Firefox Voice: An Open and Extensible Voice Assistant Built Upon the Web

Julia Cambre
jcambre@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Alex C. Williams
acw@utk.edu
University of Tennessee, Knoxville
Knoxville, Tennessee, USA

Afsaneh Razi
afsaneh.razi@knights.ucf.edu
University of Central Florida
Orlando, Florida, USA

Ian Bicking
ian@ianbicking.org
Mozilla
Minneapolis, Minnesota, USA

Abraham Wallin
abewallin@gmail.com
Mozilla
San Francisco, California, USA

Janice Tsai
janicetsai@acm.org
Mozilla
Seattle, Washington, USA

Chinmay Kulkarni
chinmayk@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Jofish Kaye
acm@jofish.com
Mozilla
Mountain View, California, USA

ABSTRACT

Voice assistants are fundamentally changing the way we access information. However, voice assistants still leverage little about the web beyond simple search results. We introduce Firefox Voice, a novel voice assistant built on the open web ecosystem with an aim to expand access to information available via voice. Firefox Voice is a browser extension that enables users to use their voice to perform actions such as setting timers, navigating the web, and reading a webpage's content aloud. Through an iterative development process and use by over 12,000 active users, we find that users see voice as a way to accomplish certain browsing tasks efficiently, but struggle with discovering functionality and frequently discontinue use. We conclude by describing how Firefox Voice enables the development of novel, open web-powered voice-driven experiences.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; User studies.

KEYWORDS

voice assistant; conversational user interface; CUI; browser extension; open source

ACM Reference Format:
Julia Cambre, Alex C. Williams, Afsaneh Razi, Ian Bicking, Abraham Wallin, Janice Tsai, Chinmay Kulkarni, and Jofish Kaye. 2021. Firefox Voice: An Open and Extensible Voice Assistant Built Upon the Web. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3411764.3445409

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445409

1 INTRODUCTION

Over the last decade, voice assistants like Amazon Alexa, Google Assistant, and Apple's Siri have gained widespread adoption, with over 50% of adults in the United States reporting that they use have used a voice assistant [52]. While the embodiments of voice assistants and the reasons for their adoption are wide-ranging, their usage shows a consistent pattern: across several studies, searching the web for content like music and recipes and performing simple informational queries consistently ranks among the most frequent use-cases for voice assistants [3, 64]. In particular, estimates suggest that as many as a third of all web searches are now invoked by a voice query rather than typed [26, 29–31]. Taken together, these studies suggest that while voice assistants are interfaces to computing in general, they functionally serve as an interface to the web.

Given the importance of voice as a modality for interacting with the web, many companies have created "voice apps" for commercially successful voice assistants: as of January 2019, Google Assistant and Alexa had 4,253 and 56,750 apps or skills respectively [60]. However, even these thousands of apps represent a tiny fraction of the more than 1.5 billion websites that are estimated to be online today [34]. More generally, even though commercial voice assistants access some information from the web, information is typically drawn from only a handful of knowledge-base websites such as Wikipedia. In other words, voice assistants draw on and provide access to only a minuscule fraction of content available across all websites on the internet.

The fundamental challenge motivating our work is that extending a voice assistant to leverage more of the Internet's content or creating new voice services currently requires developers to invest significant resources (and acquire specialized developer and user experience skills) [5, 46], and is effectively controlled by a small number of commercial entities. Each commercial assistant's developer platform controls what content may be presented, what voice may be used to present that content, and even the duration of a conversational turn, which can severely limit opportunities

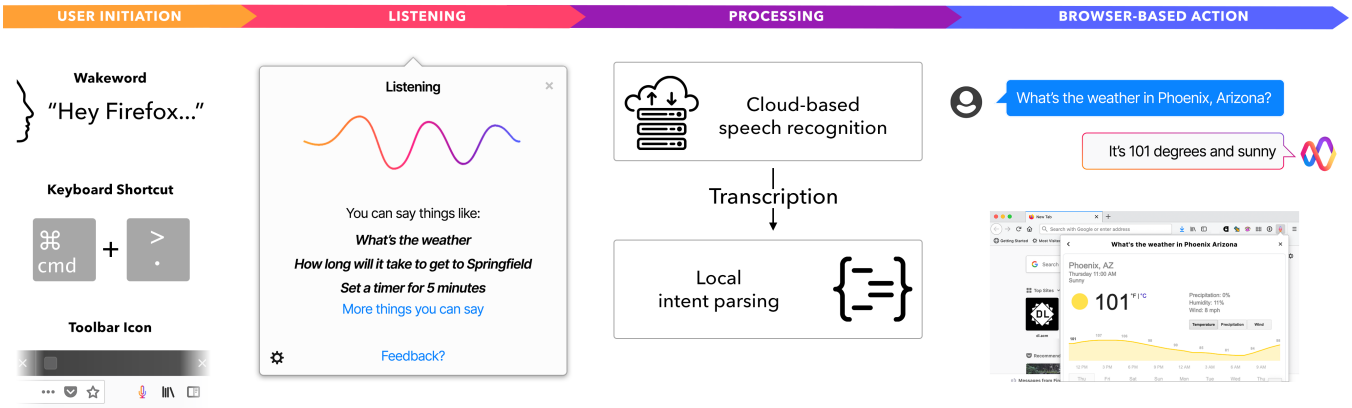

Figure 1: Overview of the Firefox Voice system. Users can invoke the voice assistant through a hands-free wakeword, by using a keyboard shortcut, or by clicking an icon in the browser toolbar. Once invoked, Firefox Voice displays a small tooltip UI that overlays their current browser tab and listens for the user's query. When the user stops speaking, the audio is sent to a cloud speech recognition engine to produce a transcription, which is then parsed into an intent locally within the extension. Firefox Voice then acts upon the relevant intent, and can optionally respond with speech where appropriate.

for innovation, experimentation, and exploration. In addition, each commercial platform has different specifications, making interoperability a significant challenge [10, 13]. Together, these challenges may ultimately hurt user adoption: according to one report [74], less than 3% of Alexa skills are used after two weeks.

This paper presents an alternative vision to these systems in the form of Firefox Voice, an open-source browser extension that enables users to navigate the web and control browser utilities through voice commands. It demonstrates that it is possible to build sophisticated voice services by leveraging what is already on the internet and to expand voice interactions to the billions of existing websites without requiring extra developer effort. To enable a fully hands-free experience akin to that of a smart speaker, Firefox Voice supports a locally-listening wakeword (e.g. "Hey Firefox"), and can optionally respond to many queries through spoken, text-to-speech output. Through Firefox Voice, users can achieve much of the same functionality as commercially-popular voice assistants, such as setting timers, asking for the weather, and rendering an informational card in response to certain questions. However, Firefox Voice goes beyond the web-based capabilities of existing voice assistants by allowing users to navigate directly to deep sub-pages of the web via voice command (e.g. "Go to the At Home section of the New York Times") and perform in-page browsing actions such as following links, taking screenshots, or controlling slideshows.

The implementation of Firefox Voice relies on a mix of existing APIs, regular expression-based parsing, and simple heuristics about the structure of websites. Except for our wakeword, it also requires no specialized machine learning techniques for natural language processing or information extraction. As we demonstrate in this paper, this simple approach yields surprisingly powerful assistant-like capabilities, but has the added benefit of easy extensibility, thus allowing for experimentation with voice interaction by developers and researchers. All code for Firefox Voice is available open source at https://github.com/mozilla-extensions/firefox-voice, and we eagerly welcome the community to fork, remix, and contribute. Thus far, this open-source codebase has received contributions from 50 external contributors who have extended Firefox Voice with more than 286 pull requests.

The design and development of Firefox Voice reflects a focus on powerful voice assistance through bricolage and extensibility. Starting in May 2019, we conducted a large-scale "Needfinding" survey (N=1,002) to understand users' needs for a voice assistant, and how users might perceive the unique value and use cases for a browser-based assistant relative to other voice assistants. Informed by these user needs, we conducted iterative, in-person "First impressions" user studies (N=21) using working prototypes of Firefox Voice. This yielded insights that we built into the design of Firefox Voice. We then ran a "First-use" survey (N=217) with participants of an external public beta test to validate and refine this interaction design.

Finally, we released Firefox Voice in a wide-scale public deployment, yielding a total of over 30,000 installs and more than 12,000 all-time active users. During this time, we also collected reasons why people stopped using Firefox Voice as a part of our "Uninstall" survey (N=698) to help us improve our system.

Our findings suggest that participants find value in a browser-based assistant, noting that voice often feels more efficient for searches or navigational tasks that would otherwise require typing a long phrase or clicking through several interstitial pages before arriving at the desired webpage. In addition, we find that our architecture allows for easy extensibility. However, participants nevertheless struggled with several of the challenges known to plague speech interfaces and voice assistants, such as discoverability [17, 49, 83], speech recognition failures [14, 68], and lapsed use [7, 18].

Based on our experience iteratively designing, developing, deploying, and evaluating Firefox Voice at large scale with real-world use, this paper makes three key contributions:

(1) An illustration of how a fully-featured voice assistant can be designed with open web technologies, and how such

an approach yields unique benefits over other closed voice ecosystems.
(2) A case study offering insight into how a novel consumer voice product was developed over several rounds of user research, and into the unique considerations of introducing voice assistance in a web context.
(3) The Firefox Voice system itself, which introduces a novel, open-sourced resource for researchers and the broader community to experiment with voice assistant technology.

2 RELATED WORK

Firefox Voice builds upon several areas of prior work on voice assistants, end-user conversational scripting in the browser, and existing systems that incorporate voice on the web.

2.1 Current Benefits and Challenges with Voice Assistants

As voice interactions have become more commonplace, a growing body of work [19] has helped to illustrate when, where, why, how, and for whom voice can be most useful, and the key challenges and limitations of voice assistants.

Several studies have considered how today's commercially popular assistants such as Alexa and the Google Assistant have become integrated into users' everyday lives, and how users' relationships with these assistants change over time [7, 43, 55]. In particular, voice interactions are especially useful in contexts in which a user's hands and eyes are busy [66], such as while cooking [76] or driving [65]. Through analyses of usage logs [3, 64] and self-report data [43, 73], common patterns have emerged around the most commonly used assistant features: users frequently ask their voice assistants to answer search or informational-type queries, control music, set timers, and manage paired smart home devices [3, 43, 64, 73]. At their best, these voice assistants can offer helpful and efficient hands-free convenience for multitasking [43, 65, 78], provide entertainment through deliberate jokes and games or amusing breakdowns [6], and empower users in many diverse populations including the elderly [57, 63] and individuals with disabilities [2, 58].

However, voice assistants also face several well-documented challenges that have complicated social consequences, and that hinder their usability. In line with the theory that Computers Are Social Actors (CASA) [50] which suggests that people transfer and attribute human social characteristics to computers, several studies point to the ways in which users anthropomorphize voice assistants, and how this affects their interaction. On the one hand, this tendency to anthropomorphize can have positive effects, such as by leading to a sense of attachment or companionship with the assistant [56, 59]. On the other hand, it can have problematic effects as well: prior research suggests that users routinely over-estimate the intelligence of voice assistants, in part because of the convincingly human-like voices, names, and personalities given to the assistants [4, 18, 25, 43, 47]. Recent work has also critiqued the default voices given to many of the commercially popular voice assistants, suggesting that they may reinforce harmful stereotypes [12, 23, 67, 69, 79].

