USER ACCEPTANCE OF PROACTIVE VOICE ASSISTANT BEHAVIOR

Maria Schmidt[1,2]*, Wolfgang Minker*[2]*, Steffen Werner*[3]
[1]*Mercedes-Benz AG,* [2]*Ulm University,* [3]*Daimler Trucks North America LLC*
maria.m.schmidt@daimler.com

Abstract: Using Personal Assistants (PAs) via voice becomes increasingly popular and available in multiple environments, thus we aim to provide proactive PA suggestions to car drivers via speech. Since these suggestions should enhance the user experience while avoiding obtrusiveness and cognitive load, we assess these factors in a usability study. We investigate how 42 participants perceive proactive voice output in a Wizard-of-Oz study in a driving simulator. We varied traffic density during a highway drive and included six in-car-specific use cases. The latter were presented by a proactive voice assistant and in a non-proactive control condition. We assessed the users' subjective cognitive load in a DALI questionnaire during the interaction with both PA variants. Additionally, the users rated their experiences in a SASSI questionnaire. The results show that proactive assistant behavior is rated similarly positive as non-proactive behavior. In line with previous research, the most driving-relevant use cases receive the best ratings.

1 Introduction

These days more and more people use Personal Assistants (PAs) via voice, such as Google Assistant and Amazon Alexa at home [1], Apple Siri and Microsoft Cortana on the smartphone [2], or Mercedes-Benz MBUX Voice Assistant and BMW Intelligent Personal Assistant in the car [3]. These are available on many different devices and offer convenient functionalities in different environments, such as setting reminders, navigating through traffic, or sending messages to friends and colleagues. While serving the users' needs, PAs constantly collect personal data in order to personalize their services and adapt their behavior. Adaptation needs not only to be performed towards the user, but also towards the situation, in which these PAs are used. Especially, when the user drives or is busy with another task at home (e.g., cooking), the interaction with a PA is only the secondary task. Thus, user experience designers need to focus on the user's cognitive load in such settings, too. [4, 5] In order to investigate how users perceive proactive voice output while driving, we conducted a Wizard of Oz study in a driving simulator with 42 participants. We varied traffic density during a highway drive to induce different levels of cognitive load. Furthermore, we permuted six in-car specific use cases and added a non-proactive control condition with the same six use cases. By employing a subjective DALI questionnaire [6] we assessed the users' cognitive load during the interaction with the two PA variants. Additionally, we let the participants rate both PA variants through the SASSI questionnaire [7] both while driving and afterwards. The results show that the proactive assistant behavior has been rated similarly positively as the non-proactive one, where users initiated the dialog. In line with previous research, the most driving-relevant use cases were rated the best.

2 Related Work

In this work, we take a look at proactivity in PAs, but relate it to the user's cognitive load during the interaction as well. [8] focusses on the proactive actions of robots, which are sometimes

related to the user's recognized intention. But here proactivity lies in the proactive planning or execution of tasks and does not contain proactive dialog behavior. Regarding the latter, Nothdurft et al. [9] declare appropriate interaction strategies for proactive dialogue systems as an open quest. L'Abbate [10] suggested in his dissertation how to model proactive behavior of conversational interfaces: He defined that the assistant takes over the initiative in problematic and unclear situations in a virtual risk management advisor scenario. Concerning cognitive load, Lindström et al. [11] have shown that there is an effect of cognitive load on disfluencies when the user speaks to in-vehicle spoken dialog systems. In [12] the topic is discussed in a broader manner, modeling driver-behavior and assessing distraction for these in-vehicle speech systems. Radlmayr et al. [13] present how traffic situations and non-driving related tasks (such as talking to a PA) affect the take-over quality in highly automated driving, whereas the works by Villing [5] as well as Fors and Villing [14] are exactly focusing on cognitive load while driving and talking to a dialog system or voice assistant. While Hamerich [15] did not take cognitive load into account, he presented proactive dialogs relying on the context of real-time traffic situations (at that time transmitted via TMC). Semmens et al. [16] performed an empirical study on the timing when a PA would be allowed to interact with the driver via speech, but did not research the proactive utterances or use cases as such. According to previous research by Schmidt et al. [17], proactivity and certain use cases that are closely related to tasks while driving are preferred by users during in-car HCI. Based on their findings and the prior work of Hamerich [15], we designed the usability study presented in this work. To the best of our knowledge, we are the first ones to systematically combine all areas: proactive voice assistant behavior, cognitive load, and the subsequent user acceptance during the interaction (secondary task) while driving (primary task). In this work, we decided to assess our subjects' cognitive load while driving in a subjective manner. For this purpose, we rely on the DALI questionnaire as introduced in [6]. Regarding evaluation, we are assessing both the proactive and the non-proactive assistant (control condition) ratings by means of the SASSI questionnaire [7].

