At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces

A Real World Driving Study

Michael Braun
BMW Group Research
New Technologies, Innovations
LMU Munich
Munich, Germany
michael.bf.braun@bmw.de

Anja Mainz
LMU Munich
Munich, Germany
anja.mainz@campus.lmu.de

Ronee Chadowitz
BMW Group Research
New Technologies, Innovations
Munich, Germany
ronee.chadowitz@bmw.de

Bastian Pfleging
Technische Universiteit Eindhoven
Eindhoven, Netherlands
b.pfleging@tue.nl

Florian Alt
Bundeswehr University; LMU Munich
Munich, Germany
florian.alt@unibw.de

Figure 1: In a real world driving study ($N = 55$), participants experienced an in-car voice assistant in two different character manifestations: a personalized assistant based on a user's placement in the Big Five personality model, and a default character designed as generic agent to suit any driver. (Icons in this Figure © Dinosoft Labs.)

ABSTRACT

This paper investigates personalized voice characters for in-car speech interfaces. In particular, we report on how we designed different personalities for voice assistants and compared them in a real world driving study. Voice assistants have become important for a wide range of use cases, yet current interfaces are using the same style of auditory response in every situation, despite varying user needs and personalities. To close this gap, we designed four assistant personalities (Friend, Admirer, Aunt, and Butler) and compared them to a baseline (Default) in a between-subject study in real traffic conditions. Our results show higher likability and trust for assistants that correctly match the user's personality while we observed lower likability, trust, satisfaction, and usefulness for incorrectly matched personalities, each in comparison with the Default character. We discuss design aspects for voice assistants in different automotive use cases.

CHI 2019, May 4–9, 2019, Glasgow, Scotland, UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300270

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Automotive UI, Voice Assistant, Personalization, Personality.

ACM Reference Format:
Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces: A Real World Driving Study. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300270

1 INTRODUCTION

Voice assistants are becoming a pervasive means of interaction in everyday life [41]. A similar trend is apparent for automotive UIs [44]. Apart from minimizing driver distraction during manual driving [27, 39], speech interfaces also offer a more natural user experience (UX), compared to conventional UIs in cars [1], which is of particular interest in the transition towards automated driving.

Current voice assistants can understand natural language and express information through speech synthesis [41]. However, up to now, most assistants lack an inter-personal level of communication which is required to build relationships. Related research suggests that to become more widely accepted, such systems need to satisfy the expectations users have towards social interaction [28, 34]. It is so far unclear how personalized voice characters affect trust, UX, acceptance, and workload in the real world. To close this gap, we designed a set of personalized voice assistants and tested them in a real-world driving study ($N = 55$).

Our results show that personalized assistants are accepted when they match users' preferences. A personalized assistant leads to more trust and higher likability than the default character. For driving-related use cases, assistants require a more serious tone than for entertainment use cases. Therefore, we endorse personalization for in-car voice interfaces, as the driver workload was comparable among all assistant characters and we see a potential for improvements on trust and UX, both for manual and automated driving. Our work is complemented by a discussion of design considerations, meant to support the design of future voice assistants.

Contribution Statement. We contribute insights from a real world driving study (N=55) where we compared the influence of personalized voice assistants on trust, UX, acceptance, and workload to a non-personalized voice assistant.

2 RELATED WORK

This work builds upon research on natural language interfaces and their potential benefits on automotive UIs.

Natural Voice User Interfaces

Conversational interfaces, or voice assistants, that is interfaces which are operated by voice input and respond with synthesized speech [41], have become ubiquitous through smart phones and home automation [41]. Prominent examples are Amazon Alexa[1], Apple's Siri[2], or Google's Assistant[3].

Such systems are easy to use as no commands have to be learned. They are more efficient than conventional screen-based interfaces, less error-prone [33], and can offer accessibility to users with disabilites [42]. However their unlimited range of possible inputs also facilitates a gulf between user expectations and actual capabilities [5, 24].

Assistant behavior can be classified as reactive (listen and answer), proactive (initiate conversation, e.g., for reminders), and social (e.g., talk to someone else on your behalf) [47, 53].

Voice Assistants in The Car

Voice interfaces have been shown as valuable alternative input modalities for automotive user interfaces [40, 44, 45]. Drivers mainly utilize visual and manual cognitive resources for the driving task, without extensively straining vocal and auditory channels [52]. This can be used to optimize voice interfaces for limiting overall cognitive load, preventing effects thereof, such as inattentional blindness [21, 23, 54]. In-car voice assistants however also bare risks when demanding verbal interactions coincide with critical driving tasks [50, 51].

In the near future, the omission of the steering wheel in automated vehicles might lead to an increased distance between driver and dashboard, making haptic interactions less convenient than speech. Social implications could also favor voice assistants, as users can hand over responsibilities [46].

The current landscape of commercial in-car voice assistants comprises Google's Assistant via Android Auto, Apple's Siri via CarPlay, Mercedes-Benz' MBUX[4], NIO NOMI[5], BMW IPA[6], and many more upcoming systems, e.g. by Ford, Toyota[7], Honda, Hyundai, or Peugeot.[8]

Personalization of User Interfaces

Many interfaces today, and especially intelligent assistants, incorporate features of personalization. These range from knowing the user's name, to content customizations based on models of needs and behaviors [19, 49]. Such systems can also act as social players by proactively pointing out information [36, 47] or by helping users to accept new technologies, e.g., by mimicking their behavior in automated driving [37]. Personalization can also help maintaining attachment to cars when ownership and driving are things of the past [7].

[1]https://developer.amazon.com/alexa, last access: 2018-09-15
[2]https://www.apple.com/siri/, last access: 2018-09-15
[3]https://assistant.google.com, last access: 2018-09-15
[4]https://www.mercedes-benz.com/a-class/com/en/mbux, 2018-09-15
[5]https://www.nio.io/es8, last access: 2018-09-15
[6]https://www.press.bmwgroup.com/global/article/detail/T0284429EN/
[7]https://www.toyota.com/concept-i, last access: 2018-09-15
[8]https://soundhound.com/powered-by-houndify, 2018-11-26

Designing Personality For Digital Agents

Humans are quick on first impressions, be it with humans or with digital systems [35]. Immediate assessments of personality helps us decide whether we aim to converse with an opponent and it allows us to adjust our expectations [8]. Digital assistants benefit from a consistent personality as it helps users to predict their behavior [2].