At the same time, the design guidelines and best practices for designing for voice are still in their earliest stages [9, 48], and there are a number of open questions and technical concerns for voice assistants. For example, voice interfaces—particularly those without a screen—often suffer from a lack of discoverability: in contrast to graphical user interfaces, where the possibilities for interaction are mostly visible to the user, users often do not know what voice interfaces are capable of, or what they are able to say [20, 83]. In addition, voice assistants also frequently suffer from speech recognition errors resulting in inaccurate transcriptions of what the user said [49]. These errors can be especially common depending on the user's accent or use of specialized vocabulary and proper nouns [14, 68].

There are also a number of barriers to users' adoption of voice assistants in the first place, and challenges around their long-term use. Prior work has frequently documented privacy concerns towards voice assistants among both users and non-users who feel uncertain about how their data is being collected and used, and who express discomfort with the idea that the assistant is always listening [21, 39]. Among those who do choose to use an assistant, there is also growing evidence that their engagement tends to decline over time. In one study, Bentley et al. [7] conducted a longitudinal analysis of users' query logs with the Google Home, and found that users quickly settled into using the assistant for roughly three different domains of commands (e.g. information, music, home automation, etc.), and rarely explored new functionality after two weeks of use. Other studies have similarly found that overall use of voice assistants tends to decline after an initial period of exploration, at times to a point of abandonment, suggesting that voice assistants may struggle to retain users who were once actively engaged [18, 64]. These declines in use are largely driven by the reactive rather than proactive nature of how voice assistants interact with users, and by a lack of discoverability to engage and educate users about new features [18].

2.1.1 *Voice Within Desktop Applications.*

While smartphones and smart speakers are common form factors for voice interaction [28], desktop applications also create meaningful use cases for voice interaction. In particular, multimodal interfaces that combine both speech and graphical input and output, as Firefox Voice does, are useful for contexts such as hands-free video navigation [17] and for assistance with complex creative applications [27, 36, 38]. User studies with these voice-driven controls have found that they can help to preserve task focus by reducing the need to context switch [27, 54], and can be especially helpful in identifying less familiar elements of the interface [36]. Analogously, we anticipate that Firefox Voice could streamline the browsing experience by reducing the number of clicks and keystrokes necessary to reach an end goal, and by enabling users to more easily access browser features that are difficult to discover.

A few voice assistants exist within a desktop context: Apple introduced its Siri assistant with the release of its MacOS Sierra desktop operating system[1] in 2016, Microsoft introduced its Cortana assistant in 2015 with Windows 10[2], and the BBC created a custom voice assistant named Beeb, which was released in beta as a

[1] https://support.apple.com/en-us/HT206993
[2] https://blogs.windows.com/windowsexperience/2015/02/10/how-cortana-comes-to-life-in-windows-10/

Windows-based desktop application in June 2020[3]. However, to our knowledge, the desktop-based versions of these assistants have not been well-studied, and most recent research centers on smartphone or smart-speaker assistant experiences. One notable exception to this is a large-scale analysis of usage logs from the desktop-based Cortana [45]. In their work, Mehrotra et al. found the variety of requests to the desktop Cortana resembled the types of requests made to other assistants (e.g. controlling alarms, performing a general web search, and asking for weather information). However, the most common use category was for searching a user's local files, accounting for over 40% of searches [45]. This insight informed and shares parallels with usage patterns on Firefox Voice: like Cortana on desktop, Firefox Voice supports most of the standard baseline features that a user would expect from a voice assistant, but its deep integration with the platform (the operating system in the case of desktop Cortana, and the browser and web context in the case of Firefox Voice) creates opportunities for new kinds of interactions.

2.2 Existing Approaches to Voice on the Web

Web content is primarily built to be *seen*, and to be interacted with through a keyboard, mouse, and touch [8, 84], rather than conversed with. Prior work has addressed this challenge in several ways.

2.2.1 Accessibility Tools. Firefox Voice is not designed as a tool for accessibility, but it is nevertheless informed by prior work on screen reader and other voice browsing technology, and by how individuals with vision impairments interact with voice assistants while using the web.

In addition to the published studies below, we consulted extensively with an expert in screen readers and accessibility for blind people, including monthly check-ins to discuss feature development and debugging. While we deliberately departed from screen reader norms (like identifying links read out loud as "Link"), many of our features were developed to be aware of and compatible with browser use practices of blind and visually impaired users.

Commercial voice assistants are largely accessible "by accident" [58] because they were not primarily designed as assistive technology. Nonetheless, voice assistants are popular with blind and low vision individuals [58, 78]. In many cases, voice assistants are complementary to other tools such as screen readers in addressing blind users' needs, particularly in providing convenient access to the types of information retrieved through a web search, and through integrations with third-party applications and control for IoT devices [1, 58]. However, blind users have also noted that voice assistants can often fail to provide an appropriate amount of information in response to a query [1, 77]. Prior work also highlights the strong desire among blind users for more productivity-oriented features (e.g. for writing emails) through voice [2].

Motivated by the observation that blind users alternate between voice assistants and other tools such as screen readers to leverage the particular strengths of each, Vyturina et al. [78] created the VERSE system as a design probe that augments a voice assistant with additional information retrieval functionality informed by screen readers, such as reading a Wikipedia page aloud, listing alternative search results following a simple answer, or reading headings aloud and allowing users to skip between them. In a user study with VERSE, participants found speaking to often be faster than typing, provided they did not encounter speech recognition errors [78].

While Firefox Voice implements some of the same features as VERSE, Firefox Voice differs in its use of a multimodal (visual and audio) interface for output, and in its scope; whereas VERSE draws information from the Bing Search API and from Wikipedia, Firefox Voice allows navigation to the broader web, and to browser-based utilities.

2.2.2 Enabling Technology for Web-based Voice Interaction. Beyond the accessibility domain, there have also been a number of recent approaches to introduce voice interactivity to the web, or based upon web content. For example, Lee et al. [41] demonstrated the feasibility of running a lightweight keyword-spotting model implemented in Tensorflow.js within the browser, which is able to distinguish between a relatively small, fixed set of spoken keywords.

Others are beginning to introduce new ways for developers to author voice-based content and conversational interactions. As one example of this, Schema.org, an organization that defines standards for structured web data, has introduced a "speakable" tag[4] that will allow web developers to indicate which content on a page would be suitable for synthesized voice output (e.g. by a voice assistant[5]). The Geno system [61] scaffolds the process of adding voice input and output to websites by providing authoring tools and APIs to developers. To add voice functionality with Geno, users import their existing projects into Geno's custom IDE, and the system guides them through a workflow of declaring intents (e.g. specifying sample phrases and their parameters), and associating those intents with a target function, or through GUI-based demonstration. Mycroft [6] is one of the most feature complete, and has a marketplace of external skills, and runs either on custom smart speaker hardware, or on a Raspberry Pi, a commonly available development board. Stanford's Almond project [15] has similar aims to Firefox Voice, including leveraging the power of the open web, but their focus has been on using natural language processing to understand queries [16] and interfacing with IoT devices through their tool Thingpedia and a collaboration with the popular open-source Home Assistant project[7].

2.2.3 Understanding Voice Search. A growing body of work has investigated the space of voice-based search and its particular challenges. For example, Schalkwyk et al. [62] describe Google Search by Voice (an early version of Google's in-product voice search features around 2010), and detail the many complex technical challenges involved, from accurate speech recognition to multimodal interface design. In an analysis of voice search log data, they found that voice searches were more likely than desktop queries to relate to topics like food and drink, and were more likely to start with a "wh" or "how" question. Similarly, Guy [31] conducted an analysis of search logs on the Yahoo mobile search interface and compared queries issued by voice to those that were typed, and found that compared to text-based search, voice searches were substantially longer, more

[3] https://www.bbc.co.uk/blogs/aboutthebbc/entries/45d6b6c9-9f40-4f90-8616-37d5a294490f

[4] https://schema.org/speakable

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

[6] http://www.mycroft.ai

[7] https://home-assistant.io

likely to include natural language phrasings (e.g. "I'm looking for"), and more likely to trigger a card-based response, suggesting that they may have been relatively more direct informational queries, such as recipes or maps. Others have pointed to other unresolved challenges with voice search, such as a need to support complex, exploratory searches [44] and to provide appropriate responses based on the user's context at the time of issuing the query [75].

Together, this prior work informs the types of search tasks that might be common with Firefox Voice, and consequently, the functionality it must support.

2.3 Programmatic Control of the Browser

Firefox Voice can be seen as a web automation tool, operated through a speech interface: it translates a spoken command into a sequence of browser actions that are performed on the user's behalf. In this sense, Firefox Voice shares some similarities with a broader set of systems for conversational scripting and end-user programming within the browser. Examples of such systems include CoScripter [42] and CoCo [40], which allow users to demonstrate a "macro" or sequence of web-based actions, and later invoke that macro through written natural language commands.

Firefox Voice follows the legacy of these web macro tools that recognized the benefits of automating or building shortcuts for browsing tasks. At present, Firefox Voice supports a lightweight version of end-user conversational scripting similar to that of CoScripter and CoCo through a *routines* feature, which enables users to create an alias for a sequence of commands defined in natural language.

3 AN ITERATIVE, HUMAN-CENTERED DESIGN PROCESS

Building upon this related work and our own prior explorations in the voice products space [3, 11, 72, 80–82], we took an iterative approach to the development of Firefox Voice. In this paper, we present four key studies, which also map onto key moments in a user's potential lifecycle with Firefox Voice:

(1) "Needfinding" survey (N=1,002) with potential users to inform the space of browser-based voice assistants.
(2) "First impressions" user studies (N=21) interacting with working prototype versions of Firefox Voice for the first time.
(3) "First-use" survey (N=217) distributed to public beta testers at the end of their first week using Firefox Voice.
(4) "Uninstall" survey (N=698) presented to users in the public beta and public release immediately upon uninstalling the Firefox Voice extension.

Figure 2 provides a visual representation of the development, user studies, and major milestones in Firefox Voice's history that are covered in this paper, which represents a timeframe from May 2019 through August 2020. We discuss the two formative studies "Needfinding" survey and "First impressions" in sections 3.1 and 3.2 that follow, and discuss the "First-use" survey (section 6.1) and "Uninstall" survey (section 6.4) after introducing the Firefox Voice system and its implementation.

In all of our user research and data collection, we obtained approval via our organization's user research review process, and we were as conservative as possible in the amount of sensitive and potentially identifying information that we collected from users and study participants. This is both based on principle to respect each individual's privacy, and is a key practice and policy within our organization. As such, we collected and report demographic data such as gender only where relevant in the user studies presented throughout the paper.

3.1 Formative Studies: Large-Scale Needfinding Survey on Voice Assistants

To validate an interest in using a voice assistant built into the browser, and to gather feedback on the use cases that are most important to people, we conducted a "Needfinding" survey in May 2019. We recruited English-speaking participants (N=1,002) in the United States, Canada, and the United Kingdom by showing a banner with the headline "Voice control the Internet" on the new tab page of the Firefox web browser.

