When Users Assist the Voice Assistants: From Supervision to Failure Resolution

Dounia Lahoual
EDF R&D - Palaiseau, France
dounia.lahoual@edf.fr

Myriam Fréjus
EDF R&D - Palaiseau, France
myriam.frejus@edf.fr

ABSTRACT

We conducted an in situ study of six households in domestic and driving situations in order to better understand how voice assistants (VA) are used and evaluate the efficiency of vocal interactions in natural contexts. The filmed observations and interviews revealed activities of supervision, verification, diagnosis and problem-solving. These activities were not only costly in time, but they also interrupted the flow in the inhabitants' other activities. Although the VAs were expected to facilitate the accomplishment of a second, simultaneous task, they in fact were a hindrance. Such failures can cause abandonment, but the results nevertheless revealed a paradox of use: the inhabitants forgave and accepted these errors, while continuing to appropriate the vocal system.

CCS CONCEPTS

Voice User Interfaces, User Studies, Human-Machine Interaction

KEYWORDS: Voice control; Smart home; smart speakers; video-ethnography of users; family life; users' skills; driving.

1 INTRODUCTION

The massive arrival of voice assistants (VAs) has made voice the next form of natural interaction for human-machine interfaces (HMIs) [2,6]. At the heart of key services related to comfort, safety, entertainment and health, the development of voice control poses real challenges for designers,

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK.

ACM ISBN 978-1-4503-5971-9/19/05. https://doi.org/10.1145/3290607.3299053

ACM Reference format: D. Lahoual & M. Fréjus. 2019. When users assist the voice assistants: from supervision to failure resolution. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA. 8 pages.
https://doi.org/10.1145/3290607.3299053

especially in terms of improving the usability and appropriability of the devices [8, 5]. According to a recent study by the consulting firm Ovum, nearly 2,770,000 digital VAs have been installed worldwide, including smartphones, tablets and smart speakers[2] [3]. Yet, although advertising campaigns encourage user hopes and expectations about the fluidity, simplicity and speed of use, VAs clearly remain problematic, and several studies have detailed the set of difficulties that users encounter.

Maintaining the illusion of fluid and natural interactions has tended to create a gap between user expectations and the real capabilities of these devices, which are often the source of experiences of suboptimal use and frustration. Not only are users unable to use natural language effortlessly, they are also required to do additional work to understand the reasons for malfunctions and attempt to repair control failures [8,6].

To shed more light on this problem in a variety of situations, this article focuses on this additional form of user activity in domestic and driving situations: *What types of activities are carried out to supervise and control VAs? What are the impacts on the other activities that are also being carried out? How do users perceive this supervision activity? What recommendations for interaction design can we offer?*

To answer these questions, we present a qualitative and user-centered study oriented toward understanding the contexts of VA use (Siri, Google Assistant, Mini Google Home, Amazon Echo Show and Echo).

2 USING VOCAL INTERACTIONS: A VARIETY OF DIFFICULTIES ENCOUNTERED

As vocal interactions are closely related to human conversation, they create expectations and beliefs among users, who may, for example, attribute intelligence and abilities to VAs that they do not have. Researchers have thus demonstrated that VAs produce an illusion of natural conversation that is translated into activity through difficulties in interaction: users expect that VAs will infer the context for these interactions based on previous commands and actions [6,8]. To overcome the frequently encountered voice recognition errors, users engage in processes of "repairing and constructing meaning in the interaction" [8], which results in a continuous investment in adapting to the system through, for example, reformulation, simplification or hyper-enunciation [7,6,8]. The failures to interact fluidly and efficiently with the system are likely to weaken the trust relationship that is being constructed and cause users to reduce their scope of action to simple tasks with low risk of failure [6].

