Talking to Stupid?!? Improving Voice User Interfaces

Insights on user behavior, pain points and improvement opportunities for interaction and dialog design

Maresa Biermann
Usability Engineering
designaffairs GmbH
Munich, Bavaria, Germany
maresa.biermann@designaffairs.com

Evelyn Schweiger
Interaction Experience
designaffairs GmbH
Munich, Bavaria, Germany
evelyn.schweiger@designaffairs.com

Dr. Martin Jentsch
Concept Development
designaffairs GmbH
Munich, Bavaria, Germany
martin.jentsch@designaffairs.com

ABSTRACT

Worldwide, voice user interfaces (voice UIs) are on the rise by being integrated in smartphones, computers, smart home devices and many more consumer goods. Although users praise the possibilities certain functions offer, many users show reservations when it comes to using speech assistants regularly in daily life.

A research and conception project at the strategic design agency designaffairs in Munich demonstrates that users especially praise devices with voice UIs for the functionality they offer, such as music or smart home appliances. Reasons for disliking or not using voice systems are mostly problems with speech recognition and limited functionality. Based on the insights from user research, important factors for the development of a voice UI concept are identified in a workshop with subject matter experts (SMEs). The focus is set on user interaction and dialog design. These factors are validated with users by conducting a conjoint analysis. Here, users prefer gesture input as activation method and the dialog behavior to consist of no humor and no answer variety. A software prototype with this dialog behavior is developed in order to test it with real users in a user test. Although the prototype is rated better than other devices on the market, users' comments indicate that the dialog behavior is not considered ideal for all users. Some users view their voice UI rather as a neutral assistant which should therefore be efficient, brief and concise in answering requests. By contrast, other users wish for a friendly companion which is assisting in a human-like, sympathic way. Thus, developing one speech assistant appropriate for all users calls for the possibility to customize certain features, such as dialog behavior.

The project results show that a user centered design approach is helpful in developing usable products. Further concept improvement must be made especially in the field of speech recognition & processing.

KEYWORDS

Assistive Technology, Voice User Interfaces, UX Research, User Testing, User Centered Design, Voice UI, Voice Assistant

1. Introduction

The number of people using voice user interfaces (voice UIs) is on the rise, especially in the USA where the possession of a speech assistant with smart home integration increased by 14% in only 6 months from January to August 2018 [1]. In comparison to the US market, speech assistants are still less common in Germany [2], but a growth in market share can be observed worldwide [3]. Especially the younger target group (18–39 years) views voice UIs as useful for information services, media consumption, smart home appliances or future use cases such as learning languages [4]. In particular, home appliances with voice integration are viewed as useful, show a good usability and a positive user experience [5]. Here, the degree of usage is dependent on the system's main usage scenarios and use cases [6]. These are either driven by daily routines, such as listening to the daily news podcast while getting dressed in the morning, or situation driven, e.g. looking for a cooking recipe or setting a timer. While smart home devices with voice UIs are location-bound and therefore mainly used stationary, voice UIs in mobile devices, such as Google assistant, Siri or Bixby are also used on the go (ibd.).
According to Cohen et al., "a voice user interface [...] is what a person interacts with when communicating with a spoken language application." [7, p. 5]. As for speech assistants such as Amazon Alexa, Google Assistant or Siri (among others), the voice UI is integrated into a device and the user interacts using voice commands directly addressed to the device. The system processes the user's speech input in order to return an answer which fulfills the intent formulated by the user. What makes the design of voice UIs challenging is the auditory modality: once an information is given, it is gone and not kept present as for a visual interface where the information in form of text or images stays available [7]. Also, the pace of the interaction is driven by the system. The user experience is mainly based on the perception of the voice, the

Veröffentlicht durch die Gesellschaft für Informatik e.V. und die German UPA e.V. 2019 in S. Hess & H. Fischer (Hrsg.):
Mensch und Computer 2019 – Usability Professionals, 08.-11. September 2019, Hamburg

https://doi.org/10.18420/muc2019-up-0253

informational content and associations linked to the speech output because it is the main touchpoint with the system. Its characteristics such as voice, intonation, pace, dialog flow, used phrases/wording and ultimately the personality it shows through these factors play the main part in shaping the user experience of the system [7].

Currently, voice assistants are gaining more and more relevance in the industrial context: Voice assistant dialogs are developed for manufacturing practices, such as teaching robots via voice control [8]. Also, in the medical sector it is likely that voice assistants will generate impact such as providing hands-free applications to support surgeries [9].

