Poster: Predicting Opportune Moments for In-vehicle Proactive Speech Services

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
Daejeon, South Korea
uclee@kaist.ac.kr

ABSTRACT

Auditory-verbal or speech interactions with in-vehicle information systems have became increasingly popular. This opens up a whole new realm of possibilities for serving drivers with proactive speech services such as contextualized recommendations and interactive decision-making. However, prior studies have warned that such interactions may consume considerable attentional resources, thus degrade driving performance. This work aims to develop a machine learning model that can find opportune moments for the driver to engage in proactive speech interaction by using the vehicle and environment sensor data. Our machine learning analysis shows that opportune moments for interruption can be conservatively inferred with an accuracy of 0.74.

CCS CONCEPTS

• **Human-centered computing → User interface management systems**; **Ubiquitous and mobile computing**.

KEYWORDS

Interruptibility; In-vehicle information system; Human-vehicle interaction; Auditory-verbal interface; Speech-based interaction; Speech interface

ACM Reference Format:
Auk Kim, Woohyeok Choi, Jungmi Park, Kyeyoon Kim, and Uichin Lee. 2019. Poster: Predicting Opportune Moments for In-vehicle Proactive Speech Services. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 9–13, 2019, London, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3341162.3343841

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6869-8/19/09.
https://doi.org/10.1145/3341162.3343841

1 INTRODUCTION

The popularity of auditory-verbal interfaces and proactive intelligent agent open up new opportunities for drivers to receive personalized *proactive speech services* while driving [5]. Although such services benefit drivers, they can negatively affect driving performance. For example, Faure et al. showed that the use of speech interfaces can degrade driving performance, particularly when a driver is cognitively overloaded with driving [10]. This is due to a concurrent execution with a limited amount of cognitive resources. In driving contexts, for their safety, drivers must concurrently execute driving task and secondary task (e.g., proactive services) [10] and share their limited capacity of cognitive resources for the both tasks (or dual-task) [14]. If the residual cognitive resources are not available for the dual-task, it cause performance decrements in either one of, or both tasks (i.e., *dual-tasking interference*). This emphasizes the importance of finding opportune moments in driving contexts, in which people can safely engage in proactive speech services while driving.

Finding such opportune moments has been one of the active research areas. Traditionally, studies have considered mostly computing environments (e.g., desktop and mobile devices), where task-switching is feasible [2, 12]. However, the findings in these studies cannot be directly applicable to the driving contexts. As task-switching is not feasible while driving. In driving contexts, task-switching induces various unfavorable effects (e.g, eye-off-road and hand-off-the-wheel) [10]. Relatively, a small number of studies has investigated opportune moments in driving contexts (in-vehicle opportune moments) [11, 12]. These studies utilized subjective measures (e.g., driver preference and availability) for the investigation. For example, Semmens et al examined opportune moments by asking "Is now a good time?" (driver availability) [11]. Such moments, however, may not represent safe moments for drivers to perform a secondary task. Prior research suggest that drivers tend to overestimates their driving capability and carelessly engage in any tasks [10, 13]. This highlights that when finding in-vehicle opportune moments, it is important to systematically consider driving safety.

Despite the popularity of speech interfaces and proactive services [5], there are still insufficient systematic studies of what constitutes in-vehicle opportune moments and of how they can be measured and predicted for delivering in-vehicle proactive speech services. In this study, we first defined opportune moments for speech interactions, and then iteratively developed an experimental framework through a series of simulation and 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 collected data in a real-world field trial with 29 drivers (IRB Approval No. KH2016-49). Finally, we built machine learning models for predicting driver interruptibility.

2 DEFINING IN-VEHICLE OPPORTUNE MOMENTS

Our goal is to develop a driver interruptibility classifier that infers in-vehicle opportune moments to engage in proactive speech services. The finding in prior studies (e.g, dual-tasking interference) implies that driving safety must be considered. In this work, we propose using following dimensions to define in-vehicle opportune moments. (1) *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. Therefore, engagement should not negatively affect driving performance for safety reasons. (2) *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. (3) *overall perceived difficulty* is particularly important to assess perceived interruptibility from user perspective, as typically considered by prior works [12]. The driver should not feel a considerable burden when performing a dual-task.

Depending on the purpose of services, it is possible to use a combination of the dimensions both disjunctively and conjunctively. However, driving safety must be always included in the combination as it is always the top priority for any in-vehicle system. In this study, although it is very conservative, we considered the conjunctive form of the three dimensions to label interruptibility for the classifier.

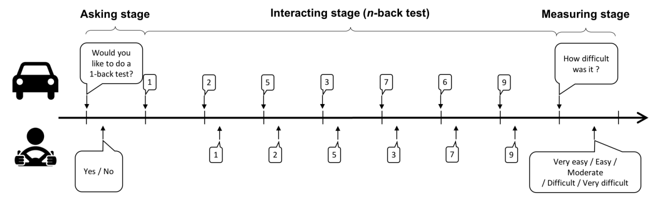

Figure 1: Procedure of a secondary task (N-back test type = 1-back test).

