Case for a Voice-Internet: Voice Before Conversation

John Zimmerman
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA
johnz@cs.cmu.edu

Abstract

This sample paper describes the formatting requirements for SIGCHI Extended Abstract Format, and this sample file offers recommendations on writing for the worldwide SIGCHI readership. Please review this document even if you have submitted to SIGCHI conferences before, as some format details have changed relative to previous years. Abstracts should be about 150 words and are required.

CHI 2020 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.
© 2020 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6819-3/20/04.
DOI: https://doi.org/10.1145/3334480.XXXXXXX
*update the above block & DOI per your rightsreview confirmation (provided after acceptance)

Author Keywords

Voice activated personal agents; Conversational Interaction; Conversational user interface; accessibility; universal design.

CSS Concepts

- **Human-centered computing~Human computer interaction (HCI);** *Interaction paradigm; Natural language interfaces.*

Introduction

This position paper makes the case for HCI's active participation in constructing a voice-Internet. There is an alignment of technical, social-behavioral, and economic forces that make this the right time for a voice-Internet to be created.

Not long ago, I was working to improve the experience of screen reader users who struggle to navigate tabled information of web pages. Hearing the linearized content is both tedious and confusing. A participant in a study asked why they could not access the table conversationally, sending our work in a new direction. We carried out two studies that led to our reframing of the problem of web tables and screen readers as an opportunity for a voice-Internet. First, we investigated how well current voice activated personal agents (VAPAs) answer questions where the desired answer resides in a web table. They did really well, correctly answering more than 70% of the questions. Second, we deployed Google Homes in the homes of people who regularly use a screen reader and probed them on what this device could and should do. Participants found the devices useful, especially for local information and planning. However, they often felt limited to a thin layer of information.

HCI research shows that VAPAs are unintentionally beneficial to people with disabilities, especially people that use screen readers [4]. Research also shows the experience of using a screen reader is often pretty terrible [3]. The web uses visual semantics (e.g., color, typeface, co-location of information) to communicate relationships. It also uses lots of images. It is a media designed to be seen, not heard. It does not seem like reading content, made to be seen and not heard, will ever be a good solution.

The Internet has a history of creating new forms of complimentary information. When mobile devices arrive with apps, these apps required that enterprises do more than shovel their web content into an app. Companies had to rethink their message, crafting materials for this new form. Today, they often leverage APIs to share some of the same content between web pages and apps. VAPAs in the form of smart speakers, smartphone agents, and now AirPods and ear buds offer a new platform for accessing the Internet; one that privileges speaking over seeing. Technically, the hardware is largely already in place for a voice-Internet. In addition, many companies are starting work on chatbots, indicating a willingness to create new conversational content, when it attracts new users or saves the cost of paying human service agents.

My idea of a voice-Internet is different than a conversational Internet. I am talking about people creating content that is meant to be heard, not seen. And I am talking about the development of apps that support voice interaction, not apps where a screen reader reads a visual interface. In my experience, the technology is not in place for a truly conversational Internet. Natural language processing (NLP) technology has good syntactic and word-level semantic information. But it does not understand sentences. It can construct human-understandable responses, but most often lacks an understanding of its own response. I believe a voice-Internet is the next logical step in the evolution of conversational interaction.

The case for voice accessible information

The development of the web resulted in a huge and valuable information resource and in a transactional platform. It reduced many of the limitations imposed by time and place, making information and transactions available from anywhere. The web reduced the demands on many call centers by providing information that was browsable and searchable, allowing more and more people to serve themselves. Screen readers made much of this information available to people who are blind or low vision. However, since their development, screen readers have offered only limited access to all information on the web. Much of the content on the web is images and visual semantics (e.g. typographic hierarchy, co-location of information, color as a categorial feature), most of which get lost when the web is converted into spoken word.

In work to improve screen readers’ ability to navigate and make sense of tables on web pages, we started exploring if VAPAs like Alexa and Google Home might

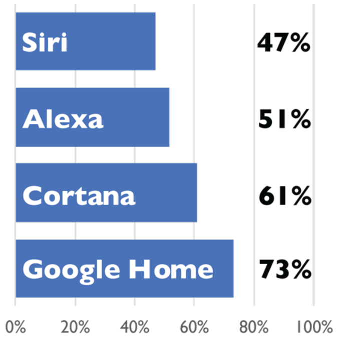

Figure 1: Ability of the four VAPAs to correctly answer the crowdworkers' questions. This was calculated by dividing the sum of a VAPA's score by 1500 (the sum a VAPA would have had if it earned a 3 for each question).

offer some help. We wanted to know how well they answer questions when the answer can be found within tabled information on the web. We wanted a performance baseline before committing to the development of a new technical system. To generate a corpus of questions, we asked crowdworkers to search their web browser history for instances where they had turned to the web to answer a question and the answer had appeared within a web table. We collected 2500 questions. We then curated this down to a set of 500. We filter out non-questions, unintelligible questions, and repetitive questions (e.g., weather, sport scores). We wanted a corpus that covered the breadth of questions, not one that captured the frequency of different kinds of questions. We then tested all questions in our corpus using Siri, Cortana, Alexa, and Google Home. The VAPAs all performed well, with Google Home doing the best (See Figure 1.).

This study gave us hope that VAPAs might offer some solution to the challenge of web tables. When users knew what they were looking for, a conversational question could often help them find it. But this offered no solution to situations where users did not know what they wanted before browsing a table. We also were unsure if the needs of crowdworkers were a good enough representation of why screen reader users might turn to the web.