[1] This number includes the following digital assistants: Apple's Siri, Google Assistant, Amazon's Alexa, Microsoft's Cortana, Samsung's Bixby and the Chinese assistants, Baidu and iFlytek

Table 1: **Participant characteristics and collected data** (* **participants interviewed)**

	Family composition	**Socio-professional category**	**Smart speakers & frequency of use**	**Data collection**
H1	- Husband (33)* - Wife (29)* - Daughter (18 M)	- Computer executive (M) - Housewife (F)	- Google Home (GH) Mini - JBL speaker designed to be compatible avec GH ➔ the couple uses them regularly on a daily basis (lighting, music, news, pollution checks)	Interviews: -Semi-directive (2h01) - Self-confrontation (1h48) - Videos (28h10)
H2	- Husband (52)* - Wife (44)* - Daughter (16 Y)	- Director of information systems (M) - Strategic analyst (F)	- Amazon Echo Show - Amazon Echo ➔ the couple uses them regularly on a daily basis (lighting, entertainment, scheduling, weather, Wikipedia)	Interviews: - Semi-directive (1h58) - Self-confrontation (1h25) - Videos (16 h)
H3	- Husband (41)* - Wife (42)* - Son (10 Y)* - Daughter (9 Y)*	- IT project manager (M) - Marketing director (F)	- Siri ➔ the husband uses Siri occasionally in his car (only) for managing music and calls and dictating text messages	Interviews: - Semi-directive (2h38) - Self-confrontation (1h55) - Videos (45h30)
H4	- Husband (46)* - Wife (44)* - Son (14 Y)* - Daughter (9 Y)	- IT project manager (M) - Pre-school assistant (F)	- Google Assistant ➔ only the adolescent uses it often on a daily basis for starting his applications, entertainment (music, humor)	- Semi-directive interview (2h31)
H5	- Husband (54) - Wife (50)* - Son (26 Y) - Son (23 Y)	- Project manager (M) - Computer engineer (F)	- GH Mini - Google Assistant ➔ the couple rarely uses GH (time, web searches, news), the wife uses the assistant often to dictate text messages	- Semi-directive interview (1h03)
H6	- Husband (38)* - Wife (34) - Daughter (3 Y) - Son (1 ½ Y)	- Computer scientist (M) - Jurist (F)	- Siri ➔ only the husband uses it to manage calls/entertainment in the car an in the bathroom (controlling lights and shutters)	- Semi-directive interview (1h48)

The difficulties have been noted in various spheres of use. In driving situations, the vocal interactions that are supposed to help drivers keep their eyes on the road and their hands on the steering wheel can instead create problems. When this occurs, most drivers turn to manual control to carry out actions [9]. Far from reducing distractions while driving, we thus see that voice commands can trigger a variety of failures. On the system side, this may be a voice recognition failure requiring clarification or the choice of one of several proposed options. The device may erroneously interpret a command and trigger the wrong response. The system may also stop working if it suddenly loses or has a weak internet or Bluetooth connection. Users, on the other hand, may issue commands prematurely or neglect to respond within the allotted time [9].

Despite these problems in interaction, however, several studies have reported the overall great satisfaction of users [5]. Luger and Sellen [6] offered greater nuance to these results by pointing out that users with more technical knowledge had lower expectations of the system and were therefore more likely to forgive errors.

To shed light on the difficulties and their impact on user activities, we documented domestic situations and highlighted their dynamic, collective and diachronic aspects. This temporal approach prompted us to explore the continuity of activities in other spheres, such as driving situations. We next present our method of data collection.

3 METHOD

We first conducted semi-structured interviews with our 13 participants, 7 men and 6 women (6 households). The interviews took place after the participants had been using the VA for about 6 months, depending on the household. Their ages ranged from nine to 52 years old and the user profiles were varied (regular and occasional VA users, technical and non-technical profiles, those with and without projects to install ecosystems of connected objects in the home). Participants were recruited via ads posted on forums and social networks. It should be noted here that this sample size is standard for qualitative and exploratory studies of this type [8, 5].

We then installed mini-video cameras in several areas of three volunteer homes (H1, H2, H3): the living room, kitchen, hallway, and car interior. By doing so, we were able to observe the use of VAs and connected objects in natural situations. The recordings were made during time periods co-selected with the participants and representative of family activities and routines (mornings and evenings, in the week and on the weekend). Last, video clips considered relevant from the researcher's viewpoint (interactions with the VA, difficulties of use, disruption of activities) were selected and then viewed with the participants so that they could make first-person comments on their activities during the self-confrontation interviews. After the interview verbalizations were transcribed, we carried out inductive and thematic content analysis by categorizing the verbalizations based on recurring themes.

