Context-Specific Usability Measures for Voice Assistants

Sanjana Mendu
Pennsylvania State University
State College, Pennsylvania
sanjana.mendu@psu.edu

S. Shyam Sundar
Pennsylvania State University
State College, Pennsylvania
sss12@psu.edu

Saeed Abdullah
Pennsylvania State University
State College, Pennsylvania
saeed@psu.edu

ABSTRACT

Voice assistants (VAs) have proliferated in the last 10 years. They are now used to support a wide range of tasks and activities across different domains. Effective evaluation of VA usability is critical for sustained user adoption and acceptance. Current approaches for assessing VA usability largely make use of traditional technology-agnostic measures and heuristics. We argue that this not only overlooks the unique affordances and nuances of voice interaction usability, but also fails to provide meaningful feedback for designers and developers. That is, usability measures for VAs should focus on unique affordances of voice interfaces and conversational interactions. Given the ubiquitous adoption of VAs, there is a pressing need for standardized and context-specific usability measures for VAs. In recent years, a number of emerging systems, such as mobile health (mHealth) apps, have focused on developing domain-specific usability measures and evaluation criteria. We believe a similar approach will lead to more robust and nuanced usability measures for VAs as well.

KEYWORDS

usability evaluation; voice assistants

INTRODUCTION

Over the last decade, voice assistants (VAs) have become more ubiquitous and mainstream, increasing in both affordability and functionality [8, 13]. Google reported that nearly 1 billion devices of their consumer devices would be voice assistant compatible by early 2019, up from 500 million the previous

year [2]. Reports similarly indicate that over 100 million Amazon Alexa devices were sold around the same time period [1]. While terminology for VAs varies widely across disciplines, following Sezgin et al. [20] we define a VA as a device "programmed with some type of artificial intelligence capable of two-way dialogue, differentiating this from one-way voice technology".

Existing literature points to both the utilitarian and hedonic benefits of VAs [22]. These benefits along with other underlying factors have been shown to underpin the usability, and thereby use and adoption, of VA technologies [22]. However, research on evaluating the usability of VAs lags far behind technical developments. To date, there are no standard or well-defined metrics for evaluating the usability of VAs [10, 22]. The lack of any standardized measures for evaluating the usability of VAs significantly inhibits researchers' ability to meaningfully evaluate these systems.

In this paper, we argue that existing approaches to evaluating VA usability are insufficient for providing meaningful feedback to designers and need to more directly account for the unique affordances of voice interaction. We present a brief review of existing usability evaluation work in this domain, and highlight limitations which we propose future work should address. Finally, we contend that context-specific measures developed for usability assessment of other emerging technologies (e.g., mHealth apps) serve as a useful foundation upon which researchers can develop robust, multidimensional scales for assessing the quality of VA systems.

CURRENT APPROACHES

Leveraging Existing Usability Measures

Since its inception, the System Usability Scale (SUS) [3] has been widely used by the HCI community to assess usability of a wide range of technologies. While prior studies have used SUS to evaluate usability of VAs [6], others have questioned the validity of using SUS for VAs. Zwakman et al. [22] argue that the evaluation criteria targeted by SUS are too heavily oriented towards graphical interfaces to appropriately translate to voice interface evaluation. To mitigate this gap, they proposed an adapted version of SUS developed specifically for VAs. Holmes et al. [7] contend that SUS lacks sensitivity to issues specific to voice interfaces . Apart from SUS, a number of other usability evaluation tools are also available, including The Usability Metric for User Experience (UMUX) [5], The Computer System Usability Questionnaire (CSUQ) [12], Post-Study System Usability Questionnaire (PSSUQ) [11], Software Usability Measurement Inventory (SUMI) [9], and many others. However, to our knowledge, none of these scales have been applied widely to assess usability of VAs.

Heuristics

Recent work has also used expert-based heuristics to evaluate usability of VAs [17]. Heuristics consist of sets of guidelines that can support expert evaluators in their task and improve their ability to spot

flaws in the target interface. An evaluation using heuristics or ergonomic criteria is both analytical and predictive, comparing the characteristics of an interface to principles or desirable ergonomic dimensions to hypothesize its likely use [19]. Expert-based usability evaluations allow for efficient assessment of an interface at several phases of its conception and can be used in a diverse range of evaluative scenarios (e.g., during the design phase, to audit a pre-existing interface, or to improve a product before moving forward with user tests). Heuristic and guideline evaluations are widely used thanks to their intuitiveness and ease of implementation.