Despite all the functions and possibilities voice UIs offer, only about half of the German users see a real benefit in using speech assistants [4]. Reasons for this are the predominating concerns about data security, reservations against talking to a machine and not knowing for what to use the devices [10]. Another reason for reservations against using voice UIs is the special communication style required for addressing the system: users have to reduce their talking pace and consider how to formulate their intent so that the assistant will understand the input [6].

Hence, it can be observed that there is a certain tension field of acknowledging the potential of voice UIs while at the same time still having concerns to use them in daily life. Even more, voice assistants will be integrated in more and more usage contexts such as the working area. Thus, the systems need to be designed in a way in which working efficiency and productivity won't decrease when applying this technology. In order to target these challenges and to explore opportunities for improving voice UIs, a user research and conception project was set-up with a German speaking sample comprising 5 phases: initial research, conception, concept evaluation, prototyping and user testing (Figure 1).

The voice UIs or speech assistants considered in this article use speech as main interaction medium, which means that devices with other interaction possibilities such as typing commands or receiving the system's response on a display are out of the project scope.

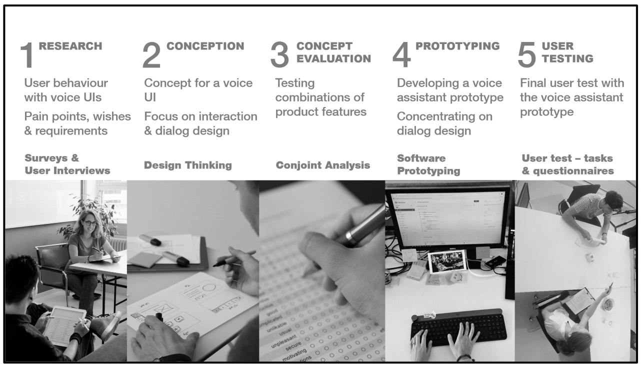
1 RESEARCH
User behaviour with voice UIs
Pain points, wishes & requirements
Surveys & User Interviews
2 CONCEPTION
Concept for a voice UI
Focus on interaction & dialog design
Design Thinking
3 CONCEPT EVALUATION
Testing combinations of product features
Conjoint Analysis
4 PROTOTYPING
Developing a voice assistant prototype
Concentrating on dialog design
Software Prototyping
5 USER TESTING
Final user test with the voice assistant prototype
User test – tasks & questionnaires

Figure 1. Five phases of the research and conception project.

An initial research phase with surveys was undertaken to gain first insights on voice UIs, the users' behavior with voice assistants, as well as needs and wishes. Based on the knowledge gained in this phase, a focus was set on two opportunity areas: interaction and dialog design of voice UIs. In a Design Thinking workshop [11], ideas were generated of how to improve the interaction with the system and how to form a natural dialog between human and machine. These concept ideas were then further refined: By conducting a conjoint analysis, the users' preferences for dialog design and activation method of the assistant were explored and used for concept detailing. Based on the preceding findings, a software prototype of a voice assistant was developed containing two skills. Finally, a user test was conducted testing these two skills and rating the prototype in order to identify which final concepts should be focused on in later realization phases.

2. Methodology & Results

2.1 Initial Research

2.1.1 Methodology. Following the user centered design approach, the project started with gaining user insights regarding the usage of voice assistants to utilize this understanding for further concept development and testing [12]. For the initial research, the goal was to find out which voice assistants are currently used the most, which are the main pain and gain points and where users see room for improvement. Therefore, a questionnaire was conducted with 35 employees of the company designaffairs Munich (16 female, 17 male, 2 diverse), mean age *M* = 37 years (*SD* = 19 years). The questionnaire contained mainly open questions as well as standardized questionnaire elements like the "subjective assessment of speech system interfaces" (SASSI; [13]). With this mixture of qualitative and quantitative data it could be guaranteed that not only numeric data such as frequency of usage of voice assistants and tendencies in technical affinity are uncovered, but also the reasons for these circumstances. Especially the "why" is crucial for creating concepts and designs which really meet the users' needs and are superior. Furthermore, user behaviors and patterns can be revealed and an understanding of user wishes regarding a product or system is created [14].

The questionnaire had the following structure: starting with the subjects' demographic data, next the technical affinity was queried with three chosen items regarding the willingness and motivation to use more technical products [15]. Afterwards, the subjects were asked to state if they possess a specific voice assistant and to rate how often they use the following voice assistants present on the market on a six-point scale varying from daily up to never: Siri (iPhone), Alexa (Amazon Echo), Google Home, Windows Cortana, Bixby (Samsung) and cars with voice assistance. For each of these evaluations the subjects had the chance to name a top and flop feature and to give a reason if a particular voice assistant was never used. Also, the participants could add feedback on further voice assistants which had not been already considered by the questionnaire. Then, the participants should evaluate which of the mentioned devices they use the most and describe it with three positive and negative aspect and features. Additionally, they should give feedback on which functions they used the most. The next set of questions taken from the SASSI [13] should assess their mostly used voice assistant regarding learnability, pleasure, amount of concentration needed and an overall evaluation. For this, a four-point scale was used varying from "I strongly disagree" to "I strongly agree". Then, the participants should mention which three functionalities they would like to use with the voice UI in the future.

