Human-Computer Interaction -- INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 856-859

The Ears Have It: A Task by Information Structure Taxonomy for Voice Access to Web Pages

Manuel A. Pérez-Quiñones, Robert G. Capra & Zhiyan Shao

Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
{ perez | rcapra | zshao }@vt.edu

Abstract: We present a taxonomy of *task* by *information structure* for voice interfaces to web pages in an analogous effort to Shneiderman's taxonomy for information visualization. Our goal is to develop guidelines for the development of voice navigation of web spaces not as a replacement for visual web browsing, but instead to support focused information seeking tasks such as known-item search and directed browsing. We describe high-level user tasks (Situate, Navigate, Query, and Details) and information structures (regions, menu/lists, text areas, repeated/structured information) that comprise the axes of our taxonomy and show how voice interfaces can support these tasks and structures.

Keywords: voice user interface, VoiceXML, WWW tasks, voice navigation, taxonomy

1 Introduction

From a technology point of view, voice navigation of web spaces is a reality today. VoiceXML [2, 6] is a language for building voice user interfaces, in particular, ones to be used over a phone. Using VoiceXML, application designers can create effective voice interfaces that integrate with web-based information sources. Deploying such interfaces is made relatively easy through voice services such as BeVocal[1] or TellMe[2].

However, two challenges remain for voice access to web information that are harder to solve. The first challenge is how to provide voice access to the large body of HTML content that already exists. In other words, how can HTML-coded information be converted for presentation in VoiceXML? The web has been designed with visual attributes so intertwined with content that it is difficult to think of one without the other. Widespread use of technologies such as CSS, storing content in XML, and those promoted by the Semantic Web initiative[3] may change this, but by far, the web is still coded with HTML, using layout features (e.g. tables, frames, color highlights, images) that are woven together with the content of the page.

One solution is the use of transcoding and annotations. Transcoding [4] is a process by which content and presentation available in HTML for large displays is *transformed* to a more suitable representation that could be rendered on other devices (e.g. text-only, small screens, voice). Annotations are *hints* that are added to the original HTML with the purpose of helping the transcoding, thus improving the quality of the resulting interface. Transcoding and annotations have been used to modify HTML for display on small screens (i.e. PDAs) [4], to provide increased non-visual access [1], and to convert HTML to VoiceXML [3].

The second challenge, and the one that is addressed in this paper, is to decide how to present web-based information in an *effective* and *useable* voice user interface. How should we navigate web spaces using a voice-*only* user interface, in particular for web spaces that were designed to be visually presented? The approach presented in this paper is a qualified answer to this question. We assume that, for most users, access to the web over a telephone-based voice user interface is not going to be a

[1] www.bevocal.com
[2] www.tellme.com
[3] www.w3.org/2001/sw/

substitute to using the web from a personal (desktop or laptop) computer. We further assume that telephone-based access will be done while away from the computer and that the primary tasks will be to search for some known item (e.g. what will be the *weather* for my town tomorrow?) or to conduct focused-browsing (e.g. what is the latest sports news?).

Overall, the goal of our approach is not to reproduce the visual representation of page over a voice interface, but instead to support limited, directed information seeking tasks.

In this paper, we present a taxomony of *task* by *information structure* of web pages, defined with the goal of producing voice user interfaces for directed information seeking. This effort is analogous to the work described by Shneiderman [8] for information visualization. Our work in this area is just beginning; in this paper we present the development of the taxonomy, including descriptions of the tasks and information structures, and highlight existing successes and challenges for voice interfaces to support elements of the taxonomy.

We currently have an implementation of an HTML-to-VoiceXML transcoder that uses annotations to realize a voice dialogue for the information structures that we describe in this paper. It is implemented with Web Intermediaries[4] within the IBM's WebSphere Transcoding Publisher[5].

