, showed the best performance of 0.87. The model with interruptibility definition of *driving_safety & auditory-verbal_performance*, which is appropriate the scenarios of delivering critical information that requires a drivers' accurate responses, showed the second highest performance of 0.79. Finally, the model with interruptibility definition of *driving_safety & auditory-verbal_performance & perceived_overall_difficulty*, which is very conservative compared to the other models, showed the slightly lower performance of 0.74.

6.2.6 Driver Variance. The interruptibility of a secondary task could be varied by individual differences, because driving and auditory-verbal performance for a driver is highly correlated with the driver's capabilities [22, 65]. Because of this, we considered individual differences in our models. We built user-specific models and compared performance of user-specific models and a general model. For the performance of the user-specific models, we used an average value for the models as there were multiple models (one for each user). The general model was built using all of our drivers' data. For the user-specific models, each model was individually trained and tested with specific user data. We used a 10-fold cross-validation for the test. Similar to the general model, we dealt with an unbalanced dataset by applying oversampling (SMOTE) to the training data. For both types, we included all features for model building. When compared to the performance of the general model, the user-specific models

Table 10. Performance of user-specific models. For the user-specific models, the value of F-measure shows the average of value among the models.

	Model type	
	User-specific	**General**
Average F-measure	0.71	0.74

showed slightly lower performance, with a decrease of 0.03 (the average of user-specific models = 0.71, compared to 0.74 for the general model). However, there was no statistically significant performance difference between the two models according to the unpaired t-test ($t_{(38.010)} = 0.397, p = 0.70$). The confusion matrix for the user-specific models shows that the number of true positives was 749 among 939 interruptible cases, and the number of true negatives was 261 among 449 uninterruptible cases (vs. the general model: number of true positives = 786, number of true negatives = 309). Slightly lower performance in the user-specific model may be due to insufficient training data for uninterruptible cases when compared to that in the general model.

7 DISCUSSION

7.1 Prediction of Opportune Moments

Our ML analysis showed that the opportune moments of driver interruption could be inferred as accurately as 0.87 in F-measure when we only consider driving safety. The model accuracy varied across types of interruptibility definitions (range = 0.74 – 0.87), types of model building strategies (ensemble method = 0.73), and types of learning models (user-specific model = 0.71, general model = 0.74). When considering the type of features, we found that the features describing the amount of *vehicle state variation* related to vehicle maneuvering are important for predicting interruptible moments. For example, when making turns, there is high state variation, and therefore, it would be less appropriate for interruption, regardless of the current speed. This observation complements the prior study that simply reported interruptible moments were likely to happen when the speed of vehicles or the curvature of the road was low [37].

Driver interruptibility research should carefully consider potential impacts of *compensatory behaviors* such as reducing speed and increasing inter-vehicular distance that occur *when users engage in any types of secondary tasks while driving*. This type of compensatory behavior is well known in the driver distraction field [27], but it is still foreign to our community. We recommend that compensatory behaviors should be handled with care for model building. For example, if interruptible moments are defined as the moments of performing user-initiated secondary tasks such as eating [37], it is likely that drivers' compensatory behaviors will be captured in the sensor data, and thus, model building would be biased. In fact, the system needs to predict interruptibility before a driver engages in a secondary task (at that point, compensatory behaviors are not observed yet), as training a model based on a biased dataset would be misleading. As recommended in our model building, we need to use a window of sensor data collected before a driver preforms a secondary task for interruptibility prediction in driving contexts.

7.2 Subjective Interruptibility and Overestimation

In our experimental framework design, we introduced the asking stage that allows the drivers to make an early decision on whether they perform a proactive secondary task. This decision can be viewed as a driver's subjective response of interruptibility, measured in the beginning of a test. This approach is widely used in traditional interruptibility research. In our field data collection, we instructed that the participants can freely make their decisions. We initially wanted to incorporate this measure into the interruptibility dimension (e.g., by treating 'no' as uninterruptible). Interestingly, our participants almost unanimously said "yes" in this stage. Answering "no" happened only a few instances (occurred in the excluded data). For this reason, we did not consider that in our interruptibility measure. What is remarkable is that despite answering "yes", our participants had quite a significant faction of instances labeled as uninterruptible, particularly when we use a conservative criterion (i.e., considering driving safety, task difficulty, and task performance).