Statistical data analysis was conducted using MS Excel. Qualitative data, such as stated positive and negative features, were clustered via card sorting by two subject matter experts (SMEs, background in usability engineering and psychology). Then, identified clusters were double checked for face validity by another SME.

2.1.2 Results. Analysis showed that Siri is the voice assistant possessed the most frequently (n = 25) in the sample followed by Alexa (n = 17) and Cortana (n = 16). The least possessed voice devices were Bixby (n = 5) and Google (n = 4), see Figure 2.

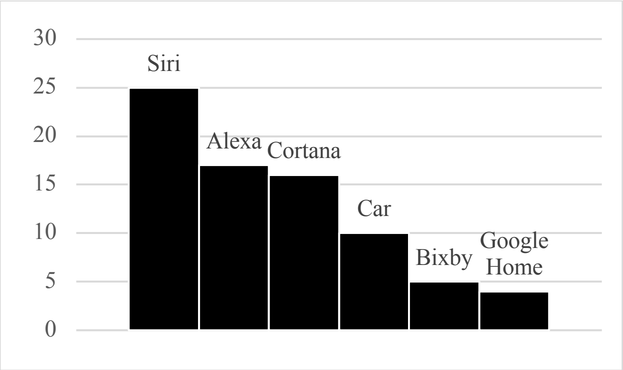

Figure 2. Number of voice UI systems in possession.

Data analysis shows that there is almost no moderate usage of voice assistants. Mainly, the subjects state to use a voice assistant either daily or never. This reveals that voice assistants polarize in this sample regarding the usage frequency and overall usage which is matching with previous studies [4]. Analyzing the SASSI Score for every voice assistant indicates that the car is rated best followed by Alexa and Siri (Figure 3). Here, small sample size should be considered especially for the rating of Car, Bixby and Others. Nevertheless, the good rating of the car could imply that voice UIs are better usable in special, clearly defined contexts such as in-car navigation, where also the design of the dialogs is kept short and simple [16].

In total, the sample showed a mean technical affinity of M = 3.86 (SD = 1.03). The mean score for technical affinity for male participants was M = 4.44 (SD = .81) which is very high, considering a maximum value for technical affinity of 5.0. For female subjects the score was M = 3.25 (SD = 1.0). There is a significant difference in technical affinity between men and women in this sample (t(30.42) = -3.43, p < .005). Also, subjects with a higher technical affinity rated their device better in the SASSI (t(27) = -2.438, p < .05). For analyzing this, a two-level categorical variable was created for technical affinity by median split (M = 3.86; median = 4): subjects with a score of 4 and less were assigned to the group "low", subjects with a technical affinity of > 4 were assigned to "high". Also descriptive data shows a tendency that higher technical affinity is connected to possessing more different devices with voice UIs.

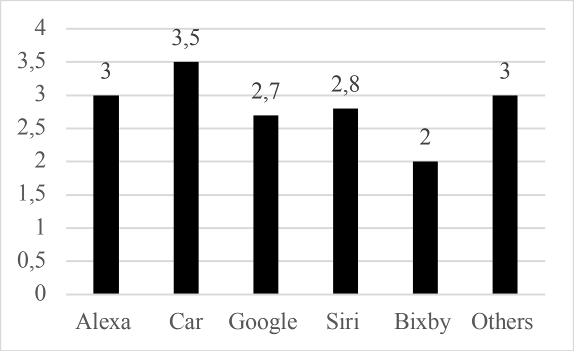

Figure 3. Mean SASSI score per device.

The most frequently used voice assistant in this sample is Alexa. N = 9 subjects stated that they use Alexa daily. The second most frequently used voice assistant is Google: n = 4 subjects said to use it daily. Although Siri is the device which is possessed the most, it is also used the least frequently: N = 15 subjects stated that they never use Siri, n = 10 reported to just use it rarely. Reasons mentioned for this were that Siri is not considered as useful, it doesn't perform well and it is not efficient to use (see also later).

Cortana as part of Windows laptops is owned by n = 16 subjects, but no one indicates that (s)he uses this voice assistant at all. Reasons for this circumstance were given and were for almost all subjects the same: The subjects do not want to give commands to their computer in the work environment. Also, subjects don't see an additional value in using Cortana for work related tasks.