We asked "What features would be most important to you when using voice controls in your browser?" where respondents could select all the features that interested them, or none of the above. The results are depicted in Figure 3.

We found that the most-requested feature was "Search the internet" (65%), which could arguably be better handled in a browser-based assistant, as compared to voice assistants in a smart speaker form factor. Other popular features were based on functionality specific to the browser such as "Control audio playback in the browser" (59%) and "Open a webpage" (57%). Still others were features expected of voice assistants in general, which could also be implemented in the browser, such as "Play music" (56%), "What's the weather?" (52%), and "Set a timer" (46%).

The results of the "Needfinding" survey suggested that there was potential value and interest in controlling the internet via voice, and in building an assistant with some browser-exclusive capabilities, and we proceeded to build a working prototype of Firefox Voice.

3.2 Formative Studies: First Impressions User Studies with Early Working Prototypes

We began developing Firefox Voice in June 2019, and improved the system over the course of the next year. The earliest versions of the system supported a relatively small set of key features (e.g. navigating directly to a page, playing YouTube videos, finding tabs) and were not robust to a variety of phrasings for a given intent. Similar to previous work on voice assistance, these prototype versions of the system had a "roughness" of functionality which afforded users more license to critique and propose features than they might have experienced with a more polished version of the system [14].

To this end, we conducted in-person user studies (N=21) with early versions of Firefox Voice during this design and iteration period from June 2019 through October 2019, split across four sessions. Figure 2 includes additional details on the dates and locations of the user studies. We recruited five to six participants per user study session via snowball sampling, resulting in a total of 21 participants. Participants were selected to have a rough balance of voice interface users (e.g. people who reported owning a smart speaker and using voice assistants regularly) and infrequent or non-users of voice, and to be well-balanced in representation of different genders,

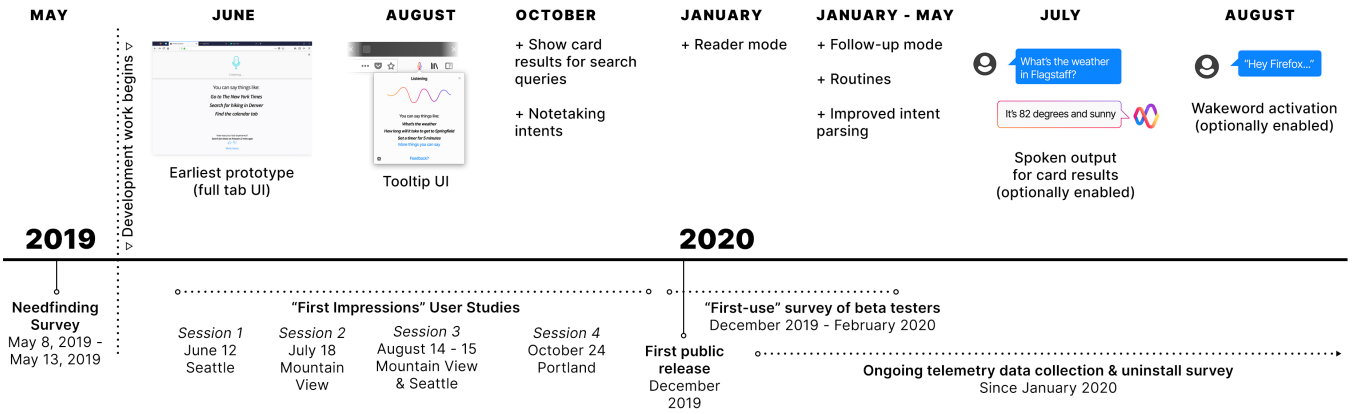

Figure 2: A timeline representing Firefox Voice's iterative development and user research process. Major feature releases are presented above the line, with important milestones in user research and deployments below the line. Throughout this entire process, there were also countless improvements to the underlying architecture of Firefox Voice, and to the range of intents it supports.

ages, and job roles and fields (e.g. tech, construction, therapy, retail). Across the four sets of user studies, participants ranged in age from 21 to 66 (Median=35; SD=11.40), with two ages unreported. Nine participants identified as female and 12 participants identified as male, with no other gender identities reported.

Each study shared the same general procedure and semi-structured interview script. After receiving participants' informed consent, we began with a series of introductory questions about the participant and their familiarity and experience with voice interfaces (both voice assistants, and other forms of voice input such as voice-controlled television remotes). We then presented participants with the Firefox Voice prototype, and (to reduce potential response bias due to experimenter demand characteristics [24]), told them that we were evaluating it on behalf of a team that had built it within our organization. Participants were asked to explore Firefox Voice's functionality for a few minutes, and were given a short list of sample commands that they could try. We concluded the study by asking participants about their impressions of Firefox Voice, any usability challenges they encountered, and about their interest in using such a system in the future. Each study lasted 30 minutes, and participants were compensated with a $35 Amazon gift card for their time. All sessions were recorded and transcribed for later analysis.

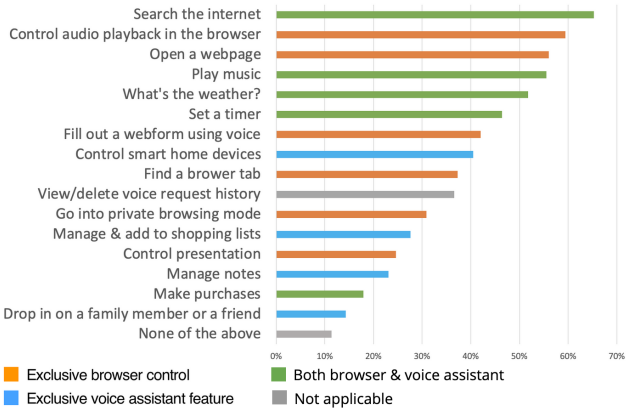

Figure 3: A chart summarizing the top features that participants in a needfinding study (N=1,002) indicated were most important to them for voice in the browser, categorized by whether the feature is exclusive to the browser, or to a voice assistant, or possible on both the browser and voice assistant. Participants could select multiple responses, and the x-axis represents the percentage of total responses that selected each feature.

3.3 Design Considerations

Based on the "Needfinding" survey and "First impressions" user studies, we derived the following set of insights that guided our further development on Firefox Voice.

3.3.1 Privacy. A theme apparent in our research results was concern about privacy. Respondents expressed particular concerns about existing smart speaker devices. In particular, participants routinely mentioned feeling hesitant to use Alexa, Siri, and the Google Assistant because of privacy and trust issues, such as a sense of discomfort in knowing that the assistant was "always listening" (e.g. P3) or concerns about the companies' data use (e.g. P10). This concern for privacy becomes particularly apparent in the web browser context: while smart speaker usage is primarily focused around playing music, one-off searches and IoT control [3], web browsers have a far wider scope of use:

> *"Big companies (Facebook, Google, Amazon) have gotten some heat for privacy concerns. If I'm searching for more personal stuff, information being drawn up from these queries, I'd want to make sure the voice assistant is keeping my data private and secure.* (P7)

This strongly suggested to us that privacy would be a key aspect of our system, and we made sure that choices regarding data collection were clearly presented during installation.

3.3.2 Supporting Hands-Free Interactions. Voice assistants navigate a difficult trade-off between convenience and potential privacy concerns. Because privacy was a central concern, we initially designed Firefox Voice without a wakeword: users had to either click the toolbar icon or use a keyboard shortcut to activate the assistant. However, many participants in these formative user studies expressed a strong desire to activate Firefox Voice by voice so the experience could be hands-free. Some participants described existing pain-points in trying to use their browser during tasks in which their hands are dirty, such as cooking or painting, and suggested that Firefox Voice would be helpful in effectively turning their laptop into a smart speaker in those contexts:

> *"If I'm in the studio, I have paint all over. It'd be nice to say 'What is the Myth of Judith?' "* (P2)

This and similar feedback informed our decision to introduce a "Hey Firefox" wakeword in later versions of the system. To accommodate those who specifically do not want an "always listening" assistant, the wakeword is optional, and disabled by default. The development of our wakeword system is discussed at length in [70].

3.3.3 Multimodal UI should be unobtrusive. Early prototype versions of the Firefox Voice system that we user tested opened and focused a new tab as the "listening" interface on each request (see Figure 2, leftmost image). However, participants found this to be distracting because it broke their task focus on their current tab, and contributed to a sense of latency:

> *"I don't know if it'd be any faster, but there's something about having it open a new tab that makes it feel slower."* (P6)

This suggests that a browser-based assistant should occupy as little screen "real estate" as possible to avoid interfering with the user's current context.

3.3.4 Guide users on what the system can do, and how it differs from other voice assistants. As users interacted with the Firefox Voice prototype for the first time, we found that their first impressions were often strongly influenced by—and evaluated against—their prior experiences with other voice assistants. For example, P18 shared:

> *"First thought is, is it more of a Cortana that lets me search?"* (P18)

In many cases, participants tried issuing requests to Firefox Voice that are typical of other assistants, such as saying "Good morning" (P15), which will trigger a configured morning routine on other assistants such as Alexa and the Google Assistant [3]. At the same time, we rarely saw participants exploring the more unique, browser-specific features of Firefox Voice unless we gave them concrete examples or nudged them to do so.

While some of these comparisons centered on specific functionality, others reflected on how their impression of other commercial voice assistants would more generally shape their expectations for Firefox Voice:

> *"I'm more like a slow adopter of something like this given previous experiences with Alexa and Siri. Turn it on for a little bit, turn it off for a little bit. See how organically it becomes part of my daily usage."* (P13)

Other participants were surprised by their experience of interacting with earlier versions of Firefox Voice, which did not support speech-based output and therefore did not align with their mental model of voice assistants:

> *"If it's going to be a voice interaction, it would talk to me."* (P11)

Taken together, these findings suggest that users will inevitably transfer their expectations from their experience with other voice assistants (both positive and negative), consistent with prior work on other novel voice tools [36]. With this in mind, we believe it's necessary to educate users on what Firefox Voice is capable of, and how it differs from other assistants. As such, we redesigned our interface to display examples of supported utterances within the onboarding webpages a user sees upon installing the app, as well as directly within the pop-up user interface the first several times a user interacts with Firefox Voice.

3.3.5 Focus on functionality and contexts in which using voice on a desktop browser is most appropriate. When asked under what circumstances they might imagine themselves using a system like Firefox Voice, several participants mentioned that they would be unwilling to interact with a browser-based voice assistant in a shared public setting like an office or coffee shop because they might disrupt others:

> *"If I was in the office, people would probably stare if I'm talking to my computer"* (P1)

This finding aligns with prior work, which has also found that users are reluctant to interact with voice assistants in public or shared spaces because of social norms [2, 21, 36]. While there may be a longer-term opportunity to mitigate these concerns through whispering [53] or silent speech interfaces [35], this suggests that a desktop-based assistant like Firefox Voice will primarily be used in private settings such as the home or individual office spaces.

Instead, voice can provide some enhanced functionality, especially around navigation in the browser, and in streamlining browser-focused multi-tasking. In particular, participants repeatedly emphasized that whatever functionality Firefox Voice provided would have to be easier or more efficient than accomplishing their goals by using a keyboard and mouse.