4 RESULTS

4.1 Users willing to assist the voice assistants

Far from being fluid and simple to use, VAs provoke control and "assistance" activities when the system malfunctions, which can occur for several reasons. First, the system may fail in voice recognition (accents, background noise, ambiguous commands). It may be unable to respond to a complex command or be limited in its understanding of keywords and syntax at the beginning of use. Loss of the wi-fi/internet connection can also be extremely disruptive when it is not identified or identifiable, with users continuing their vain attempts at voice command. All of these challenges require users to engage in additional VA control activities: supervision and verification, followed by diagnosis and problem-solving.

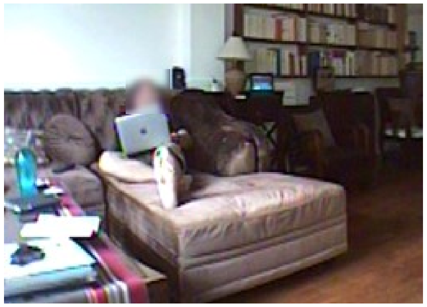

Figure 1. At the end of a weekday, an inhabitant turns to Alexa and asks it to repeat a definition while he's working on his computer (52-year-old man using Amazon Echo Show in the living room, Home 2).

Supervision and verification activity: This activity has several steps. It emerges during voice commands and consists of visual and auditory information-gathering and physical involvement with the VA throughout the activity. In the first step of visual and auditory verification, users check whether the assistant is activated to be sure of the right moment for making a request: they observe whether the VA is "listening" and processing the voice command. They then visually check whether the VA has finished “speaking” to decide whether or not to continue the command (Fig. 1). If the VA's response suggests voice recognition failure or a wrong interpretation of the command, the users begin the second step of language rectification. This step is usually the second or even third attempt (or more) to improve voice recognition: they may have to repeat, rephrase, clarify by changing phrases or keywords, enunciate more carefully, speak more loudly or more slowly, or shorten the command. The third step of physical involvement emerges when users engage their bodies, through movement, posture and gaze: they turn toward, approach and look closely at the VA, seeking new clues when the two previous steps have not been sufficient to complete the voice command (Fig. 2). This third step suggests an upcoming diagnostic activity because it has caused an interruption in the course of domestic activity. It requires a complete reorientation of activity in order to focus exclusively on the failure. These steps show how the device brings about activities that are no longer exclusively vocal (two tasks at once are no longer possible here) and requires an exclusive focus on supervision and verification.

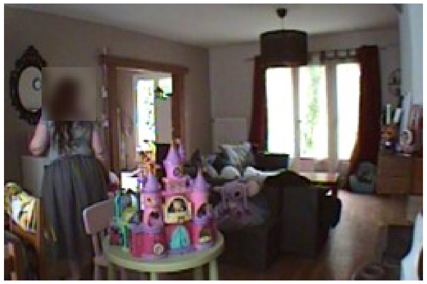

Figure 2. In full morning preparation in front of a mirror (weekday), a resident turns completely to the speaker. She is about to approach it to collect clues to help her understand the failure of a command for music (*29-year-old woman using a JBL speaker in the living room, Home 1*).

Diagnosis and problem-solving activity: This activity occurs when the supervision and verification activity has failed to achieve the desired voice command and/or the users have failed to identify the source and nature of the error. This is therefore a post-command activity of making a diagnosis to better understand and explore the types of failure that may arise within the system, in the environment, or with the users themselves. It is also a problem-solving activity as users explore potential solutions. It may involve other devices or other forms of interaction, such as using a smartphone to check the history of the conversation, check that the router is working, or renew the request via a phone application (Fig. 3). This activity thus consists of making hypotheses, testing them and perhaps looking for new solutions. Note that with these activities, users develop knowledge-in-action and gain a form of expertise about their devices coupled to the environment.

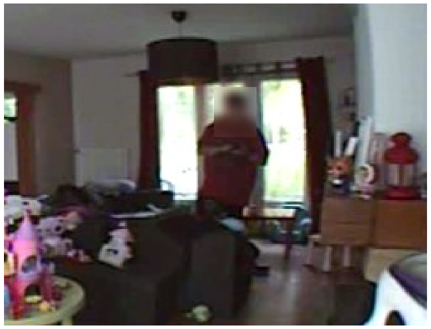

Figure 3: In the early evening (weekday), an inhabitant stops vacuuming when his request for music fails. He immediately uses his smartphone to start the music with Google Play *(33-year-old man using a JBL speaker in the living room, Home 1).*