In our post-interview, we asked the reasons behind their decisions, by showing the task failure cases and associated driving situations. Qualitative content analysis revealed that our drivers often overestimated their capability of dual-tasking, stating that they thought that they could easily perform the secondary tasks. Test failure cases are mostly the situations where our drivers could not fully pay attention to both driving and *n*-back tasks (e.g., turning). For safety reasons, our drivers tended to pay more attention to their driving task as opposed to the secondary tasks. Interestingly, they usually did not even realize how many mistakes they made. For example, *P*7 stated, "I thought I could do, [...] I didn't know I had made that many errors." This means that a simple subjective measure in the beginning of interruption may not truly reflect whether a driver is interruptible in driving contexts, because drivers tend to overestimates their driving capability and to engage in dual-tasking [83], which is unfortunately one of the major causes of car accidents [3]. In a safety critical scenario like driving, it is very important to include *objective measures* such as steering wheel control and task performance.

7.3 Length of Auditory-verbal Secondary Task and 15-second Rule

The average length of secondary tasks was 26.6 seconds (SD = 0.6). The length was different depending on how the drivers quickly answered in the asking stage and measuring stage. The length was still longer than those recommended for interactive sequences by guidelines such as the NHTSA and SAE (i.e., SAE Recommended Practice J2364 [52], well-known as the 15-second rule [25]). However, note that this rule is specifically designed for visual-manual operations. SAE J2364 "specifically states [the scope of the rule] is limited to navigation systems" or interfaces "concerned with the operation of controls that require visual guidance" [26]. At this point, to our knowledge, there is no guideline about the length of auditory-verbal interactions. In addition, speech interactions tend to take much longer time than traditional visual-manual operations [1, 34], and thus, the current task duration would be appropriate for auditory-verbal task.

7.4 Prediction of Opportune Moments for Real-world Auditory-verbal Tasks

We employed *n*-back tests [50] that induce consistently structured and controlled levels of cognitive demand (0-back: a very mild level, 1-back: a moderate level, and 2-back: a high level). Each type of *n*-back test can be used as a reference task for investigating the interruptibility of real-world proactive auditory-verbal interactions. We can match the cognitive demand of such interactions with *n*-back tests. According to the literature, a phone contact calling via a Toyota Corolla system requires similar cognitive demand to that of a 1-back test, whereas an address entry task requires similar cognitive demand to that of a 2-back test [48]. We can determine whether this type of proactive auditory-verbal service/task requires a similar cognitive demand as an *n*-back test. If a task is comparable to a 1-back test, the service will be delivered if the system finds interruptible moments for successful 1-back task execution. In practice, this type of matching may not necessarily be required. The system could always start with the most conservative level (e.g., 2-back tests). We then adjust the interruptibility criteria by possibly using automatic workload assessment models that can identify the cognitive demands of arbitrary secondary tasks [71].

7.5 Use of Alternative Data Sources for Prediction

Vehicle data was an important data source for our classifier, as in prior studies [37, 71]. We collected the vehicle data using two different approaches: 1) OBD-II PIDs and 2) CAN-bus signal decoding. This type of data collection requires special hardware (OBD scanner) and software (CAN-bus signal decoding). This would be particularly challenging when a classifier needs to be deployed into a portable device (e.g., checking interruptibility for delivering notifications via smartphones). We highlight that interruptibility could be estimated using alternative data sources to pure in-vehicle sensor data. Motion sensing and GPS tracking in smartphones could be used to measure the speed and direction of the vehicle [7] as well as road conditions/types [45]. Alternatively, we could monitor driver activity based on in-vehicle or smartphone cameras. For example, Veeraraghavan et al. [80] proposed a driver activity classifier using computer vision, which provides useful information for interruptibility classification. Similarly, we can use a portable motion sensor to monitor driver behaviors; e.g., Liu et al. used a motion sensor attached to the steering wheel to estimate wheel rotation [45].

