VERSE: Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search

Alexandra Vtyurina*
University of Waterloo
Waterloo, Ontario, Canada
sasha.vtyurina@uwaterloo.ca

Adam Fourney
Microsoft Research
Redmond, WA, USA
adamfo@microsoft.com

Meredith Ringel Morris
Microsoft Research
Redmond, WA, USA
merrie@microsoft.com

Leah Findlater
University of Washington
Seattle, WA, USA
leahkf@uw.edu

Ryen W. White
Microsoft Research
Redmond, WA, USA
ryenw@microsoft.com

ABSTRACT

People with visual impairments often rely on screen readers when interacting with computer systems. Increasingly, these individuals also make extensive use of voice-based virtual assistants (VAs). We conducted a survey of 53 people who are legally blind to identify the strengths and weaknesses of both technologies, and the unmet opportunities at their intersection. We learned that virtual assistants are convenient and accessible, but lack the ability to deeply engage with content (e.g., read beyond the first few sentences of an article), and the ability to get a quick overview of the landscape (e.g., list alternative search results & suggestions). In contrast, screen readers allow for deep engagement with content (when content is accessible), and provide fine-grained navigation & control, but at the cost of reduced walk-up-and-use convenience. Based on these findings, we implemented VERSE (Voice Exploration, Retrieval, and SEarch), a prototype that extends a VA with screen-reader-inspired capabilities, and allows other devices (e.g., smartwatches) to serve as optional input accelerators. In a usability study with 12 blind screen reader users we found that VERSE meaningfully extended VA functionality. Participants especially valued having access to multiple search results and search verticals.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Virtual assistants; voice search; screen readers; accessibility

*Work done while at Microsoft Research.

ASSETS'19, October 28–30, 2019, Pittsburgh, PA, USA

ISBN 123-4567-24-567/08/06. . . $15.00
DOI: http://dx.doi.org/10.475/123_4

INTRODUCTION

People with visual impairments are often early adopters of audio-based interfaces, with screen readers being a prime example. Screen readers work by transforming the visual content in a graphical user interface into audio by vocalizing on-screen text. To this end, they are an important accessibility tool for blind computer users – so much so that every major operating system includes screen reader functionality (e.g., VoiceOver[1], TalkBack[2], Narrator[3]), and there is a strong market for third-party offerings (e.g., JAWS[4], NVDA[5]). Despite their importance, screen readers have many limitations. For example, they are complex to master, and depend on the cooperation of content creators to provide accessible markup (e.g., *alt text* for images).

Voice-activated virtual assistants (VAs), such as Apple's Siri, Amazon's Alexa, and Microsoft's Cortana, offer another audio-based interaction paradigm, and are mostly used for everyday tasks such as controlling a music player, checking the weather, and setting up reminders [47]. In addition to these household tasks, however, voice assistants are also used for general-purpose web search and information access [31]. In contrast to screen readers, VAs are marketed to a general audience and are limited to shallow investigations of web content. Being proficient users of audio-based interfaces, people who are blind often use VAs, and would benefit from broader VA capabilities [36, 2].

In this work, we explore opportunities at the intersection of screen readers and VAs. Through an online survey with 53 blind screen reader and VA users, we investigated the pros and cons of searching the web using a screen reader-equipped web browser, and when getting information from a voice assistant. Based on these findings, we developed VERSE (Voice Exploration, Retrieval, and SEarch) – a prototype that augments the VA interaction model with functionality inspired by screen

[1] https://www.apple.com/accessibility/mac/vision/
[2] https://support.google.com/accessibility/android/answer/6283677
[3] https://www.microsoft.com/en-us/accessibility/windows
[4] https://www.freedomscientific.com/Products/Blindness/JAWS
[5] https://www.nvaccess.org/

readers to better support free-form, voice-based web search. We then conducted a design probe study of VERSE, and identified future directions for improving eyes-free information-seeking tools.

This work addresses the following research questions:

- **RQ1:** What challenges do blind people face when: (a) seeking information using a search engine and a screen reader versus (b) when using a voice assistant?
- **RQ2:** How might voice assistants and screen readers be merged to confer the unique advantages of each technology?
- **RQ3:** How do blind web searchers feel about such hybrid systems, and how might our prototype, VERSE, be further improved?

In the following sections we cover prior research, the online survey, the functionality of VERSE, and the VERSE design probe study. We conclude by discussing the implications of our findings for designing next-generation technologies that improve eyes-free web search for blind and sighted users by bridging voice assistants and screen readers paradigms.

RELATED WORK

This work builds on several distinct threads of prior research, as detailed below.

Web Exploration by Screen Reader Users

Accessing web content using a screen reader can be a daunting task. Though the Web Content Accessibility Guidelines (WCAG [6]) codify how creators can improve the accessibility of their content, many websites fail to adhere to these guidelines [13, 22]. For example, Guinness et al. report that, in 2017, alternative text was missing from 28% of the images sampled from the top 500 websites indexed by alexa.com [22]. More generally, poor design and inaccessible content are the leading reasons for frustration among screen reader users [27], despite nearly two decades of web accessibility research. In fact, many of the challenges described by Jonathan Berry in 1999 [10] are still relevant to this day [25, 14, 42]. Consequently, screen reader users learn a variety of workarounds to access inaccessible content: they infer the roles of unlabeled elements (e.g., buttons) by exploring the nearby elements, they develop "recipes" for websites by memorizing their structure, and they use keyword search to skip to relevant parts of documents [15]. Even with these mitigation strategies, comparative analysis has shown that blind users require more time per visited web page compared to sighted users, signalling that more accessibility research is needed to to close this gap [40, 12].

Web search engines pose additional unique challenges to screen reader users. Sahib et al. [40] found that blind users may encounter problems at every step of information seeking, and showed lower levels of awareness of some search engine features such as query suggestions, spelling suggestions, and related searches, compared to sighted users. Although these features were accessible according to a technical definition, using them was time consuming and cumbersome [35]. Likewise, Bigham et al. [12] found that blind participants spent significantly longer on search tasks compared to sighted participants, and exhibited more probing behaviour (i.e., "a user leaves and then quickly returns to a page" [12]) showing greater difficulty in triaging search results. Assessing trustworthiness and credibility of search sources can also pose a problem. Abdolrahmani et al. [3, 1] found that blind users use significantly different web page features from sighted users to assess page credibility.

In this paper, our survey lends further support to these prior findings on web accessibility, and extends them to include challenges encountered when using voice-activated virtual assistants.

Novel Screen Reader Designs

Traditional screen readers provide sequential access to web content. Stockman et al. [43] explored how this linear representation can mismatch the document's spatial outline, contributing to high cognitive load for the user. To mitigate this issue, prior research has explored a variety of alternative screen reader designs [39], which we briefly outline below.