However, these evaluations are traditionally based on subjective opinions which contributes to variability when comparing similar systems. Furthermore, they are limited by their relative inaccessibility to non-experts [16]. Moreover, Scapin & Bastien [19] explicitly warned that their ergonomic criteria might not be valid to assess an interface based on new technology since their inspection only reflected interactive system features for which ergonomic knowledge existed at the time. In other words, naively using generic heuristics and ergonomic criteria may not be adequate to assess the usability of emerging technologies including VAs given their unique affordances.

UNIQUE AFFORDANCES OF VAS

Advances in artificial intelligence and deep learning technologies such as natural language processing and speech recognition provide more flexible opportunities for users to interact with VAs through dialogues and conversations [20]. In contrast to traditional technologies that rely on graphical interfaces to facilitate user interaction (e.g., scrolling, swiping, and clicking), VAs operate by awaiting a keyword to “wake," before capturing the user’s voice input, conducting natural language processing to interpret user input, and responding back with a dialogue or completed task. Prior work has shown that voice-interfaces are distinctly different from their graphical counterparts and give rise to unique usability issues, such as the ability to understand non-conversational cues (i.e., pauses in the middle of a conversation), difficulty with back and forth navigation, and increased cognitive workload due to the absence of visual feedback [4, 10, 14].

Observations of users’ interactions with VAs show that talking to these sytems leads users both to change their attitudes to accommodate an inanimate object and to give human traits and interpretations to some of the VAs features [17]. For example, users take turns in talking, simplify their commands and decrease background noise when uttering a command to maximize the chances of success. On the contrary, when a VA takes too much time to answer, the silence is interpreted as a sign of trouble and leads users to describe the system as “not liking something” [18]. Because VAs leverage a form of modality that is primarily used for human-to-human interaction, users may be inclined to project anthropomorphic qualities onto their interaction with VAs, potentially eliciting emotional responses such as distrust in users when the VAs response misaligns with their expectations [17]. Nass et al. [15] found that experienced computer users consistently applied social rules to their

interactions with voice interfaces. These observations indicate that as technology evolves to increase the accuracy of both understanding and uttering human speech, the use of voice interface is creating a new kind of interaction with unique expectations from users, which the designers of VAs need to address.

For example, Nowacki et al. [17] proposed an improved set of ergonomic criteria for voice user interfaces, building upon the heuristics established by Scapin & Bastien [19] while also incorporating advice from 26 design guidelines for voice user interfaces from both academic and professional sources. They added and modified criteria to directly address the interface between VAs and social context and conversations, which have criteria that apply specifically to voice interfaces due to the social context (e.g. compatibility, personality). They also de-emphasized criteria that more heavily focused on visual cues in the original guidelines (e.g. legibility, significance of codes). This work highlights the importance of establishing new expectations and standards as VAs add more functionalities and become more ubiquitious.

TOWARDS BETTER USABILITY MEASURES FOR VOICE ASSISTANTS

Thus far, usability evaluation for VAs has primarily consisted of either adapting widely-accepted usability measures or developing domain-specific heuristics. We argue that there is a need for novel VA usability evaluation tools that balance the affordances of both approaches while still appropriately accounting for unique affordances of voice interaction. Toward this goal, we think the designers of VAs can leverage recent work on domain specific usability measures for emerging technologies.

For example, Stoyanov et al. [21] aimed to assess multidimensional measure of mHealth app quality by developing the Mobile Application Rating Scale (MARS). They focused on providing a more comprehensive assessment of user experience for mHealth apps. They conducted a systematic review of existing applications, which lead to 349 identified evaluation criteria. A panel of experts then grouped these criteria into six categories. These six categories provided both broad usability measures and domain specific heuristics. While the scale development process relied heavily on expert knowledge, the resulting scale is easy-to-use and requires minimal training to be used by non-experts. However, the scale is still able to provide useful insights into the unique affordances and user expectations into mHealth apps.

We believe that a similar approach that focuses on both the broader usability concerns and unique affordances of VAs can lead to better usability assessment methods. That is, we should collect data about different tasks and activities supported by VAs to identify needs and requirements that vary across contexts. We can then use the collected data to define broad usability aspects of voice interactions as well as nuanced domain specific needs (e.g., information retrieval, entertainment, interaction support for children).