For example, one participant (P6) directly suggested that Firefox Voice could help in supporting "browser-specific things," noting that the assistant could help in finding features (e.g. clearing their cache) that they perform relatively infrequently. Similarly, some participants drew an analogy to keyboard shortcuts, echoing a concept explored in prior work on photo editing software [36]. For instance, one participant who noted that they frequently made use of keyboard shortcuts mentioned they would want to use Firefox Voice for *"anything that I'd have to take my hands off the keyboard to the mouse to click... [I'd want to use] voice to eliminate that step"* (P4).

To support these shortcut-like experiences through voice commands, Firefox Voice implements a wide range of intents for browser features, which we enumerate in Appendix A.

Participants in the study also repeatedly mentioned that Firefox Voice should make their browsing more efficient, and noted that they particularly appreciated being able to navigate directly to

a page of interest with a voice command. Offering customizable defaults and integrating with existing web service was also seen as important:

> *"If I could set defaults ahead of time so I could say "Play Green Day" and it'd go straight to Spotify, that'd be a home run. If I can connect it to what I'm already using, that'd be a huge value"* (P16)

Because Firefox Voice is situated within the browser, we realized it had the unique ability to both leverage people's mental models of existing voice assistants, and accomplish more with a multimodal (including visual) interface and browser infrastructure, such as cookies providing stateful interactions to the web or saved passwords facilitating low- or no-overhead logins.

4 FIREFOX VOICE: A BROWSER-BASED VOICE ASSISTANT

Firefox Voice is an open-source voice assistant designed to leverage the infrastructure of the browser and the larger Internet. Users interact with Firefox Voice through spoken requests such as "Find me a recipe for chocolate chip cookies from Smitten Kitchen." Firefox Voice responds through a multimodal experience that renders the corresponding webpage or glanceable card of information, and synthesized speech when appropriate. Users can trigger Firefox Voice through a locally listening "Hey Firefox" wakeword, so their experience can be fully hands-free.

4.1 Turning Websites into Voice Services

Because Firefox Voice is embedded directly into the browser, browser features such as session cookies, cached responses, and bookmarks are available to it automatically. This allows Firefox Voice to support a range of web services by voice with no additional configuration.

This stands in contrast to other voice assistants, which rely upon third-party applications for integration. For example, for an Alexa user to view their order history at a retailer such as Starbucks, they must first enable the retailer's skill (presuming they offer one, which may be unlikely), link their Starbucks account to their Alexa one – a process that requires the user to log in through a smartphone, and restart the skill. This process is cumbersome to set up, error-prone, and problematic from an accessibility perspective [58].

By leveraging existing browser login and session mechanisms, Firefox Voice removes the need for this configuration. For users logged into Starbucks' website, a query such as "Hey Firefox, show me my Starbucks order history," displays their order history directly.

Firefox Voice similarly detects relevant elements of a website's DOM through query selectors, and issues JavaScript events that simulate the user taking the intended action, such as clicking on a button. This allows Firefox Voice to support a number of popular features such as controlling music and videos, on both ad-supported platforms such as YouTube, and on subscription platforms like Spotify. For these services, Firefox Voice implements a number of intents to provide richer access for playback control such as pausing, resuming, muting, unmuting, skipping songs, and so on.

4.2 A Voice Interface to Browser Interactions

As web-based applications become more popular, they are supplanting many applications that were previously run on users' computers as separate programs (e.g. word processing, spreadsheets, calendars, email, and more). Consequently, user interactions with browsers are increasingly similar to their interactions with operating-systems (e.g. "open Gmail", or "Play the next song on Spotify".)

Firefox Voice supports OS-like commands such as "Find the time entry tab." with data from the content of the window. Firefox Voice collects the title, URL, and text content of each open tab, and performs a fuzzy search to find the best matching tab (if any), allowing more conversational interactions with ambiguous inputs. Similarly, users can follow links by describing their text content (e.g. "Click the show more button"). Other Firefox Voice commands emulate keyboard and mouse-based interactions within the context of a website. For example, users can ask Firefox Voice to "scroll down" (resulting in a page scroll event).

Firefox Voice also allows faster interaction with the browser itself. For instance, Firefox Voice also offers a shortcut into a lesser-known Firefox feature which reads webpages using a text-to-speech voice (with "read this page"). Similarly, Firefox Voice supports other browsing tasks such as finding and switching to a tab, bookmarking a page, taking a full-page screenshot, or clearing browsing history. Firefox Voice also makes it possible for users to define voice *routines* (similar to OS-level scripting), which are named shortcuts for a sequence of actions. For example, as users start their day and open the key websites they need ("Open Gmail," "Open GitHub," and "Open my calendar"), and then asking Firefox Voice to group and name those commands as a routine (e.g. "Name the last three commands *let's get to work*"). When the user says in the future, "Hey Firefox, let's get to work," Firefox Voice will perform the routine, opening the three websites in new tabs. Routines are also editable through a graphical authoring interface.

4.3 A Platform for Voice Interface Experimentation

Firefox Voice is open-source, and built with common browser-based technology, allowing experimentation that is not possible with other voice assistants and voice prototyping tools. Existing voice assistant ecosystems like Alexa and the Google Assistant, for example, hard-code key aspects of the interaction, such as how many seconds the microphone will remain open while expecting a response from the user [13]. In contrast, Firefox Voice is modular and open-source, providing developers and researchers with a platform to experiment with all aspects of voice interaction, from the timeout duration before closing the microphone, to which speech recognition engine is used.

As one example of these extension capabilities, the Voice team experimented with *follow-ups*, such that Firefox Voice will listen for an additional command upon finishing a previous request. This functionality changes the default behavior of closing the microphone and the popup interface after each command to instead keep the microphone and popup open for eight seconds so the user can continue interacting with it without the need to re-invoke the extension.

In sum, Firefox Voice leverages the web to makes a range of voice-driven experiences available to the user with a smaller implementation effort. It replicates the most popular functionality of smart speakers or smartphone-based voice assistants, and also introduces new forms of voice-driven, multimodal interactions that are uniquely possible within the context of a web browser.

5 IMPLEMENTATION

We implemented Firefox Voice as a browser extension that runs within the Firefox desktop web browser. The system is built entirely with web-based technologies such as JavaScript, with user interfaces built in React.js, and is available as an open source project on GitHub at https://github.com/mozilla-extensions/firefox-voice.

5.1 Invoking Firefox Voice

There are three ways that a user can prompt the Firefox Voice system to begin listening for a command. One option is to click on the extension's icon—a stylized microphone—from within the browser toolbar. Alternatively, the user can issue a keyboard shortcut. For a hands-free experience, users may optionally enable a "Hey Firefox" wakeword. The wakeword functions by keeping the microphone open through a tab, which listens locally for "Hey Firefox" through streaming recognition in TensorFlow.js. To ensure that the model is lightweight and able to run inference in the background without a significant impact to the browser's performance, the wakeword is trained as a small residual network (ResNet) on approximately 5.4k audio clips containing some or all of our vocabulary for the wakeword. All aspects of the wakeword are available open source, including the training code, the runtime inference code, and the data used to train the model. Additional details regarding the technical implementation of the wakeword are described in [70].

5.2 Recognizing and Parsing Users' Natural Language Commands

Once triggered, Firefox Voice must listen and transcribe the user's speech, and map the transcription onto one of the system's supported intents or to a fallback. To accomplish this, Firefox Voice uses a cloud-based automatic speech recognition service, followed by local processing to match the transcription to an intent.

5.2.1 Listening and Transcribing User Speech. While listening, Firefox Voice's primary user interface—a small popover window that descends from the toolbar icon—becomes visible, and a brief audio chime sounds to indicate that the microphone is active and Firefox Voice is awaiting the user's utterance. While the user is speaking, we display an animated line that oscillates in a wave-like pattern, where the amplitude of the wave increases and decreases to reflect the audio input volume. We use local voice activity detection (VAD) to differentiate between background noise and speech, and the microphone remains open until the system detects that the user has stopped speaking for more than one and a half seconds (1500 milliseconds). Firefox Voice does not currently support long-form dictation, and times out after 15 seconds of speech.

For speech recognition, we leverage the Google Cloud Speech-to-Text[8] engine, proxied through a server operated by our organization to better protect users' privacy. This server returns a transcription of the user's speech, which we briefly display within Firefox Voice's popup interface before rendering the resulting information or action.

Firefox Voice also allows users to type their commands rather than speaking them. If the user begins typing after Firefox Voice has been invoked, the microphone will close, and the popup window instead displays a text box with the user's typed input. For typed queries, no audio content is sent to remote servers for processing.

5.2.2 Intent Parsing with Simple Slot-Filling Heuristics. To map a user's command to a corresponding, supported intent within the system, we implement a simple parser that uses a slot-filling approach. Because Firefox Voice is a general-purpose voice assistant, the nature of requests it must support have an open vocabulary (for instance, a user may ask about arbitrary people or place names, or refer to newly emergent artist and song titles [68]). In our early explorations of intent parsing with Firefox Voice, we found that a relatively simplistic pattern-matching approach performed better for our purposes over more sophisticated machine learning models, or existing open source systems like Rasa[9].

Firefox Voice's intent parsing works as follows: intents are declared through one or more text-based patterns, which specify variations in how a user might phrase a particular command. These phrases can contain alternative or optional words as well as slots, which function as a wildcard that captures one or more words from the user's utterance. There is also the option to include a *typed* slot within an utterance, which will expand the phrase list to match against a list of pre-defined services or sets of words (e.g. the music services that Firefox Voice supports, or numbers).

At runtime, the intent parser compiles all possible phrases, and compares the user's utterance against each phrase to find a match. To be more robust to speech transcription errors and slight variations in how a user might phrase a given intent, the parser also accounts for commonly mis-transcribed words through a list of aliases (e.g. "coffee" instead of "copy"), for repeated words (e.g. "next next"), and for adding flexibility around stopwords, using a list adapted from the SpaCy natural language processing library[10]. In cases where there are multiple matching intents, Firefox Voice will prefer the intent that is more precise (e.g. if both "play [query] on Spotify" and "play [query]" match, then the former will be preferred). If there is no matching intent for the user's utterance, Firefox Voice defaults to performing a search with the user's default search engine.

5.3 Resolving Intents into Actions

Firefox Voice resolves intents into primarily three kinds of actions: 1) navigating to a webpage; 2) invoking a browser action or script on the page, and 3) extracting information from the page and optionally speaking it aloud.

To do so, each intent in Firefox Voice registers its own handler code that runs in the background of the browser. These handlers have access to certain browser-level functionality (e.g. creating tabs, navigating to a particular URL, switching the tab in focus).

[8] https://cloud.google.com/speech-to-text

[9] http://www.rasa.com

[10] https://github.com/explosion/spaCy/blob/e0cf4796a505bd26e8fcb95d23b89fd7eca3be0a/spacy/lang/en/stop_words.py

Handlers can also optionally inject content scripts into a tab to access or manipulate data on a webpage. When a given intent is matched by the intent parser, its corresponding handler function is run with a `context` variable that contains the parsed intent's slots.