One approach is to use concurrent speech, where several speech channels simultaneously vocalize information [21, 52]. For example, Zhu et al.'s [52] Sasayaki screen reader augments primary output by concurrently whispering meta information to the user.

A method for non-visual skimming presented by Ahmed et al. [4] attempts to emulate visual "glances" that sighted people use to roughly understand the contents of a page. Their results suggest that such non-visual skimming and summarization techniques can be useful for providing screen reader users with an overview of a page.

Khurana et al. [26] created SPRITEs – a system that uses a keyboard to map a spatial outline of the web page in an attempt to overcome the linear nature of screen reader output. All participants in a user evaluation completed tasks as fast as, or faster than, with their regular screen reader.

Another approach, employed by Gadde et al. [20], uses crowdsourcing methods to identify key semantic parts of a page. They developed DASX – a system that transported the users to the desired section using a single shortcut based on these semantic labels; as a result, they saw performance of screen reader users rise significantly. Islam et al. [24] used linguistic and visual features to segment web content into semantic parts. A pilot study showed such segmentation helped the user navigate quickly and skip irrelevant content. Semantic segmentation of web content allows clutter-free access, at the same time reducing the user's cognitive load.

Our work builds on these prior systems by employing elements of summarization and semantic segmentation to allow people to quickly understand how search results are distributed over verticals (e.g., web results, news, videos, shopping, etc.)

Virtual Assistant Use by People Who Are Blind

A number of recent studies have explored user behaviors with VAs among the general population [29, 30], as well as elderly users, children, and, in particular, people with disabilities [17, 49, 2, 36, 50]. Voice assistants, and more generally voice

[6] https://www.w3.org/WAI/standards-guidelines/wcag/

interfaces, can be a vital productivity tool for blind users [7]. Abdolrahmani et al. [2] explored how this population uses voice assistants, as well as the main concerns and error scenarios they encounter. They found that VAs can enable blind users to easily make use of third party apps and smart home devices that otherwise would cause problems, but that VAs sometimes return suboptimal answers (either too verbose or too limited), and that there are privacy concerns around using VAs in public. Further, they found that VAs and screen readers can interfere with each other, complicating interactions (e.g., the screen reader can trigger a VA by reading a wake word that appears on the screen, or both may start speaking at the same time). Pradhan et al. [36] analyzed Amazon reviews of VAs purchased by people with disabilities and conducted semi-structured interviews with blind VA users. Their findings were similar to those of Abdolrahmani et al. [2], providing further evidence of the utility of VAs for people with visual impairments.

Our survey lends further support to findings regarding the use of VAs by people who are blind, and adds new information specifically about search tasks and around users' mental models regarding the roles of screen readers versus VAs.

Voice-controlled Screen Readers

Prior work has also explored the use of voice commands to control screen reader actions. Zhong et al. [51] created JustSpeak – a solution for voice control of an Android OS. JustSpeak accepts user voice input, interprets it in the context of metadata available on the screen, tries to identify the requested action, and finally executes this action. The authors outline potential benefits of JustSpeak for blind and sighted users. Ahok et al. [6] implemented CaptiSpeak – a voice-enabled screen reader that is able to recognize commands like "click <name> link," "find <name> button," etc. Twenty participants with visual impairments used CaptiSpeak for the task of online shopping, filling out a university admissions form, finding an ad on Craigslist, and sending an email. CaptiSpeak was found to be more efficient than a regular screen reader. Both JustSpeak and CaptiSpeak reduce the number of user actions needed to accomplish a task by building voice interaction into a screen reader. In this paper we investigate a complementary approach, which adds screen-reader-inspired capabilities to VAs, rather than adding voice control to screen readers.

Voice Queries and Conversational Search

Finally, prior research has explored voice-based information retrieval systems. For example, Guy [23] investigated how voice search queries differ from text queries across multiple dimensions, including context, intent, and the type of returned results. Trippas et al. [45, 44] studied user behaviour during conversational voice search for tasks with differing complexity. In their other work, Trippas et al. [46] studied audio and text representation of web search results, and found that users prefer longer summaries for text representation, while preferences for audio representation varied depending on the task. Radlinski et al. [38] proposed a theoretical model for a conversational search system. They outlined possible scenarios and the desired system behavior for producing answers in a natural and efficient manner. This research activity shows that voice-based web search and browsing is not aimed exclusively at people who are blind, but is also of interest to a wider population.

In summary, past research has characterized the challenges people face when browsing the web with screen readers, and has sought to improve these accessibility tools through advances in semantic segmentation, summarization, and voice control. At the same time, VAs have emerged as a popular tool for audio-based access to online information, and, though marketed to a general audience, confer a number of accessibility and convenience advantages to blind users. Our work explores augmenting a VA interaction model with functionality inspired by screen readers. In so doing, we hope to broaden the range of online content that can be accessed from virtual assistants – especially among people who are already skilled at using screen readers on other devices.

ONLINE SURVEY

To better understand the problem space of non-visual web search, we designed an online survey addressing our first research question: What challenges do blind people face when (a) seeking information using a search engine and a screen-reader versus (b) when using a voice assistant?

Survey Design and Methodology

The survey consisted of 40 questions spanning five categories: general demographics, use of screen readers for accessing information in a web browser, use of virtual assistants for retrieving online information, comparisons of screen readers to virtual assistants for information seeking tasks, and possible future integration scenarios (e.g., voice-enabled screen readers). The survey questions are included as supplementary material accompanying this paper. When asking about the use of screen readers and virtual assistants, the survey employed a recent critical incident approach [19], in which we asked participants to think of recent occasions they had engaged in web search using each of these technologies. We then asked them to describe these search episodes, and to use them as anchor points to concretely frame reflections on strengths and challenges of each technology.

We recruited adults living in the U.S. who are legally blind and who use both screen readers and voice assistants. We used the services of an organization that specializes in recruiting people with various disabilities for online surveys, interviews, and remote studies. While we designed the online questionnaire to be accessible with most popular web browser/screen reader combinations, the partner organization worked with participants directly to ensure that content was accessible to each individual. In some cases, this included enabling respondents to complete the questionnaire by telephone. The survey took an average of 49 minutes to complete, and participants were compensated $50 for their time. The survey received an approval from our organization's ethics board.

A total of 53 people were invited to complete the survey. Since the recruiting agency was diligent in following up with respondents, there were no dropouts. The survey included numerous open-ended questions. Though answer lengths varied, the median word count for open-ended questions was 18 words (IQR = 19.5).

Two researchers iteratively analyzed the open-ended responses using techniques for open coding and affinity diagramming [28] to identify themes.