3 Driving Simulator Study

In this study (cf. [18]), 42 subjects completed the entire experiment in the driving simulator. The distribution of sexes was almost even with 22 male (52.4%) and 20 female (47.6%) subjects. Their age averaged out on 43.7 years (range: 22 to 65 years). Table 1 shows the subjects' age distribution and their yearly driven kilometers. We balanced the distribution of yearly kilometrage among participants because driving was the primary task in the experiment. Driving habits could have influenced the subjects' perceived cognitive load, though we did not induce challenging driving maneuvers. As shown in Figure 2, the setup of the driving experiment consisted of a fixed-base simulator with a 180° screen in a room with controlled light and temperature conditions. The operator desk was located in the same room, but could not be observed while the participants sat on the driver's seat. Methodically, the study was designed as a two factor within-subject experiment. Figure 1 illustrates that each subject interacted with both a proactive (P) as well as a non-proactive (NP) voice assistant, separated by a short driving break in which the first assistant was rated. In between the interaction with each of the assistants, the traffic density was varied from low to high or vice versa. Consequently, every subject interacted with both assistants and experienced both traffic conditions during the respective interaction phases. The order in which the assistants and traffic conditions were presented was permuted, so that we created the following four different experiment procedure variants:
Variant 1: starting with NP and low traffic, switching to high traffic; switching to P while remaining in high traffic, ending with P and low traffic
Variant 2: starting with NP and high traffic, switching to low traffic; switching to P while re-

maining in low traffic, ending with P and high traffic
Variant 3: starting with P and low traffic, switching to high traffic; switching to NP while remaining in high traffic, ending with NP and low traffic
Variant 4: starting with P and high traffic, switching to low traffic; switching to NP while remaining in low traffic, ending with NP and high traffic
The subjects only knew that they were interacting with assistant A or B, but they did neither know about the current interaction type (NP or P), nor about the traffic condition.

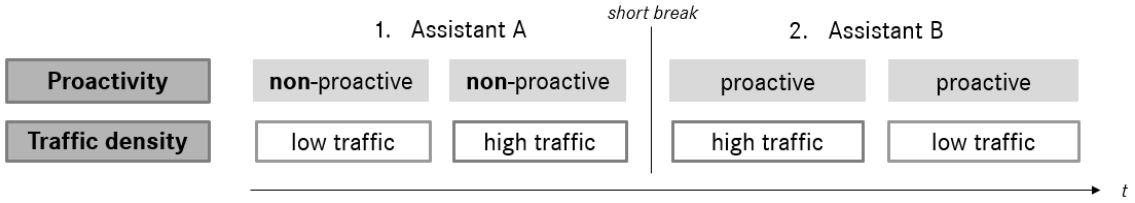

Figure 1 – Exemplary experiment procedure [18]

age / sex	10k	<20k	20k	>20k	>50k	km
20-29	2	2	3	1	0	16,5k
30-39	1	3	3	1	0	17,1k
40-49	1	2	3	3	0	23,8k
50-59	3	2	1	5	1	28,0k
60-65	2	2	0	0	1	28,2k
m	2	5	6	8	1	27,2k
f	7	6	4	2	1	18,1k

Table 1 – Subjects' kilometrage per year for age groups and gender.

Figure 2 – Driving simulator with operator desk.

The tested voice assistant variants were operated in a Wizard of Oz (WOz) setup, in which all potential dialog paths were modeled in a rule-based manner. We prepared up to four different possible responses, depending on the subject's input, and were able to repeat selected phrases, if subjects requested for it. In addition to a synthesized female assistant voice (same as in current Mercedes-Benz models), we also integrated a synthesized male voice to ask for ratings on subjectively perceived cognitive load (DALI) and system behavior (SASSI). *He* acted as a standardized virtual co-examiner. To further establish controlled conditions, we chose a WOz setup so that speech recognition performance cannot negatively influence our results on proactive voice assistant behavior. Furthermore, we avoided a proactive voice assistant in the wild, because we did not know whether it would be too cognitively demanding for any subject.