Users were asked to state positive and negative aspects and features for each of the rated voice assistants. The most positive and negative ones were given for Siri and Alexa which were also the devices mostly possessed in the sample. Overall, the number of positive and negative aspects was quite equal per each device. For no device there were clearly more positive or more negative statements made. Positive aspects stated for the devices were mainly specific functions: listening to music, smart home, weather information and countdown or timer. Negative aspects focused on the interaction with the voice assistant: speech recognition, starting the assistant and naturalness of dialog. Also, the limited functionality was stated.

Regarding the most frequently used voice assistant three positive and negative feature clusters could be identified (Table 1). They match well to the overall evaluation of voice assistants: One cluster was about specific functions, such as music, smart home, entertainment and calendar or reminder functions. The second cluster concerned the interaction. Participants mentioned the ease of use, comfort, efficiency or being practical as positive features of voice assistants. The third cluster formed positive emotion about voice assistants such as being cool, interesting and innovative.

Table 1. Positive and negative features mentioned for voice UIs.

positive features	number of mentions	negative features	number of mentions
Specific functions	29	Speech recognition & dialog	22
Interaction	15	Trust and security	7
Positive emotion	3	System and functionality	6
∑	47	∑	35

The first negative cluster, and most important for the participants, was speech recognition and dialog. Many subjects complained about poor speech recognition and the communication itself being not natural enough yet. The second cluster focused on functionality and system such as poor functionality and the compatibility with other devices. The third cluster identified was security and trust, such as data security concerns and talking in public to the voice assistant.

Regarding the most frequently used voice assistant, participants should also state which functions or skills they mostly use. The most frequently used function is Music (n = 16) followed by setting a countdown (n = 7) and retrieving weather information (n = 6). Other used functions focus on communication (like sending & reading out messages, calling other persons), smart home (light control & smart home) and organizational tasks (e.g. asking for time, setting a reminder or appointment).

Considering new functions and useful skills for voice UIs the following were mentioned: N = 7 new functions were stated by the subjects. They varied from concrete functions such as answering medical questions or reading out loud recipes to vague ideas such as a Xbox skill. The rest of the mentioned new functions concentrated on system control (n = 9), improving existing functions (n = 6), the speech recognition and dialog (n = 5), and personalization and configuration (n = 4). For the existing functions, subjects wished for a better Spotify connection or good radio functionality. Regarding the system's control, statements were about being able to have the complete control over all systems but wishing for a learning system and wireless usage. For the speech recognition and dialog, the participants desired a better speech recognition in general but also being able to have more natural and complex conversations as well as context-free commands. For the cluster personalization and configuration, functions like an avatar and an artificial intelligence with personality were requested.

2.2 Conception

After analyzing the collected data from the initial research phase, prototype concepts for a voice UI were developed in a one-day design-thinking workshop. The focus was set on improving the interaction and the dialog flow between user and system since this topic outnumbered the other aspects that caused negative feedback. At the end of the workshop with 4 SMEs (background in usability engineering, interface design and psychology), a list of 10 requirements for a speech assistant was developed (Table 2).

Table 2. List of requirements for a device with voice UI.

Factor	Example
1. Brand	Apple, Amazon, NoName...
2. Price	80€, 100€, 120€...
3. Microphone muting	via button, none...
4. Humor / empathy	No humor, medium, very funny...
5. Voice	Male, female, robot-like, human...
6. Answer variety	None, always using different expressions, ...
7. Answer length	Short, medium, long...
8. Language style	Colloquial, formal, dialect, accent...
9. Activation method	Gesture, button, speech etc.
10. Device	Smartphone, smart home device, laptop...

Three factors with the highest potential to improve the voice UIs dialog were chosen to be further examined by dot-rating: (9.) activation method of the speech assistant, (4.) level of humor or empathy, and (6.) answer variety. Each of the selected factors were specified with three factor levels, respectively (see Table 3).

2.3 Conception Evaluation—Conjoint Analysis

2.3.1 Methodology. A rating-based traditional conjoint analysis [17] was conducted to find out which factor was the most important in influencing the rating and which combination of factor levels (e.g. gesture activation—medium humor—high answer variety) was the most desirable for the respondents.

The advantage of using conjoint analysis is that the participants are not explicitly asked for the most important factor but instead it reveals the users' latent preferences which they normally do not articulate directly to the observer. Compared to other approaches, this analysis method avoids leading and influencing the users to ensure that they give feedback which reflects their real intent and preferences. Especially for developing innovative products and services this is crucial [18].

Table 3. Factors for conjoint analysis and the respective factor levels.