7.6 Facilitating In-vehicle Auditory-verbal Interactions

Our results could also be used to build a virtual agent that can proactively interact with a driver to engage in various tasks while driving, ranging from simple context-aware information delivery to interactive decision-making. For example, an agent may initiate an auditory-verbal interaction with drivers to overcome boredom and shake-off drowsiness in long-distance driving [19, 43]. Such interactions could also be helpful for drivers, as the Yerkes-Dodson Law of Arousal [86] suggests that the performance of a task is best when an operator performs a task at intermediate levels of workload compared to that at an extreme level. For example, during the period of underflow, the auditory-verbal interaction with such an agent may help drivers improve their performance. Our models could be also applicable to auditory-verbal interfaces in existing portable or in-vehicle interfaces. For safe driving reasons, our model can be used to warn and proactively limit user interactions (e.g., checking schedule changes) when drivers are classified as uninterruptible [13, 35, 36, 38].

We analyzed the post-interview data collected after a real-road study. Our drivers often stated that the *n*-back test differed compared with their ordinary conversations, mainly because they cannot pause the conversation or ask for a repeat. This suggests that it could be helpful to provide a pause option in the auditory-verbal interfaces that do not require visual and manual operations, especially when the drivers are actively interacting with the interfaces. They also commonly mentioned a verbal command of "*hold on*," when they faced difficulty during driving, hoping that they could resume their conversational interactions later on. Enabling adaptive pace controlling mechanisms would be very useful for drivers to better manage system-initiated conversational interactions.

7.7 Limitations

This study provides a basis for further investigation of predicting opportune moments for proactive in-vehicle auditory-verbal tasks, but there is limited generalizability of our model. Despite cognitive workload matching between *n*-back tests and real auditory-verbal tasks, there are still auditory-verbal tasks that partly involve visual or/and manual operations as well [2, 49]. In recent years, conversation interactions have continued to gain popularity, and future interfaces will increasingly rely on auditory-verbal interactions with minimal visual-manual operations for driver safety. Indeed, the U.S. National Highway Traffic Safety Administration encourages locking up any interaction tasks that require unreasonable visual-manual distraction [2]. This trend partly affirms our approach of focusing on auditory-verbal tasks. Whereas *n*-back tests have been widely used in in-vehicle system research, we also think that it is desirable to compile a well-defined verbal task set for conversational agents. In practice, however, it is difficult to design standardized task sets, owing to contextualized responses and multi-turn-taking patterns. There should be follow-up studies on these issues, beyond large-scale data collection and validation. Note also that the current system design collect multiple sensor data ranging from in-vehicle control to dashcam videos, and there should be further studies on the privacy concerns on sensor data collection and usage [55].

8 CONCLUSION

We proposed an experimental framework for proactive, system-initiated auditory-verbal tasks, by exploring multiple dimensions of driver interruptibility, i.e., driving safety, task performance, and perceived difficulty. Our framework was iteratively developed through an extensive literature review, simulator-based user studies (n = 11, 17), and real-road pilot studies (n = 4, 9). We then integrated our experimental framework into an open-source navigation app, and collected a real-road dataset (n = 29). This dataset was used to develop an automatic interruptibility classifier. Our evaluation results showed that our model can achieve a reasonable accuracy (F-measure = 0.74), even when conservatively considering all three dimensions.

Recent advances in intelligent agents will enable a variety of proactive auditory-verbal services, even in driving contexts. In this growing field, driver interruptibility covers important research topics such as driver safety and user experience. We hope that our study serves as another step towards investigating driver interruptibility, as well as enabling various proactive conversational services in driving contexts.

ACKNOWLEDGMENTS

This research was supported in part by Hyundai NGV and by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017M3C4A7065960).

REFERENCES

[1] National Highway Traffic Safety Administration. 2012. Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices. (2012).

[2] National Highway Traffic Safety Administration. 2014. Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices. (2014).

[3] National Highway Traffic Safety Administration. 2017. Distracted Driving 2015 - Traffic Safety Facts: Research Note. (2017).

[4] Christoph Anderson, Isabel Hübener, Ann-Kathrin Seipp, Sandra Ohly, Klaus David, and Veljko Pejovic. 2018. A Survey of Attention Management Systems in Ubiquitous Computing Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 58 (July 2018), 27 pages.