In the following we describe an exemplary experiment procedure of our driving simulator study (cf. [18]). It had three parts: briefing, main experiment (the drive), and debriefing. First, the examiner welcomed the subject and led them to the briefing room. Then the subject was informed about audio- and video-taping which was agreed by signing a respective form. Following this, the subjects should fill out a general questionnaire on experience with PAs, technical affinity, their own car etc. Afterwards, the subjects were led to the cabin and introduced to the car for the main experiment. The examiner informed about the video camera and the two-way intercommunication system inside the car. Furthermore, they offered assistance in case the subject needs help at any point during the study. The subjects were given driving instructions: stay on the right lane, drive around 110 km/h and follow the lead car, do not overtake. The examiner gave the Empatica E4 wristband to the subject and checked that it was worn correctly. After answering potential questions, the examiner took a seat at the operator's desk. They assured

Assistant Type	Interlocutor	Sample Dialogs
Non-Proactive	examiner	*Please express your request to refuel.*
	customer	Hey Mercedes, I need to refuel.
	vehicle	The next gas station is located at a highway service area in 10 kilometers. Should I navigate you there?
	customer	Yes, please.
	vehicle	Ok, I set the gas station as an intermediate stop.
Proactive	vehicle	Your remaining fuel range is 150 kilometers. Should I already search for a gas station for you?
	customer	Yes, please.
	vehicle	Ok, the next gas station is located at a highway service area in 30 kilometers. Should I navigate you there?
	customer	Yes, please.
	vehicle	Ok, I set the gas station as an intermediate stop.

Table 2 – Sample Dialogs [18]

that the subject can hear them (and vice versa), that the Empatica E4 was recording properly. The simulated car was situated in a service area next to a three lane highway. When the subjects were asked to start driving, they entered the highway with no other traffic (neither same nor opposite direction). After around one minute, the subject closed up to the lead vehicle, which they should follow at all times. It drove with a constant speed of 110 km/h. After around two minutes of the baseline drive the examiner reassured that the subject feels well (no motion sickness due to graphic projection). After this point the controlled experiment started and only the WOz assistant(s) talked to the subject for the remaining drive (exception: in the middle of the drive, when the subject stopped at a service area, the examiner checked again for the subject's well-being). Following the baseline drive (around five minutes), the traffic simulation started and cars in the same and opposite direction were shown. After the drive was finished, the examiner prepared the cabin and the simulation setup for the next subject. They took back the Empatica E4 wristband and led the current subject to the debriefing room. The examiner asked the subject to fill out a final short questionnaire about the usefulness of the presented use cases, and then saw them off. To manipulate the subjects' cognitive load, we varied the traffic density during the experiment. After the baseline drive without any traffic, the neural network traffic simulation was being started. Depending on the variant, it started with a low or high traffic condition. In the low traffic condition, 10 cars were simulated per 1 km on the three-lane highway. In the high traffic condition, 40 cars were simulated per 1 km on the three-lane highway. We determined these numbers experimentally, taking the average speed and speed variations during these situations into account which influence the subjects' level of exposure and the total time spent driving. If we would have increased the number of cars from 10 to more than 50, there would have appeared highly demanding braking situations when traffic slows down, comparable to a real "stop and go" traffic. Because this might have caused many motion sick subjects, we limited the high traffic condition to 40 cars per 1 km. Additionally to the traffic density, the traffic simulation included different types of drivers (excluding very aggressive ones). As described above, our subjects interacted with two different assistants In order to be able to compare both interactions to each other, we controlled the experiment by applying the same six use cases to the P and NP assistant, respectively (see examples in Table 2). Overall we presented the subjects five driving-related and one not driving-related use cases. Most driving-related use cases were close to the navigation domain, such as refueling or rerouting. The order in which the use cases were presented was permuted across subjects and variants. After three use cases, i.e. when either the assistant or the traffic condition was changed, the virtual co-examiner posed the same five SASSI and six DALI questions.

