Interactive Visual Description of a Web Page for Smart Speakers

Peggy Chi
Research Scientist
Google Research
1600 Amphitheatre Pkwy
Mountain View, CA 94043, USA
peggychi@google.com

Irfan Essa
Research Scientist
Google Research
1600 Amphitheatre Pkwy
Mountain View, CA 94043, USA
irfanessa@google.com

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
Smart speakers are becoming ubiquitous for accessing lightweight information using speech. While these devices are powerful for question answering and service operations using voice commands, it is challenging to navigate content of rich formats–including web pages–that are consumed by mainstream computing devices. We conducted a comparative study with 12 participants that suggests and motivates the use of a narrative voice output of a web page as being easier to follow and comprehend than a conventional screen reader. We are developing a tool that automatically narrates web documents based on their visual structures with interactive prompts. We discuss the design challenges for a conversational agent to intelligently select content for a more personalized experience, where we hope to contribute to the CUI workshop and form a discussion for future research.

Author Keywords
Smart Speakers, Text-to-Speech, Voice User Interfaces, Web Design, Screen Readers, Content Analysis

CCS Concepts
•Human-centered computing → Human computer interaction (HCI);

Figure 1: Designed for smart speaker users, our pipeline automatically creates narration for visually describing a web page based on the document structure. It connects key DOM elements in a narrative form that helps listeners better capture the content.

Introduction

Smart speakers provide lightweight access to digital content for everyday users via a hands-free voice user interface. Supportive AI-powered platforms–including Google Assistant, Amazon Alexa, and Apple's Siri–enable users of a wide range of age groups to obtain real-time information (e.g., weather and time) and service controls (e.g., music playback, product orders) [9, 7]. The responses are often limited to curated content in short, simple sentences, such as "*Today in San Francisco it's predicted to be 60 degrees and clear*". To navigate content of rich formats, such as Web documents, assistants rely on Text-to-Speech (TTS) voice readers [8]. Designed for accessibility purposes, TTS reads aloud text elements from a document, often by traversing its Document Object Model (DOM) structure. Recent work has shown how AI can power the selection process based on text content[1].

While the visual structure is critical for web designs [5, 6], it is rarely preserved by TTS, including the *hierarchy and grouping* (e.g., headings or containers), *spatial relations* (e.g., the largest banner or grids of similar elements), and *visual cues* (e.g., small text or color indicators). These visual design decisions are critical for web users to efficiently consume content beyond text. For examples, the hierarchy presents what is important[2] , and density improves scanning content[3] . Although users of smart speakers may be familiar with visual cues when navigating a website with their computers or phones, they are not able to capture such information with existing voice readers. Furthermore, it is challenging to form a conversation with the agent to interactively navigate a rich document.

[1] https://blog.google/products/assistant/ces-2020-google-assistant/
[2] https://material.io/design/usability/accessibility.htmlhierarchy
[3] https://material.io/design/layout/applying-density.htmlusage

Researchers have suggested that "*a good image description is often said to "paint a picture in your mind's eye.*"" [2]. We hypothesize that with an understanding of visual structure in a web document, we could create narration in a dialog form that preserves the design intentions for listeners to better capture the web content. To this end, we conducted a formative study that investigated whether reading web pages with a narrative structure better supports users following the content than conventional TTS readers. Building on this study, we propose a technique that automatically presents a web document for smart speakers based on the page's visual structure as a dialog (see Figure 1). We discuss the design challenges and opportunities to enable users to navigate a rich page through a conversational UI.

Narrative Design Guidelines

We propose four design guidelines for interactive web page narration on smart speakers:

G1. *WYSIWYG* ("What you see is what you get"): Image captioning research suggests that describing the salient objects and structure can help audience reconstruct the scene [2]. In our context, narration should focus on the dominant elements that infer important information shown on a web page. Small or hidden content is not always suitable to include.

G2. *Reveal content ordering and hierarchy*: Web structures often maintain design intention and usability rationale [5, 6]. It is critical to reveal the ordering (from top to down) and unfold the hierarchy of web elements (e.g., headings and grouping) for better content following [10].