Factor	Level 1	Level 2	Level 3
Activation method	Gesture	Speech	Tapping/ clicking a button
Humor/ empathy	None: neutral formulation	Medium: giving slightly subjective evaluations	High: giving subjective evaluations with slightly humoristic elements
Answer variety	None: always giving the exact same answer to the same request	Medium: using small variations in wording to answer the same request	High: using strongly different syntax, words and phrases for answering the same request

An orthogonal experimental design with 8 stimuli (or factor combinations) was created from the original full factorial design with 27 stimuli by using the R software [19] with the R package "Conjoint" [20]. Together with 2 initial training stimuli these were then presented to a sample of 21 designaffairs employees in an online-survey (14 male, 6 female, 1 diverse; average age 35 years, *SD* = 6 years). After reading a general description of the survey and giving demographic data, participants received a scenario description and an instruction for the following stimulus rating. Each stimulus was presented with an image of the activation method of the system, a textual description of the level of answer variety, and an audio output with the system's response showing different levels of humor. After ticking a box that they had listened to the audio file, participants were asked to rate the stimuli on a scale from 1 to 10 for desirability.

Two participants had to be excluded from data analysis because their answering times were too short to have listened to all audio files.

2.3.2 Results. The conjoint analysis showed that the humor/empathy level was the most important factor, followed by the activation method and the answer variety (Table 4). The most desirable factor combination was gesture activation with no humor/empathy and no answer variety. The linear regression model calculated was significant ($F(6, 145) = 2.494$, $p < .05$). However, the adjusted R-squared with a value of $r^2 = .06$ can be considered very low. The analysis computed a regression model for each of the 21 participants first and then aggregated the single models over the whole sample. Therefore, it is plausible that differences between individuals cause most of the variance in rating, as is typical for this kind of conjoint analysis [21].

Table 4. Results of conjoint analysis: factor importance & partworth utilities for factor levels.

Factor	Factor importance [%]	Factor Level	Partworth utility
Activation method	32.4	**Gesture**	**0.60**
		Speech	-.55
		Button	-.04
Humor/ empathy	48.9	**None**	**.52**
		Medium	.32
		High	-.83
Answer variety	18.8	**None**	**.17**
		Medium	.04
		High	-.21

2.4 Prototyping

Based on the findings of the preceding phases, a software prototype was created with the software Dialogflow [22], a tool provided by Google to implement voice UIs based on natural language processing (NLP). For the design of the system prompts the optimal dialog behavior identified in the conjoint analysis was used: no humor/empathy and no answer variety. The prototype contained two skills (Figure 4): (1) a weather information service providing the weather forecast for a certain location, (2) a concert organizer skill giving concert suggestions for a certain music genre with the option to then purchase the ticket for the selected event. For each skill possible trigger words expressed by the user and the respective systems answer were stored in the dialogflow data base in advance. Since the users prefered no humor and empathy the pre-defined sytsem answers were formulated in a short and fact-based manner. When the user articulated a trigger word, the respective system answer was issued. With the NLP of dialogflow also similar expressions to the trigger words activated the system answers.

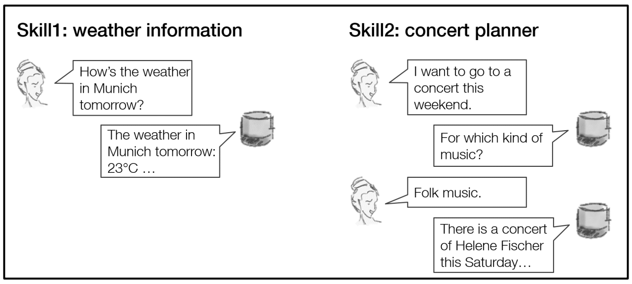
Skill1: weather information
How's the weather in Munich tomorrow?
The weather in Munich tomorrow: 23°C …
Skill2: concert planner
I want to go to a concert this weekend.
For which kind of music?
Folk music.
There is a concert of Helene Fischer this Saturday…

Figure 4. Exemplary dialog flow between the user and the voice UI prototype.

2.5 User Testing

2.5.1 Methodology. The prototype's two skills were evaluated in a 20-minute user test at the Usability lab of designaffairs office in Munich. The subjects were instructed to interact normally with the voice UI which was started on a laptop using also the laptop's microphone and speaker. In order to focus on a natural dialog between the user and the system, users did not have to activate the voice assistant each time, but the test manager ensured that the system was ready for speech in- and output. First, users were asked to test the weather skill by asking for the forecast of two locations. Then, they were instructed to plan their concert visit at the weekend with the voice assistant. After testing the two skills, users answered a questionnaire on demographic data and then rated the prototype with the SASSI [13]. Also, users rated answer length, naturalness of the dialog, level of humor, and probability of using the system on a 4-point Likert scale and were asked to mention positive and negative aspects of the dialog.