4 Results & Evaluation

4.1 User Satisfaction & Cognitive Load

In this section we present the ratings of user satisfaction as well as subjective cognitive load, elicited by means of SASSI and DALI questionnaire items. Figure 3 illustrates the mean SASSI ratings across the four auditory items *having fun* using the system, finding the system *useful*, finding it *boring*, or *feeling tensed* while using the system, per variant. It shows that the negatively connoted items *boring* and *tensed* got relatively low ratings on the 7-point Likert scale from *I do not agree at all* to *I totally agree*. Coherently, the positive items *fun* and *useful* were rated relatively high. Generally speaking, there are no noteworthy effects, but there is a significant rise of *fun* between variant 2 and 3 ($p < 0.05$), which is reversely reflected in the negative item *boring*. Figure 4 shows the SASSI item *useful* sorted by assistant/traffic condition. Apart from minor variations among variants, only the very positive rating of P high traffic in variant 3 is notable. It seems contradictory that subjects find it more *useful* using the proactive system during high traffic, but as we observed while conducting the study and as the DALI results show, most subjects were not highly loaded during high traffic. Furthermore, participants might favor a proactive assistant during high traffic over a non-proactive assistant because of its more efficient way of interaction. We conclude that especially participants in variant 3 liked the proactive assistant. As already indicated, the DALI questionnaire did not give us striking results. Both the mean ratings across variants (cf. Figure 5) and when breaking it down to assistants/traffic conditions do not show big differences between the ratings. Especially for the latter we would have expected more distinct results.

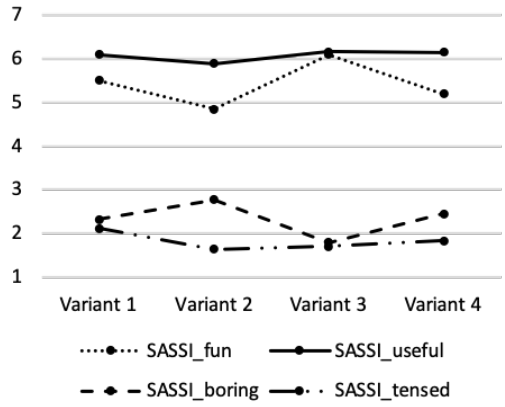

Figure 3 – SASSI mean ratings.

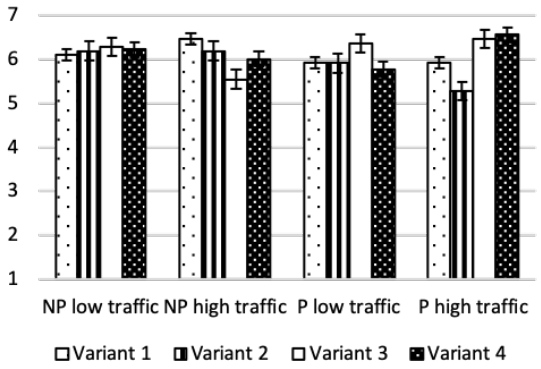

Figure 4 – SASSI item "useful" rated according to assistant type and traffic condition.

4.2 Use Cases

As shown in Figure 6, the subjects had a clear preference towards certain use cases. The two best rated use cases *Rerouting (4.86)* and *Refueling (4.69)* got no low ratings, i.e. 1 or 2 on a 5-point Likert scale. *Appointment (4.67)* was rated slightly lower and got rated between 2 and 5. While all three remaining use cases *Parking (4.10)*, *Break (3.57)*, and *News (2.81)* got rated on the full scale from 1 to 5, *Parking* was clearly the preferred use case among those three. While the suggestion to *take a break* because of car-detected tiredness of the driver was still perceived as a somewhat positive feature (probably because of safety reasons as shown in [17]), informing about *news* was not rated as positively with an average below scale mean. We assume the reasons for this are the following: first, especially in the non-proactive case, when the subjects should "inform [themselves] about news", it seems that the subjects expected a different kind of system behavior that has not been met by the assistant.

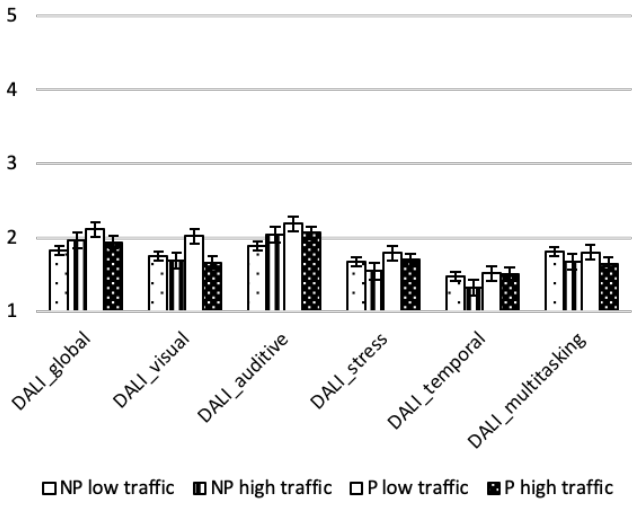

Figure 5 – DALI items rated according to assistant type and traffic condition.

	freq.	%
not important	4	9.5
rather not important	2	4.8
neutral	8	19.0
rather important	22	52.4
extremely important	6	14.3

Table 3 – Importance that a voice assistant makes proactive suggestions.