In some cases, handling an intent is as simple as calling a method implemented by the standard WebExtension APIs[11]. To handle more complex intents, many of the intent handlers in Firefox Voice share a common pattern to shortcut directly to highly specific content that users are interested in on the web. Firefox Voice accomplishes this by making use of Google search's "I'm Feeling Lucky" functionality, which automatically redirects a user to the top matching result for a query, rather than displaying a list of search results. Google currently makes it possible to set a flag to enable "I'm Feeling Lucky" as a URL query parameter, thus making it a straightforward task to reformulate the user's request to Firefox Voice into a Google search URL that, when opened in a tab, will take the user directly to the top resulting page. On its own, this approach makes it possible for Firefox Voice to support a diverse and open-ended set of web navigation requests, such as "Show me the Twitter feed for SIGCHI," which will bring the user directly to https://twitter.com/sig_chi in a new tab.

Other intents build upon the affordances of the Google "I'm Feeling Lucky" redirect by using the search engine to directly navigate to a piece of content of interest, waiting for the page to load, and performing further actions in-page to accomplish a particular task (e.g. clicking a button to trigger music playback).

As a fallback when no intents match, Firefox Voice defaults to performing a search. In many cases, these are informational queries for which Firefox Voice does not have a built-in intent (e.g. "What time is it in Madrid?"). To address these requests, Firefox Voice will perform a Google search in a hidden tab. If the search result contains a card (such as a Wikipedia snippet, calculator result, or weather card), Firefox Voice extracts a screenshot of that card, and renders it within the popover user interface.

5.3.1 Generating Speech Output. Firefox Voice can respond to a subset of commands through synthesized speech output. Two types of commands yield speech output. When a user asks that Firefox Voice reads a page aloud, Firefox Voice changes the page to the built-in Firefox reader mode, and triggers the audio playback feature within the page.

Firefox Voice is also capable of responding with voice output for many informational queries. For example, if a user says "What's the weather?" Firefox Voice responds by stating the temperature and conditions (e.g. "It's 80 degrees and sunny").

To accomplish this, we leverage the information embedded within the Google search card result that is returned for a particular query. If one of the known card types (e.g. sports scores, translation, generic fact) is found, Firefox Voice will then extract relevant entities from the card using query selectors and incorporate them into template phrases (e.g. "It's `{temperature}` degrees and `{conditions}`"). The compiled phrase is then spoken aloud using the `SpeechSynthesis`[12] interface of the Web Speech API browser standard. By default, Firefox Voice uses the default system voice from the user's operating system, though the user can select their preferred voice from any available system voices in an options panel.

While powerful as a means of bootstrapping spoken output, we acknowledge that such an approach is limited: it is brittle to any potential changes to the structure or identifiers used within the card output, and also currently lacks the nuance to differentiate and provide appropriately detailed responses to queries that might yield the same card. Future work could explore alternative approaches to question answering that are more robust to these cases.

6 REAL-WORLD DEPLOYMENT

Following an iterative product roll-out process, Firefox Voice was first made available to a small number of internal beta testers beginning in September 2019, and an initial public beta version of the extension was made available starting in December 2019. The final version of the system described in this paper—including the key features of a wakeword and text-to-speech output—was released in August 2020 (see Figure 2).

Due to the COVID-19 pandemic and the logistical challenges it entailed, we were forced to forgo an in-person user study with the final version of the system. Instead, we describe data and insights from both the beta release period and the system's public deployment. Qualitative insights and quotes reported here are drawn from the "First-use" survey (N=217) sent to public beta study participants to gather their initial impressions of Firefox Voice after their first week of using the assistant, and the "Uninstall" survey (N=698) that individuals are invited to complete immediately upon removing the Firefox Voice extension from Firefox. Open-ended survey responses were open coded by a member of the research team.

For the quantitative findings describing the use of the system, we focused on analyzing the telemetry data from a two-month window, July 1, 2020 until August 31, 2020. During this time, the extension received some publicity, resulting in unique usage patterns that also shed light on how common challenges of voice interfaces (such as discoverability and failures of speech recognition) manifest in a browser-based context.

6.1 First Use: Voice-Enabled Efficiency and Continued Use

Results from the "First-use" survey (N=217) distributed to public beta test participants suggested that the majority of participants (70%) had a positive first impression of Firefox Voice, based an analysis of the coded responses to the open-ended question "What are your first impressions of Firefox Voice?" In general, participants found Firefox Voice to be empowering, that it felt faster and easy to use, and allowed them to perform tasks beyond search. When asked to rate how likely they were to continue using Firefox Voice on a 5-point Likert scale, 54% indicated that they were likely or very likely to continue using Firefox Voice.

As part of this survey, participants were invited to describe what they enjoyed the most about using Firefox Voice. Of the 185 participants who responded to this question, 28% (N=52) cited functionality for controlling their browser. For example, participants appreciated tab management:

[11] https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions
[12] https://developer.mozilla.org/en-US/docs/Web/API/SpeechSynthesis

> *"Changing tabs - I find it difficult to find the tab I want, so doing it by voice command was very quick and efficient"*

> *"It was a good surprise to see actual functionality tied behind Firefox Voice, such as closing tabs. This would be significantly less useful for me if it were simply a 'Siri'-like voice search service."*

Others cited Firefox Voice's ability to make certain browsing tasks more efficient as its key strength. Many found it more convenient (26%, N=48) and faster (15%, N=27) to use Firefox Voice than completing the same task by the hand with other methods:

> *"Not having to type my requests. It is faster. For example, searching Amazon is nice and fast."*

While we had some negative feedback around voice recognition issues and technical problems, the positive first impressions and request for features ("What do you want to be able to do with Firefox Voice that you can't yet do?") in the "First-use" survey helped us to prioritize continued feature development (e.g. a wakeword and speech output).

6.2 Overall Usage Patterns

As of August 31, 2020, a total of 14,064 users had Firefox Voice installed as an extension in their Firefox web browser. To understand how people are actually using Firefox Voice, we focus our analysis in this section specifically on a two-month window (July 1 - August 31, 2020) during which we implemented and released the final version of Firefox Voice. During this two-month window, 8,637 unique users (representing 61.4% of the install base) interacted with Firefox Voice, issuing a total of 57,424 requests to the assistant. Queries issued to Firefox Voice were on average 3.47 words long (median = 3, SD = 2.41), and comprised a vocabulary of 16,495 unique terms.

Figure 4 plots the number of daily active users (DAU) for each day of this two-month window. We consider a user "active" on a day if they issue at least one request to Firefox Voice within a 24-hour period from 00:00 UTC to 23:59 UTC. Daily usage patterns during this period fell into three distinct windows of interest: prior to August 3, the number of DAU was relatively low, fluctuating between 16 and 113 users per day. From Monday, August 3 to Sunday, August 9, usage increased dramatically, with DAU numbers between 862 and 1262 users each day. On August 10, daily usage dropped considerably, and fell for the next several days, before leveling off again around 200 DAU. The seven day period from August 3 - 9 corresponded to a week in which Firefox Voice was promoted on the new tab page of the Firefox browser, and the extension therefore saw a large number of new users each day.

These usage patterns suggest that, while initial curiosity about Firefox Voice is high, a significantly smaller percentage of those who install it actually use it. Within this two-month July and August window, we found that 1,774 users used Firefox Voice exactly one time, and 7,533 individuals (53.7% of the user base) had the extension installed, but had not used it within that timeframe.

6.3 Search-Dominant User Experiences

Of the features that Firefox Voice makes available to users, those that are uniquely well-suited to the web and browser were, by far, the most commonly used. Table 1 presents the top 10 most frequently invoked intents, along with the number of times it was matched and the percentage of overall use of Firefox Voice that the intent represents. Search was the most frequently resolved intent (perhaps unsurprisingly, because it was also the fallback when no other intent matched), comprising 53% of all user interactions with Firefox Voice. Users also used Firefox Voice to navigate directly to a webpage on 16% of all queries. Other requests to Firefox Voice followed a long-tail distribution, with relatively fewer requests to intents involving music, tab control, and more.

Num. requests	Percent	Intent
30290	52.75	Search (card result, if applicable or search page)
9457	16.47	Navigate directly to a page
3031	5.28	Play music
1890	3.29	Search within a page for a query
1459	2.54	Read an article aloud
1176	2.05	Find and focus a tab
1094	1.91	Focus on a music player page
888	1.55	Close current tab
794	1.38	Perform a Google search
766	1.33	Go to the next search result
57424	100.00	Total

Table 1: Top ten most commonly-invoked intents on Firefox Voice between July 1 and August 31, 2020.

Because these intent classifications were made programmatically by our intent parser, we note that the intent label ascribed to a particular utterance may not correspond to the user's actual intent. In particular, the "search" intent actually encompasses a much broader set of functionality through the card-based results that are returned for certain queries. For example, if the user asks "How many tablespoons in a cup?" Firefox Voice will surface the appropriate answer through a card ("16 tablespoons"), but the utterance will nevertheless be labeled as *search*, rather than a more accurate and more nuanced label such as "unit conversion."

Despite this coarse labeling and heavy reliance on search as the fallback behavior, the majority of user feedback was positive in response to Firefox Voice's response on utterances resolved as search. Within the popup interface, Firefox Voice displays a small banner at the bottom that asks "Did we get this right?" with an option for the user to respond with a smiling or frowning face, and optionally provide written feedback describing their choice. Of those responses, 67.8% (N=6,720) were positive (32.2% (N=3,198) negative), and many of the reasons for providing negative feedback attribute them to errors with speech recognition.

However, this feedback prompt initially appears at the end of the "processing" phase and the prompt phrasing is ambiguous, which may have led some users to believe we were asking about the speech recognition's accuracy, rather than that of the complete interaction and resolved search-based result.