In total, 15 participants (7 female, 8 male) with a mean age of M = 35 years (SD = 5 years) participated in the user test. All of them were employees of designaffairs Munich.

2.5.2 Results. Overall, users rated the tested prototype better than the devices in the first survey: the mean SASSI score in the first survey over all devices with a value of M = 2.76 was surpassed by the mean of M = 3.35 for the developed prototype in the user test, a score of 4 being the maximum (SD = .42 & SD = .54, respectively; $t(22) = -3.60$, $p < 0.01$). Users stated to be pleased with the answer length (M = 3.6, SD = .88) and that they would like to use the dialog behavior in their device (M = 3.7, SD = .85). Rating for naturalness of dialog (M = 2.5, SD = 1.15) and degree of humor (M = 2.3, SD = .94) polarized when looking at the users' comments (Table 5).

Table 5. Positive and negative comments on the prototype's dialog behavior by users during the user test.

Positive comments	Efficiency ("it's fast", "really efficient")
	Neutrality ("neutral is nice", "I like the scarcity", "it's good and simple")
	Humor ("too neutral", "a bit more funny would be nice")
Negative comments	Friendliness ("not charming", "it could be more friendly", "not very polite")
	Machine-like character ("it's not authentic", "I miss some human warmth")

3. Summary & discussion

The initial research phase provided valuable insights on user behavior with voice UIs. Speech assistants in the sample were either used never, seldomly or regularly with daily periodicity. The best rated voice UI was the in-car system, followed by Alexa, Siri and Google. Users especially valued the functionality and ease of interaction voice UIs offer, whereas the speech recognition as well as trust and security were the strongest pain points. Future wishes to the system were more reliability, a better speech recognition, personalization and certain new functions.

The conjoint analysis revealed that the factor humor/empathy was more important to the users than the answer length and the activation method. Users in this survey preferred a system activation via gesture and a rather machine-like dialog where the assistant showed little answer variety and a low level of humor and empathy. However, many users confronted with the corresponding dialog behavior during the user test stated the system to be too neutral, lacking authenticity, friendliness and sympathy. Thus, it can be summarized that the demands on the ideal level of humor/empathy polarize: on the one extreme, users wish for a personal, empathic friend-like assistant. On the other extreme, users prefer a neutral task-oriented assistant.

One finding which came as a surprise to the SMEs in the conjoint analysis was the preference of a gesture as activation method. The main advantage of voice UIs is them being "hands-free, eyes-free" [7, p. 11]. This advantage is weakened if a gesture activation of the system is implemented which again most probably needs involvement of hands—as for the activation via tapping a button on a wearable. In addition, the recognition of the gesture would require the installation of sensors or a camera which would in turn raise questions about privacy and data security (as for the microphone of Amazon's Alexa and the uncertainty if it records and stores user data [23]). Considering these limitations, voice activation may remain the most practible option if hands-free activation is a key requirement for the system.

During the user test, it could be observed that users already show learned behavior towards voice UIs and have adopted a certain interaction style. Even though they were not instructed to do so, some participants used key words or called the prototype's name each time when addressing the speech assistant. Others took some consideration time to mentally pre-formulate their prompts before stating them out loud to the voice UI. This demonstrates that many users expect natural language not to be understood by the system so that they adapt special communication strategies—as could also be observed in previous studies [6, 10]. Another observation made by the SMEs during the user tests was that participants showed some hesitation or uncertainty when talking to a voice UI in front of the test manager. They were hesitant in starting to talk to the speech assistant, took more time in searching for the right words and reformulated their sentences more often as opposed to talking to the test manager. This indicates that talking to a machine is not a completely natural act yet and that resistance and anxiety concerning human-machine communication should be reduced.

Overall, the presented research implies that different functions of voice UIs are used in various contexts and users have different wishes and demands to the ideal voice UI or speech assistant. So, in order to create the perfect voice assistant for all types of users, customization is a key requirement. Having a learning system which refines its character during usage in order to fit to the user's dialog preferences would be one solution for this. Also, allowing users to adjust certain characteristics of their voice UI, such as answer variety or level of humor or empathy, could increase the fit between human and machine. Knott & Kortum [24] already demonstrated that building a personal relationship to a voice UI increases the probability of interacting with it. For example, users interacted more and gave more specific information when a voice UI introduced itself by name opposed to giving no initial introduction (ibid.). Even more, the system should be context sensitive: by detecting the user's tone of voice it could adapt its behavior to the current situation and mood of the user—just as we change our behavior when we realize that our counterpart is stressed or in a hurry.