3 EXPERIMENTAL FRAMEWORK

There are still insufficient systematic studies of how to collect interruptibility dataset under various interruption contexts, such as varying level of secondary task demand and diverse driving contexts [11, 12]. Thus, we iteratively develop a framework, via two simulator and two real-road pilots studies, that enables such collection. The final version of the framework consists of (1) secondary-task procedure (2) task-triggering method, and (3) interruptibility metrics.

Secondary Task Procedure: As shown in Figure 1, to systematically induce varying levels of cognitive demand, we employed *n*-back tests [6], which has a similar cognitive engagement to externally-paced speech interactions. The entire procedure of the task consists of three stages: (1) *Asking stage* is when a driver is asked if he is willing to engage in the remaining stages. If the driver answers "Yes" within 2.5 seconds, the remaining stages are presented. Otherwise, the current task ends immediately. (2) *Interacting stage (n-back test)* is when the driver is asked to perform the n-back test. 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) (3) *Measuring stage*, in which the procedure finishes, is when the driver is asked to verbally indicate the overall perceived difficulty (very easy: 1 - very difficult: 5) of the dual-task during the previous stage (interacting stage).

Task Triggering Method: To trigger secondary tasks in diverse driving contexts, we designed the triggering method that presents the tasks based on a hybrid approach of two triggering methods: (1) location-based triggering method - presenting tasks at specific locations, associated with various driving environmental factors (see Figure 2), and (2) random-interval triggering method - presenting at random intervals between 30 and 90 seconds, if the distance to the next predetermined location is sufficiently long.

Interruptibility Metrics: The metrics consist of three metrics to measure three interruptibility dimensions in our definition. Each metric indicates interruptibility as a binary outcome (i.e., interruptible or uninterruptible). (1) *Driving safety*

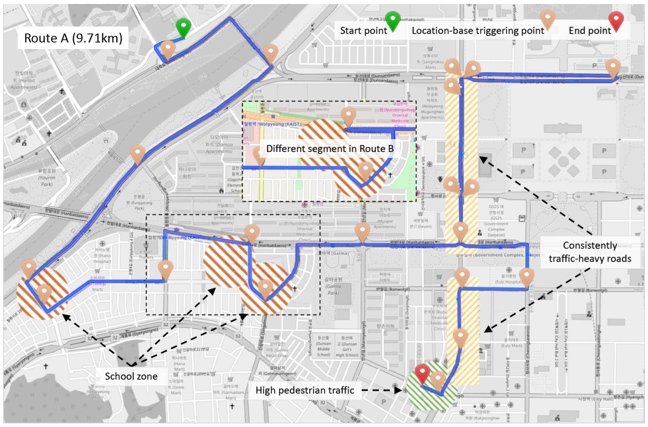

Figure 2: Route A in the round-trip driving course. Route B (returning route) is slightly different from Route A (see the dotted box for a different route).

measures how safely a user drives a vehicle. It indicates interruptible if driving performance when dual-tasking of driving and secondary tasks (n-back test) was not degraded when compared to the performance when performing driving task alone. For the driving performance metric, steering wheel reversal rate [7] was used. (2) *Auditory-verbal performance* measures how well a user performs an auditory-verbal task based on the accuracy of the n-back test. It indicates interruptible if a driver correctly answered all the items in a given test. (3) *Overall perceived difficulty* measures how difficult it is to perform a dual task. It is based on the rating in the measuring stage (see Figure 1). It indicates interruptible if the normalized value of the rating was less than or equal to 0.5. We re-scale the rating to be range of [0, 1] using min-max normalization, because we found that in iterative development stages, each driver had a different range of values.

4 ON-ROAD DATA COLLECTION AND RESULTS

We used our framework to collect the real-road driving dataset, which was used to build interruptibility classifier. For the data collection, we recruited 29 drivers (Age: M = 43.2, SD = 11.4). Drivers were asked to drive a round-trip driving course (route A and B) that is designed to reflect various driving environmental complexity (various maneuvers, differing levels of traffic density, etc.). The vehicle was instrumented with OBDII and dashcams to collect following data: (1) vehicular data (e.g., steering wheel angle, speed, etc.), (2) environmental information (maneuver type, the distance to adjacent cars, and the number of nearby cars in the front, left/right lane of the vehicle), and (3) task-related data (self-reported overall difficulty level, accuracy of given n-back tests).

The experiment took approximately 4.5 hours (driving: 30 mins x 4, break: 10 mins x 2 + 30 mins). Drivers drove the driving course for twice (baseline and secondary-task driving session). Prior to each driving session, campus drive was provided for the adaptation to driving. The baseline session involved no secondary tasks and was used as a baseline to compare driving performances. For secondary-task session, a training session of n-back tasks was additionally provided.