[5] Brian Bailey and Joseph Konstan. 2006. On the need for attention award systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior Computers* 22, 4 (2006), 709–732.

[6] Adriana Barón and Paul Green. 2006. *Safety and usability of speech interfaces for in-vehicle tasks while driving: A brief literature review.* Technical Report. University of Michigan, Transportation Research Institute, Ann Arbor, MI, USA.

[7] Luis Bergasa, Daniel Almeria, Almazan. Javier, Javier Yebes, and Roberto Arroyo. 2014. DriveSafe: An app for alerting inattentive drivers and scoring driving behaviors. In *Proceedings of the 2014 IEEE Intelligent Vehicles Symposium.* IEEE, 240–245.

[8] Marie-Pierre Bruyas, Laëtitia Dumont, and Bron France. 2015. Sensitivity of Detection Response Task (DRT) to Driving Demand and Task Difficulty. *Accident Reconstruction Journal* 25 (2015), 12–15.

[9] Joseph Campbell. 1997. Speaker recognition: a tutorial. *Proc. IEEE* 85, 9 (1997), 1437–1462.

[10] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3, Article 27 (May 2011), 27 pages.

[11] Woohyeok Choi, Aejin Song, Darren Edge, Masaaki Fukumoto, and Uichin Lee. 2016. Exploring User Experiences of Active Workstations: A Case Study of Under Desk Elliptical Trainers. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16).* ACM, New York, NY, USA, 805–816.

[12] Louis Coraggio. 1990. *Deleterious Effects of Intermittent Interruptions on the Task Performance of Knowledge Workers: A Laboratory Investigation.* Ph.D. Dissertation. The University of Arizona.

[13] Anna L Cox, Sandy JJ Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* ACM, 1389–1397.

[14] Roderick Currie. 2017. *Hacking the CAN Bus: Basic Manipulation of a Modern Automobile Through CAN Bus Reverse Engineering.* Technical Report. The SANS Institute.

[15] Dick De Waard. 2002. *Human factors for highway engineers.* Pergamon, Netherlands, Chapter Mental Workload, 161–175.

[16] Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* 7 (Jan 2006), 1–30.

[17] Thomas G. Dietterich. 2000. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems.* Springer Berlin Heidelberg, Berlin, Heidelberg, 1–15.

[18] Thomas Dingus, Sheila Klauer, Vicki Neale, Andy Petersen, Suzanne Lee, J. Sudweeks, Miguel Perez, Jonathan Hankey, D. Ramsey, Santosh Gupta, C. Bucher, Z. Doerzaph, J Jermeland, and R. Knipling. 2006. The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment. (2006).

[19] Shingo Douno and Masami Kato. 2005. A Technique for Preventing an In-Car Conversational Agent's Responses from Becoming Monotonous. In *Proceedings of the 12th World Congress on Intelligent Transport Systems (ITS '05)*, Vol. 6. Intelligent Transportation Society of America, Washington, DC, USA, 3484–3494.

[20] ETS2MODS.LT. 2018. ETS2 mods - Euro truck simulator 2 mods.

[21] Humayun Faheem, Sahibzada Mahmud, Gul Khan, Maria Rahman, and Haseeb Zafar. 2013. A Survey of Intelligent Car Parking System. *Journal of Applied Research and Technology* 11, 5 (2013), 714 – 726.

[22] Vérane Faure, Régis Lobjois, and Nicolas Benguigui. 2016. The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. *Transportation Research Part F: Traffic Psychology and Behaviour* 40 (2016), 78–90.

[23] Davis Fred. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (1989), 319–340.

[24] Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences* 180, 10 (2010), 2044 – 2064.

[25] Paul Green. 1999. Estimating Compliance with the 15-Second Rule for Driver-Interface Usability and Safety. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 43, 18 (1999), 987–991.

[26] P Green. 2000. Potential expansion of the 15-second rule. In *NHTSA Driver Distraction Expert Working Group Meetings, Summary and Proceedings*. Washington, DC, USA.

[27] D. Haigney, R. Taylor, and S. Westerman. 2000. Concurrent mobile (cellular) phone use and driving performance: task demand characteristics and compensatory processes. *Transportation Research Part F: Traffic Psychology and Behaviour* 3, 3 (2000), 113–121.