Regarding the sampling, a heterogenous sample composition could be achieved in terms of gender, but not for age nor technical affinity. Nontheless, the results of the initial research phase fit quite well to the findings of other studies regarding often used features and pain points [6, 1, 4]. This indicates a good representativity of the sample. However, future research should also target interaction behavior with voice UIs in other age groups as were assessed in this study, such as the millenials or the elderly. Swoboda et al. [25] for example show that voice UIs hold high potential for the elderly, especially to assist people with visual impairment or movement disabilities in the hands (e.g. due to arthritis or Parkinson disease).

The conjoint analysis was successful in revealing the preferences of the users for the examined three factors. However, it would have been interesting to include further aspects of voice UIs in the analysis, such as gender of voice, speaking pace, physical design of the device, price, and many more as identified throughout the Design Thinking workshop (Table 2).

One limitation to the user testing phase of this study was mistakes in the system's speech recognition, such as incorrectly identified text input. By using the already quite powerfull NLP tool "dialogflow" by Google, failures of the system were kept low, but did happen so that users sometimes had to repeat commands. In this study, users mainly attributed these mistakes to the prototype status of the system. However, failing voice recognition or misinterpretation of the users' prompt has been identified as a major issue creating user frustration and negative user experience in other studies with devices which are already on the market [26, 6]

Linked to this is the issue of devices still showing a better usability for native speakers of the set language than for people speaking with accent [27]. However, it has become common that several languages are being spoken in one household by people of different nationalities, as is also the case for the participants of this study. Thus, the recognition and processing of different languages and accents must be enhanced. The aim is to enhance a natural conversation with authentic, spontaneous and natural speech between the human and the machine. So overall, further technical improvement in language processing is a key requirement for improving usability and user experience of voice UIs.

4. Conclusion & future work

In this study, valuable insights from user research could be retrieved on which voice UI devices are used how frequently, which functions are used mostly, and what users like and dislike about their devices. The conjoint analysis gave an indication for the desired interaction style of voice UIs. However, the final user test demonstrated that the implemented dialog behavior was not ideal for all users which is why future systems call for customization.

In conclusion, further technical development must take place to ensure a more reliable speech recognition and functioning dialog. Here, building useful skills which map the whole workflow of a task from beginning to end with one single device should better meet the user's need (e.g. asking for concerts up to ordering the ticket). This requires a better connectivity of the systems involved: a seamless interaction and data transmission must be achieved without menacing the user's privacy or raising concerns of data security.

To make full use of their potential, voice UIs should be integrated in contexts where their main benefit of being hands- and eyes-free can be fully utilized. This is whenever a task requires manual activity which cannot be interrupted that easily, such as in the car or in the medical operation theater. Also a future integration of voice UIs in virtual/augmented or mixed reality seems a reasonable use case.

ACKNOWLEDGMENTS

A special thanks goes to our project team members Melissa Ruppel, Christoph Baur and Balazs Kocsis for a fun, enriching and successful team-work. We also thank all employees at designaffairs Munich who participated in our study and thereby made our project possible. Thanks goes to Werner Spicka, Nico Michler and Michael Lanz for funding and supporting us in our research interest.

REFERENCES

[1] G. Abramovich, "Study Finds Consumers Are Embracing Voice Services. Here's How.," Adobe, 10 September 2018.

[Online]. Available: https://www.cmo.com/features/articles/2018/9/7/adobe-2018-consumer-voice-survey.html#gs.4rb2l3. [Accessed 29 March 2019].

[2] M. Brandt, "Wenig Echo in Deutschland," 13 February 2018. [Online]. Available: https://de.statista.com/infografik/12884/smart-speaker-besitz-in-deutschland-und-den-usa/.

[3] Statista, "Intelligente Lautsprecher - Absatz weltweit nach Hersteller bis Q4 2018 | Statistik," Statista, 2019. [Online]. Available: https://de.statista.com/statistik/daten/studie/818995/umfrage/absatz-der-hersteller-von-intelligenten-lautsprechern-weltweit-pro-quartal/. [Accessed 09 04 2019].

[4] F. Hüber, "Alexa, Cortana, Siri und Co.: Mehrheit sieht Datenschutz kritisch, will sie aber nutzen.," 03 May 2018. [Online]. Available: https://www.computerbase.de/2018-05/sprachassistenten-studie-etablierung-datenschutz/. [Accessed 01 April 2019].

[5] A. Pyae and T. N. Joelsson, "Investigating the usability and user experiences of voice user interface: a case of Google home smart speaker.," in the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, 2018.