A widely recognized approach for the classification of personalities is the Big Five model by McCrae and Costa, consisting of openness, conscientiousness, extraversion, agreeableness, and neuroticism (OCEAN) [29]. Extraversion is the most prevalent dimension in HCI studies as it has high informative value and is easy to observe [18].

For the design of artificial personalites, we can build upon the similarity-attraction hypothesis. It states that humans like to interact with others of similar personality [35]. Furthermore, digital assistants and users need a shared understanding of acceptable behavior [16] and the assistant must not be too human-like, to avoid uncanny experiences [32].

In related work, Bickmore & Picard show a relational agent capable of social and emotional interaction, which was evaluated with high ratings for trust and likability [4]. Nass et al. applied a similar concept to a simulated driving context and found increased driving performance and attention if an emotional voice assistant is matched to drivers in a similar state [34]. We are building on the results of this work by exploring personalized voice assistants in real world settings.

3 RESEARCH APPROACH

To investigate personalized voice assistants for driving scenarios, we first collect requirements for personality traits of in-car assistants. We then designed a set of assistants which we finally evaluated in a real-world driving study.

4 CHARACTER DESIGN

In the following section, we report on the design process of the four different voice assistants for in-car speech interfaces that we evaluated in a subsequent study.

Our initial design of eight assistants is based on Argyle's two-dimensional 'Model of Attitudes towards Others' [3]. The model includes the dimensions *Dominant – Submissive* and *Hostile – Friendly* as basic pillars of inter-personal communication. We designed assistants to cover the different dimensions of the model. Following a technique proposed by Luria [26], assistants' personalities are based on fictional characters. In particular, the characters and their respective placement[9] were: Sherlock Holmes (D, H), Sheldon Cooper (D, mH), HAL9000 (mH, mS), Marvin the Android (mH, S), Jack Sparrow (D, F), Baloo the Bear (mD, F), Donkey from Shrek (mS, F), and Ron Weasley (S, F).

[9]D = dominant, S = submissive, H = hostile, F = friendly, m = medium

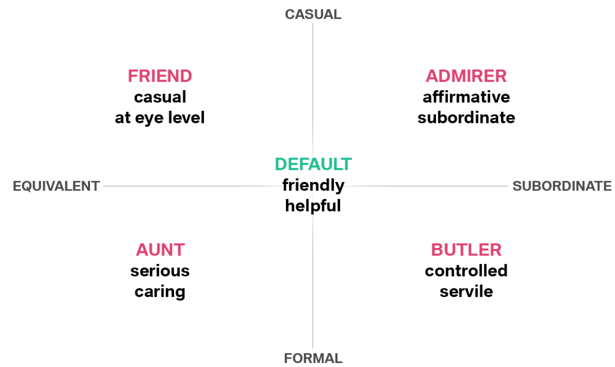

Figure 2: We derived a two-dimensional model from prior work and our pre-study. Each one assistant was designed to match the four dimensions.

Pre-Study

We conducted a pre-study to obtain a better understanding about the required and undesirable personality traits for an in-car assistant. We invited 19 participants (12 male) aged 19-53 years (M=35, SD=11). Most were non-HCI researchers.

Apparatus. We designed six scenarios, three related to driving and three related to entertainment. Each scenario contained a specific task, such as asking the assistant for the nearest gas station. Participants could engage in a dialogue with each of the eight speech assistants. The assistants' responses were pre-recorded by a voice actress for each assistant and reflected the respective placement in the model.

Procedure. Participants were invited to our lab. They experienced all six scenarios with each voice assistant, adding up to 48 total interactions. The order was counter-balanced. After each assistant, we had participants fill in a MeCue and an Acceptance Scale questionnaire [20, 31]. At the end we conducted a semi-structured interview with participants.

Results. Results from the MeCue and Acceptance Scale questionnaires [20, 31] and personal interviews identify unfriendly behavior and excessive talking as negative traits, while assistants with a perceived friendly attitude were liked by most participants. The data shows a dissent on the desired levels of distance between assistant and user, the assistant's professionalism (i.e., how respectful it behaves towards the user), and the balance of power within the conversation.

From the feedback of our pre-study, we can assume a hostile assistant as unsuitable. As a result, we excluded such characters for further investigation. On the dominance scale (balance of power), we identified notions of preferences towards *equivalent* and *subordinate* characters. We also decided to introduce the dimension 'professionalism', with *casual* and *formal* being the two ends of the scale.

The resulting model facilitated the design of four plausible characters for an in-car voice assistant (see Figure 2).

Final Characters

All characters were designed in cooperation with a screenwriter. We focused on credible and distinguishable features. The traits are expressed through choice of words and intonation, while content and extent of speech output were identical for all characters to ensure comparability. All audio snippets used in the study were recorded by a professional voice actress. Audio samples are available online.[10]

We provide a brief description of each assistant.

Friend. This character was designed to exhibit a casual conversational tone while at eye level with the user. She has fun being a co-driver and lightens the mood with her wittiness.

Admirer. She also has a casual conversation tone but is designed to be subordinate towards the user. This character is affirming and almost praises the user's decisions.

Aunt. The aunt character is a rather formal instance, who behaves familiar with the user. She cares deeply about the user's well-being and takes things serious.

Butler. This character is designed to be subordinate and neutral. She delivers facts and follows orders.

In addition to the four characters that fit the dimensions in Figure 2, we designed a fifth character, Default, as a trade-off. This character is supposed to be suitable, albeit not perfect, for a majority of users. We use the Default as a comparable standard assistant with the same capabilities but less distinct personality features than the previously defined characters.

Default. She is correct in what she does but not too technical, and neither like a subordinate nor like a friend.

Assigning Assistants

As a final step, we came up with an approach to match users to one of the four characters. We conducted a user study ($N = 31$), where participants (aged 19–62 years, $M = 35$, $SD = 12.45$) first underwent a 60-item Big Five Inventory questionnaire [12]. We then introduced them to the four characters we had designed in the previous step. They were told about the character's qualities by the examiner and could listen to a set of recorded voice samples. Finally, they voted on their favorite character and how they would place a perfect voice assistant within our model dimensions.