[28] Nils Hammerla and Thomas Plötz. 2015. Let's (Not) Stick Together: Pairwise Similarity Biases Cross-validation in Activity Recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 1041–1051.

[29] Joanne Harbluk, Ian Noy, Patricia Trbovich, and Moshe Eizenman. 2007. An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. *Accident Analysis and Prevention* 39, 2 (2007), 372–379.

[30] Sandra Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908.

[31] Haibo He and Edwardo Garcia. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 21, 9 (2009), 1263–1284.

[32] Jun-Ho Heo and Seon-Bong Lee. 2015. A Study on Parking Guideline Generation Algorithm. *Journal of the Korea Academia-Industrial cooperation Society* 16, 5 (2015), 3060–3070.

[33] Shamsi T. Iqbal, Yun-Cheng Ju, and Eric Horvitz. 2010. Cars, Calls, and Cognition: Investigating Driving and Divided Attention. In *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1281–1290.

[34] John D. Lee Chun-Cheng Chang Vindhya Venkatraman Madeleine Gibson Kaitlin E. Riegler Daniel Kellman James W. Jenness, Linda Ng Boyle. 2016. In-vehicle voice control interface performance evaluation (Final report. Report No. DOT HS 812 314). (2016).

[35] Inyeop Kim, Gyuwon Jung, Hayoung Jung, Minsam Ko, and Uichin Lee. 2017. LetâĂŹs FOCUS: Mitigating Mobile Phone Use in College Classrooms. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 63.

[36] Jaejeung Kim, Chiwoo Cho, and Uichin Lee. 2017. Technology Supported Behavior Restriction for Mitigating Self-Interruptions in Multi-device Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 64.

[37] SeungJun Kim, Jaemin Chun, and Anind Dey. 2015. Sensors Know When to Interrupt You in the Car. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 487–496.

[38] Minsam Ko, Seungwoo Choi, Koji Yatani, and Uichin Lee. 2016. Lock n' LoL: group-based limiting assistance app to mitigate smartphone distractions in group activities. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 998–1010.

[39] Georgios K. Kountouriotis, Panagiotis Spyridakos, Oliver M.J. Carsten, and Natasha Merat. 2016. Identifying cognitive distraction using steering wheel reversal rates. *Accident Analysis & Prevention* 96 (2016), 39–45.

[40] Timo Lajunen and Heikki Summala. 1995. Driving experience, personality, and skill and safety-motive dimensions in drivers' self-assessments. *Personality and Individual Differences* 19, 3 (1995), 307 – 318.

[41] David R. Large, Gary Burnett, Ben Anyasodo, and Lee Skrypchuk. 2016. Assessing Cognitive Demand During Natural Language Interactions with a Digital Driving Assistant. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (Automotive'UI 16)*. ACM, New York, NY, USA, 67–74.

[42] Eun-Kyu Lee, Mario Gerla, Giovanni Pau, Uichin Lee, and Jae-Han Lim. 2016. Internet of Vehicles: From intelligent grid to autonomous cars and vehicular fogs. *International Journal of Distributed Sensor Networks* 12, 9 (2016), 1550147716665500.

[43] Uichin Lee, Kyungsik Han, Hyunsung Cho, Kyong-Mee Chung, Hwajung Hong, Sung-Ju Lee, Youngtae Noh, Sooyoung Park, and John M. Carroll. 2019. Intelligent positive computing with mobile, wearable, and IoT devices: Literature review and research directions. *Ad Hoc Networks* 83 (2019), 8 – 24.

[44] Zhen Li, Zhiqiang Wei, Lei Huang, Shugang Zhang, and Jie Nie. 2016. Hierarchical Activity Recognition Using Smart Watches and RGB-Depth Cameras. *Sensors* 16, 10, Article 1713 (2016).

[45] Luyang Liu, Hongyu Li, Jian Liu, Cagdas Karatas, Yan Wang, Marco Gruteser, Yingying Chen, and Richard Martin. 2017. BigRoad: Scaling Road Data Acquisition for Dependable Self-Driving. In *Proceedings of the 15th Annual International Conference on Mobile Systems,*

Applications, and Services (MobiSys '17). ACM, New York, NY, USA, 371–384.

