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, as well as 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 Wikipedia), and the ability to get a quick overview of the landscape (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 increased complexity, and reduced walk-up-and-use convenience. In this demonstration, we showcase VERSE, a system that combines the positive aspects of VAs and screen readers, and allows other devices (e.g., smart watches) to serve as optional input accelerators. Together, these features allow people with visual impairments to deeply engage with web content through voice interaction.

CCS CONCEPTS

• Information systems → Search interfaces; **• Human-centered computing → Accessibility systems and tools**; **Natural language interfaces**; **Sound-based input / output**.

KEYWORDS

Virtual assistants, voice search, screen readers, accessibility

ACM Reference Format:
Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, Ryen W. White. 2019. Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308558.3314136

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3314136

1 INTRODUCTION

People with visual impairments are often expert users 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 people who are blind – 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). This includes the myriad of web page owners who host documents on the Internet.

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 [6]. In addition to these household tasks, however, voice assistants are also used for general purpose web search and information access [4]. Here, 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 [1, 5]. Additionally, extended VA functionality could be useful for a wider audience during activities such as driving and cooking.

In this work, we explore augmenting a VA interaction model with basic functionality of screen readers to better support free-form, voice-based web search. Through an online survey with 53 blind screen reader and VA users, we investigated what challenges people experience when searching the web with a screen reader and when getting information from a voice assistant. Based on these findings, we developed VERSE (Voice Exploration, Retrieval, and SEarch) – a prototype that employs a VA model, yet provides rich functionality for web exploration. In the remainder of this document we: (1) present a quick overview of the survey findings that motivate the

[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/

design of VERSE, (2) present an overview of the VERSE system, and (3) conclude by outlining the hardware and logistics requirements for demonstrating VERSE at TheWebConf'19.

2 ONLINE SURVEY

To better understand the problem space of non-visual web search with screen readers and VAs, we designed an online survey consisting of 44 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-controlled screen readers).

We recruited adults living in the U.S. who are legally blind and who use both screen readers and voice assistants. Through a collaboration with an online survey organization that specializes in recruiting people with various disabilities, we ensured that all participants could successfully access the survey. The survey took an average of 49 minutes to complete, and participants were compensated $50 for their time. Two researchers iteratively analyzed the open-ended responses using techniques for open coding and affinity diagramming [2] to identify themes.

2.1 Participants

A total of 53 respondents completed the survey (28 female, 25 male). Participants were diverse in age, education level, and employment status. 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 participants were experienced users of voice assistant technology. 35 respondents (66%) reported having more than three years of experience with such systems, 17 participants (32%) – between one and three years, and one participant (2%) reported having used VA technology for less than a year.

2.2 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 transitioning between VAs and browsers introduced additional challenges and opportunities for future integration. Each of these trade-offs is codified by a theme below.

2.2.1 ***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 (suitable for simple question answering but not for exploratory search tasks [8]), that may be short and provide limited insight. This aspect was pointed out by 27 respondents. For example P24 noted that: *"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"*. These concerns were echoed concisely by P37: *"you just get one answer and sometimes it's not even the one you were looking for"*, and 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 reported that using a search engine via a screen reader-equipped browser affords access to multiple sources, and access to more details if needed (e.g., P46: *"you can study detailed information more thoroughly"*). But those details come at a price – when using a screen reader a user has to cut through the clutter on web pages before getting to the main content, as mentioned by 8 survey respondents (e.g., 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"*).

2.2.2 ***Theme 2:*** *Granularity of Control vs Ease of Use.* Our survey participants widely recognized that VAs were a convenient tool for performing simple tasks (22 people), but greater control was needed for in-depth exploration (e.g., 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 management (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.*")

Additionally, 15 respondents reported that screen readers are advantageous in that they provide a greater number of navigation modes, each operating at different granularities. For example, P18 reports: *"[with screen readers] you can navigate by heading, landmark or words"*. Similar sentiments are reported by P24: *"It's easier to scan the headings with a screen reader when searching the web"*, and P31: *"one is able to navigate through available results much faster than is possible with virtual assistants."*

Finally, screen readers afford greater control by allowing users to customize multiple settings (speech rate, pitch) to fit people's preferences – a functionality not yet available in voice assistants (P29: *"sometimes you can get what you need quicker by going down a web page, rather then waiting for the assistant to finish speaking"*). The desire for customization of VAs was mentioned by only one participant in our survey, but has been identified as a limitation of VAs in prior work [1].

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 tradeoff was mentioned by 3 participants.

2.2.3 ***Theme 3:*** *Transitioning between systems.* Another prominent theme detailed the frequent need for users to transition from a VA to a screen reader-equipped web browser. Out of 53 survey respondents, 39 recalled recent situations where they began a session with a VA but then had to switch to using a web browser. Reasons for switching mentioned in participants' incident descriptions included: failure of automatic speech recognition (4 people), a VA response that lacked sufficient details (11 people), or the lack of any relevant response (14 people).