Participants

A total of 53 respondents completed the survey (28 female, 25 male). Participants were diverse in age, education level, and employment status. Ages were distributed as follows: 18-24 (9.4%), 25-34 (32%), 35-44 (22.6%), 45-54 (16.9%), 55-64 (11.3%), 65-74 (7.5%). Participants' highest level of education was: some high school, no diploma (1.8%), high school or GED (7.5%), some college, no diploma (32%), associate degree (13.2%), bachelor's degree (22.6%), some graduate school, no diploma (1.8%), graduate degree (20.7%). Occupation statuses were: employed full-time (39.6%), employed part-time (13.2%), part-time students (7.5%), full time student (11.3%), not currently employed (18.8%), retired (5.6%), not able to work due to disability (5.6%).

All participants reported being legally blind, and most had experienced visual disability for a prolonged period of time ($\mu = 31.6$ years, $\sigma = 17$ years). As such, all but three respondents reported having more than three years of experience with screen reader technology. Likewise, most of the participants were early adopters of voice assistant technology. 35 respondents (66%) reported having more than three years of experience with such systems. Of the remaining respondents, 17 (32%) had between one and three years of experience, and only one (2%) reported being new to VA technology (i.e., having less than one year of experience).

More generally, our respondents were active users of technology. 40 participants (75%) reported using three or more devices on an average day including: touchscreen smartphones (53 people; 100%), laptops (46 people; 87%), tablets (29 people; 55%), desktop computers (27 people; 51%), smart TVs (21 people; 40%) and smartwatches (11 people; 21%).

Findings

We found that respondents made frequent and extensive use of both virtual assistants and screen-reader-equipped web browsers to search for information online, but both methods had shortcomings. Moreover, we found that transitioning between VAs and browsers introduces its own set of challenges and opportunities for future integration. In this section we first detail broad patterns of use, then present specific themes around the technologies' advantages and challenges.

General Patterns of Use

Most of the respondents were active searchers: when asked how often they searched for answers or information online using a web browser and screen reader, 41 people said multiple times a day (77.3%), 9 searched multiple times a week (16.9%), 2 only once a day (3.7%), and 1 only searched multiple times a month (1.8%). The most popular devices for searching the internet were touchscreen smartphones (45 people) and laptops (41 people), as well as touchscreen tablets (23 people) and desktop computers (23 people).

Respondents also reported avid use of voice assistant technology. When asked how often they used voice assistants to find answers and information online, over half (29) reported using VAs multiple times a day, 7 said once a day, 11 said multiple times a week, and 6 said once a week or less often. VAs were accessed from a variety of devices including: smartphones (53p, 100%), smart speakers (34p, 64%), tablets (18p, 33.9%), laptops (15p, 28.3%), smart TVs (13p, 24.5%), smartwatches (7p, 13.3%), and desktop computers (5p, 9.4%). The most popular assistant used on a smartphone was Siri (used by 51 people), followed by Google Assistant (23 people) and Alexa (18 people). Fewer people used assistants on a tablet, but a similar pattern emerged, with Siri the most popular (18p), followed by Alexa and Google Assistant (8 people each). Amazon Echo was the most popular smart speaker among our respondents (29p), followed by Google Home (14 people) and Apple Home Pod (1p). The most popular assistant on laptops and desktops was Cortana (17p), followed by Siri (8p). Siri and Alexa were the most popular assistants on smart TVs (the Apple TV and Amazon Fire TV, respectively).

In sum, respondents made frequent and extensive use of both virtual assistants and screen-reader-equipped web browsers to search for information online. In open-ended responses, respondents also provided important insights into the trade-offs of each technology. Each trade-off is codified by a theme below.

Theme 1: *Brevity vs. Detail*

The amount of information provided by voice assistants can differ substantially from that returned by a search engine. VAs provide a single answer that is suitable for simple question answering, but less suited for exploratory search tasks [48]. This dualism clearly emerged in our data. 27 respondents noted that VAs provide a direct answer with minimal effort (P12: *"The assistant will read out information to me and all I have had to do is ask"*, P45: *"[VAs] are to the point and quick to respond"*, P40: *"when you ask Siri or Cortana, they just read the answer for you if they can, right off."*). On the other hand, 27 respondents complained that VAs provide limited insight. For example, P24 noted: *"a virtual assistant will only give you one or two choices, and if one of the choices isn't the answer you are seeking, it's hard to find any other information"*. Likewise, P37 explained: *"you just get one answer and sometimes it's not even the one you were looking for"*. A similar sentiment was expressed by P30: *"a lot of times, a virtual assistant typically uses one or two sources in order to find the information requested, rather than the entire web"*.

In contrast, 20 respondents thought that search engines were advantageous in that they allow the user to review a number of different sources, triage the search results, and access more details if needed (P9: *"information can be gathered and compared across multiple sources"*, P46: *"you can study detailed information more thoroughly"*). But, those details come at a price – using a screen reader a user has to cut through the clutter on web pages before getting to the main content – a sentiment shared by 8 respondents (P18: *"you don't get the information directly but instead have to sometimes hunt through lots of clutter on a web page to find what you are looking for"*, P19: *"the information I am seeking gets obfuscated within the overall web design of the Google search experience. Yelp, Google, or other information sites can be over designed or poorly designed while not taking any of the WCAG standards into consideration"*).

Theme 2: *Granularity of Control vs. Ease of Use*
Our survey participants widely recognized (22 people) that VAs were a convenient tool for performing simple tasks, but greater control was needed for in-depth exploration (P38: *"They are good for specific, very tailored tasks."*). This trade-off in control, between VAs and screen-reader-equipped browsers, was apparent at all stages of performing a search: query formulation (P30: *"[with VAs] you have to be more exact and precise as to the type of information you are seeking."*), results navigation (P22: "*[with screen readers] I can navigate through [results] as I wish*"), and information extraction and reuse (P51: "*If I use a screen reader for web searching I can bookmark the page and return to it later. I cannot do it with a virtual assistant.*") In regards to the latter stage, eight participants noted that information found using a VA does not persist – it vanishes as soon as it is spoken (P47: *"With a virtual assistant, I don't know of a way to save the info for future release. It doesn't seem efficient for taking notes."*). Additionally, sharing information with third party apps is impossible to achieve using a VA (P47: *"[with the screen reader] I can copy and paste the info into a Word document and save it for future use."*).

Additionally, 15 respondents reported that screen readers are advantageous in that they provide a greater number of navigation modes, each operating at different granularities (P24: *"It's easier to scan the headings with a screen reader when searching the web"*, P31: *"one is able to navigate through available results much faster than is possible with virtual assistants."*, P40: *"With something like Siri and Cortana you <...> have to listen very carefully because they won't go back and repeat unless you ask the question again, or use VoiceOver or Jaws to reread things."*) Likewise, users can customize multiple settings (speech rate, pitch) to fit their preferences – a functionality not yet available in voice assistants (P29: *"sometimes you can get what you need quicker by going down a web page [with a screen reader], rather then waiting for the assistant to finish speaking"*). While the issue of VAs' fixed playback speed was only mentioned by one participant, previous research suggests it may be a more common concern [2].