[46] Sean MacLain. 2017. Toyota's Talking Car Wants to Be Your Clingy BFF. (30 October 2017).

[47] Angela Mahr, Michael Feld, Mohammad Moniri, and Rafael Math. 2012. The ConTRe (Continuous Tracking and Reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, New York, NY, USA, 88–91.

[48] Bruce Mehler, Bryan Reimer, Jonathan Dobres, and Joseph Coughlin. 2015. *Assessing the Demands of Voice Based In-Vehicle Interfaces-Phase II Experiment 3-2015 Toyota Corolla (2015b) (Technical Report 2015-14)*. Technical Report. MIT AgeLab, Cambridge, MA, USA.

[49] Bruce Mehler, Bryan Reimer, Jonathan Dobres, Hale McAnulty, Alea Mehler, Daniel Munger, and Joseph Coughlin. 2014. *Further evaluation of the effects of a production level "voice-command" interface on driver behavior: replication and a consideration of the significance of training method (Technical Report 2014-2)*. Technical Report. MIT AgeLab, Cambridge, MA, USA.

[50] Bruce Mehler, Bryan Reimer, and Jeffery Dusek. 2011. MIT AgeLab Delayed Digit Recall Task (n-back). (2011).

[51] Christian Meurisch, Maria-Dorina Ionescu, Benedikt Schmidt, and Max Mühlhäuser. 2017. Reference Model of Next-generation Digital Personal Assistant: Integrating Proactive Behavior. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '17)*. ACM, New York, NY, USA, 149–152.

[52] Society of Automotive Engineers. 2004. SAE Recommended Practice J2364: Navigation and Route Guidance Function Accessibility While Driving. (2004).

[53] Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K. Dey, and Hideyuki Tokuda. 2015. Reducing Users' Perceived Mental Effort Due to Interruptive Notifications in Multi-device Mobile Environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 475–486.

[54] Joakim Östlund, Björn Peters, Birgitta Thorslund, Johan Engström, Gustav Markkula, Andreas Keinath, Dorit Horst, Susann Juch, Stefan Mattes, and Uli Foehl. 2005. *Driving performance assessment - methods and metrics*. Technical Report. Information Society Technologies (IST) Programme.

[55] Sangkeun Park, Joohyun Kim, Rabeb Mizouni, and Uichin Lee. 2016. Motives and Concerns of Dashcam Video Sharing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4758–4769.

[56] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 897–908.

[57] Rahul Rajan, Ted Selker, and Ian Lane. 2016. Task Load Estimation and Mediation Using Psycho-physiological Measures. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. ACM, New York, NY, USA, 48–59.

[58] Michael Rakauskas, Leo Gugerty, and Nicholas Ward. 2004. Effects of naturalistic cell phone conversations on driving performance. *Journal of Safety Research* 35, 4 (2004), 453–464.

[59] Thomas Ranney, Scott Baldwin, Ed Parmer, John Martin, and Elizabeth Mazzae. 2012. Distraction effects of in-vehicle tasks requiring number and text entry using auto alliance's principle 2.1 b verification procedure. (2012).

[60] Thomas Ranney, Scott Baldwin, Larry Smith, Elizabeth Mazzae, and Russell Pierce. 2014. Detection Response Task (DRT) Evaluation for Driver Distraction Measurement Application. (2014).

[61] Thomas Ranney, Elizabeth Mazzae, Riley Garrott, and Michael Goodman. 2000. NHTSA Driver Distraction Research: Past, Present, and Future. (2000).

[62] Nornadiah Mohd Razali and Yap Bee Wah. 2011. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *Journal of statistical modeling and analytics* 2, 1 (2011), 21–33.

[63] James Reason, Antony Manstead, Stephen Stradling, James Baxter, and Karen Campbell. 1990. Errors and violations on the roads: a real distinction? *Ergonomics* 33, 10-11 (1990), 1315–1332.

[64] Michael Regan, Charlene Hallett, and Craig Gordon. 2011. Driver distraction and driver inattention: Definition, relationship and taxonomy. *Accident Analysis and Prevention* 43, 5 (2011), 1771–1781.