A decision tree analysis [43] was used to identify a distribution of assistant characters towards user personality traits. Neuroticism had the highest information gain and was thus used as first layer, branching off to conscientiousness, agreeableness, and openness, which were further separated in a third layer. To name an example, neurotic users' most probable match is the Aunt, as they might prefer the caring attitude. Very open participants likely match with the Friend, as they might appreciate its casual way of communicating. Interactions between dimensions are of course more complex than these examples and can lead to divergent assignments.

The presented approach was used in the subsequent study to assign participants to assistants.

5 REAL WORLD DRIVING STUDY

We performed a driving study in realistic traffic conditions ($N = 55$) to investigate voice assistant personalities. The main question we answer with this experiment is whether the personalization of voice assistants, based on user personality traits, has advantages in comparison to a default characters, as is common in today's vehicles.

Hypotheses

We assume positive effects of personalization on user experience with limitations caused by high workloads in certain driving situations. Our assumptions are based on literature and can be summarized in seven hypotheses. Following best practices discussed at CHI '18 [10], we pre-registered our hypotheses prior to the study.[11]

H1: A default one-fits-all character is generally accepted by users.

H2: A personalized character scores higher ratings for trust and user experience, compared to the default assistant (cf. [37]).

H3: A personalized character leads to more expressions of positive emotions compared to the default assistant.

H4: A less serious character leads to higher workloads than a serious character (cf. [50, 51]).

H5: Users would pick the character adapted to their personality over other characters or the Default(cf. [37]).

H6: Users prefer less emotional characters for driving-related tasks, and emotional ones for non-driving-related tasks (cf. [51]).

H7: User extraversion correlates with the estimation of extraversion of their preferred character [35].

Study Design

This study used assistant personality as a between-subject independent variable. Four personalities (Friend, Admirer, Aunt, Butler) were tested against a baseline character (Default). Additionally, three use case clusters (Driving Related, Proactive Assistant, Connected Car) were deployed as independent variables within each ride.

As dependent variables we collected user personality traits (Big Five Inventory [12]) and subjective ratings on user experience (UEQ modules Attractiveness and Stimulation [22],

[10] http://www.drivingstudy.de/audiosamples

[11] Pre-registered online at https://aspredicted.org/blind.php?x=5p2ta6

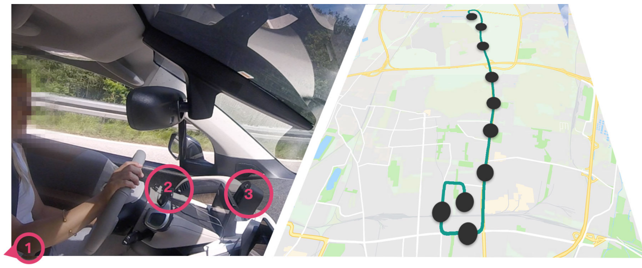

Figure 3: Left: the voice interface was controlled by a wizard on the backseat (1), two cameras streamed data to an emotion recognition software and provided the wizard with a view of the instrument cluster (2), the voice feedback was output through loudspeakers (3). Right: Experiment route. Map data ©2018 GeoBasis-DE/BKG, map view ©2009 Google

one-item likability scale), acceptance (Acceptance Scale [20]), trust (one-item, c.f. [14]), and workload (Driving Activity Load Index [38]). We furthermore analyzed the driver's facial expressions regarding displayed emotions using the Affdex SDK [30] and assessed the propriety of the used matching algorithm, i.e., whether or not the recommended character matched with the preferred character. At the end of the experiment, participants answered questions on the experienced characters in a semi-structured interview and assessed the perceived personalities using a semantic differential rating.

Participants

55 participants aged 23–60 years (M=41.3, SD=10.2) took part in the study (45 male, 10 female). None of them had expert knowledge on digital assistants, though more than half have used one in their everyday life (6 daily, 8 often, 14 rarely, 26 never). The majority of the participants were familiar with voice controls, in general, and within an in-car systems.

Apparatus

We conducted the experiment as a Wizard-of-Oz study [11] with the participants driving a real car and the operator sitting in the back. The operator watched the surroundings and the driver's behavior in order to trigger the experiment use cases in appropriate situations.

The operator was equipped with a tablet computer running the experiment interface from which audio files could be played on loudspeakers in the front of the car. Two cameras were positioned on top of the steering wheel (see Figure 3). One was connected to the system interface and allowed the operator to watch the instrument cluster and thus, react to, e.g., the activation of the distance control system. The second camera was directed at the driver's face and supplied a data stream to the emotion recognition system.

The experiment took place in an upper-middle class car on a road section in Munich, Germany. The route allowed for a ca. 5 minute familiarization ride on private grounds before entering public streets. Thus, participants could become accustomed with the car itself before being introduced to the assistant. The road started in an inner-city area, passing a construction site with speed limits, and several lateral parking spots, before leaving the city on a state road. The route also provided sightseeing spots at both ends. Speed limits on the route ranged from 40 to 100 km/h and participants took between 16 and 24 minutes for a one-way ride, depending on traffic. The total distance for both ways was 22 km. The construction site, parking spaces, and sightseeing spots were used as triggers for the experiment use cases.

Procedure

Several days before the experiment took place, participants answered an online Big Five Inventory questionnaire which was used to select a fitting assistant character (see Section 4).

At the site, we introduced participants to the concept of an intelligent voice assistant and to the procedure. They answered general questions on demographics and signed a declaration of consent. After a familiarization ride on private property, the experiment commenced on public roads. Each participant experienced two rides: one with a personalized assistant (Friend, Admirer, Aunt, or Butler) and one with the Default assistant. The order of experienced characters was alternated between participants to prevent sequence effects.

On the road, 12 use cases were triggered by the operator at certain locations. The use cases were split into three clusters: Driving Related (e.g., using the automatic parking function when passing an empty lateral spot), Proactive Assistant (e.g., offering sport mode when leaving the city limits), and Connected Car (e.g. getting information on a sightseeing spot nearby), as well as a welcome and farewell message by the assistant. After each use case, participants rated the interaction verbally as good, neutral, or bad.