The increased dexterity of screen readers comes at a price of having to memorize many keyboard commands or touch gestures, whereas VAs require minimal to no training (P38: *"[with VAs] you don't have to remember to use multiple screen reader keyboard commands"*). This specific trade-off was mentioned by three participants.

Theme 3: *Text vs. Voice*
According to 24 of our respondents, speaking a query is often faster than typing it (P9: *"typing questions can take more time"*), less effortful (P32: *"It is easier to dictate a question rather than type it."*), and can help avoid spelling mistakes (P53: *"You do not know how to spell everything"*). That said, speech recognition errors are a major problem (mentioned by 39 respondents) and can cancel out the benefits of voice input (P48: *"I can type exactly what I want to search for and don't have to edit if I'm heard incorrectly by the virtual assistant."*) and even lead to inaccurate results (P23: *Virtual assistant often 'mishears' what I am trying to say. The results usually make no sense.*) Especially prone to misrecognition are queries containing *"non-English words, odd spellings, or homophones"* (P19). Environmental conditions can create additional obstacles for voice input and output (P3:*"it [voice interaction] is nearly impossible in a noisy environment, such as a crowded restaurant. Even when out in public in a quiet environment, the interaction may be distracting to others."*). Environmental limitations of voice assistant interaction were pointed out by six of our respondents and have also surfaced as a user concern in prior work for phone-based [18] and smart-speaker-based assistants [2].

Theme 4*: Portability vs. Agility*
Assistants are either portable – such as Siri on an iPhone (P46:*"Its in your pocket practically all the time"*), or are always ready to use – like smart speakers (P15: *"I can be on my computer doing an assignment and ask Alexa"*). On the other hand, to use a screen reader one needs to spend time setting up the environment before performing the search (P37: *"It takes more time to go to the computer and find the browser and type it in and surf there with the results"*). This fact was noted by 20 respondents.

Eight respondents also emphasized the hands-free nature of interaction with VAs as an opportunity for physical multitasking (P33: *"[VAs are] especially helpful if I have my hands dirty or messy while cooking"*, P45: *"using [VAs] without having to touch anything is awesome."*).

Theme 5: *Incidental vs. Intentional Accessibility*
One of the major obstacles for screen reader users is inaccessible content due to poor website design [13, 22] and the lack of compliance with WCAG guidelines. Such content can be difficult or impossible to access using screen readers (for example, text embedded in pictures). On the other hand, the content provided by VAs is audio-based, making their content inherently accessible through an audio channel (P38: *"You don't have to worry about dealing with inaccessible websites."*). Such an approach *"levels the playing field, as it were (everyone searches the same way)."* (P42). The notion of accidental accessibility of VAs was previously discussed in Pradhan et al. [36].

Theme 6: *Transitioning between Modalities*
Another theme worth noting is transitioning from a VA to a screen reader. To study this part of respondents' experience, we used a recent critical incident approach and asked participants to describe a case when they started by asking a VA a question, but then switched to using a search engine with a screen reader. 39 respondents said they needed to do this switch at some point. Reasons for switching mentioned in participants' incident descriptions included VAs returning a non-relevant answer or no answer at all (14 people), VAs not

providing enough details in the answer (11), and failure of speech recognition (5), especially when non-trivial words were involved. When asked about the ideal scenario for a transition between a VA and a screen reader, respondents suggested persisting VAs' responses by sending an email, supporting smooth transitions to continuing in-depth search with a screen reader (P24: *"A virtual assistant could give you basic information and then provide a link to view more in depth results using a screen reader."*), and performing more in-depth voice-based search upon a user's request (P21: *"[VA] would ask you if you wanted more details. If you replied yes, it would open a web page such as google and perform a search"*).

VERSE

Inspired by our survey findings and the aforementioned related work, we created VERSE (Voice Exploration, Retrieval and SEarch), a prototype situated at the intersection of voice-based virtual assistants and screen readers. Importantly, VERSE serves as a design probe, allowing us to better understand how these technologies may be merged (RQ2), and how such systems may impact VA-based information retrieval (RQ3). In this section we described VERSE in detail. Later, we present the results of a design probe study.

Overview

When using VERSE, people interact with the system primarily through speech, in a manner similar to existing voice-based devices such as the Amazon Alexa or Google Home Assistant. For example, when asked a direct question, VERSE will often respond directly with a concise answer (Figure 1a). However, VERSE differs from existing agents in that it enables an additional set of voice commands that allow users to more deeply engage with content. The commands are patterned on those found in contemporary screen readers, for example, allowing navigation over a document's headings.

As with screen readers, VERSE addresses the need to provide shortcuts and accelerators for common actions. To this end, VERSE optionally allows users to perform gestures on a companion device such as a phone or smart watch (see Table 2). For most actions, these companion devices are not strictly necessary. However, to simplify rapid prototyping, we limited microphone activation to gestures, rather than also allowing activation via keyword spotting (e.g., "Hey Google"). Specifically, microphone activation is implemented as a double-tap gesture performed on a companion device (e.g., smartphone or smartwatch). Although hands-free interaction can be a key functionality for VA users [30], a physical activation is a welcomed ancillary, and at times, a preferred option [2]. There are no technological blockers for implementing voice-only activation in future versions of VERSE.

The following scenario illustrates VERSE's capabilities and user experience.

Example Usage Scenario

Alice recently overheard a conversation about the Challenger Deep and is interested to learn more. She is sitting on a couch, her computer is in another room, and a VERSE-enabled speaker is on the coffee table. Alice activates VERSE and asks "What is the Challenger Deep?". The VERSE speaker responds with a quick answer – similar to Alice's other smart speakers – but also notes that it found a number of other web pages, Wikipedia articles, and related searches (Table 1a). Alice decides to explore the Wikipedia articles ("Go to Wikipedia"), and begins navigating the list of related Wikipedia entries ("next") before backtracking to the first article, this time rotating the crown on her smartwatch as a shortcut to quickly issue the *previous* command (Table 1b).

Alice decides that the first Wikipedia article sounded good after all, and asks for more details ("Tell me more"). VERSE loads the Wikipedia article and begins reading from the introduction section (Table 1c), but Alice interrupts and asks for a list of section titles ("Read section titles"). Upon hearing that there is a section about the Challenger Deep's history, Alice asks for it by section name ("Read history").