Transitions between system are not well-supported at present, and respondents suggested numerous ways in which this could be improved. For example P24 notes: *"A virtual assistant could give you basic information and then provide a link to view more in depth results using a screen reader."* Likewise P21 suggested that, upon performing in-depth search, the VA "*(could) ask you if you wanted more details.*

If you replied yes, it would open a web page such as Google [in a browser] and perform a search". Such a strategy would save people from having to re-input their query and begin a completely new search session.

2.2.4 ***Theme 4:*** *Incidental vs. Intentional Accessibility.*

Finally, one of the valuable features of voice assistants is their audio-first design. Thus, while targeting a general audience, VAs are immediately, and incidentally, accessible to people with visual impairments. This was mentioned by 7 participants. For example, P38 reports: *"You don't have to worry about dealing with inaccessible websites"*, while P42 notes that such an approach *"levels the playing field, as it were [since] everyone searches the same way."*

3 VERSE

Inspired by our survey findings, we created VERSE (Voice Exploration, Retrieval and SEarch), a prototype situated at the intersection of voice-based virtual assistants and screen readers. People interact with VERSE primarily through speech, in a manner similar to existing voice-based agents such as Amazon Alexa or Google 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 access different online sources and search engine features (such as related searches), as well as to engage more deeply with content for select sources (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 [3], a physical activation is a welcomed ancillary, and at times, a preferred option [1]. There are no technological blockers for implementing voice-only activation in the future versions of VERSE.

The following scenario, and the video accompanying this paper [7], illustrate VERSE's capabilities, and indicate how VERSE could be demonstrated at TheWebConf'19.

3.1 Example Demonstration Scenario

Alice recently overheard a conversation about the Challenger Deep and is interested to learn what it is. 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 VERSE 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 smart watch 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 section").

Finally, Alice wonders if there may be other useful resources beyond Wikipedia, and decides to return to the search results ("Go to search 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") and the chosen web page opens on her iPhone (Table 1d) where Alice can navigate it using the phone's screen reader.

3.2 VERSE Design Elements

The design of VERSE was informed by a select set of themes that emerged during the analysis of survey responses. Below we discuss how VERSE addresses each of the themes.

3.2.1 *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[6] 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 (see 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 screen reader-like 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 about the article (e.g., the number of sections and words), then hear narration of the first section. At any 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.

3.2.2 *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.

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

Table 1: Example usage scenario

(a) Initial search

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.

(b) Results navigation

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 ...

(c) Wikipedia exploration

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 section.
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...

(d) Transition between modalities

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)*

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 sources.

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 & 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.

3.2.3 VERSE: Transitioning between modalities. VERSE addresses *Theme 3* 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 built-in screen reader (in our case, VoiceOver). From this point onward, all gestures are routed to the default screen-reader until a "scrub" gesture is performed[7], or a new voice query, is issued. Gesture parity between VERSE and the screen reader ensures a smooth transition. This interaction is illustrated in Table 1d.

3.2.4 VERSE: Incidental vs Intentional Accessibility. Finally, as already noted, VERSE submits user queries, and retrieves results, via 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 other visual elements. This behaviour is consistent with the brief one or two sentence summaries narrated by existing virtual assistants, but allows convenient and efficient access to the entire article content.

4 EQUIPMENT AND LOGISTICS

We are proposing to run a demonstration that showcases the VERSE prototype system. The demo consists of a few smart speakers, smart watches, and laptops which we will supply. Our demonstration requires that internet connectivity – preferably wireless – be available. The demonstration should not require much physical space (e.g., one table), and can be scaled down as needed. Given the potential for a noisy environment, we will provide headsets with microphones, and will also modify the prototype to accept typed input as a contingency. We will also prepare a video showcasing the system [7]. Exhibit visitors will be free to conduct their own search and browsing sessions – we will not prescribe any particular flow or golden path through the system. Our hope is that the demo will inspire those researching web browser standards to consider new ways to support voice-based navigation of the web.

5 CONCLUSION

We have investigated the challenges that people who are blind experience when searching for information online using screen readers and voice assistants (e.g. Siri, Alexa). 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

[7] 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 crown	Next/previous element
"Tell me more"	Double tap with one finger	n/a	Continue reading the most recently mentioned answer / result

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.

REFERENCES

[1] Ali Abdolrahmani, Ravi Kuber, and Stacy 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.

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

[3] 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.

[4] Rishabh Mehrotra, A Hassan Awadallah, Ahmed E 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)*.

[5] 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.

[6] 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.

[7] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W. White. 2019. Demo video of VERSE. https://aka.ms/verse_demo_www2019

[8] 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.