G3. *Provide concise information with options*: Smart speaker users look for direct access to concise information [9]. Duplicated messages (e.g., alt text) and general

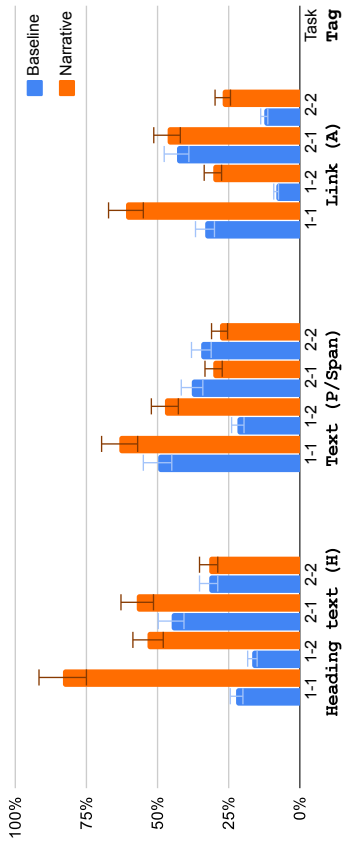


Figure 2: Participants captured the headings and links more from the narrative condition than the baseline for all tasks.

statements (e.g., disclaimer) should be avoided. Detailed content can be provided as a conversational option for users to follow if appropriate.

G4. *Present as a narrative*: Content in a descriptive, contextual form could improve comprehension and usability [8]. Designed for smart speakers, narration should be constructed as if a human agent were describing the document in a conversational context.

Formative User Study

In the first step, we hypothesize that users can capture the key messages of a web page from a voice reader that *describes a document with narrative principles* than a conventional reader that *reads only the text elements* without a narrative structure. We formulated a within-subject study to compare two conditions:

The *baseline* is the TTS functionality provided by operation systems. We chose the built-in function from MacOS' VoiceOver by recording the TTS output and polished the transcript by removing unnecessary readouts. The *narrative condition* is the manual-modified transcripts from the baseline based on our proposed design guidelines. We created a set of rules to convert the TTS output to a narrative form. Specifically, we added context (page title), content categories (menu, heading, image, and link) and grouping, and connection. Headers and footer (e.g., copyright and legal statements) were removed.

We selected pages where their design styles covered our editing rules and synthesized all the tasks with the same male voice at the speech rate 0.50 using Google Cloud's TTS API [4]. The voice outputs were recorded as WAV audio files for playback. We recruited 12 sighted participants (5 females and 7 males) from our company in the U.S. Participants were all native or professional working proficiency in English and had used at least one voice-based AI assistant product prior to the study.

Each 30-minute in-lab session included 2 conditions in the random order (within-subjects), each consisted of 1 warm-up task and 2 experimental tasks presented in the following procedure: In the warm-up task, participants followed an audio recording file given the instruction: "*Please capture the content of the web page. Note that the recording will be only played once.*" They were then asked to describe the document verbally. The original web page was then shown to visualize the content and the reader's logic. Participants experienced the 2 tasks in the same format without seeing the websites. Once the 3 tasks were completed, a survey with 5 Likert-scale and 1 multiple-choice questions was provided.

Results For each task, we extracted the web page's component by tags into 5 groups: headings (`H1-H3`), standout links (`A` that are not embedded in a paragraph), regular text (`P`, `SPAN`, or non-annotated text in a `DIV`), header or footer, and others. We recorded and mapped participants' verbal descriptions to the components. Below we summarized the results and discussed how they support our design guidelines.

Describing content hierarchy helps capturing. (G2,G3) Participants strongly agreed that content was well-structured in the narrative condition (Median=4.5) and disagreed in the baseline (Median=2.5). Among the 12 participants, eight found the structural description helpful. We observed that in the narrative condition, participants wrote more structured notes and verbally described in the similar way (using key words such as "*First*", "*Next*", "*There are 5 sections*".) P4 explained, "*I like it. It provides the visuals that I could picture nicely,*" and P11, "*It is much*

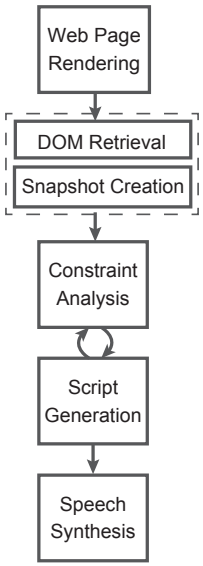


Figure 3: Our script creation pipeline. Given a URL, our system renders the web page to retrieve its hierarchical structure and a snapshot. It then analyzes the content based on the constraints and creates a script that can be synthesized for a speech output.

easier to know what's important."