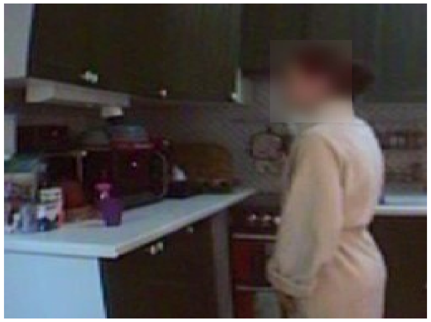

Figure 4: While preparing breakfast (weekend), an inhabitant moves closer to Alexa, which has stopped playing the radio. Not understanding what happened, she decides to turn it off and turn on her regular radio that is just near it (*44-year-old woman using Amazon Echo, Home 2*).

This enables them to map the system capabilities and limitations in order to delineate their scope of action in future uses or even to make the decision to change the device to better meet their needs (Fig. 4): ***"I noticed, finally we investigated and figured out there was interference. When Alexa was near the microwave, there, it systematically crashed (...) While I was heating up things for breakfast, it would keep stopping. So I gave up and in the morning, I now use my regular radio"*** (Interview French, H2).

It should also be noted that these diagnostic activities may fail, weakening the trust relationship and the ongoing appropriation process. In the absence of support during diagnosis, users can be left with the feeling that all is obscure. They may well lack adequate information about the system's functioning, capabilities and limitations: "***The first time we asked a trick question, it was a dumb thing, my husband wanted to know the results of the Formula 1 race and the VA couldn't answer (...) But when we tried again some time later, two or three months, and well, there we had the answer (...) So maybe they take advantage of questions they're asked to improve themselves. Is that it or was it the wording that wasn't good? It's hard to know...***" (Interview excerpt, 50-year-old woman using the Google Home Mini, H5).

Among the users observed and interviewed, we noted that diagnosis most often was followed through to the end with a positive outcome when it was a collective activity (4/6 households: H1,2,3,5). In fact, exchanging ideas and testing actions with one or more members of the family increases the chances of identifying the nature of the error or succeeding at voice control. In an excerpt of the interaction between a father and Siri (Fig. 5, H3), we observe his two children (9 and 10 years) intervene and help him in the activities, with 2 minutes 24 seconds of collective supervision and diagnosis before the command was successfully made. The father later explained that he was used to Siri failing at times and that it works better with others. He speculated that the malfunction was due to a poor wi-fi connection but that his children had experienced the whole episode as a family game.

4.2 Impacts of these assistance activities: obstacles, disruptions and abandonment

As opposed to the expectations of ease of use, vocal interactions lead to unanticipated control and assistance activities. These activities disrupt and interrupt the flow of activity, leading to control that is no longer exclusively vocal but instead tends to involve a wider variety of supports and multimodal interactions (smartphones; older dedicated devices like conventional radios or paper shopping lists; and visual, auditory and physical interactions). Users begin to focus exclusively on the control activity, making it the top priority. The disadvantage of assistance activities is thus that they prevent the users from carrying out another task simultaneously, although this has been touted as one of the best attributes of VAs. Assistance activities can even become real risks in driving situations, when attention to the road is disrupted because the drivers have tuned in to visual information as well or they have resumed manual control on their smartphones. Some users (4/6 households: H2,3,5,6) perceive this experience as quite negative: stressful, annoying, trying. It can be discouraging, as, for example, when the VA shows no improvement (and therefore learning) in response to a difficulty recognizing accents during voice commands in English.

Figure 5: Collective diagnosis excerpt in the family (Home 3)

- **(Father)** I have to call mom. “*Hey Siri, call P. on her cellphone*” (Siri beeps) **14 seconds later**: “*Call P. on her cellphone*”
- **(Son)** 4 seconds later: “Siri, call P. on her cellphone”
- **(Daughter)** 10 seconds later: “*But you have to press the button*” (button vibrating on the interface). “*Siri, call P.*” (Siri beeps again) “*Siri, call P. on her cellphone*”
- **(Father)** 6 seconds later: “*Be quiet*” (speaking loudly, carefully enunciating, with an impatient tone of voice and a sharp intonation): “*Hey Siri! Call P. on her cellphone!*” 4 seconds later (Siri beeps again)
- **(Daughter)** 3 seconds later: “*Have to, have to close the window maybe*” (10 seconds later): “*So go on, do it again*”
- **(Father)** 3 seconds later: “*Hey Siri!*” 4 seconds later
- **(Siri)** “*Sorry I missed that*”
- **(Father)** 4 seconds later: “*Well, sorry, no harm done*” (4 seconds later) (with a disillusioned tone of voice and slow and careful enunciation): “*Hey Siri, call P. on her cellphone*” (4 seconds later Siri beeps again) 3 seconds later: “*Or not*” (the father gives up and lets his children keep trying) (...)”