After each ride, participants answered the questionnaires described above, giving their feedback for the experienced character. At the end of the final ride, participants listened to key statements from all five characters and decided which character they would like to use in future conversations. In the end, the examiner held a short semi-structured interview (ca. 5 minutes) focusing on appropriate and inappropriate situations with respect to the three use case clusters.

Limitations

All study participants were BMW Group employees. This guaranteed that participants were covered by insurance. We expect no major effects on the results as the demographic data was reasonably heterogeneous.

The small number of women in the study is a limitation which could not be avoided during recruiting. Related work suggests that women are on average demonstrating higher levels of interpersonal sensitivity than males [9]. However this does not necessarily affect personality preferences [25].

Task	Cluster	Content
-	On-boarding	Assistant welcomes driver
1	Proactive Assistant	Speed limit at construction site
2	Driving Related	Automatic parking function
3	Connected Car	Set a reminder
4	Driving Related	Activate distance control
5	Driving Related	Deactivate distance control
6	Proactive Assistant	Assistant offers sport mode
7	Driving Related	Deactivate sport mode
8	Connected Car	Check lights at home
9	Connected Car	Information on nearby area
10	Proactive Assistant	Assistant marvels at surroundings
11	Connected Car	Parking information
12	Proactive Assistant	Reminder: bag in trunk
-	Off-boarding	Farewell

Table 1: Four tasks in each ride belonged to one cluster (Driving Related, Proactive Assistant, Connected Car).

The matching of personalities based on the user's personality led to an uneven distribution of assigned characters. This could not be avoided as recruiting based on a personality test would have violated union rules on personal data privacy. We could also not let participants chose a character by themselves as this would have given them too much information in advance and thus jeopardized the data integrity. We excluded one between-group because of the low number of participants (3), the other groups consist of 21, 16, and 15 data sets, which qualify as solid basis for statistical analysis.

6 RESULTS

We used a decision tree to assign characters: Of the 55 participants, 21 were matched to the character Friend, 16 to the Butler, and 15 to the Aunt, leaving three participants who experienced the Admirer. As such a small sample cannot be used to generate statistically meaningful assertions, we are omitting the results for this character in direct comparisons.

We divide the data set into two groups based on whether the matching algorithm suggested the same character as the participant chose as his/her favorite at the end of the experiment. In total, 16 of 55 participants chose the suggested personalized character, 27 preferred the Default, and 12 decided for another personalization than proposed by the matching algorithm. All values reported in Table 2 were collected on a scale from -3 (low) to +3 (high).

Trust

Trust in digital systems is an essential prerequisite for their long-time adoption, especially in the automotive context where automation is expected to disrupt the industry [14]. Pairwise comparisons show the Default character as significantly more trustworthy than Friend, Aunt, and Butler (T-test, $p < .05$). The Friend character was judged least trustworthy, Aunt and Butler only marginally differ (Table 2).

	Friend		Aunt		Butler		Default	
	M	SD	M	SD	M	SD	M	SD
Trust	0,00	1,73	1,07	1,44	1,56	1,09	**2,21**	0,85
Attractiven.	0,38	0,88	**0,44**	0,83	0,35	1,49	**0,44**	0,82
Stimulation	1,12	0,95	1,16	0,91	0,89	1,08	**1,17**	0,91
Likability	0,05	1,93	0,93	1,49	0,00	2,19	**1,94**	1,21
Usefulness	0,86	0,59	1,23	0,77	1,29	0,87	**1,70**	0,81
Satisfaction	-0,24	0,56	0,79	1,04	0,62	1,35	**1,63**	0,91
Mental W.	**1,32**	1,49	0,50	1,40	1,00	1,32	1,15	1,50
Auditory W.	**1,26**	1,52	0,71	1,07	1,06	1,34	1,19	1,47
Interference	**1,89**	0,94	1,43	1,40	1,25	1,06	1,65	1,14
Stress	1,79	1,44	1,21	1,31	2,00	1,03	**2,06**	0,98

Table 2: User ratings for experienced assistant characters Friend, Aunt, Butler, and Default. Mean values and standard deviations on a scale from -3 to +3.

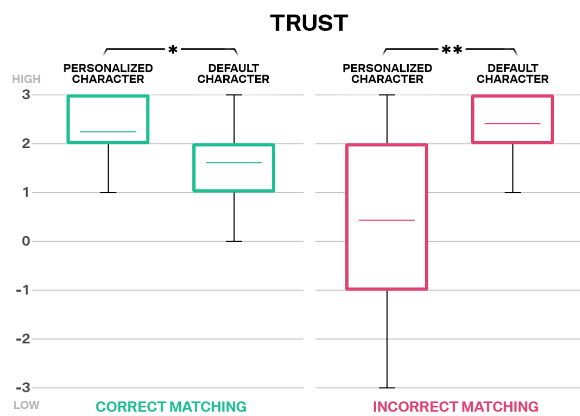

Figure 4: Personalized characters were trusted more than the Default assistant when the matching algorithm was right and less when it was assessed as wrong. $^{*}p < .05$; $^{**}p < .01$

Separated by efficacy of the matching algorithm, we can see in Figure 4 that the personalized characters were trusted more than the Default assistant when the matching was correct and less when the algorithm was wrong.

User Experience (UX)

Participants provided subjective UX evaluations via a questionnaire incorporating the modules Attractiveness and Stimulation of the UEQ, and a one-item scale on likability from [22].

Attractiveness & Stimulation. All evaluated characters were assessed as medium attractive to use without noteworthy preferences. Stimulation was rated neutral (Butler) to good (Friend, Aunt, Default, see Table 2).

Likeability. In direct comparison and over all participants, the Default character is rated most likable ($p < .05$), with no significant differences between all other characters. In case of correct matching however, personalized characters were rated significantly more likable than the Default, as shown in Figure 5. Vice versa, the characters were rated poorly when incorrectly matched.