Finally, Alice wonders if there may be other useful resources beyond Wikipedia, and decides to return to the search results ("Go to web results"). As before, Alice rotates the crown on her smart watch to quickly scroll through the results. Alice identifies an interesting webpage from the list VERSE reads out to her, and decides to explore it more deeply on her phone ("Send this to my phone"); the chosen web page opens on her iPhone (Table 1d), where Alice can navigate it using the phone's screen reader.

VERSE Design Elements

The design of VERSE was informed by the themes that emerged in the survey. Below we discuss how VERSE directly addresses four of the six themes. The remaining two themes – Text vs. Voice, and Protability vs. Agility – are not directly relevant to VERSE's current focus on smart-speaker-based voice interaction.

VERSE: Brevity vs. Detail

The design of VERSE addresses *Theme 1* by providing users with an instant concise answer (in a manner similar to VAs) as well as an opportunity to explore multiple sources returned by a web search engine (breadth), and then to engage with individual content items and web sources (depth).

With respect to breadth, VERSE leverages the Bing.com search API[7] to collect results across numerous search verticals: web pages, facts, entities, dictionary definitions, Wikipedia articles, news stories, images, videos, and related search queries. If facts, entities, or definitions are present, VERSE reads them out similar to existing VAs, then follows by summarizing the results available in other verticals (Table 1a).

With respect to depth, VERSE allows voice- and gesture-based navigation of Wikipedia articles. We chose Wikipedia as it has rich data, is often included among the top results, and has a consistent page structure that facilitates semantic navigation. When a desired Wikipedia article is selected, the user can say "tell me more," or perform an alternative gesture (Table 2) to get a quick summary of the article (e.g., the number of sections and words), then hear narration of the first section. At any

[7] https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/

Alice: *(activation)* What is the Challenger Deep?

VERSE: The Challenger Deep is the deepest known point in the Earth's seabed hydrosphere. With the depth 10898 to 10916m by direct measurements from submersibles and slightly more by sonar bathymetry. I also found 2 entities, 9 web pages, 8 related search queries, 10 videos, 10 Wikipedia articles.

(a) Initial search

Alice: *(activation)* Go to Wikipedia.

VERSE: I found 10 Wikipedia articles. The first one is Challenger Deep, Wikipedia. The Challenger Deep is the...

Alice: *(activation)* next article

VERSE: Article 2. Deepsea Challenger, Wikipedia. Deep Sea Challenger, DCV1, is 7.3 metre, 24 feet...

Alice: (*rotate smartwatch crown to return to previous element*)

VERSE: Article 1. Challenger Deep, Wikipedia. The Challenger Deep is the deepest ...

(b) Results navigation

Alice: *(activation)* Tell me more.

VERSE: I parsed the Wikipedia article called Challenger Deep. It has 16 sections and 3127 words. Here's the first section. Introduction. The Challenger Deep is the deepest...

Alice: *(activation)* Read section titles.

Verse: Introduction. Section 1: History of depth mapping from the surface. Section 2: Descents. Section 2.1...

Alice: *(activation)* Read History.

VERSE: Okay, here's the section number 1, titled History of depth mapping from the surface. Over many years, the search for the point of maximum depth...

(c) Wikipedia exploration

Alice: *(activation)* Go to web results.

VERSE: I found 9 web pages. The first one is Challenger Deep, Wikipedia.

Alice: *(rotates smartwatch crown to advance to next result)*

VERSE: Page 2. Deepest ocean Challenger Deep, Extreme Science. Challenger Deep got its name from...

Alice:*(activation)* Send this to my phone.

VERSE: Opening this link in your phone app. *(Web page opens within VERSE app)*

Alice: *(uses VoiceOver gestures to navigate on the page)*

(d) Transition between modalities

Table 1: Example usage scenario

time, the user can ask for an overview of the article's sections (*"read section titles,"*), and can ask for a given section by name (*"read <section name>"*). This interaction is illustrated in Table 1c.

VERSE: Granularity vs. Ease of Use

To address *Theme 2* from the survey findings, VERSE allows users a quick and easy way to navigate between search results using either voice commands or touch gestures. By saying *"next"* or *"previous,"* the user is able to move on to the next element in the selected search vertical (Table 1b). A similar effect is achieved by swiping right and left on a companion device (Table 2). These gestures mirror those used by screen readers on popular smart phones.

To switch between different search verticals, a user can say *"go to <search source>"* (e.g., *"Go to Wikipedia."*). VERSE will respond with the number of elements found in the new vertical and start reading the first element (Table 1b). Alternatively, the user can swipe up or down to move along the available search verticals.

Finally, when exploring Wikipedia articles, VERSE also supports screen-reader-inspired navigation modes (by headings, sentences, paragraphs, and words). The navigation mode then impacts the granularity of navigation commands and gestures, such as *"next"* and *"previous"*. Without loss of generality, one can switch modes by saying *"navigate by headings"*, or can swipe up or down on a companion device to iterate between modes – again, these gestures are familiar to people who use screen readers on mobile devices.

VERSE: Incidental vs Intentional Accessibility

VERSE addresses *Theme 5* by submitting user queries, and retrieving results via the Bing.com search API. This allowed us to design a truly audio-first experience consistent with existing VAs, rather than attempting to convert visual web content to auditory format. Likewise, our treatment of Wikipedia allows VERSE to focus on the article's main content rather than on visual elements. This behaviour is consistent with the concept of semantic segmentation [24]. It also mirrors the style of the brief summaries narrated by existing virtual assistants, but allows convenient and efficient access to the entire article content.

VERSE: Transitioning between Modalities

Finally, VERSE addresses *Theme 6* by giving users an opportunity to seamlessly transition between voice-based interaction and a more traditional screen-reader-equipped web browser. If the user requests an in-depth exploration of a web resource that is not Wikipedia, VERSE will open its url within the VERSE phone application. The user can then explore the web page using the device's screen reader. From this point onward, all gestures are routed to the default screen-reader until a "scrub" gesture is performed[8] or a new voice query is issued. Gesture parity between VERSE and popular screen readers ensures a smooth transition. This interaction is illustrated in Table 1d.

[8] A standard VoiceOver gesture for "go back".

Table 2: Mapping of voice commands and corresponding gestures in VERSE

Voice commands	**Phone gestures**	**Watch gesture**	**Action**
(Activation gesture)	Double tap with two fingers	Double tap with one finger	VERSE opens mic
"Cancel"	One tap with two fingers	One tap with one finger	Stop voice output
"Go to <source>"	Up/down swipe	Up/down swipe	Previous/next search source
"Next"/"Previous"	Right/left swipe	Right/left swipe or rotate digital crown	Next/previous element
"Tell me more"	Double tap with one finger	n/a	Provide details if available or open link in the phone app

DESIGN PROBE

After developing the initial prototype and receiving an approval from our ethics board, we invited 12 blind screen reader users to use VERSE, and to provide feedback pertaining to our second and third research questions. In the following sections we detail the procedure, describe the participants, then present participant feedback.

