Evaluating Smartphone Voice Assistants: A Review of UX Methods and Challenges

Noor Ali-Hasan
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
nalihasan@google.com

Abstract
Voice assistants are becoming increasingly popular on smartphones. Despite the growing popularity of these assistants, voice user experience continues to be a nascent field. In this paper, I describe two of the research methods that some user experience researchers at Google have employed in evaluating smartphone voice assistants in a product development setting and present some challenges we have observed in users’ adoption of smartphone voice assistants.

Author Keywords
Voice interfaces; voice interaction design; conversational agents; speech; personal assistants; voice assistants; user experience research.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction
Voice assistants, where a user interacts with a device or app using their voice to complete a task or retrieve information, are becoming increasingly popular both on smartphones (such as Apple’s Siri [2], Samsung’s Bixby [6], and Google’s Assistant [3]) and in the home (in the
form of devices like the Amazon Echo [1] and Google Home [4]). Despite the growing popularity of these apps and devices, voice user experience continues to be a nascent field, especially when compared to other input methods like touch or keyboard/mouse. In this paper, I describe two research methods that some of my colleagues and I have employed in evaluating smartphone voice assistants in a product development setting at Google and present some challenges we have observed in users' adoption of smartphone voice assistants.

**UX Research Methods for Evaluating Voice Assistants**

In my research with voice assistants at Google, I have employed two primary user experience research methods to evaluate and understand usage of our products: traditional usability studies and longitudinal studies. UX researchers at Google have adapted both of these approaches to voice interactions and found advantages and disadvantages to both methods.

*Usability studies*

We have found that traditional usability studies (in a user experience lab) [5] to be useful in gathering quick feedback in a product development setting. Doing so has enabled us to quickly iterate on designs and to help our engineering and design teams improve the usability and desirability of these products. We have employed this method utilizing different design fidelities such as static mocks, functional prototypes, and launched products. It may seem counterintuitive to evaluate a voice interface with a static mock but we have used this method to learn how a user may phrase a query or command. In some instances, we were also able to understand how they may interact with touch components of the voice interface. For instance, we could determine if a user was more likely to utilize a hotword (such as "OK Google") or tap on an interface's microphone icon. By using static mocks early in the product development cycle, we were able to evaluate concepts that user experience designers on our team were considering, without needing to involve engineering (which can be costly both in terms of time and effort). Doing so also helped our teams iterate early in the development cycle.

During later stages of the development cycle, I have conducted more traditional usability studies with functional prototypes and launched (or close to completed) products. In these instances, I would provide participants with a list of tasks to complete using our app or device. We would assess how participants in our studies phrased their queries or commands, how well the system recognized what they said, and how well it performed in the eyes of these participants (for instance, did the system deliver the results that the user expected or did it perform the requested action or command?). When planning these studies, we have had to account for research participants' primary language, regional accents and dialects, and their familiarity and proficiency with voice assistants and interfaces.

As much value as we have seen in these studies to deliver timely usability and performance feedback, we have found limitations with this approach. Lab based evaluations do not provide us with real life situations of users interacting with the system.
**Longitudinal studies**

To account for the limitations in usability evaluations of voice interfaces, I started conducting longitudinal studies with our voice assistants. In these studies, I supplied research participants with smartphones that feature our voice assistant and asked them to use these smartphones as their primary phones for a period of several weeks (my studies have ranged from 5 weeks to two months). I collected data from these participants through interviews (both in-person and remotely through video chat), periodic surveys, and a digital diary. I found the digital diary to be the most compelling source of data. I asked participants to complete a diary of their interactions with the voice assistant using diary documents I created using Google Docs. As part of each diary entry, I encouraged participants to take screenshots of their interactions with the system (see Figure 1 for an example).

Figure 1: A participant’s screenshot from a 2014 study. Here we can see the actual voice query that the participant said ("how many calories are in cherries"), along with how the system responded to that query.

These screenshots were critical to our analysis in a number of ways:

- Helping us understand the types of commands and queries that participants expected to be able to complete with the system.
- Seeing how participants phrased their commands and queries.
- Visually seeing how the system responded to their commands and queries.
- Helping the engineering team to identify the root cause of instances where the system did not perform as expected.
- Providing research stakeholders with compelling anecdotes and examples to bolster our research findings and conclusions.

Along with understanding usage of these voice assistants over time, the longitudinal approach provided us with real life scenarios and situations for using voice assistants that we either could not replicate in the lab or had not thought of before. For instance, many participants reported back their experiences of using the voice assistant in a vehicle to get directions or send text messages. But participants also told us about times when they used the voice assistant while active:

- One participant recounted how he used the voice assistant while skiing to coordinate meeting up with his friends who were in other areas of the ski resort.
- A participant told us she used the voice assistant to play music when she went running.
Another participant told us about how he used the voice assistant while commuting on his bike to and from work every day. Along with playing music or podcasts, he also used the voice assistant to look for information like stock prices.