All of these experiences push users to turn conventional device: turning on the radio instead of Amazon Echo, drawing up a shopping list on paper instead of putting the list on Google Home, using the remote control to open the garage rather than Siri, and so on. These activities thus weaken the trust relationship that is rather painfully being built with these new devices. The process of appropriation can even be impeded to the point where the use of a type of service or a device drops off or is outright abandoned. This was the case, for example, when a user perceived the disadvantage of being forced to visually check a message on her phone while driving because the system did not offer message replay: ***“I don’t use it anymore in the car, it got me a nice ticket (laughs). Because the problem with having a dictation assistant is that we tend to write messages at traffic lights. The light turned green and I kept going, I got caught. The cops saw that I was dictating an email on the phone”*** (Interview excerpt, 50-year-old woman using Google Assistant, H5).

4.3 Paradox of use: forgiving and accepting mistakes as a condition for appropriation

Nevertheless, most users accept these failures and tend to forgive system errors (8/13 participants: 4 men, 3 women and 1 adolescent). We identified seven reasons: (1) The error is local and isolated to a specific type of service and this is counterbalanced by good system functioning for other services. (2) Users accept the failures as they expect to see future improvement. (3) They have incorporated the VA into their collective routines so fully that they have adapted to technical failures: getting rid of the VA would be more restrictive, transforming the appreciated routines: “***When we went on vacation I took a speaker. So she (his wife) was talking to the speaker and she told it to play some music when it couldn’t, since it was not connected to the internet (...) It becomes a bit of an automatic reflex***” (Interview excerpt, 33-year-old man using a JBL speaker, H1).

Errors and their correction can also become an integral part of the relationship with the VA: (4) Users explain that when they first started using the VA, they accepted its errors more easily when they were able to understand what was not working correctly in the system. (5) This indulgent attitude is reinforced when a space for regaining control is possible, especially via the conversation history. A 14-year-old user says he is not bothered when his smartphone VA malfunctions. He can instantly fix errors by looking at the dialogue bubbles. Moreover, he states that even when he loses all his data because he has bought a new smartphone – and thus has lost a personalized service based on the history of interactions – this failure does not bother him because he sees it as a new opportunity to develop his VA. (6) Error acceptance can therefore go as far as becoming involved in training the VA: “***The more I use it, the more it understands what I’m doing and then there isn’t this problem anymore (...) There it started up again from zero with my new phone. And I like to know that I made it do more, I made it grow. It's kind of a game***” (Interview excerpt, 14-year-old boy using Google Assistant, H4).

(7) Moreover, recognizing that the VA has a form of intelligence with the ability to evolve also contributes to error acceptance because the system is perceived as being in the process of learning: “***We’ve seen progress in understanding commands, it understands French better (...) Even if we give a command in shorter form or a little differently (...) I think that it’s got a bigger vocabulary, a bigger lexicon (...) It’s also more precise***” (Interview excerpt, 44-year-old woman using the Amazon Echo Show and Echo in French, H2).

5 DISCUSSION: INTELLIGIBILITY, NATURAL INTERACTION AND SYSTEM INTELLIGENCE

This study has shown that the participants' difficulties using the VA were mainly related to intelligibility issues. The results demonstrate the importance of strengthening the system's ability to identify and detail the problems in vocal interaction and especially to make them visible to users. It is therefore important to design vocal interactions around problem-solving during use. This would prevent spending too much time on assistance activity and turning to other media or forms of interaction. For example, the system might indicate that it no longer has a wi-fi/internet connection or that it has not understood when the user repeats or reformulates a request that fails.