2 Approach

We set out to examine the question of how to present web-based information in a voice interface, especially for directed information seeking tasks. We printed 40+ web pages, covering different types and styles of web sites: commercial sites, university pages, personal home pages, news and portal pages. Our research group discussed possible voice interfaces that would provide access to information on the pages printed. We paid particular attention to how to navigate within a page, an issue that is not too difficult for graphical web pages because they rely heavily on visual scanning, but one that can be very challenging in the serial auditory channel of a voice interface. We were not particularly concerned about how the voice-only interface would be built, but focused on how it would be used by a user.

As a result of our analysis, we identified major information structures that are present on web pages and user tasks that would be afforded by those structures. In this sense, what we found is similar to the "Eyes Have It" paper by Shneiderman [8], where each type of data affords a type of visualization and interaction. Once we identified the types of information structure, which we discuss below, we then focused on the user needs while navigating these spaces. What we came up with is a taxonomy of information structure by user task. The rest of the paper discusses the user tasks and information structures we found on the web, the taxonomy we propose, and comments about how various aspects of the taxonomy can be realized in effective voice interfaces.

3 User Tasks

In [8], Shneiderman proposes an information visualization mantra, "Overview, zoom & filter, details on demand," as the most important guideline for visualization design. The approach described by Shneiderman, and used by many in the field, makes extensive use of the tremendous visual perceptual abilities that humans have. The bandwidth of communication using the visual domain is so high that a single interface with an easy to use interaction model is effective for a large number of tasks. For example, the Dynamic HomeFinder [10] can answer a large number of different queries about real estate, using just a few very simple interactions. This is the advantage of information visualization; the combination of large visual displays, with the high bandwidth communication afforded by visual perception, coupled with an easy to use direct manipulation interface.

A voice-only interaction channel lacks the visual bandwidth to transfer large amounts of data in parallel. The serial and temporal nature of voice communication requires a different style of interface; one that leverages and supports human capabilities for auditory and language processing. The interface should support providing users an understanding of its structure, how to navigate that structure, how to filter information, and how to focus in on detailed information. Thus, the information seeking approach for voice could be characterized as:

Situate, Navigate, Query, Details on demand

where ***situate*** means helping the user identify his/her location within the information space; ***navigate*** refers to the user's movement within the information space or control of the interface; ***query*** means to issue specific information requests; and ***details-on-demand*** refers to users task where s/he requests more details on specific item(s).

[4] www.almaden.ibm.com/cs/wbi/

[5] www-3.ibm.com/software/webservers/transcoding/

These four high-level user tasks form one axis of our taxonomy. They address the serial nature of the voice interface channel, and the ease with which a user could get lost in a space that is presented only using voice. Furthermore, they address the need to quickly navigate to the desired information, and to support known-item search and focused-browsing.

4 Information Structures

In this section, we outline four high-level information structures we identified, how they can be realized in a voice user interface, and how they are related to the user level tasks mentioned above.

Top-Level Visual Regions: many web pages have a visual structure that divides the page into visual regions of related, but independent, content.

Voice interface: the titles or headings of these visual regions can become items in the top-level menu of a voice interface. For example, the corresponding voice interface for the page in Figure 1 would have as choices at the top level: Weather, and Headlines

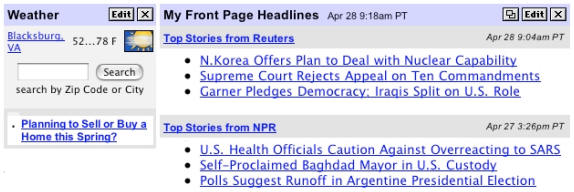

Figure 1: Part of a My Yahoo Page as an example of top-level visual regions on a web page

Menu/List Structures: many web pages present horizontal or vertical lists of information in a structured manner that yields a hierarchical or pseudo-hierarchical organization. In Figure 1 above, the Headlines region of the page is further decomposed into news categories (e.g. Top Stores from Reuters, Top Stores from NPR).

Voice interface: the menu/list structures together with the top-level regions of the page create a hierarchical menu structure that can be used to allow users to quickly and simply navigate to information. This supports our design goal of directed-browsing and known-item search by supporting the user task of *Navigate*. Such menus should afford the user a way to skip levels, a way to backup to previous levels, and should use incremental and tapering prompting techniques where applicable [11] (both techniques are examples of *Details-on-demand*).