Specifying element types reduces guessing. (G1,G2)
To understand how narrative supports users gathering information, we coded whether each web page component was mentioned by the participants. In the narrative condition, participants captured content from the headings and links significantly more ($P < 0.001$ and < 0.01 respectively), whereas there is no significant difference for regular text (see Figure 2). In the baseline, all the participants used at least one assumption word in the tasks (e.g., "*I think*", "*Maybe*", "*I guess (it was a button)*"). In the narrative condition, participants showed more confidence. 9 out of 12 participants found the description of element types helpful.

Narrative makes content easy to follow. (G4)
Participants agreed that the narrative was easy to follow (Median=4) and was neutral using the baseline (Median=3). In Task 2-2 with long content, P1 specifically shared that "*I was kind of lost in the later part about the exhibitions, but when I heard the word "Finally", I was immediately back on to the ending.*"

Automatic Dialog Generation for Navigation

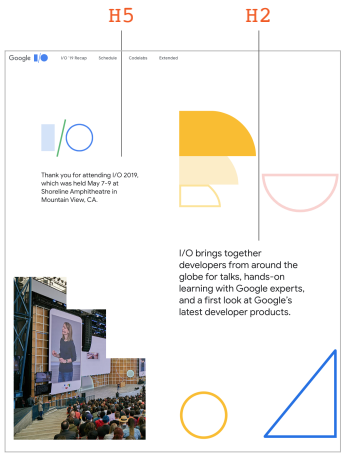
Our study shows that, content of a web page can be highlighted as a narrative for listeners to capture effectively. In order to bring such benefits to smart speaker users for accessing web content at the scale, we are developing an automatic pipeline that generates interactive narration and voice output of a web document.

We target on web pages that contain certain degree of visual elements, such as commercial or organization websites. Content from these sites are often well-structured to present headings (to introduce products, campaigns, or missions), text snippets (detailed information), and images (of products or services). Our system is not designed for long articles, such as news or blogs when text summarization is critical.

System
In our proposed pipeline, the system first renders a web document from a URL and acquires the document's hierarchical HTML structure, visual style, and content (including text and multimedia) of a rendered snapshot. It analyzes its DOM elements, selects and orders the content based on constraint settings. Finally, our pipeline creates an interactive narration script that can be used to synthesize speech output and respond to user's command (see Figure 3).

Narrative Construction
Based on a set of selected DOM objects, our system wraps the nodes with a narrative template, which adds sentence connections and contextual information. For example, a heading can be described as "*Below it shows a title, . . .*" (which describes the vertical ordering), "*It has a title, . . .*" (which only points out the type), or simply the text content like existing TTS outputs. Our text template also includes follow-up prompts, such as "*would you like me to read more?*" Finally, the text components from the script are synthesized to a speech output for playback.

Preliminary Results
We tested our current pipeline with 35 web pages. Specifically, we noticed how hierarchy analysis supports narration ordering. For example, the Google I/O page starts from a lower-level `H5` subtitle with smaller font size on the top left, followed by a higher-level `H2` heading on the lower right (see Figure 4). We tested the page with VoiceOver, which described from the subtitle and the main heading. However, similar to poster design, the most important message with a more significant visual occupancy is often



It shows a title: "I/O brings together developers from around the globe for talks, hands-on learning with Google experts, and a first look at Google's latest developer products."

It then describes: "Thank you for attending I/O 2019, which was held May 7-9 at Shoreline Amphitheatre in Mountain View, CA."

Figure 4: By analyzing hierarchy and DOM annotations, our prototype result describes web content based on the design intents (from higher-level heading) instead of rendered ordering (from the top-left corner) in a `section` container.

placed further from the top. Our work was able to narrate from the main message (the `H2` heading) based on the hierarchy and annotations.

We invited 5 sighted participants (1 female) who did not attend our formative study and provided the same amount of incentive to follow a set of our system output. We received consistent results: Participants well captured the content, explicitly highlighted the titles and links, and provided structured responses with confidence. Participants strongly agreed that the voice output was well-structured (Median=5). They agreed that the speech was easy to follow (Median=4) and they could capture the information (Median=4).