Procedure

Participants completed consent forms, provided demographic information, then listened to a scripted tutorial of VERSE's voice commands and gestures. Each participant was asked to use VERSE to complete two search tasks, and to think aloud as they engaged with the system. One of the tasks was pre-specified and the same for all participants; specifically, participants were asked to find two or three uses for recycled car tires. This task has previously been used in investigations of conversational speech-only search systems [45], is characterized as being of intermediate cognitive complexity, and occupies the "Understanding" tier of Krathwohl's Taxonomy of Learning Objectives [5]. Completing the task requires consulting multiple sources or documents, [8], and is thus difficult to perform with contemporary VAs. In a second task, participants were asked to express their own information need by searching for a topic of personal interest. Half the participants began with the fixed task, and half began with their own task. Each task had a time limit of 10 minutes.

This design was not meant to formally compare search outcomes on tasks of different difficulties – indeed, we had no control over the difficulty of self-generated tasks. Rather, the fixed task ensured that we observed a variety of strategies for a moderately complex information need, whereas the self-generated task ensured that we observed a variety of information needs for which we had no advance knowledge. Together, this provided a varied set of experiences with the system that would provoke interesting opportunities for observation and comment.

Regardless of task order, the first search session required participants to use a smart phone for gesture input, while the second session used a smart watch. This order of introduction reflects anticipated real-world use where phones would be the primary controller, with watches an optional alternative.

Throughout the tasks, participants were encouraged to think aloud. Following the completion of both tasks, participants completed the System Usability Scale (SUS) questionnaire [16]. Finally, the interviewer conducted an exit interview, prompting participants to provide open-ended feedback and suggestions. Participants' comments during the study, and their responses to the interview questions, were transcribed and analyzed by two researchers using a variation of open coding and affinity diagramming [28].

Participants

We recruited 12 blind screen reader users (4 female, 8 male) through a mailing list in the local community. Participants were reimbursed $50 for their time. We also offset their transportation costs to our laboratory by up to $50. The study lasted about an hour.

Participants' average age was 36.6 years old ($\sigma = 13.8$ years). Seven reported being totally blind and five were legally blind but had some residual vision. Ten participants had their vision level since birth, and two reported having reduced vision for 15 or more years. Participants had an average of 18.5 years of experience with screen readers ($\sigma = 7.6$ years), and 5.7 years of experience with VAs ($\sigma = 2.5$ years). For comparison, at the time that the study was conducted, Apple's Siri VA had been available on the market for 6.9 years, suggesting that our participants were indeed early adopters of this technology.

SYSTEM USABILITY

All participants successfully completed the fixed search task, which required that they identify at least two uses of used car tires. Though it was difficult to apply a common measure of completeness or correctness for user-chosen queries, we report that participants indicated satisfaction with VERSE's performance, as is reflected in open-ended feedback, and in responses to items on the System Usability Scale.

VERSE received a mean score of 71.0 ($\sigma = 15.5$) on the System Usability Scale. To aid in interpretation, we note this score falls slightly above the average score of 68, reported in [41], and just below the score of 71.4, which serves as the boundary separating systems with "Ok" usability from those with "Good" usability, according to the adjective rating scale developed by Bangor et al. in [9]. Breaking out individual items, we found that most participants found VERSE to be "*easy to use*" (median: 4, on a 5-point Likert scale), and its features were "*well integrated*" (median: 3.5). Likewise, participants "*felt very confident using the system*" (median: 4), and reported that they would "*use the product frequently*" (median: 4). These results suggest that the VERSE prototype reached a sufficient quality to serve as a design probe, and to ground meaningful discussions of VERSE's capabilities.

PARTICIPANT FEEDBACK

Participants commented on VERSE throughout use, and answered questions about the prototype in an exit interview. Here, participants' feedback was generally positive, and largely aligned with responses to SUS items, described above. For instance, participants reported that the system was easy to learn, given prior experience with screen readers ("*if we're talking about screen reader users, they kind of know what they are doing, I think it would be fairly easy*," P4). In this capacity, VERSE's gesture accelerators were especially familiar ("*the touch experience doesn't feel that different from VoiceOver (...) I think I would have probably figured them out on my own*," P3; "*[Y]ou're just using the same gestures as VoiceOver, and that, in itself, is comprehensive.*," P5).

Participants also found that VERSE extended VAs in meaningful ways, increasing both the depth and breadth of exploration. For instance, P4 reported:

> *"The information it gives is quite a bit more in-depth. [...] There was one time I asked Siri something about Easter eggs. Siri said 'I found this Wikipedia article, do you want me to read it to you?' [...] It only read the introduction and then stopped, and I think [VERSE] could come in so that you can read whole sections."*

Likewise, P7 reported:

> *"[VERSE] gives you a lot more search options like web pages, or Wikipedia. Even though the smart speaker I use [Echo] has some ability to read [Wikipedia], I can't get back and forth by section and skip around. In that way, it's an improvement. I like it."'*

However, participants were more mixed about how VERSE compared to traditional screen readers. For instance, P7 noted "*screen readers are a lot more powerful*", whereas P6 noted "*I like it better than desktop screen readers, but I would probably prefer phone screen readers*." VERSE was never intended to replace screen readers, and was instead focused on extending the web search and retrieval capabilities of VAs with screen-reader-inspired functionality. This point was immediately recognized by P5, who noted:

> *"I think [VERSE and Screen Readers] are fundamentally different. There's just no way to compare them. Screen readers aren't for searching for stuff, they are about giving you control."*

Restricted to the domain of web search and retrieval, VERSE was found to confer numerous advantages. P10 commented that, compared to accessing web search with a screen reader, VERSE was "*Much better. This gives you much more structure.*" P3 elaborated further:

> *"Most screen readers and search engines do use headings, [...] but it's hard to switch [search verticals]. This is different and kind of interesting. It seems to put you at a higher level."*

This sentiment was echoed by P5, who explained:

> *"One thing that immediately caught my eye was that different forms of data were being pulled together. When you go to Google and you type in a search you just get a stream of responses. [VERSE] gathers the relevant stuff and groups it in different ways. I really did like that."*

Additionally, participants expressed a strong interest in voice, often preferring it to gestural interaction. For instance, P8 stated "*Just using voice would be fine with me.*", while P7 noted:

> *"I preferred voice integration. There were times where it's just going to be faster to use my finger to find it, but mostly [I preferred] voice."*

Other participants offered more nuanced perspectives, noting that gestures were advantageous for high-frequency navigation commands. ("*I liked being able to use the gestures. [With voice] it would have been 'next section', 'next section.'* ", P6; "*I liked the gestures. I will spend more time with gesture, but getting this thing started with voice is beautiful.*", P9).