As to the users, they might directly resolve the problem by interacting with the system: "What did you understand of my request?" Identifying the type of failure is not only important to ensure the users' appropriation and better use of the device: it is also decisive for ultimately promoting the continuous improvement in the learning algorithms. Furthermore, intelligibility involves providing users with sufficient information, appropriate to the context and at the right time, on system operation [1, 4] (specific lights according to the type of failure, voice messages, message on the application, etc.). This information concerns the state of the system, how it works, what it understands or doesn't about commands, what it knows about a situation, how it makes decisions, what it keeps "in memory" of past interactions and then reuses (understanding of the context and users), and what it is unable to do and can potentially do. Thus, associating intelligibility and system intelligence would ultimately improve natural interactions.

Another important step is to provide users with visual confirmation before an action is sent or launched to reassure them of the system's effectiveness during critical or complex tasks. In contexts where vocal interaction would be the ideal modality, as in voice dictation tasks, the VA might orally repeat the dictated message to ensure them that the dictation is error-free, thus keeping the interaction entirely vocal, for example, when the users are driving. Last, we found that users were more likely to forgive voice system errors when they had a better understanding of how the system worked and when they had a variety of options for taking charge. And we especially observed that a relationship of trust and a process of appropriation both tend to be established over time when users perceive signals of VA intelligence. The examples given here mainly concern the system's evolution and learning. But this might also be extended to its ability to understand the user context, anticipate needs, and act proactively.

6 Conclusion

Our study has shown that the use of VAs leads to additional supervision, verification, diagnosis and problem-solving activities that most often cause an interruption in domestic activity, prompting users to leave the exclusively vocal modality (5/6 households: H1,2,3,4,5). These support activities can indeed give the impression of blind navigation. They required the users' visual engagement and physical movement, as well as multimodal interaction and support. They had a strong impact on the ongoing activity to the point of disrupting its flow, which sometimes led the users to put the VA aside and even to completely abandon it.

It is interesting to note that, despite the difficulties that the users encountered and the extra activity that was generated, most of the participants we interviewed continued to use and accept the system. Moreover, they developed resources and contextualized knowledge as they did so. In addition, this study shows us that the profiles of experienced users (with greater knowledge and technical skill) are more frequently associated with error acceptance. However, these individuals can also be much more demanding about this type of technology and may categorically refuse to integrate a VA that is not secure (system on the cloud) and still not mature in terms of services and voice recognition.

This exploratory study has a few limitations: our participants were mainly technology-oriented and this allowed for forms of VA acceptability and appropriability that would likely differ for users having less "technical" profiles. In fact, our study, which is part of a broader research project, has included a wider variety of profiles (which we have not presented here), and this will enable us to more finely detail these results in future publications.

This less than smooth appropriation process and the paradox of use – revealing the tension between usability and appropriability – should be further investigated in future studies. These issues indeed open up a set of perspectives to be pursued, particularly regarding the progress in the development of ecosystems of connected objects (integrating VAs) and their use in the home.

REFERENCES

[1] V. Bellotti and K. Edwards. 2001. Intelligibility and accountability: human considerations in context-aware systems. *Hum.-Comput. Interact.*16, 2 (December 2001).

[2] S. Brennan. 1990. Conversation as direct manipulation: An iconoclastic view. In The Art of Human-Computer Interface Design. B.K. Laurel (Ed.), Reading, MA : Addison-Wesley.

[3] R. De Rennesse. Virtual digital assistants to overtake world population by 2021. https://ovum.informa.com/resources/product-content/virtual-digital-assistants-to-overtake-world-population-by-2021

[4] W. K. Edwards and R. E. Grinter. 2001. At Home with Ubiquitous Computing: Seven Challenges. In Proceedings of the 3rd international conference on Ubiquitous Computing (UbiComp '01), Gregory D. Abowd, Barry Brumitt, and Steven A. Shafer (Eds.). Springer-Verlag, Berlin, Heidelberg, 256-272.

[5] I. Lopatovska and H. Williams. 2018. Personification of the Amazon Alexa: BFF or a mindless companion? In Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18).

[6] E. Luger and A. 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).

[7] C. Myers, A. Furqan, J. Nebolsky, K. Caro, and J. Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA.

[8] J. Velkovska and M. Zouinar. 2018. The illusion of natural conversation: interacting with smart assistants in home settings. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18).

[9] J. Wu, C.-C. Chang, L. Ng Boyle, and J. Jenness. 2015. Impact of In-vehicle Voice Control Systems on Driver Distraction. Insights From Contextual Interviews. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol 59, 1583-1587.