During the secondary-task session, the drivers performed an average of 47.86 (SD = 6.83) secondary tasks, of which 40 cases were performed at predetermined locations (20 for each route). In total, drivers performed 1,413 cases, of which 25 cases were excluded due to a recording problem. Thus, the final dataset consists of 1,388 cases.

5 INTERRUPTIBILITY PREDICTION

For interruptibility prediction, we considered both general and user-specific models. We first illustrate how we label interruptibilty and generate features. We then selected the best machine learning algorithm and window size for our general model, and then evaluated the user-specific models.

Interruptibility Labeling and Feature Generation: We labeled the interruptibility of each secondary task as a binary outcome (i.e., interruptible or uninterruptible). A secondary task was labeled as interruptible when all three dimensions indicated interruptible (see metrics in Section 3). Among 1,388 cases, 939 were interruptible and 449 were uninterruptible. 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. We generated features for each time window by taking each of following mathematical operators for each type of data: mean, SD, maximum, minimum, median, and skewness. In addition, we also included a type of an incoming *n*-back test as a feature. Since it is system-initiated interaction, we assume that the cognitive demand (*n*-back type) of an incoming interaction can be measured a priori. In total, we had 171 features for model building.

Table 1: 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

Selection of Best-performing ML Algorithm and Window Size: We first examined the general models that used an aggregated dataset of all drivers. 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. As shown in Table 1, we considered four popular ML algorithms. We trained each ML model using windows with a size varying from 1 to 5 seconds. The performance of each model varied with the ML algorithms, regardless of its window size. Namely, the random forest model achieved the best performance greater than 0.70 among the four ML algorithms.

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 [10]. Because of this, we also considered individual differences in our models. We built user-specific models and compared performance of user-specific models and a general model. For the user-specific models, each model was individually trained and tested with specific user data. 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 average performance (F-measure) of the models was 0.71, which is similar to the performance of the general model (0.74).

6 CONCLUSION

In this study, we proposed an interruptibility framework for proactive speech tasks, by exploring multiple dimensions of driver interruptibility (i.e., driving safety, auditory-verbal performance, and overall perceived difficulty). We then collected a real-road dataset and showed the interruptibility can be reasonably predicted even when conservatively considering all three dimensions.

Beyond managing system-initiated interactions, our model can be also 1) used to warn and proactively limit user-initiated interactions when drivers are classified as uninterruptible [3, 4] or 2) implemented in systems that mediate risking behaviors of drivers [1]. Our 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 [8, 9].

Recent advances in intelligent agents enable a variety of proactive speech services, even in driving contexts. In this growing field, we hope that our study serves as another step towards investigating driver interruptibility, as well as enabling various in-vehicle proactive speech services.

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] Hyojin Chin, Hengameh Zabihi, Sangkeun Park, Mun Yong Yi, and Uichin Lee. 2017. WatchOut: Facilitating Safe Driving Behaviors with Social Support. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 2459–2465. https://doi.org/10.1145/3027063.3053188

[2] Woohyeok Choi, Sangkeun Park, Duyeon Kim, Youn-kyung Lim, and Uichin Lee. 2019. Multi-Stage Receptivity Model for Mobile Just-In-Time Health Intervention. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 39 (June 2019), 26 pages. https://doi.org/10.1145/3328910

[3] Inyeop Kim, Gyuwon Jung, Hayoung Jung, Minsam Ko, and Uichin Lee. 2017. Let&Rsquo;s FOCUS: Mitigating Mobile Phone Use in College Classrooms. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 63 (Sept. 2017), 29 pages. https://doi.org/10.1145/3130928

[4] Jaejeung Kim, Joonyoung Park, Hyunsoo Lee, Minsam Ko, and Uichin Lee. 2019. LocknType: Lockout Task Intervention for Discouraging Smartphone App Use. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 697, 12 pages. https://doi.org/10.1145/3290605.3300927

[5] 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), 14.

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

[7] 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.

[8] Sangkeun Park, Emilia-Stefania Ilincai, Jeungmin Oh, Sujin Kwon, Rabeb Mizouni, and Uichin Lee. 2017. Facilitating Pervasive Community Policing on the Road with Mobile Roadwatch. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3538–3550. https://doi.org/10.1145/3025453.3025867

[9] Sangkeun Park, Sujin Kwon, and Uichin Lee. 2018. CampusWatch: Exploring Communitysourced Patrolling with Pervasive Mobile Technology. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 134 (Nov. 2018), 25 pages. https://doi.org/10.1145/3274403

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

[11] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. Is Now A Good Time?: An Empirical Study of Vehicle-Driver Communication Timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 637, 12 pages.

[12] 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.

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

[14] Christopher D. Wickens. 2008. Multiple Resources and Mental Workload. *Human Factors* 50, 3 (2008), 449–455.