Nevertheless, participants reported concerns that voice commands were difficult to remember (e.g., "*I didn't find the system complicated. I'd say the most complicated part is the memorization of [...] the voice commands.*", P3). To this end, participants expressed a strong desire for improvements to conversation and document understanding. For instance, P3 expressed "*I should just have the ability to use [a] more natural voice like I'm having a conversation with you.*" Likewise, P5 explained:

> *"I'm most passionate about the whole language understanding part, where I [would like] to say 'read the paragraph that talks about this person's work' and it should understand."*

Recent results in machine reading comprehension and question answering [34] may provide a means of delivering on this promise; this remains an important area for future work.

Finally, all 12 participants preferred using the phone over the watch. Several factors contributed to this preference including: the limited input space of the watch ("*I've got fat fingers [...] and on that device feels very cumbersome*", P9), a power-saving feature that caused the screen to occasionally lose focus ("*It was a little annoying [when] I lost focus on the touch part of the screen*", P3), and latency incurred by the watch's aggressive powering-down of wireless radios ("*The watch wasn't bad, but it lagged a little. That was my chief complaint." P7*).

In sum, participants were generally positive about the VERSE prototype, and expressed interest in its continued development or public release. The design probe further revealed that participants were especially positive about voice interaction, and the expanded access to web content afforded by VERSE. While we hypothesized that watch-based interaction would be an asset (given that watches are always on hand), their appeal is diminished by the limitations of current form factors and hardware. Conversely, extending the conversation and document understanding capabilities of VERSE is a desirable avenue for future work.

DISCUSSION

Three over-arching questions motivated this research: What challenges do blind people face when seeking information via a web browser or voice assistant? How might voice assistants and screen readers be merged? And, how do blind web searchers feel about such hybrid systems? We addressed each question in turn by: conducting an online survey with 53 blind web searchers, developing a prototype system that adds screen-reader-inspired capabilities to a VA, and then collecting feedback from 12 blind screen reader users.

From the survey, we found that screen readers and VAs present a series of trade offs spanning dimensions of brevity, control, input modality, agility, incidental accessibility, and paradigm transitions. We found that transitions between the technologies can be especially costly. In the prototype, we worked to eliminate these trade offs and costs, by adding screen reader-inspired capabilities to a VA. An alternative approach would have been to augment a screen reader with voice and conversational controls, which, as noted earlier, has been explored in prior literature [6, 51]. We opted for the former since VAs are an emerging technology that open a new point in the design space, while also avoiding challenges with legacy bias [32]. For example, VERSE redefines search results pages by adding summaries, and by mapping screen reader navigation modes to search verticals. These features were received positively by design probe participants. In the future, we hope to run a controlled lab study comparing VERSE to screen readers (or VAs), to determine if participants' stated preferences are reflected in measurable reductions in task performance time or other performance metrics. Additionally, coexistence and complementary nature of VAs and screen readers bring up new research questions raised by our survey findings such as whether these two technologies should remain separate, be merged into a single technology, or be more carefully co-designed for compatibility.

We used a survey as a data collection tool to inform the design of VERSE. We recognize that there are many ways for collecting high-quality qualitative feedback, including interviews and contextual inquiries. In this work, we opted to collect such data using a "recent critical incident" approach [19], paired with open-ended survey questions, which provided us with rich data and allowed us to reach a large and geographically diverse audience.

We also recognize that contemporary VAs are often co-resident with other applications and software on computers or smart phones, and are used for tasks beyond web search and retrieval. In these settings, a similar set of VA limitations are likely to arise. For example, a VA might read recent messages, or help compose an email, but is unlikely to provide granular navigation of one's inbox folders. Generalizing VERSE to scenarios beyond web search is an exciting area of future research.

Likewise, we recognize that other user communities may also benefit from VERSE. For instance, sighted users may wish to have expanded voice access to web content when they are driving, cooking, or otherwise engaged in a task where visual attention is required – especially if VERSE were enriched with the document and conversation understanding capabilities discussed earlier. VERSE may also benefit other populations with print disabilities, such as people with dyslexia, who also have challenges using mainstream search tools [33]. Furthermore, all our survey participants were based in the U.S. Understanding the voice search needs of people from other regions [11, 37] is a valuable area of future work.

Finally, rather than accessing raw HTML, VERSE leverages APIs for Bing and Wikipedia to provide an audio-first experience. This is similar to other smart speaker software applications known as "skills." For general web pages, VERSE encounters the same challenges with inaccessible content as traditional screen readers. Given the broad appeal of smart speakers, it is possible that experiences such as VERSE could motivate web developers to consider how their content would be accessed through audio channels. For example, a recent proposal[9] demonstrates how web developers can tag content with Schema.org's `speakable` HTML attribute to help direct the Google Assistant to the parts of an article that can be read aloud. We are excited to explore a future where web developers consider smart speakers and audio devices as just one more point on the responsive design spectrum[10], thereby improving accessibility for everyone.

CONCLUSION

We have investigated the challenges that people who are blind experience when searching for information online using screen readers and voice assistants. To identify the gaps and opportunities for improvement, we ran an online survey with 53 screen reader and voice assistant users. Based on the findings from the survey, we created VERSE – a system prototype for non-visual web search and browsing. Design of VERSE combines the advantages of both screen readers and voice assistants, and allows voice-based as well as gesture-based interaction. We reported an design probe study of VERSE with twelve blind participants, and presented clear directions for future work for non-visual web access systems.

REFERENCES

1. Ali Abdolrahmani and Ravi Kuber. 2016. Should I trust it when I cannot see it?: Credibility assessment for blind web users. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 191–199.

2. Ali Abdolrahmani, Ravi Kuber, and Stacy M Branham. 2018. "Siri Talks at You": An Empirical Investigation of Voice-Activated Personal Assistant (VAPA) Usage by Individuals Who Are Blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 249–258.

3. Ali Abdolrahmani, Ravi Kuber, and William Easley. 2015. Web Search Credibility Assessment for Individuals who are Blind. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, 369–370.

[9] https://developers.google.com/search/docs/data-types/speakable

[10] https://en.wikipedia.org/wiki/Responsive_web_design

4. Faisal Ahmed, Yevgen Borodin, Andrii Soviak, Muhammad Islam, IV Ramakrishnan, and Terri Hedgpeth. 2012. Accessible skimming: faster screen reading of web pages. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 367–378.

5. Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. 2001. A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, abridged edition. *White Plains, NY: Longman* (2001).

6. Vikas Ashok, Yevgen Borodin, Yury Puzis, and IV Ramakrishnan. 2015. Capti-speak: a speech-enabled web screen reader. In *Proceedings of the 12th Web for All Conference*. ACM, 22.