Furthermore, the examiners could observe that the subjects expected news about politics or other domains, but no news about Mercedes-Benz (new model or new battery factories). Finally, once subjects asked the assistant for news and got some answer, they assumed the respective assistant to be capable giving much more information and answering questions – which was not the case due to the WOz setup. In order to approve the subjects' preference of specific use cases, we performed the Wilcoxon Signed Ranks test for crossfold validation. The following use case relations are rated significantly different with $p < 0.00\overline{3}$ (calculated Bonferoni adjustment for 95% confidence interval): Rerouting to News, Parking and Break. Refueling to Break, News, and Parking. Appointment to News and Break. Parking to News, and Break to News.

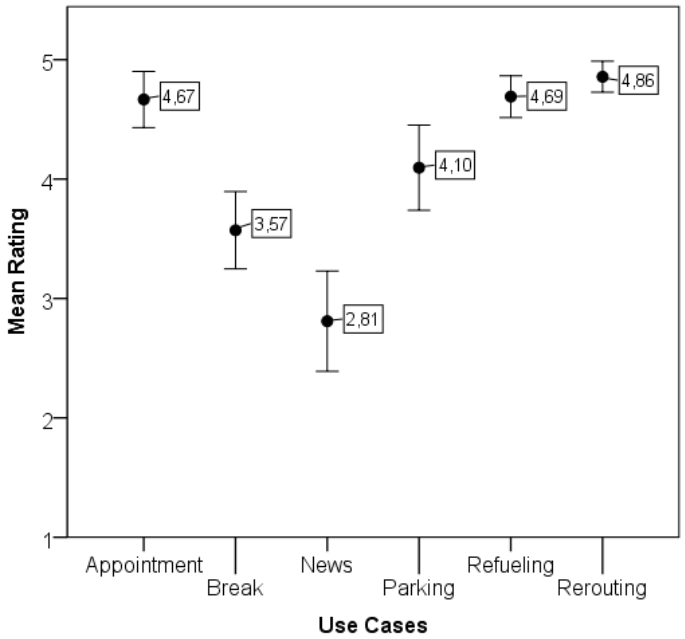

Figure 6 – Average ratings of the six different use cases.

4.3 Proactivity in General

In general, participants are satisfied with proactive suggestions by voice assistants. We can already derive this from the positive SASSI ratings presented beforehand, which in some cases

even were more positive than the ratings for the NP variants. To get a clear picture of the subjects' opinion on proactivity, we posed the following direct question in addition: *One of the two assistants you have experienced, has spoken to you unrequestedly (proactively). How important is it to you that a voice assistant makes suggestions by its own accord (proactively)?* The result is shown in Table 3: the majority of subjects responded that proactive suggestions are rather or extremely important to them. In the free text areas in the questionnaire, a few subjects wrote that proactive suggestions are the actual benefit for them and the assistant appears intelligent through these. As proactivity is a polarizing topic, we also asked the participants whether they wish to be able to deactivate proactivity in a voice assistant. The results show that only four participants do not wish for this option. Five participants wish to have a complete deactivation of proactive suggestions. The vast majority of 33 participants wishes to selectively switch proactivity on or off depending on the respective content, such as appointments, navigation etc.

5 Conclusion

In this work, we presented how users perceive proactive dialogs in a driving simulator WOz setting. As drivers are already cognitively occupied with the primary task of driving, proactively triggered interaction by the voice assistant has to remain unobtrusive to regard road safety. While the basic preconditions stayed the same among subjects, the order in which they were confronted with high or low traffic density varied. We assessed the users cognitive load by means of subjective DALI ratings as well as their user satisfaction by means of SASSI questionnaire items. The results show that proactivity in this context is at least equally likable as non-proactive interaction behavior while driving. At the same time the study subjects significantly rate that they would like to be able to deactivate proactivity for specific functionality (e.g., appointments, navigation etc.). The cognitive load measured by means of DALI items was not diverging at all between variants or assistant/traffic conditions. We conclude from these findings that though users want to deactivate proactivity, the majority sees it very positively while driving in a controlled condition with several different traffic densities. For future experiments we plan to implement a proactive assistant for a driving task taking place in the wild.