Despite all the value I have found in this approach, I have also seen some drawbacks. First, asking research participants to use a new smartphone (in place of their own) can be a burdensome task. There may be issues with inserting sim cards and ensuring that their carrier service works properly on the new phone. Participants may not have access to all of their data (like contacts, photos, music) on the new phone or may lose some of that data that they collected during the research period when switching back to their own phones. Some participants may face difficulties adapting to a new smartphone operating system (if their current phone features a different operating system than the phone provided by the researchers). I have also found that the digital diary format can be challenging for some participants to complete if they are unfamiliar with word processing software or do not own a computer at home.

Moreover, as with any longitudinal research in a real-world setting, situations may occur that prevent a participant from fully completing the research. In one of my studies, a participant accidentally dropped the research phone in the bathtub while giving his child a bath (rendering the phone inoperable). Another participant was uncharacteristically unresponsive during the middle of the research period. I learned later that she had been involved in a serious car accident.

### Challenges to Adopting Smartphone Voice Assistants

We have found that voice assistants can be a convenient way to interact with smartphones for many users. Voice assistants can bridge technological gaps for users of various abilities such as low literacy users, children who may be still learning how to read, slow touch typists, and those with vision or motor impairments. But there are some barriers to adopting voice assistants, which we will describe below.

**Cultural considerations**
In some cultures, it may be impolite or inappropriate to speak loudly to a phone in public. For instance, we learned that in Japan commuters are expected to refrain from speaking on their phones while on public transportation (see Figure 2). Such cultural expectations would make it nearly impossible to fully adopt a smartphone voice assistant.

**Performance anxiety and social costs**
Similarly, we have observed times in the research lab where participants felt a bit of performance anxiety interacting with a voice assistant. They may feel shy doing so in front of a researcher or feel pressure to quickly and correctly say their query or command to the voice assistant (especially if the voice assistant times out or misinterprets a gap in speech or stutter). Some participants also recounted times when they wanted to show off a voice assistant to family and friends but then felt embarrassed if the system did not recognize their command or misinterpreted it.
Exceedingly high expectations
We have found that some users have very high expectations for voice assistants, far beyond the capabilities of current technology. These users want to be able to communicate with voice assistants in true natural language, much like they would a human personal assistant. While natural language processing continues to improve, users at times still need to learn exact commands and triggers. Moreover, by calling these agents voice assistants, some users expect voice assistants to be capable of performing far more complicated tasks and queries than is currently technically feasible. It seems that some of these users are imagining voice assistants similar to Apple’s Knowledge Navigator [7] concept video from the 1980s but the technology has yet to catch up to users’ imaginations.

Breakdowns in the car user journey
One of the primary use cases for voice assistants that users have reported to us is using a voice assistant while driving a car. Users perceive interacting with a smartphone’s voice assistant to be more convenient and safer than doing so using touch. But there are times when a user may need to initiate interacting with the voice assistant (turning on the microphone) by physically touching a button on a phone (rather than saying a hotword). In other situations, a user may be unable to fully complete a task using their voice and may need to use touch to complete the task. Finally, users have reported times when in-car Bluetooth systems get in the way of a smartphone’s voice assistant, creating complicated scenarios where a user is unsure which microphone is listening (the car’s or the phone’s) or which speaker (the car’s or the phone’s) will the voice assistant use to communicate with the user.

Summary
Voice user experience is an exciting yet relatively new field for user experience practitioners. In this paper, I have described two user experience research methods that some UX researchers at Google have employed in our work with voice assistants: usability studies and longitudinal studies. We also discussed four primary barriers we have observed in users adopting voice assistants: cultural considerations, performance anxiety and social costs, exceedingly high expectations, and breakdowns in the car user journey.
Acknowledgments
The author would like to thank Shumin Zhai for feedback on prior versions of this work, Nikhil Sharma for our many conversations about voice UX and research methods for voice UX, and Julie Aranda, Glen Murphy, and Dave Burke for supporting this work.

About the Author
Noor Ali-Hasan is a senior user experience researcher at Google, where she currently works on Android and the Google Pixel phone. Noor’s prior projects at Google included user research for Google Now and Chromecast. Prior to Google, Noor conducted user research at Samsung Electronics and Microsoft. She holds a master’s of science in human-computer interaction from the University of Michigan.

References