7. Vikas Ashok, Yevgen Borodin, Svetlana Stoyanchev, Yuri Puzis, and IV Ramakrishnan. 2014. Wizard-of-Oz evaluation of speech-driven web browsing interface for people with vision impairments. In *Proceedings of the 11th Web for All Conference*. ACM, 12.

8. Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User variability and IR system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 625–634.

9. Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

10. Jonathan Berry. 1999. Apart or a part? Access to the Internet by visually impaired and blind people, with particular emphasis on assistive enabling technology and user perceptions. *Information technology and disabilities* 6, 3 (1999), 1–16.

11. Apoorva Bhalla. 2018. An exploratory study understanding the appropriated use of voice-based Search and Assistants. In *Proceedings of the 9th Indian Conference on Human Computer Interaction*. ACM, 90–94.

12. Jeffrey P Bigham, Anna C Cavender, Jeremy T Brudvik, Jacob O Wobbrock, and Richard E Ladner. 2007. WebinSitu: a comparative analysis of blind and sighted browsing behavior. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 51–58.

13. Jeffrey P Bigham, Ryan S Kaminsky, Richard E Ladner, Oscar M Danielsson, and Gordon L Hempton. 2006. WebInSight:: making web images accessible. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 181–188.

14. Jeffrey P Bigham, Irene Lin, and Saiph Savage. 2017. The Effects of Not Knowing What You Don't Know on Web Accessibility for Blind Web Users. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 101–109.

15. Yevgen Borodin, Jeffrey P Bigham, Glenn Dausch, and IV Ramakrishnan. 2010. More than meets the eye: a survey of screen-reader browsing strategies. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*. ACM, 13.

16. John Brooke and others. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

17. 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*. ACM, 43.

18. Aarthi Easwara Moorthy and Kim-Phuong L Vu. 2015. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human-Computer Interaction* 31, 4 (2015), 307–335.

19. John C Flanagan. 1954. The critical incident technique. *Psychological bulletin* 51, 4 (1954), 327.

20. Prathik Gadde and Davide Bolchini. 2014. From screen reading to aural glancing: towards instant access to key page sections. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. ACM, 67–74.

21. João Guerreiro and Daniel Gonçalves. 2016. Scanning for digital content: How blind and sighted people perceive concurrent speech. *ACM Transactions on Accessible Computing (TACCESS)* 8, 1 (2016), 2.

22. Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption Crawler: Enabling Reusable Alternative Text Descriptions using Reverse Image Search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 518.

23. Ido Guy. 2016. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 35–44.

24. Muhammad Asiful Islam, Faisal Ahmed, Yevgen Borodin, and IV Ramakrishnan. 2011. Tightly coupling visual and linguistic features for enriching audio-based web browsing experience. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2085–2088.

25. Melody Y Ivory, Shiqing Yu, and Kathryn Gronemyer. 2004. Search result exploration: a preliminary study of blind and sighted users' decision making and performance. In *CHI'04 extended abstracts on human factors in computing systems*. ACM, 1453–1456.

26. Rushil Khurana, Duncan McIsaac, Elliot Lockerman, and Jennifer Mankoff. 2018. Nonvisual Interaction Techniques at the Keyboard Surface. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 11.

27. Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What frustrates screen reader users on the web: A study of 100 blind users. *International Journal of human-computer interaction* 22, 3 (2007), 247–269.

28. Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.

29. Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2018. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* (2018), 0961000618759414.

30. Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5286–5297.

31. Rishabh Mehrotra, A Hassan Awadallah, AE Kholy, and Imed Zitouni. 2017. Hey Cortana! Exploring the use cases of a Desktop based Digital Assistant. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, Vol. 4.

32. Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, m. c. schraefel, and Jacob O. Wobbrock. 2014. Reducing Legacy Bias in Gesture Elicitation Studies. *interactions* 21, 3 (May 2014), 40–45.

33. Meredith Ringel Morris, Adam Fourney, Abdullah Ali, and Laura Vonessen. 2018. Understanding the Needs of Searchers with Dyslexia. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Article 35, 12 pages.

34. Kyosuke Nishida, Itsumi Saito, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2018. Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 647–656.

35. Helen Petrie and Omar Kheir. 2007. The relationship between accessibility and usability of websites. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 397–406.

36. Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. Accessibility Came by Accident: Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 459.

37. Aung Pyae and Paul Scifleet. 2018. Investigating differences between native English and non-native English speakers in interacting with a voice user interface: A case of Google Home. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*. ACM, 548–553.

38. Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, 117–126.

39. Iv Ramakrishnan, Vikas Ashok, and Syed Masum Billah. 2017. Non-visual Web Browsing: Beyond Web Accessibility. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 322–334.

40. Nuzhah Gooda Sahib, Anastasios Tombros, and Tony Stockman. 2012. A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers. *Journal of the American Society for Information Science and Technology* 63, 2 (2012), 377–391.

41. Jeff Sauro. 2011. *A practical guide to the system usability scale: Background, benchmarks & best practices*. Measuring Usability LLC Denver, CO.

42. Andreas Savva. 2017. *Understanding accessibility problems of blind users on the web*. Ph.D. Dissertation. University of York.

43. Tony Stockman and Oussama Metatla. 2008. The influence of screen-readers on web cognition. In *Proceeding of Accessible design in the digital world conference (ADDW 2008), York, UK*.

44. Johanne R Trippas. 2015. Spoken conversational search: Information retrieval over a speech-only communication channel. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1067–1067.

45. Johanne R Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, 325–328.

46. Johanne R Trippas, Damiano Spina, Mark Sanderson, and Lawrence Cavedon. 2015. Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 991–994.

47. Janice Y. Tsai, Tawfiq Ammari, Abraham Wallin, and Jofish Kaye. 2018. Alexa, play some music: Categorization of Alexa Commands. In *Voice-based Conversational UX Studies and Design Wokrshop at CHI*. ACM.

48. Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.

49. Linda Wulf, Markus Garschall, Julia Himmelsbach, and Manfred Tscheligi. 2014. Hands free-care free: elderly people taking advantage of speech-only interaction. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. ACM, 203–206.

50. Svetlana Yarosh, Stryker Thompson, Kathleen Watson, Alice Chase, Ashwin Senthilkumar, Ye Yuan, and AJ Brush. 2018. Children asking questions: speech interface reformulations and personification preferences. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*. ACM, 300–312.

51. Yu Zhong, TV Raman, Casey Burkhardt, Fadi Biadsy, and Jeffrey P Bigham. 2014. JustSpeak: enabling universal voice control on Android. In *Proceedings of the 11th Web for All Conference*. ACM, 36.

52. Shaojian Zhu, Daisuke Sato, Hironobu Takagi, and Chieko Asakawa. 2010. Sasayaki: an augmented voice-based web browsing experience. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 279–280.