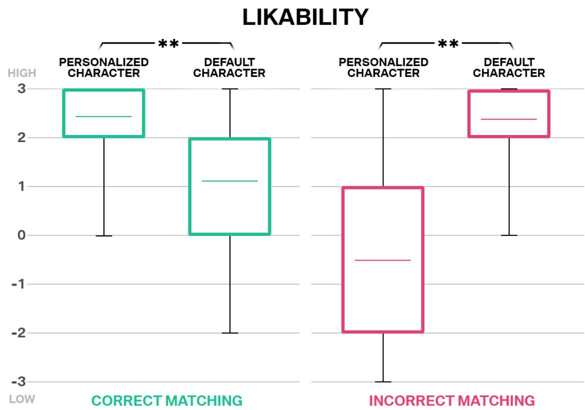

Figure 5: Likability scores were at a maximum both for personalized characters when the matching was correct, and for the Default character when participants felt incorrectly matched. $^{**}p < .01$

Acceptance

The Van der Laan Acceptance Scale [20] was used to measure perceived usefulness and satisfaction of voice assistant personalities. None of the characters was rated as useless ($M < 0$) by participants and only the Friend personality is seen as mildly unsatisfying (see Table 2). The Default character scores best in direct comparison, Aunt and Butler are assessed similarly.

Figure 6 shows that personalization has no benefit on usefulness or satisfaction when characters are correctly matched. However, in a mismatched situation, personalized characters where evaluated as significantly less useful and less satisfying than the Default assistant.

Workload

We used the Driving Activity Load Index [38], a specialized version of the Nasa Task Load Index [13] for the automotive domain, to assess participants' required mental and auditory workload towards the voice assistant, the interference of speech interaction with the driving task, and perceived stress caused by the system. An analysis of variance showed no significant results for any character (mean values and standard deviations are listen in Table 2). This suggests that personalized characters were perceived as comparably suitable for in-car use cases as the Default voice assistant.

Emotion Recognition

We evaluate the output of the Affdex facial expression detection system [30] for a 10-second time frame after each interaction. This duration was chosen as literature states emotional responses to auditory stimuli are processed against preexisting expectations within less than 10 seconds after

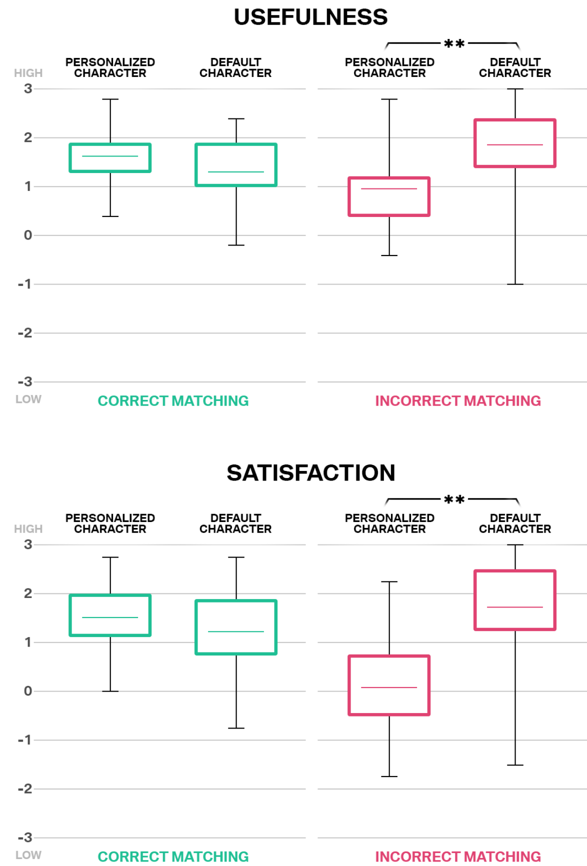

Figure 6: Incorrectly matched characters were rated as less useful and satisfying as Default. Values for correct matchings show no differences between the characters. $^{**}p < .01$

	Friend		Aunt		Butler		Default	
	M	SD	M	SD	M	SD	M	SD
Valence	-5.17	22.53	-5.86	21.92	-5.28	21.66	-5.67	21.81
Joy	5.63	15.01	5.29	14.46	5.60	14.74	5.61	14.82
Anger	0.40	1.76	0.40	1.76	0.34	1.38	0.35	1.43
Surprise	10.47	14.92	9.76	13.34	9.92	13.37	9.94	13.61
Engag.	35.08	28.88	34.37	28.32	34.75	28.01	35.17	27.97

Table 3: Emotion detection values for a 10-second time frame after experiencing the characters.

interaction [6, 17]. An ANOVA and pairwise comparisons (T-test) showed significant differences for measurements of joy between personalized characters and the Default ($p < 0.05$) when the matching was incorrect (increased joy with a personalized character). We observed a comparable tendency for correctly matched characters. However, Table 3 shows that all measures except those for Engagement are very low (scales 0–100) and with high deviations. Thus, we refrain from drawing conclusions based on these measurements.

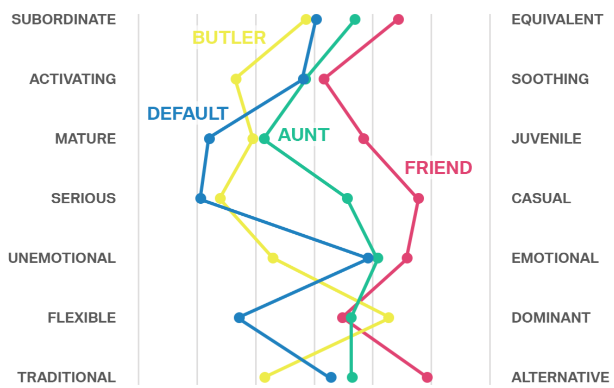

Figure 7: Semantic differential scale of perceived assistant personalities, showing participants were able to correctly distinguish between the different characters.

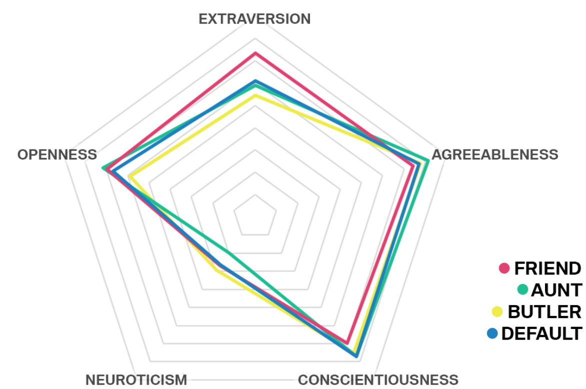

Figure 8: Mean Big Five scores for participants, split by their preferred character. Users who chose the Friend were significantly more extraverted than all other groups.