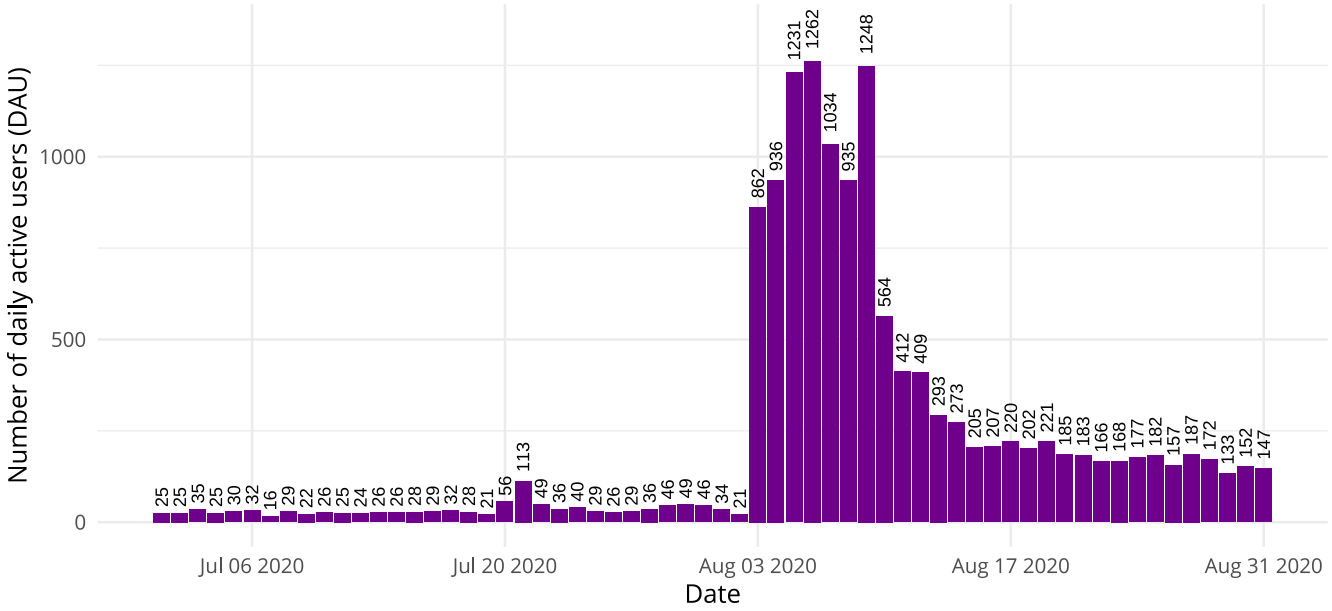

Figure 4: Number of unique daily active users (DAU) of Firefox Voice per day from July 1, 2020 to August 31, 2020. A user is considered "active" if they invoke the Firefox Voice extension at least once in a 24 hour period from 00:00 UTC to 23:59 UTC. The period from August 3 to August 9 represents a week in which Firefox Voice was featured in the new tab page of the Firefox browser, and thus saw an influx of new users.

6.4 Uninstall Survey

To better understand the challenges of maintaining a user base for Firefox Voice, we deployed an "Uninstall" survey (N=698). When people removed or deleted Firefox Voice as an add-on from Firefox, a new tab would open with the "Uninstall" survey. This is a relatively unusual practice, but we felt it was valuable to get specific and explicit feedback from people who had decided Firefox Voice was not useful to them. This survey presented a list of potential reasons for uninstalling (as 5-point Strongly Disagree - Strongly Agree Likert scale items), as well as an open response field for further comments. We present analysis of this survey from the beginning of the public beta deployment (December 2019) through August 31, 2020.

Participants were most likely to agree (Agree and Strongly Agree) with the statements: "Voice interaction could have been faster" (49.6%) and "I'm able to do what I want to do better in a different way" (41.7%). In contrast, they agreed least with "I don't typically use my laptop in an environment I'm comfortable talking out loud." (31.0%) and "I didn't know what to say" (26.5%).

We also open-coded the textual responses. The most common reason (34%, N=246) for uninstalling Firefox Voice dealt with technical difficulties. This includes various reasons such as microphone and permission issues, conflicts with other extensions, and a perception that the extension was slowing down the browser or causing it to crash:

> *"It didn't work. There was a message 'error getting stream.' It could not access my microphone even after I set it to allow in the privacy settings."*

Next, 15% (N=114) reported that they did not use Firefox Voice or had no use for it. Approximately 9% (N=67) of users mentioned issues with speech recognition failures, commonly due to Firefox Voice's inability to parse accents. Other issues included lack of multilingual support (4%, N=27), the lack of a hands-free experience (3%, N=22) (the wakeword was not implemented until the latest version of Firefox Voice), and vehement objections to the reliance on Google's services, both for search and for speech recognition (4%, N=31):

> *"I was hoping it would use [an open source speech recognition system]. I was hoping to use my default search engine (DDG [DuckDuckGo]). As it is, everything was from Google with no hope to change it. If I wanted a browser fully integrated with proprietary software pushing Google services on me, I would install Google Chrome."*

Finally, there was a "long tail" of other reasons for uninstalling Firefox Voice, including security fears, and that it simply "caused problems". Overall, the "Uninstall" survey played an important role in the development of Firefox Voice. It helped us to prioritize bug fixes to improve general functionality, helped to highlight the need to develop onboarding resources to explain and highlight relevant user scenarios (e.g. where voice interactions would be faster or

more efficient), and alerted us to necessity of implementing feature discovery and system reminder interfaces.

7 EXTENSIBILITY AND OPEN-SOURCE COMMUNITY ENGAGEMENT

As an open source project available through GitHub, Firefox Voice received a considerable amount of engagement from contributors external to our team[13]. Since the project was first posted to GitHub in August 2019, a total of 50 external contributors contributed code that was merged into the master branch, in a total of 286 pull requests.

Open source experience and contribution was a major factor in most contributors' involvement. A combination of familiar but modern technologies (such as React.js) made the project technically attractive. Contributions involved many small changes—documentation improvements, small CSS fixes, specific phrase matching improvements—to the development of many new intent handlers. In a very small number of cases contributors made large contributions, such as rewriting a view layer or adding local history.

Many of the intents authored by external contributors illustrate how Firefox Voice's extensibility makes it relatively straightforward to contribute new and powerful intents. As one example, a contributor wrote an intent to find an article by description, navigate to it, and read it aloud (e.g. "Read the 'Reduce Your Stress in Two Minutes A Day' article to me"). Building upon the infrastructure already in place for reading the current tab aloud, this contributor was able to extend Firefox Voice to add the new intent with only 11 new lines of code.

The open-source wakeword model and inference engine were also contributed by external collaborators who are academic researchers in machine learning, which also underscores Firefox Voice's contribution as a platform for modular experimentation in the voice research space.

Thus far, the Firefox Voice repository has been forked 125 times, and has over 280 stars on GitHub, suggesting that the larger community of developers may be adapting and learning from Firefox Voice's implementation.

8 LIMITATIONS

There are two key limitations both to Firefox Voice as a system, and to the research that we have presented in this paper. At present, Firefox Voice is available only in the English language. Because the system relies upon automatic speech recognition, it is unfortunately also prone to the biases inherent in the algorithms and datasets that underlie speech recognition engines [37, 71]. As a consequence, we found evidence that Firefox Voice may not perform as well for non-native speakers of English, and for those with accents other than the default of "standard American English." While not unique to Firefox Voice, this limitation troubles us, as it also suggests that the set of users who continue to engage with the assistant represent a more homogeneous population than we wish to reach. Improving localization to support a far broader population of users remains an important area of future work both for Firefox Voice, and for the language technologies community as a whole.

[13]We note that a large portion of this activity was driven by an application process to participate in an open source development internship with the Firefox Voice team.

We also recognize that this paper is limited given its focus on a shipping system. Not all of these studies are as thorough and in depth as we may have wanted, as our focus is to learn sufficiently from any given study to enable us to take the next step, rather than to necessarily create a novel research artifact that can stand alone. For instance, the needfinding survey was constrained in that it asked users about features of interest, but only offered options that were feasible for us to implement rather than options that are not currently available. While we believe the real-world, large-scale nature of the system's deployment and the data and insights derived from it are a valuable contribution, further qualitative evaluations (e.g. diary studies involving longer-term use) are an interesting opportunity for future work.

9 DISCUSSION

Our research presents an iterative exploration in building a voice assistant for the modern web. Through our formative studies, we contribute a characterization of the needs that people have when using voice assistants in the browser context, alongside an evaluation of a prototype that supports these discovered needs. Our evaluation highlights the strengths and shortcomings of Firefox Voice in a real-world deployment with more than 12,000 active users and active development from over 50 open source contributors. We now discuss the implications of Firefox Voice and our findings as they relate to future voice assistant research.

9.1 Voice assistants as unfamiliar interfaces in a familiar environment

A key challenge for voice interaction in the browser is that it is an unfamiliar interface modality in a well-trodden environment with deeply ingrained habits. Put another way, it is hard to change the browsing habits people have developed for as much as twenty years. Changing these browsing habits will require solving both problems of translation, and problems of discovery.

As others have noted, translating traditionally inaudible concepts—particularly those with familiar visual affordances—into voice is a non-linear process [84]. We began our exploration by investigating translation that was technical feasible. For example, Firefox Voice provides one answer to questions like *How can a voice assistant facilitate navigation within a list of search results? How can it audibly communicate affordances that are inherently visual? What does a link sound like?* It is likely that the best answers are yet to be found. Similarly, discovery for voice assistants remains unsolved, both in the browser and in the dedicated smart speaker. We found that users were often unsure of the utterances that Firefox Voice supported, often struggling with knowing whether a query would work correctly and how a particular query should be worded.

That said, our studies suggest that certain interactions are particularly well-suited to voice assistants in the browser. For example, "go to my calendar tab" is an intuitive way to find your calendar tab, and there is real value in querying Firefox Voice while engaged in other tasks by asking "Hey Firefox, what time zone is Minneapolis in?" without leaving the email you're currently writing. Solving the problems of translation and discovery will make these voice interactions even more appealing.

9.2 The unarticulated realities of commercial deployment

The lifecycle of a shipping system is complex. Publicity through existing products (such as banners on the new tab page of the Firefox desktop web browser in our case, or as banners in search result pages) bring many new users to the product. At the same time, not everyone who downloads the tool installs it; not everyone who installs the tool uses it. Both the sudden adoption and the non-use of technology in real-world deployments is often invisible in academic literature.

However, such deployment outcomes reveal important research questions. Firefox Voice, like many prior voice assistants [7, 18, 28], underwent both a feast and famine of users. In August 2020, Firefox Voice received publicity via a banner on the new tab page of the Firefox desktop web browser, resulting in approximately 30,000 downloads during that week alone. Its daily active users (DAU), a commonly used metric of usage, increased from an average of 28 in the first week of July to an an average of 1,072 in the first week of August, to an an average of 161 in the last week of August (Figure 4). Our uninstall survey is one of the few instruments that investigates the reasons behind these changes in usage, and what it might mean for future voice assistant design.

Firefox Voice is free to install, and has only the smallest indicator of its presence in the form of a small browser extension icon. As such, it is easier to adopt (and perhaps easier to abandon) than assistants based on dedicated hardware. Its lessons may be particularly important as voice assistant devices become more inexpensive, with lower barriers for adoption.

9.3 A Platform for Voice Assistant Experimentation and Research

It is fundamentally limiting that inquiries in voice assistant research (e.g. [32, 33, 51]) rely on product ecosystems that rigidly constrain researchers in the questions they can address, unless they choose to engineer their own voice assistant ecosystem.

This paper demonstrates how Firefox Voice's open source nature and use of existing hardware—the laptops and computers that are already in many people's homes—enables wide-open exploration of new features and new voice interactions. This exploration goes beyond functionality available through commercial voice assistant ecosystems that facilitate feature extensions through the constrained notion of "skills," running on specific hardware devices with limited functionality. Our work presents a pathway for browser-based voice assistant experimentation that allows practitioners and researchers alike to explore new voice interactions in an open ecosystem.

More broadly, we believe that Firefox Voice's open source nature adds diversity to an otherwise limited selection of voice assistant infrastructure and enables significant opportunity for exploring novel interactions on the web. In addition, we hope that teachers, researchers, and practitioners across HCI (and beyond) can use this work as a model for future systems research, whether it be voice-oriented or otherwise, as other systems contributions [22] have done so before us.