Design Challenges

As we are developing our technology to automatically generate interactive narrative for web document navigation, we organize a set of challenges of designing such a conversational user interface to be discussed with the workshop participants:

A web page commonly contains a variety of information in multiple categories. It is critical to select suitable content to initiate a conversation with the user. Designed for smart speakers, it should eventually enable conversational navigation in a multi-round dialog or through user demonstration [1] for more personalized experiences. One common design is to incorporate questioning and answering [9]. In our scenario, we develop an interactive script with follow-up options for users to choose from. Such narrative options include, to continue reading the remaining content in a section, to follow a specific topic or a link to another page, and to answer relevant questions (e.g., "*Find me the store location and hours on the web page*"). The challenges are to understand the adequate length of information per conversational round, and to provide a reasonable list of options for users. How much information should we condense? How do we prioritize the choices?

Ideally, a virtual assistant should understand users' needs [3]. It can be challenging to provide contextual response dynamically based on users' interests, time, and location under privacy concerns. For example, before dinner time, illustrating the menu or location of a restaurant might be more suitable than narrating the news on its web page. How do we respond intelligently while providing consistent information?

We also hope to see how this work can potentially bring insights to different domains, including art exploration, news consumption, and PDF reading. All in all, we aim to deep dive into CUI grand challenges in this workshop toward a future where users can universally access web information via voice-based interaction.

Conclusion

We propose a technique for creating an interactive navigation experience from a web page, in an audio form, on smart speakers, based on its visual structure. We present a formative study that verified that a narrative voice output was easier to comprehend than a conventional screen reader, which reads all the text content. We discuss design challenges of a conversational agent to select content and navigate intelligently with a user, with which we hope to contribute to the CUI workshop.

Acknowledgments

This work has been possible thanks to the support of people including, but not limited to the following (in alphabetical order of last name): Jimmy Lin, Zheng Sun, Weilong Yang, Yu Zhong, and our study participants.

REFERENCES

[1] Jeffrey P. Bigham, Jeremy T. Brudvik, and Bernie Zhang. 2010. Accessibility by Demonstration: Enabling End Users to Guide Developers to Web Accessibility Solutions. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '10)*. ACM, New York, NY, USA, 35–42. DOI: http://dx.doi.org/10.1145/1878803.1878812

[2] X. Chen and C. L. Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2422–2431. DOI: http://dx.doi.org/10.1109/CVPR.2015.7298856

[3] Ed H. Chi, Peter Pirolli, Kim Chen, and James Pitkow. 2001. Using Information Scent to Model User Information Needs and Actions and the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*. ACM, New York, NY, USA, 490–497. DOI: http://dx.doi.org/10.1145/365024.365325

[4] Google. 2019. Cloud Text-to-Speech - Speech Synthesis. (2019). Retrieved August, 2019 from https://cloud.google.com/text-to-speech/

[5] Ranjitha Kumar, Arvind Satyanarayan, Cesar Torres, Maxine Lim, Salman Ahmad, Scott R. Klemmer, and Jerry O. Talton. 2013. Webzeitgeist: Design Mining the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 3083–3092. DOI: http://dx.doi.org/10.1145/2470654.2466420

[6] Ranjitha Kumar, Jerry O. Talton, Salman Ahmad, and Scott R. Klemmer. 2011. Bricolage: Example-based Retargeting for Web Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2197–2206. DOI: http://dx.doi.org/10.1145/1978942.1979262

[7] Silvia B. Lovato, Anne Marie Piper, and Ellen A. Wartella. 2019. Hey Google, Do Unicorns Exist?: Conversational Agents As a Path to Answers to Children's Questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children (IDC '19)*. ACM, New York, NY, USA, 301–313. DOI: http://dx.doi.org/10.1145/3311927.3323150

[8] Daisuke Sato, Shaojian Zhu, Masatomo Kobayashi, Hironobu Takagi, and Chieko Asakawa. 2011. Sasayaki: Augmented Voice Web Browsing Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2769–2778. DOI: http://dx.doi.org/10.1145/1978942.1979353

[9] 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)*. ACM, New York, NY, USA, 857–868. DOI: http://dx.doi.org/10.1145/3196709.3196772

[10] Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in Data-to-Document Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2253–2263. DOI: http://dx.doi.org/10.18653/v1/D17-1239