[6] X. Bodendörfer, „Digitale Sprachassistenten als intelligente Helfer im Alltag," 2017. [Online]. Available: https://www.eresult.de/ux-wissen/forschungsbeitraege/einzelansicht/news/digitale-sprachassistenten-als-intelligente-helfer-im-alltag/.

[7] M. H. Cohen, J. P. Giangola and J. Balogh, Voice User Interface Design, Boston: Addison-Wesley Professional, 2007.

[8] Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., „Fraunhofer.de," 01 März 2019. [Online]. Available: https://www.fraunhofer.de/de/presse/presseinformationen/2019/maerz/smarter-sprachassistent-steht-rede-und-antwort.html. [Zugriff am 16. April 2019].

[9] B. Kinsella, „Voicebot.ai," voicebot ai, 2019. [Online]. Available: https://voicebot.ai/2019/02/11/nhs-study-on-future-of-healthcare-foresees-voice-assistants-helping-clinicians-and-home-based-patient-care/. [Zugriff am 23 05 2019].

[10] EARSandEYES GmbH, "Gründe für die Nichtnutzung von Sprachassistenten in Deutschland 2018 | Umfrage," Statista, 2019. [Online]. Available: https://de.statista.com/statistik/daten/studie/872316/umfrage/gruende-fuer-die-nichtnutzung-von-sprachassistenten-in-deutschland/. [Accessed 09 04 2019].

[11] A. Grots und M. Pratschke, „Design Thinking - Kreativität als Methode," Marketing Rev St Gallen, pp. 2:18-23, 2009.

[12] ISO, *Ergonomie der Mensch-System- Interaktion Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme (Vol. DIN EN ISO 9241-210),* Genf, 2010.

[13] K. S. Hone, "Usability measurement for speech systems: SASSI revisited.," in SIGCHI Conference Paper, Toronto, 2014.

[14] A. Cooper, R. Reimann, D. Cronin and C. Noessel, About Face: The Essentials of Interaction Design, Indianapolis: John Wiley & Sons, Inc., 2014.

[15] T. Franke, C. Attig and D. Wessel, "Assessing Affinity for Technology Interaction – The Affinity for Technology Interaction (ATI) Scale," Unpublished manuscript, 2017.

[16] M. Hiersemann, J. Mühlstedt, H. Unger, P. Habil und B. Spanner-Ulmer, „In-Car Auditory Signals," SAE Technical Paper 2007-01-0450, 2007.

[17] H. Fiedler, T. Kaltenborn, R. Lanwehr and T. Melles, Conjoint-Analyse, Augsburg: Rainer Hampp Verlag, 2017.

[18] M. Jentsch, S. Wendlandt, N. Clausen-Stuck und G. Krämer, „What do they really want? - Reveal users' latent needs through contextual Co-Creation.," in Tagungsband HFES Europe Chapter Conference 2016. 26.-28. Oktober, Prague, 2016.

[19] R. C. Team, *R:* A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing, 2018.

[20] A. Bak and T. Bartlomowicz, "Conjoint: An Implementation of Conjoint Analysis Method. R package version 1.41," 26 July 2018. [Online]. Available: https://cran.r-project.org/web/packages/conjoint/index.html.

[21] V. R. Rao, Applied Conjoint Analysis, Berlin: Springer, 2014.

[22] Google, „dialogflow.com," 2019. [Online]. Available: https://dialogflow.com/. [Zugriff am 23 05 2019].

[23] J. Philipp, "Smart Home: Warum uns Alexa nicht abhören kann.," giga, 27 01 2018. [Online]. Available: https://www.giga.de/unternehmen/amazon/news/smart-home-warum-uns-alexa-nicht-abhoeren-kann/. [Accessed 09 04 2019].

[24] B. A. Knott and P. Kortum, "Personification of Voice User Interfaces: Impacts on User Performance.," *Human Factors and Ergonomics Society Annual Meeting Proceedings,* no. 50(4), pp. 599-603, October 2006.

[25] W. Swoboda, F. Holl, S. Pohlmann, M. Denkinger, A. Hehl, M. Brönner and H. Gewald, "A Digital Speech Assistant for the Elderly," in *MIE Medical Informatics in Europe*, Sweden, 2019.

[26] C. M. Myers, A. Furqan, J. Nebolsky, K. Caro and J. Zhu, "Patterns for How Users Overcome Obstacles in Voice User Interfaces," in CHI Conference Paper, Montreal, Canada, 2018.

[27] A. Pyae and P. Scifleet, "Investigating differences between native english and non-native english speakers in interacting with a voice user interface: a case of google home.," in *the* 30th Australian Conference on Computer-Human interaction, 2018.