10 CONCLUSION

In this paper we present the initial exploration, iterative development, and large-scale deployment of an open-source voice assistant system, Firefox Voice. We show how our initial formative studies—notably a large scale "Needfinding" survey of the public and "First impressions" user studies with a working prototype—helped us develop design considerations that motivated the subsequent development and iterative design of the shipping system. We developed features in response to user feedback, such as a wakeword system, while providing significant amounts of functionality by repurposing existing features of browsers (cookies, saved passwords, history, and named tabs) and of the open web (search engines) for voice interaction. We released Firefox Voice in a real-world deployment that resulted in over 12,000 active users. We also showed subsequent natural drop-off in usage following users' initial exploration—a common but often under-reported phenomenon—and explored reasons for non-use through an "Uninstall" survey presented to users after they chose to uninstall Firefox Voice. We characterize the successful open-source nature of Firefox Voice, and as such, propose Firefox Voice as an open, extensible platform for exploration and development of novel voice-driven experiences running on already widely available systems.

ACKNOWLEDGMENTS

We would like to thank the engineers and researchers in our team whose work contributed directly to Firefox Voice: Tamara Hills, Ellen Spertus, Daniela Mormocea, Chioma Onyekpere, Jessica Colnago, Kai Lukoff, Jim Maddock, Benoit Zhong, Jordan Wirfs-Brock and Tawfiq Ammari. Thank you to Jimmy Lin, Raphael Tang, and Jaejun Lee for their work on Honk, the open source wakeword recognizer. It took a village at Mozilla to produce Firefox Voice: thanks to Jamie Teh, Sean White, David Bryant, Miriam Avery, Ali Spivak, Andre Natal, Venetia Tay, Rebecca Weiss, Niko Matsakis, Jane Scowcroft, Megan Branson, Janice Von Itter, Jenny Zhang, Rosanna Ardilla, Lindsay Saunders, Alex Klepel, Val Grimm, George Roter, Michael Feldman, Alicia Gray, Janette Ciborowski, Gemma Petrie, Amy Huang, Harly Hsu, Kelly Davis, Reuben Morais, Eren Gölge, Michael Stegeman, Amy Tsay, Diane Tate and many others, all of whom directly contributed in different ways to making Firefox Voice possible. Thank you to the organizers of the CUI workshops, the AAAI UX of AI Spring Symposium, the UC Santa Cruz HCI Forum, the Cornell Information Science Symposium, HCIC, and Stanford's Open Virtual Assistant Workshop for early public airings and feedback, as well as our many HCI colleagues including Frank Bentley, Alexis Hiniker, Pamela Wisniewski, Wendy Ju, Svetlana Yarosh, Leigh Clark, Martin Porcheron, Cosmin Munteanu, Katherine Isbister, Barry Brown, Don McMillan, and Mark Blythe. Thank you too to the many contributors to our open source codebase on Github, including Rubén Mur Monclús, Jennifer Harmon, Fabrice Desré, Farhat Sharif, David Okanlawon, Peter deHaan, Marwen Doukh, Amaka Mbah, Ganga Chaturvedi, Wil Clouser, John Gruen and Anna Nidhin, along with many more, as well as our active alpha and beta testers who provided suggestions and feedback.

REFERENCES

[1] 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 (ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 249–258. https://doi.org/10.1145/3234695.3236344

[2] Ali Abdolrahmani, Kevin M. Storer, Antony Rishin Mukkath Roy, Ravi Kuber, and Stacy M. Branham. 2020. Blind Leading the Sighted: Drawing Design Insights from Blind Users towards More Productivity-Oriented Voice Interfaces. *ACM Trans. Access. Comput.* 12, 4, Article 18 (Jan. 2020), 35 pages. https://doi.org/10.1145/3368426

[3] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3 (April 2019), 17:1–17:28. https://doi.org/10.1145/3311956

[4] Matthew P. Aylett, Benjamin R. Cowan, and Leigh Clark. 2019. Siri, Echo and Performance: You Have to Suffer Darling. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, alt08:1–alt08:10. https://doi.org/10.1145/3290607.3310422

[5] Marcos Baez, Florian Daniel, and Fabio Casati. 2020. Conversational Web Interaction: Proposal of a Dialog-Based Natural Language Interaction Paradigm for the Web. In *Chatbot Research and Design*, Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg (Eds.). Springer International Publishing, Cham, 94–110.

[6] Erin Beneteau, Ashley Boone, Yuxing Wu, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2020. Parenting with Alexa: Exploring the Introduction of Smart Speakers on Family Dynamics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376344

[7] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (Sept. 2018), 24 pages. https://doi.org/10.1145/3264901

[8] Yevgen Borodin, Jeffrey P. Bigham, Glenn Dausch, and I. V. 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) (W4A '10)*. Association for Computing Machinery, New York, NY, USA, Article 13, 10 pages. https://doi.org/10.1145/1805986.1806005

[9] Stacy M. Branham and Antony Rishin Mukkath Roy. 2019. Reading Between the Guidelines: How Commercial Voice Assistant Guidelines Hinder Accessibility for Blind Users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 446–458. https://doi.org/10.1145/3308561.3353797

[10] Bret Kinsella. 2020. Jovo v3 Launches with Support for More Platforms, More Devices, and Custom App Experiences. https://voicebot.ai/2020/02/28/jovo-v3-launches-with-support-for-more-platforms-more-devices-and-custom-app-experiences/

[11] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376789

[12] Julia Cambre and Chinmay Kulkarni. 2019. One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 223 (Nov. 2019), 19 pages. https://doi.org/10.1145/3359325

[13] Julia Cambre and Chinmay Kulkarni. 2020. Methods and Tools for Prototyping Voice Interfaces. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 43, 4 pages. https://doi.org/10.1145/3405755.3406148

[14] Julia Cambre, Ying Liu, Rebecca E. Taylor, and Chinmay Kulkarni. 2019. Vitro: Designing a Voice Assistant for the Scientific Lab Workplace. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. ACM, New York, NY, USA, 1531–1542. https://doi.org/10.1145/3322276.3322298

[15] Giovanni Campagna, Rakesh Ramesh, Silei Xu, Michael Fischer, and Monica S. Lam. 2017. Almond: The Architecture of an Open, Crowdsourced, Privacy-Preserving, Programmable Virtual Assistant. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 341–350. https://doi.org/10.1145/3038912.3052562

[16] Giovanni Campagna, Silei Xu, Mehrad Moradshahi, Richard Socher, and Monica S. Lam. 2019. Genie: A Generator of Natural Language Semantic Parsers for Virtual Assistant Commands. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI 2019)*. Association for Computing Machinery, New York, NY, USA, 394–410. https://doi.org/10.1145/3314221.3314594

[17] Minsuk Chang, Anh Truong, Oliver Wang, Maneesh Agrawala, and Juho Kim. 2019. How to Design Voice Based Navigation for How-To Videos. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300931

[18] Minji Cho, Sang-su Lee, and Kun-Pyo Lee. 2019. Once a Kind Friend Is Now a Thing: Understanding How Conversational Agents at Home Are Forgotten. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. Association for Computing Machinery, New York, NY, USA, 1557–1569. https://doi.org/10.1145/3322276.3322332

[19] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (Sept. 2019), 349–371. https://doi.org/10.1093/iwc/iwz016 arXiv:https://academic.oup.com/iwc/article-pdf/31/4/349/33525046/iwz016.pdf

[20] Eric Corbett and Astrid Weber. 2016. What Can I Say?: Addressing User Experience Challenges of a Mobile Voice User Interface for Accessibility. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '16)*. ACM, New York, NY, USA, 72–82. https://doi.org/10.1145/2935334.2935386

[21] 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 (MobileHCI '17)*. ACM, New York, NY, USA, 43:1–43:12. https://doi.org/10.1145/3098279.3098539

[22] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.Help: Designing a Workflow-Based Scheduling Agent with Humans in the Loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2382–2393. https://doi.org/10.1145/3025453.3025780

[23] Andreea Danielescu. 2020. Eschewing Gender Stereotypes in Voice Assistants to Promote Inclusion. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 46, 3 pages. https://doi.org/10.1145/3405755.3406151

[24] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours Is Better!": Participant Response Bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1321–1330. https://doi.org/10.1145/2207676.2208589

[25] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[26] Dustin Coates. 2020. 5 Voice Search Trends to Look out For.

[27] C. Ailie Fraser, Julia M. Markel, N. James Basa, Mira Dontcheva, and Scott Klemmer. 2019. ReMap: Multimodal Help-Seeking. In *The Adjunct Publication of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 96–98. https://doi.org/10.1145/3332167.3356884

[28] Frank Gillett. 2020. Getting Consumers Beyond Simple Tasks On Smart Speakers Is Challenging. https://go.forrester.com/blogs/getting-consumers-beyond-simple-tasks-on-smart-speakers-is-challenging/

[29] Global Web Index. 2018. *Voice Search: A Deep-Dive into Consumer Uptake of the Voice Assistant Technology*. Technical Report. GlobalWebIndex. https://www.globalwebindex.com/reports/voice-search-report

[30] Google. 2016. *Google App Voice Search Insights*. Technical Report. Google. https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/google-app-voice-search/

[31] 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 (SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 35–44. https://doi.org/10.1145/2911451.2911525

[32] Danula Hettiachchi, Zhanna Sarsenbayeva, Fraser Allison, Niels van Berkel, Tilman Dingler, Gabriele Marini, Vassilis Kostakos, and Jorge Goncalves. 2020. "Hi! I Am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376320

[33] Danula Hettiachchi, Niels van Berkel, Tilman Dingler, Fraser Allison, Vassilis Kostakos, and Jorge Goncalves. 2019. Enabling Creative Crowd Work through Smart Speakers. In *Workshop on Designing Crowd-Powered Creativity Support Systems*. CHI '19 Workshop, 1–5. http://www.jorgegoncalves.com/docs/chiea19c.pdf

[34] Internet Live Stats. 2020. The Total Number of Websites. https://www.internetlivestats.com/total-number-of-websites/

[35] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. Association for Computing Machinery, New York, NY, USA, 43–53. https://doi.org/10.1145/3172944.3172977

[36] Yea-Seul Kim, Mira Dontcheva, Eytan Adar, and Jessica Hullman. 2019. Vocal Shortcuts for Creative Experts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300562

[37] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial Disparities in Automated Speech Recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689. https://doi.org/10.1073/pnas.1915768117

[38] Gierad P. Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar. 2013. PixelTone: A Multimodal Interface for Image Editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2185–2194. https://doi.org/10.1145/2470654.2481301

[39] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 102:1–102:31. https://doi.org/10.1145/3274371

[40] Tessa Lau, Julian Cerruti, Guillermo Manzato, Mateo Bengualid, Jeffrey P. Bigham, and Jeffrey Nichols. 2010. A Conversational Interface to Web Automation. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. Association for Computing Machinery, New York, NY, USA, 229–238. https://doi.org/10.1145/1866029.1866067

[41] Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. Universal Voice-Enabled User Interfaces Using JavaScript. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 81–82. https://doi.org/10.1145/3308557.3308693

[42] Gilly Leshed, Eben M. Haber, Tara Matthews, and Tessa Lau. 2008. CoScripter: Automating & Sharing How-to Knowledge in the Enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1719–1728. https://doi.org/10.1145/1357054.1357323