We decided to conduct a second study, placing Google Homes (GH) in the homes of people who regularly use a screen reader to access the web. We recruited 10 participants. We conducted a first interview to understand what they typically did on the internet and their familiarity with technology. During this interview, we installed the GH. Approximately two weeks later, we conducted a second interview. Prior to the interview, we reviewed logs to understand what they were doing with the GH. During this second interview, we discussed how useful the GH might be for question answering. We conducted a third interview approximately two weeks later. During this interview, we focused on what they wished the GH could do.

Participants found the GH useful. All used it to gather information from near their home. This included asking about the weather, asking about hours stores were open, asking how long it took to drive to various stores, and even asking for walking directions. Participants also asked for information about restaurants. Many wanted to get details about menus, something not available from GH. When we checked some of these menus online, we noticed they were stored as images.

Several participants attempted to use the GH as a sort of screen reader proxy; they worked to get deeper information from the device. In one example, a participant asked for a list of children's books and got the top ten from the New York Times. They then attempted to hear about other books lists and to get a longer version of the New York Times book list for children. Others wanted to get details about a company, information typically found in the *about* tab of a company's web site. The VAPA could not meet their desires as it seemed designed to only provide a quick, skim level of information.

The interaction designers for VAPAs seemed biased by a screen dominant culture. Participants who attempted to change the settings on their GH, such as adding a credit card or tying it into smart outlets, were informed that they needed to change all settings via a mobile app. In addition, many of the information seeking tasks participants took on led to the VAPA responding that it had sent relevant links to their phone.

Our study showed that VAPAs can be quite valuable for people who use a screen reader to access the web. The GH provided valuable access to local information via conversational interaction that was often significantly easier than use of a PC or smartphone. Participants mostly wanted the VAPA to do more, and they wanted designers of VAPAs to get past their screen dominant bias and create a VAPA that could be more independently controlled and modified via voice.

Voice-Internet over Conversational-Internet

My call for a voice-Internet is distinctly different than a call for a conversational-internet. I am not against a conversational-Internet. However, I do not believe the state of natural language processing (NLP) is ready to support anything approaching human-to-human conversation. I think a voice-Internet is a much more likely and possible near-term goal. This insight is partially driven by the many enterprises who are investing in chatbots, most of which function as a speaking FAQ. Consumers ask questions and the bot selects an answer from a set of pre-scripted answers. This approach is quite limited in terms of conversation. However, it keeps companies safe from the many unexpected things a content generating chatbot might say. The fact that companies are investing in chatbots shows a willingness for enterprises to create new content and a new channel of interaction for their customers.

NLP has made great strides, but it is still really dumb from a UX perspective. It has no commonsense [2] meaning it often produces unexpected and ridiculous responses. It has been largely trained only on news datasets. And it works in very few languages. NLP works well at the syntactic and word-semantic level. But it does not really understand sentences. In a recent course I taught on how to design AI products and services, the time spent on NLP mostly involved lowering students' expectations of what is possible.

It was hard for students to gain a machine learning perspective on language understanding, to see how the computer looks at language using resources like Word2Vec. They found that the language used by NLP could lead to unrealistic expectations. For example, topic modeling is a well-known NLP technique for sorting a corpus of documents into clusters. But the name "Topic Modelling" implies that NLP can recognize topics in the text and that the end result will be documents sorted into a set of topics that humans would understand. This is not what topic modeling produces. Topic modeling produces a list of words/terms that frequently show up in a small set of documents, thus forming the cluster [Chang].

In my opinion, NLP is not nearly robust enough for the HCI community to make a promise of conversation to end users. The technology, as VAPAs show, works well for simple voice command and control, for simple question answering, and for completion of simple transactions, such as ordering food. We should aggressively build on what is now possible and robust, but not make promises the technology cannot keep.

First Steps Towards a Voice-Internet

Our studies of VAPAs showed two things. First, they showed that the NLP technology inside current VAPAs works well for tasks like question answering, potentially eliminating some of the reasons people turn to the web. Second, they revealed an unmet desire by people who use screen readers for VAPA access to deeper information. I believe this desire also extends to many sighted people. This larger group often engages in tasks that require the use of hands and eyes. Many mundane and repetitive tasks like cleaning or folding laundry leave a lot of available attention that might be recaptured in an entirely new way. A voice-Internet would transform these tasks for users who want more than music, talk radio, and podcasts.

A first step towards a voice-Internet could be the development of new apps designed for voice interaction. This is different than screen reading an app made to be seen; it involves reconsidering the interaction and content from the ground up. VERSE, a voice application for web search, offers a preview of what this might be like [5]. I would love to see our community push in this direction by creating voice-based versions of popular applications, particularly applications that provide API access. We should constructively investigate voice messaging (e.g., email, Discord, Slack), social media (e.g., Snapchat, Instagram), and simple transactions such as ordering food or rides. These might be an effective means to triggering enterprises to jump into this game. Building a voice-Internet can lay the groundwork for normalizing voice-based and text-based interaction that can lead to a conversational-Internet once the underlying NLP technology has been developed.

Acknowledgements

The contents of this paper were partially developed under a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90REGE0007). This work was also supported by the National Science Foundation under Grant Number SES1734456. Any findings, conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

[1] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the Conference on Neural Information Processing System.* ACM.

[2] Karen Hao. 30 January 2020. AI still doesn't have the common sense to understand human language. MIT Technology Review. Retrieved February 11, 2022 from https://www.technologyreview.com/s/615126/ai-common-sense-reads-human-language-ai2/

[3] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What frustrates screen reader users on the web: A study of 100 blind users. *International Journal of Human-Computer Interaction* 22, 3: 247-269.

[4] 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 Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3173574.3174033

[5] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen White. 2019. VERSE: Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search. In *Proceedings of the SIGACCESS Conference on Computers and Accessibility*. https://doi.org/10.1145/3308561.3353773