Personality Interplay

In the final interview, participants assessed the experienced characters using a semantic differential scale with 7 dimensions. Figure 7 shows that most characteristics were perceived as designed, only the Butler was rated as more dominant than the designers had intended.

We also asked participants to assign an estimation of extraversion to their experienced characters. Extraversion is the personality dimension which is easiest to be observed in social interactions and literature describes it as most revealing factor of personality [18]. The Default (27.5) was assessed as least extraverted (scale 0–100), followed by the Butler (56.3) and the Aunt (66.3). The Friend (73.8) was perceived as most extraverted.

Figure 8 shows the mean personality trait scores of participants, grouped by their preferred character. In pairwise comparison, participants who chose the Friend character scored significantly higher in extraversion than all other groups ($p < .05$), confirming the similarity-attraction hypothesis mentioned by Nass et al. [35]. Other dimensions did not yield substantial differences. The low scores on neuroticism are striking but constant throughout groups.

Use Cases

Participants experienced 12 use cases from 3 clusters (Table 1) with each character. In the interview, they reflected on these use cases by listening to the relevant audio snippets again and gave feedback on how their preferred assistant should have behaved. We extracted recommendations for each use case through thematic analysis and from task ratings.

On-boarding & Off-boarding. The first and last contact with the character occurred at the beginnings and end of each drive. They were not part of any use case cluster, we instead used them to express the assistant's personality with no further information provided. Both personalized and Default characters were positively rated for these types of use cases. This suggests that an assistant can express more of its full personality during on-boarding and off-boarding, especially when the car is not in motion.

Driving Related Tasks. Tasks directly related to driving, e.g., activating certain driving modes, are highly relevant for the user's security. Participants rated the Friend character as least appropriate and unanimously called for serious and short statements. In less crucial situations like sideways parking, a supportive stance as displayed by the Aunt was sought-after. Feedback should be kept affirmative and minimal. To name an example from the study, the order to deactivate the sports mode should be answered with an "okay" instead of "deactivating sports mode".

Connected Car Tasks. Modern cars possess not only multimedia functionalities but also connect to online services like calendars or smart homes. We combine these kinds of tasks into one cluster. Participants accepted the personalized characters for these situations, opposed to driving related tasks. On direct comparison, participants preferred the Friend character for her lively presented answers. Feedback suggests the assistant should behave like a co-driver who can answer quick questions in a pragmatic way ("Did I leave the lights on at home?" – "Yes, you did." A presentation of personality is accepted, as long as it is authentic.

Proactive Assistant Tasks. When voice assistants initiate conversations without prior input, we call this proactivity. Feedback on proactive tasks was mixed. Many participants opposed unsolicited speech output. If appropriate, personalized characters were preferred. Interview feedback suggests the assistant should act like an authentic, human co-driver.

7 DISCUSSION

Our results show that the correct matching of assistant characters to the user's personality is a crucial prerequisite for positive effects of personalization. We reflect on how we can improve in-car voice interaction through personalization.

Robust State-Of-The-Art

The Default character represents a voice assistant character as implemented in state-of-the-art systems. Lots of research went into the design of these assistants and we can confirm that the character was trusted and reasonably well liked with values in the upper regions of both scales (H1). We consider it safe to say that such a neutral assistant is a good starting point to introduce assistant personality to users.

Personalization Gone Wrong, or: Let Them Choose Their Co-Driver

When assistant and user are incorrectly matched, the personalized assistant is accepted less than the Default. This should by no means surprise us: the assistant acts as a co-driver and thus can easily cause displeasure. A connection between user and assistant extraversion could be observed, which suggests to accept the similarity-attraction hypothesis (H7). The proposed matching algorithm, however, performed badly with only 16/55 participants (29%) confirming its suggestion, which refuses us from confirming H5. The matching algorithm was calculated on a data set of 30 participants, which might have been too little input for this kind of classification. We also found that certain dimensions do not significantly affect character preferences (see Figure 8), by including them we added unnecessary noise to the data.

Fortunately, the need to assign characters pre-experiment only resulted from experimental purposes, as we had to conceal the true nature of the characters from participants. In a production setting, the assistant's personality could be adapted not only by implicit system decisions but also explicitly by the user, through settings or direct commands: "Hey Assistant, stop being so nice all the time."

As novice users are prone to forgo customization as long as a system works [48], some degree of personalization initiated by the system or encouragement for users could be beneficial. Incremental adaptations by the system can go hand in hand with user-initiated personality selection.

The User Benefits of a Personalized Assistant

The display of personality features was rated most appropriate in settings where the driver was not preoccupied with the primary task of driving. We can therefore confirm the hypothesis that non-driving-related situations (infotainment, connected car) are most suitable for personalized interaction, while security-relevant driving functions need to be delivered unemotionally (H6).

Participants who experienced a correctly matched personalized character trusted and liked this character more than the Default character, which suggests to accept H2. This motivates further investigations of assistant personalization, on one hand because it can improve acceptance and with that the adaptation of voice assistants which can help reducing the cognitive load of the driver [45]. On the other hand, non-driving-related activities will increase with the introduction of automated cars and affect take-over performance [15]. Here we can design a more natural interaction for a better UX and eventually steer passengers' attention to optimize situational awareness for eventual take-over requests.

How do we Get This on the Road?

In order to pave the way for an actual realization of personalized voice assistant characters in consumer vehicles, we need to build an environment which can asses user preferences (either implicitly or through direct input) to incrementally adapt the personality, and judge context information to determine the safety risk involved in the road situation. We used an emotion detection approach to monitor the driver's emotions. However, based on the recorded data, we cannot clearly accept H3 (more positive emotions with personalized characters). An evaluation of perceived mental and auditory load, interference, and induced stress, however, showed comparable results for personalized and default assistants. Thus, we fail to confirm that driver workload and with it, driving performance, are impaired by personalization (H4).

8 CONCLUSION AND OUTLOOK

In this paper we explored the influence of personalized voice assistant characters on UX, acceptance, trust, and workload compared to a non-personalized assistant. The results show that personalization has a positive effect on trust and likability if the voice assistant character matches the user's personality. However, a mismatch can cause displeasure.