WORKSHOP OUTCOMES

Our work contributes to the discussion of conversational user interface design, focusing specifically on context-specific tools, methods, and practices. In proposing a paradigm shift in how HCI researchers and practitioners evaluate the usability of voice assistant technologies, we hope to work towards bridging the gap between academic and industry approaches in this domain while making space for the unique and multi-faceted affordances of voice-based interactions. Our fellow workshop participants' collective expertise will help us to explore the following research directions: 1) establish the current state-of-the-art of usability evaluation of VAs in academia and industry, 2) determine the criteria for VA usability evaluation informed by both experts and practitioners, and 3) identify novel methods and procedures toward developing usability measures for VAs.

REFERENCES

[1] 2019. Amazon says 100 million Alexa devices have been sold — what's next? (2019). https://www.theverge.com/2019/1/4/18168565/amazon-alexa-devices-how-many-sold-number-100-million-dave-limp

[2] 2019. Here's how the Google Assistant became more helpful in 2018. (2019). https://www.blog.google/products/assistant/heres-how-google-assistant-became-more-helpful-2018/

[3] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189 (1996).

[4] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. " What can i help you with?" infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.

[5] Kraig Finstad. 2010. The usability metric for user experience. *Interacting with Computers* 22, 5 (2010), 323–327.

[6] Debjyoti Ghosh, Pin Sym Foong, Shan Zhang, and Shengdong Zhao. 2018. Assessing the utility of the system usability scale for evaluating voice-based user interfaces. In *Proceedings of the Sixth International Symposium of Chinese CHI*. 11–15.

[7] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?. In *Proceedings of the 31st European Conference on Cognitive Ergonomics*. 207–214.

[8] Matthew B Hoy. 2018. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly* 37, 1 (2018), 81–88.

[9] Jurek Kirakowski and Mary Corbett. 1993. SUMI: The software usability measurement inventory. *British journal of educational technology* 24, 3 (1993), 210–212.

[10] Ahmet Baki Kocaballi, Liliana Laranjo, and Enrico Coiera. 2019. Understanding and measuring user experience in conversational interfaces. *Interacting with Computers* 31, 2 (2019), 192–207.

[11] James R Lewis. 1992. Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors Society Annual Meeting*, Vol. 36. Sage Publications Sage CA: Los Angeles, CA, 1259–1260.

[12] James R Lewis. 1995. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction* 7, 1 (1995), 57–78.

[13] M McTear, Z Callejas, and D Griol. 2016. The Conversational Interface: Talking to Smart Devices: Springer International Publishing. *Doi: https://doi. org/10.1007/978-3-319-32967-3* (2016).

[14] Christine Murad, Cosmin Munteanu, Benjamin R Cowan, and Leigh Clark. 2019. Revolution or evolution? Speech interaction and HCI design guidelines. *IEEE Pervasive Computing* 18, 2 (2019), 33–45.

[15] 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*. 72–78.

[16] Jakob Nielsen. 1994. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 152–158.

[17] Caroline Nowacki, Anna Gordeeva, and Anne-Hélène Lizé. 2020. Improving the Usability of Voice User Interfaces: A New Set of Ergonomic Criteria. In *International Conference on Human-Computer Interaction*. Springer, 117–133.

[18] 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*. 1–12.

[19] Dominique L Scapin and JM Christian Bastien. 1997. Ergonomic criteria for evaluating the ergonomic quality of interactive systems. *Behaviour & information technology* 16, 4-5 (1997), 220–231.

[20] Emre Sezgin, Lisa K Militello, Yungui Huang, and Simon Lin. 2020. A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. *Translational Behavioral Medicine* 10, 3 (2020), 606–628.

[21] Stoyan R Stoyanov, Leanne Hides, David J Kavanagh, Oksana Zelenko, Dian Tjondronegoro, and Madhavan Mani. 2015. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. *JMIR mHealth and uHealth* 3, 1 (2015), e27.

[22] Dilawar Shah Zwakman, Debajyoti Pal, and Chonlameth Arpnikanondt. 2021. Usability Evaluation of Artificial Intelligence-Based Voice Assistants: The Case of Amazon Alexa. *SN Computer Science* 2, 1 (2021), 1–16.