References

[1] MITTAL, Y., P. TOSHNIWAL, S. SHARMA, D. SINGHAL, R. GUPTA, and V. K. MITTAL: *A voice-controlled multi-functional smart home automation system*. In *2015 Annual IEEE India Conference (INDICON)*, pp. 1–6. IEEE, 2015.

[2] KEPUSKA, V. and G. BOHOUTA: *Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home)*. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference*, pp. 99–103. IEEE, 2018.

[3] BRAUN, M., A. MAINZ, R. CHADOWITZ, B. PFLEGING, and F. ALT: *At your service: Designing voice assistant personalities to improve automotive user interfaces*. In *2019 CHI Conference on Human Factors in Computing Systems*, p. 40. ACM, 2019.

[4] GABAUDE, C., B. BARACAT, C. JALLAIS, M. BONNIAUD, and A. FORT: *Cognitive load measurement while driving. in: Human factors: a view from an integrative perspective*. 2012.

[5] VILLING, J.: *Dialogue behaviour under high cognitive load*. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 322–325. Association for Computational Linguistics, 2009.

[6] PAUZIÉ, A.: *A method to assess the driver mental workload: The driving activity load index (dali). IET Intelligent Transport Systems*, 2(4), pp. 315–322, 2008.

[7] HONE, K. S. and R. GRAHAM: *Towards a tool for the subjective assessment of speech system interfaces (sassi). Natural Language Engineering*, 6(3-4), pp. 287–303, 2000.

[8] BUSS, M., D. CARTON, B. GONSIOR, K. KUEHNLENZ, C. LANDSIEDEL, N. MITSOU, R. DE NIJS, J. ZLOTOWSKI, S. SOSNOWSKI, E. STRASSER ET AL.: *Towards proactive human-robot interaction in human environments*. In *2011 2nd International Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 1–6. IEEE, 2011.

[9] NOTHDURFT, F., S. ULTES, and W. MINKER: *Finding appropriate interaction strategies for proactive dialogue systems – an open quest*. In *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, August 6-8, 2014, Tartu, Estonia*, no. 110, pp. 73–80. Linköping University Electronic Press, 2015.

[10] L'ABBATE, M.: *Modelling Proactive Behaviour of Conversational Interfaces*. Ph.D. thesis, Technische Universität, Darmstadt, 2007.

[11] LINDSTRÖM, A., J. VILLING, S. LARSSON, A. SEWARD, N. ÅBERG, and C. HOLTELIUS: *The effect of cognitive load on disfluencies during in-vehicle spoken dialogue*. In *Ninth Annual Conference of the ISCA*. 2008.

[12] ANGKITITRAKUL, P., D. KWAK, S. CHOI, J. KIM, A. PHUCPHAN, A. SATHYANARAYANA, and J. H. HANSEN: *Getting start with utdrive: Driver-behavior modeling and assessment of distraction for in-vehicle speech systems*. In *Eighth Annual Conference of the International Speech Communication Association*. 2007.

[13] RADLMAYR, J., C. GOLD, L. LORENZ, M. FARID, and K. BENGLER: *How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving*. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58, pp. 2063–2067. Sage Publications Sage CA: Los Angeles, CA, 2014.

[14] FORS, K. L. and J. VILLING: *Reducing cognitive load in in-vehicle dialogue system interaction*. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, SemDial*, pp. 55–62. Association for Computational Linguistics, 2011.

[15] HAMERICH, S. W.: *Towards advanced speech driven navigation systems for cars*. In *2007 3rd IET International Conference on Intelligent Environments*, pp. 247–250. 2007.

[16] SEMMENS, R., N. MARTELARO, P. KAVETI, S. STENT, and W. JU: *Is now a good time? an empirical study of vehicle-driver communication timing*. In *2019 CHI Conference on Human Factors in Computing Systems*, CHI '19. ACM, 2019.

[17] SCHMIDT, M., D. STIER, S. WERNER, and W. MINKER: *Exploration and assessment of proactive use cases for an in-car voice assistant*. In P. BIRKHOLZ and S. STONE (eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 148–155. TUDpress, Dresden, 2019.

[18] SCHMIDT, M., D. HELBIG, O. BHANDARE, D. STIER, W. MINKER, and S. WERNER: *Assessing objective indicators of users' cognitive load during proactive in-car dialogs*. In *27th Conference on User Modeling, Adaptation and Personalization Adjunct, June 9–12, 2019, Larnaca, Cyprus*. 2019.