Future voice assistants need to adapt to the user as well as to the environment. Our findings provoke the personalization of voice assistants not only on a cultural level as it is already happening, but on a context-aware basis, e.g. being short and precise in driving-related situations, and getting chatty on an empty road when the driver is bored. A neutral assistant is recommended as starting point before gradually adjusting its personality to the user's needs, either through implicit system decisions or explicit user input. Successive work should look into how an optimal assistant personality can be chosen for the driver. Furthermore, it would be interesting to investigate how to transfer our findings to driving settings with higher automation levels.

REFERENCES

[1] Ignacio Alvarez, Hanan Alnizami, Jerone Dunbar, Andrea Johnson, France Jackson, and Juan E. Gilber. 2011. Designing Driver-Centric Natural Voice User Interfaces. In *Adjunct Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. online, 156–159.

[2] Elisabeth André, Martin Klesen, Patrick Gebhard, Steve Allen, and Thomas Rist. 2000. Integrating models of personality and emotions into lifelike characters. In *Affective interactions IWAI 1999. Lecture Notes in Computer Science*, A. Paiva (Ed.). Vol. 1814. Springer, Berlin, Heidelberg, 150–165. https://doi.org/10.1007/10720296_11

[3] Michael Argyle. 1988. *Bodily communication* (second ed.). Routledge, New York, NY.

[4] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-term Human-computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327. https://doi.org/10.1145/1067860.1067867

[5] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant. *Journal of Medical Internet Research* 20, 9 (04 Sep 2018). https://doi.org/10.2196/11510

[6] Emmanuel Bigand, Suzanne Filipic, and Philippe Lalitte. 2005. The time course of emotional responses to music. *Annals of the New York Academy of Sciences* 1060, 1 (2005), 429–437. https://doi.org/10.1196/annals.1360.036

[7] Michael Braun, Anna-Katharina Frison, Sarah Theres Völkel, Florian Alt, Heinrich Hussmann, and Andreas Riener. 2018. Beyond Transportation: How to Keep Users Attached When They Are Neither Driving nor Owning Automated Cars?. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18)*. ACM, New York, NY, USA, 175–180. https://doi.org/10.1145/3239092.3265963

[8] Angelo Cafaro, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2016. First Impressions in Human–Agent Virtual Encounters. *ACM Trans. Comput.-Hum. Interact.* 23, 4, Article 24 (Aug. 2016), 40 pages. https://doi.org/10.1145/2940325

[9] Meanne Chan, Katherine H. Rogers, Kate L. Parisotto, and Jeremy C. Biesanz. 2011. Forming first impressions: The role of gender and normative accuracy in personality perception. *Journal of Research in Personality* 45, 1 (2011), 117 – 120. https://doi.org/10.1016/j.jrp.2010.11.001

[10] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More: On the Preregistration of CHI Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 141, 12 pages. https://doi.org/10.1145/3173574.3173715

[11] N. Dahlbäck, A. Jönsson, and L. Ahrenber. 1993. Wizard of Oz studies – why and how. *Knowledge-Based Systems, Special Issue: Intelligent User Interfaces* 6, 4 (1993), 258 – 266. https://doi.org/10.1016/0950-7051(93)90017-N

[12] Daniel Danner, Beatrice Rammstedt, Matthias Bluemke, Lisa Treiber, Sabrina Berres, Christopher J. Soto, and Oliver John. 2016. *Die deutsche Version des Big Five Inventory 2 (BFI-2)*. Technical Report. Zusammenstellung sozialwissenschaftlicher Items und Skalen. https://doi.org/10.6102/zis247

[13] Sandra G. Hart and Lowell E. Stavenland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[14] Sebastian Hergeth, Lutz Lorenz, Roman Vilimek, and Josef F. Krems. 2016. Keep Your Scanners Peeled: Gaze Behavior as a Measure of Automation Trust During Highly Automated Driving. *Human Factors* 58, 3 (2016), 509–519. https://doi.org/10.1177/0018720815625744

[15] Oliver Jarosch and Klaus Bengler. 2019. Is It the Duration of the Ride or the Non-driving Related Task? What Affects Take-Over Performance in Conditional Automated Driving?. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.). Springer International Publishing, Cham, 512–523.

[16] Malte F. Jung. 2017. Affective Grounding in Human-Robot Interaction. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. ACM, New York, NY, USA, 263–273. https://doi.org/10.1145/2909824.3020224

[17] Patrik N Juslin and Daniel Västfjäll. 2008. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences* 31, 5 (2008), 559–575. https://doi.org/10.1017/S0140525X08005293

[18] Lara K. Kammrath, Daniel R. Ames, and Abigail A. Scholer. 2007. Keeping up impressions: Inferential rules for impression change across the Big Five. *Journal of Experimental Social Psychology* 43, 3 (2007), 450 – 457. https://doi.org/10.1016/j.jesp.2006.04.006

[19] Joseph Kramer, Sunil Noronha, and John Vergo. 2000. A User-centered Design Approach to Personalization. *Commun. ACM* 43, 8 (Aug. 2000), 44–48. https://doi.org/10.1145/345124.345139

[20] Jinke D. Van Der Laan, Adriaan Heino, and Dick De Waard. 1997. A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies* 5, 1 (1997), 1 – 10. https://doi.org/10.1016/S0968-090X(96)00025-3

[21] 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. https://doi.org/10.1145/3003715.3005408

[22] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work*, Andreas Holzinger (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 63–76.

[23] Nilli Lavie. 2010. Attention, Distraction, and Cognitive Control Under Load. *Current Directions in Psychological Science* 19, 3 (2010), 143–148. https://doi.org/10.1177/0963721410370295

[24] Ewa Luger and Abigail Sellen. 2016. `Like Having a Really Bad PA`: The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[25] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[26] Michal Luria. 2018. Designing Robot Personality Based on Fictional Sidekick Characters. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. ACM, New York, NY, USA, 307–308. https://doi.org/10.1145/3173386.3176912

[27] Jannette Maciej and Mark Vollrath. 2009. Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention* 41, 5 (2009), 924 – 930. https://doi.org/10.1016/j.aap.2009.05.007

[28] Bertram F Malle and Stuti Thapa Magar. 2017. What Kind of Mind Do I Want in My Robot?: Developing a Measure of Desired Mental

Capacities in Social Robots. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, US, 195–196.