[43] 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 (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[44] Xiao Ma and Ariel Liu. 2020. Challenges in Supporting Exploratory Search through Voice Assistants. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 47, 3 pages. https://doi.org/10.1145/3405755.3406152

[45] Rishabh Mehrotra, A Hassan Awadallah, 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)*. SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17), 1–5. https://rishabhmehrotra.com/CAIR17-cortana.pdf

[46] Sarah Mennicken, Ruth Brillman, Jennifer Thom, and Henriette Cramer. 2018. Challenges and Methods in Design of Domain-Specific Voice Assistants. In *2018 AAAI Spring Symposium Series*. 2018 AAAI Spring Symposium Series, 1–5. https://doi.org/10.21437/Interspeech.2017-1746

[47] Roger K. Moore. 2017. Is Spoken Language All-or-Nothing? Implications for Future Speech-Based Human-Machine Interaction. In *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Kristiina Jokinen and Graham Wilcock (Eds.). Springer Singapore, Singapore, 281–291. https://doi.org/10.1007/978-981-10-2585-3_22

[48] C. Murad, C. Munteanu, B. R. Cowan, and L. Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 2 (April 2019), 33–45. https://doi.org/10.1109/MPRV.2019.2906991

[49] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 6:1–6:7. https://doi.org/10.1145/3173574.3173580

[50] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[51] Elnaz Nouri, Robert Sim, Adam Fourney, and Ryen W. White. 2020. Proactive Suggestion Generation: Data and Methods for Stepwise Task Assistance. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1585–1588. https://doi.org/10.1145/3397271.3401272

[52] NPR and Edison Research. 2020. *The Smart Audio Report (Winter 2019)*. Technical Report. NPR and Edison Research. https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/

[53] Emmi Parviainen and Marie Louise Juul Søndergaard. 2020. Experiential Qualities of Whispering with Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376187

[54] Randy Pausch and James H Leatherby. 1991. An Empirical Study: Adding Voice Input to a Graphical Editor. In *Journal of the American Voice Input/Output Society*. Citeseer.

[55] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214

[56] Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. "Phantom Friend" or "Just a Box with Information": Personification and Ontological Categorization of Smart Speaker-Based Voice Assistants by Older Adults. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 214 (Nov. 2019), 21 pages. https://doi.org/10.1145/3359316

[57] Alisha Pradhan, Amanda Lazar, and Leah Findlater. 2020. Use of Intelligent Voice Assistants by Older Adults with Low Technology Use. *ACM Transactions on Computer-Human Interaction* 27, 4, Article 31 (Sept. 2020), 27 pages. https://doi.org/10.1145/3373759

[58] 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 (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174033

[59] A Purington, J G Taft, S Sannon, N N Bazarova, and S H Taylor. 2017. "Alexa Is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. *Conference on Human Factors in Computing Systems - Proceedings* Part F1276 (2017), 2853–2859. https://doi.org/10.1145/3027063.3053246

[60] Sarah Perez. 2019. Google Assistant Actions up 2.5x in 2018 to Reach 4,253 in the US. https://techcrunch.com/2019/02/18/google-assistant-actions-up-2-5x-in-2018-to-reach-4253-in-the-u-s/

[61] Ritam Jyoti Sarmah, Yunpeng Ding, Di Wang, Cheuk Yin Phipson Lee, Toby Jia-Jun Li, and Xiang 'Anthony' Chen. 2020. Geno: A Developer Tool for Authoring Multimodal Interaction on Existing Web Applications. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1169–1181. https://doi.org/10.1145/3379337.3415848

[62] Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. 2010. "Your Word Is My Command": Google Search by Voice: A Case Study. In *Advances in Speech Recognition*. Springer, 61–90. https://link.springer.com/chapter/10.1007/978-1-4419-5951-5_4

[63] S. Schlögl, G. Chollet, M. Garschall, M. Tscheligi, and G. Legouverneur. 2013. Exploring Voice User Interfaces for Seniors. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '13)*. Association for Computing Machinery, New York, NY, USA, Article 52, 2 pages. https://doi.org/10.1145/2504335.2504391

[64] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[65] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. Is Now A Good Time? An Empirical Study of Vehicle-Driver Communication Timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300867

[66] Ben Shneiderman. 2000. The Limits of Speech Recognition. *Commun. ACM* 43, 9 (Sept. 2000), 63–65. https://doi.org/10.1145/348941.348990

[67] Marie Louise Juul Søndergaard and Lone Koefoed Hansen. 2018. Intimate Futures: Staying with the Trouble of Digital Personal Assistants through Design Fiction. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. Association for Computing Machinery, New York, NY, USA, 869–880. https://doi.org/10.1145/3196709.3196766

[68] Aaron Springer and Henriette Cramer. 2018. "Play PRBLMS": Identifying and Correcting Less Accessible Content in Voice Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173870

[69] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice As a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 603:1–603:14.

https://doi.org/10.1145/3290605.3300833

[70] Raphael Tang, Jaejun Lee, Afsaneh Razi, Julia Cambre, Ian Bicking, Jofish Kaye, and Jimmy Lin. 2020. Howl: A Deployed, Open-Source Wake Word Detection System. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. Association for Computational Linguistics, Online, 61–65. https://doi.org/10.18653/v1/2020.nlposs-1.9

[71] Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Proc. Interspeech 2017*. INTERSPEECH, 934–938. https://www.isca-speech.org/archive/Interspeech_2017/pdfs/1746.PDF

[72] Janice Y Tsai and Jofish Kaye. 2018. Hey Scout : Designing a Browser-Based Voice Assistant. (2018), 460–462.

[73] Voicebot.ai. 2019. *Smart Speaker Consumer Adoption Report.* Technical Report. Voicebot.ai. https://voicebot.ai/wp-content/uploads/2019/03/smart_speaker_consumer_adoption_report_2019.pdf

[74] VoiceLabs. 2017. *2017 VoiceLabs Voice Report, Executive Summary.* Technical Report. VoiceLabs. https://s3-us-west-1.amazonaws.com/voicelabs/report/vl-voice-report-exec-summary_final.pdf

[75] Alexandra Vtyurina. 2019. Towards Non-Visual Web Search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 429–432. https://doi.org/10.1145/3295750.3298976

[76] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173782

[77] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W. White. 2019. Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 3590–3594. https://doi.org/10.1145/3308558.3314136

[78] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W. White. 2019. VERSE: Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 414–426. https://doi.org/10.1145/3308561.3353773

[79] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. *I'd Blush If I Could: Closing Gender Divides in Digital Skills through Education.* Technical Report. UNESCO, EQUALS Skills Coalition. https://unesdoc.unesco.org/ark:/48223/pf0000367416.locale=en

[80] Alex C. Williams, Julia Cambre, Ian Bicking, Abraham Wallin, Janice Tsai, and Jofish Kaye. 2020. Toward Voice-Assisted Browsers: A Preliminary Study with Firefox Voice. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 49, 4 pages. https://doi.org/10.1145/3405755.3406154

[81] Jordan Wirfs-Brock, Janice Tsai, Abraham Wallin, and Jofish Kaye. 2019. Listening: It's Not Just for Audio. https://blog.mozilla.org/ux/2019/12/listening-its-not-just-for-audio/

[82] Jordan Wirfs-Brock, Janice Tsai, Abraham Wallin, and Jofish Kaye. 2019. People Who Listen to a Lot of Podcasts Really Are Different. https://blog.mozilla.org/ux/2019/12/people-who-listen-to-a-lot-of-podcasts-really-are-different/

[83] Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. 1995. Designing SpeechActs: Issues in Speech User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 369–376. https://doi.org/10.1145/223904.223952

[84] John Zimmerman. 2020. Case for a Voice-Internet: Voice Before Conversation. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 44, 3 pages. https://doi.org/10.1145/3405755.3406149

A SUPPORTED INTENTS

Category	Features
Bookmarks	Bookmark current page, bookmark current page to folder, open bookmarks, remove bookmark for current page
Browser pages	Open installed add-ons page, open bookmarks, open history, open browser preferences
Clipboard	Paste, copy screenshot of visible window, copy full page screenshot, copy title of current tab, copy link of current tab, copy selection, copy best image in tab, copy Markdown link of title and url for current tab, copy HTML title and link, copy specified value to clipboard
Download	Download screenshot of visible window, download full page screenshot, download webpage, show last download
Email	Create draft email message with a given subject and/or body
Find tab	Find and focus a given browser tab by description
Forms	Dictate into form field, focus next form field, focus previous form field, submit form, turn the selected text into markdown or html link
Music	Open or switch focus to a music service, play next, play previous, mute, unmute, play, pause, resume, play album, play playlist, show title of currently playing, adjust volume, show which music services are supported in a card
Muting	Mute all tabs, mute specific tab, unmute
Navigation	Navigate to a particular webpage by description, click link, go back, go forward, search within a particular web service (e.g. search Amazon or search Wikipedia), translate full webpage, translate selection, close a lightbox-style dialog box, search current url on archive.org, follow the named sign in or out link and click
Notes	Create a note anchored to a tab, add to note with given text, add link and title of current tab to note, paste clipboard to note
Pocket	Open Pocket, save page to Pocket
Print	Open print dialog for page
Read	Open reader mode and begin speaking article aloud in synthesized voice, stop reading, go forward/backwards by one paragraph
Routines	Nickname a single intent, combine last N intents into a routine, nickname a page, remove nickname for page, remove a named intent, pause a running intent, resume a running intent, specify the beginning of a for loop in a routine, specify the end of a for loop in a routine
Scroll	Scroll down, scroll up, scroll to bottom, scroll to top
Search	Search with the user's browser-configured default search engine, open default search engine (without query), search Google, follow-up on prior search to show next search result, follow-up on prior search to show previous search result, focus the hidden tab used for card-based search results, search and show card-based result
Self	Help / testing, enable "smart speaker" mode (activates speech output and the wakeword), open Firefox Voice options, tell a joke, open lexicon of example supported commands, open developer-facing intent viewer, respond to simple "hello" greeting, cancel in-progress intent,
Sidebar	Open bookmarks or history sidebar, close bookmarks or history sidebar, toggle bookmarks or history sidebar
Slideshow	Begin presenting on Google slides, open slideshow
Tabs	Open a new tab, open homepage, open a window, open a private window, close current tab, close selected tab(s), close window, undo close tab, undo close window, refresh current tab, refresh selected tabs, zoom in, zoom out, reset zoom, pin current tab, un-pin current tab, focus the previous tab, duplicate current tab, move selected tab(s) to a new window, move current tab to a new window, make tab full screen, find query within page, focus next result of query in page, focus previous result of query in page, select all tabs, select first/last numbers of tabs, select tab by description, select tabs to the left/right, save tab as PDF, collect tabs similar to active tab or by query, count total number of open tabs, focus on the previous/next tab, go to the first/last tab, read title of tab
Timer	Set a timer for a given duration, restart a timer, pause a running timer, resume a paused timer, cancel a timer
Window	Minimize, clear browser history, close window, switch focus to the previous / next window, maximize a window, open downloads, quit Firefox

Table 2: An brief enumeration of all of the features (individual intents) that Firefox Voice supports, grouped by the general category of functionality