[65] Michael Regan, John Lee, and Kristie Young. 2009. *Driver distraction: Theory, effects, and mitigation*. CRC Press.

[66] Dario D. Salvucci, Niels A. Taatgen, and Jelmer P. Borst. 2009. Toward a Unified Theory of the Multitasking Continuum: From Concurrent Performance to Task Switching, Interruption, and Resumption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1819–1828.

[67] Walter Schneider and Mark Detweiler. 1988. The Role of Practice in Dual-Task Performance: Toward Workload Modeling a Connectionist/Control Architecture. *Human Factors* 30, 5 (1988), 539–566.

[68] LTD. Seeed Technology Co. 2018. CAN-Bus shield V1.2.

[69] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[70] Victor Shcherb, Alexey Pelykh, Hardy Mueller, and Pavol Zibrita. 2017. OsmAnd - OSM Automated Navigation Directions - navigation software based on OpenStreetMap.

[71] Erin T. Solovey, Marin Zec, Enrique Abdon Garcia Perez, Bryan Reimer, and Bruce Mehler. 2014. Classifying Driver Workload Using Physiological and Driving Performance Data: Two Field Studies. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing*

Systems (CHI '14). ACM, New York, NY, USA, 4057–4066.

[72] SCS Software s.r.o. 2018. Euro Truck Simulator 2.

[73] David Strayer, Joel Cooper, Jonna Turrill, James Coleman, and Rachel Hopman. 2015. Measuring cognitive distraction in the automobile III: A comparison of ten 2015 in-vehicle information systems. (2015).

[74] David Strayer, Joel Cooper, Jonna Turrill, James Coleman, Nate Medeiros-Ward, and Francesco Biondi. 2013. Measuring Cognitive Distraction in the Automobile. (2013).

[75] David Strayer, Jonna Turrill, James Coleman, Emily Ortiz, and Joel Cooper. 2014. Measuring Cognitive Distraction in the Automobile II: Assessing In-Vehicle Voice-Based Interactive Technologies. (2014).

[76] David L Strayer and Frank A Drews. 2004. Profiles in Driver Distraction: Effects of Cell Phone Conversations on Younger and Older Drivers. *Human Factors* 46, 4 (2004), 640–649.

[77] Carlos Toxtli, Andres Monroy-Hern´andez, and Justin Cranshaw. 2018. Understanding Chatbot-mediated Task Management. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 58, 6 pages.

[78] Inc. Transcend. 2018. Transcend DrivePro 200.

[79] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2015. Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 801–812.

[80] Harini Veeraraghavan, Nathaniel Bird, Stefan Atev, and Nikolaos Papanikolopoulos. 2007. Classifiers for driver activity monitoring. *Transportation Research Part C: Emerging Technologies* 15, 1 (2007), 51 – 67.

[81] Bernhard Wandtner, Markus Schumacher, and Eike Schmidt. 2016. The role of self-regulation in the context of driver distraction: A simulator study. *Traffic Injury Prevention* 17, 5 (2016), 472–479. PMID: 27082493.

[82] Jason M. Watson and David L. Strayer. 2010. Supertaskers: Profiles in extraordinary multitasking ability. *Psychonomic Bulletin & Review* 17, 4 (01 Aug 2010), 479–485.

[83] Mathew White, Richard Eiser, and Peter Harris. 2004. Risk Perceptions of Mobile Phone Use While Driving. *Risk Analysis* 24, 2 (2004), 323–334.

[84] Eibe Witten, Ianand Frank, Mark Hall, and Christopher Pal. 2016. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

[85] Leviathan Yaniv and Matias Yossi. 2018. Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone.

[86] Robert Yerkes and John Dodson. 1908. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology* 18, 5 (1908), 459–482.

[87] Lei Yu and Huan Liu. 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. *The Journal of Machine Learning Research* 5 (2004), 1205–1224.

[88] Lawrence Zeitlin. 1993. Subsidiary task measures of driver mental workload: A long-term field study. *Transportation Research Record* 1403 (1993).

Received May 2018; revised August 2018; accepted October 2018