[29] Robert R McCrae and Paul T Costa. 1986. Personality, coping, and coping effectiveness in an adult sample. *Journal of Personality* 54, 2 (1986), 385–404. https://doi.org/10.1111/j.1467-6494.1986.tb00401.x

[30] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. 2016. AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 3723–3726. https://doi.org/10.1145/2851581.2890247

[31] Michael Minge and Laura Riedel. 2013. meCUE - Ein modularer Fragebogen zur Erfassung des Nutzungserlebens. In *Mensch & Computer 2013: Interaktive Vielfalt*, Susanne Boll, Susanne Maaßü, and Rainer Malaka (Eds.). Oldenbourg Verlag, München, 89–98.

[32] M. Mori, K. F. MacDorman, and N. Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics Automation Magazine* 19, 2 (June 2012), 98–100. https://doi.org/10.1109/MRA.2012.2192811

[33] M. Nafari and C. Weaver. 2013. Augmenting Visualization with Natural Language Translation of Interaction: A Usability Study. *Computer Graphics Forum* 32, 3pt4 (2013), 391–400. https://doi.org/10.1111/cgf.12126

[34] Clifford Nass, Ing-Marie Jonsson, Helen Harris, Ben Reaves, Jack Endo, Scott Brave, and Leila Takayama. 2005. Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1973–1976. https://doi.org/10.1145/1056808.1057070

[35] Clifford Nass, Youngme Moon, B. J. Fogg, Byron Reeves, and Chris Dryer. 1995. Can Computer Personalities Be Human Personalities?. In *Conference Companion on Human Factors in Computing Systems (CHI '95)*. ACM, New York, NY, USA, 228–229. https://doi.org/10.1145/223355.223538

[36] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[37] Dennis Orth, Nadja Schömig, Christian Mark, Monika Jagiellowicz-Kaufmann, Dorothea Kolossa, and Martin Heckmann. 2017. Benefits of Personalization in the Context of a Speech-Based Left-Turn Assistant. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '17)*. ACM, New York, NY, USA, 193–201. https://doi.org/10.1145/3122986.3123004

[38] A. Pauzie. 2008. A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems* 2, 4 (December 2008), 315–322. https://doi.org/10.1049/iet-its:20080023

[39] Matthias Peissner, Vanessa Doebler, and Florian Metze. 2011. *Can voice interaction help reducing the level of distraction and prevent accidents.* Technical Report. Nuance Communications.

[40] Bastian Pfleging, Stefan Schneegass, and Albrecht Schmidt. 2012. Multimodal Interaction in the Car: Combining Speech and Gestures on the Steering Wheel. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '12)*. ACM, New York, NY, USA, 155–162. https://doi.org/10.1145/2390256.2390282

[41] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[42] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. Accessibility Came by Accident: Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 459, 13 pages. https://doi.org/10.1145/3173574.3174033

[43] J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (01 Mar 1986), 81–106. https://doi.org/10.1007/BF00116251

[44] Andreas Riener, Myounghoon Jeon, Ignacio Alvarez, and Anna K. Frison. 2017. Driver in the Loop: Best Practices in Automotive Sensing and Feedback Mechanisms. In *Automotive User Interfaces: Creating Interactive Experiences in the Car*, Gerrit Meixner and Christian Müller (Eds.). Springer International Publishing, Cham, 295–323. https://doi.org/10.1007/978-3-319-49448-7_11

[45] Florian Roider, Sonja Rümelin, Bastian Pfleging, and Tom Gross. 2017. The Effects of Situational Demands on Gaze, Speech and Gesture Input in the Vehicle. In *Proceedings of the 9th ACM International Conference on Automotive User Interfaces and Vehicular Applications (AutomotiveUI '17)*. ACM, New York, NY, USA, 9. https://doi.org/10.1145/3122986.3122999

[46] Sonja Rümelin, Peter Siegl, and Andreas Butz. 2013. Could You Please...? Investigating Cooperation In The Car. In *Adjunct Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '13), Eindhoven, The Netherlands, October 2013. (AutomotiveUI '13)*.

[47] Maria Schmidt and Patricia Braunger. 2018. A Survey on Different Means of Personalized Dialog Output for an Adaptive Personal Assistant. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP '18)*. ACM, New York, NY, USA, 75–81. https://doi.org/10.1145/3213586.3226198

[48] Jacob Solomon. 2016. Heterogeneity in Customization of Recommender Systems By Users with Homogenous Preferences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4166–4170. https://doi.org/10.1145/2858036.2858513

[49] Shashi Thakur. 2016. Personalization for Google Now: User Understanding and Application to Information Recommendation and Exploration. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 3–3. https://doi.org/10.1145/2959100.2959192

[50] Jessica Villing. 2009. 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 (SIGDIAL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 322–325. http://dl.acm.org/citation.cfm?id=1708376.1708423

[51] Mark Vollrath. 2007. Speech and driving - solution or problem? *IET Intelligent Transport Systems* 1, 2 (June 2007), 89–94. https://doi.org/10.1049/iet-its:20060052

[52] Walter W. Wierwille. 1993. Demands on driver resources associated with introducing advanced technology into the vehicle. *Transportation Research Part C: Emerging Technologies* 1, 2 (1993), 133 – 142. https://doi.org/10.1016/0968-090X(93)90010-D

[53] Michael Wooldridge and Nicholas R Jennings. 1995. Intelligent agents: Theory and practice. *The Knowledge Engineering Review* 10, 2 (1995), 115–152. https://doi.org/10.1017/S0269888900008122

[54] Baoshi Yan, Fuliang Weng, Z. Feng, F. Ratiu, M. Raya, Y. Meng, S. Varges, M. Purver, A. Lien, T. Scheideck, B. Raghunathan, F. Lin, Rohit Mishra, Brian Lathrop, Zhaoxia Zhang, Harry Bratt, and Stanley Peters. 2007. A Conversational In-car Dialog System. In *Proceedings of Human Language Technologies: Demonstrations (NAACL-Demonstrations '07)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 23–24. http://dl.acm.org/citation.cfm?id=1614164.1614176