To support the user task of *Situate*, the voice interface should also allow the user to query where on the page s/he currently is and what options are available at a particular time. It also affords returning to a known level (top-level) of the navigation space.

Text Areas: text areas on web pages are used to convey content information. Blocks of text may include structural information, such as subheadings, and often include hyperlinks embedded in the text.

Voice interface: the text content can be presented in a voice interface through the use of a text-to-speech synthesizer. However, this is not sufficient, as the interface should support the user's *Navigate* and *Situate* tasks. It may be appropriate to allow the user to have control of the speed of presentation (e.g. slow down, speed up), and control to move around the text (e.g. backup a sentence, skip ahead). These types of controls are typical of screen readers for sight-impaired users [7].

Users need to be made aware of hyperlinks embedded in a text area and provided a way to follow the links. Presentation of links using audio is a known problem [5]. None of the solutions presented in the literature seem to be error-free [9]. The presentation of a text area should allow the user to issue an explicit query (such as "links") to be read a list of the links present in the current text area. This would support the *Details-on-Demand* task and will help the user in the *Situate* task.

The text presentation affords other user commands including summarization and data extraction. Summarization supports the *Details-on-Demand* task by allowing to user to decide if s/he wants to hear the full story or not. Data extraction provides information nuggets, such as phone numbers, addresses, and emails embedded in the text passage. It supports the *Query* user task, which addresses our goal of known-item search.

Figure 2. Repeated/Structured Information

Repeated/Structured Information: this type of information structure is one that has a clear set of information fields that are repeated for different instances of the information. The structure resembles a database listing, with fields (data labels) and records

(repeated instances of information, often in groups). Figure 2 shows an example, a movie listing for a local movie theater. Each group has movie title, rating, movie times, and the cast of the movie.

Voice interface: there are several features that can be used to help voice navigate lists of repeated/structured information. Each "record" in the list could be read completely in sequential order. *Navigation* commands could be provided to allow users to move forward and backward in the list, including the ability to "barge-in" to issue a command. Additionally, users could be allowed to request specific items (*Query*, *Details-on-Demand*). For example, a user may wish to hear information about a specific movie. While the specific movie information is being read, the user could then say a field name such as "starting times" to hear only that part of the information (*Query*). Furthermore, the group could be treated as a menu, with the user moving from a specific movie listing to other movies in the list (*Navigate*).

5 Conclusions

Users will surf the web using phone-based voice user interfaces, however we do not believe that users will stop using their Web browser. Instead, we believe users will use telephone voice interfaces to perform known-item searches or focused-browsing.

In this paper we have presented a taxonomy of information structure by user task for voice navigation of web-spaces. In particular, the type of user tasks proposed here are intended to support the voice browsing mantra of "Situate, Navigate, Query, Details-on-demand" which we propose be used as a design guideline for voice interfaces to navigate the web. We have enumerated a number of user tasks based on the types of information structures available on the web.

We are currently evaluating how these user tasks work together to allow the user to carry out information seeking or focused browsing. We are also exploring ways to automatically transcode HTML content. This will open up more of the existing web information to access by voice navigation.

The work reported in this paper was funded in part by a grant from IBM to explore the use of VoiceXML within the WebSphere Transcoding Publisher.

References

1. Asakawa, C. and Takagi, H. (2000). Annotation-Based Transcoding for Nonvisual Web Access. *The Fourth International ACM Conference on Assistive Technologies ASSETS*, 172-179. ACM.

2. Danielsen, Peter J. (2000) The Promise of a Voice-Enabled Web. *IEEE Computer 33*(8): 104-106

3. Hopson, Nichelle. (2002). WebSphere Transcoding Publisher: HTML-to-VoiceXML Transcoder. IBM WebSphere Developer Domain Library, January 2002. Accessed on April 27, 2003 from: http://www7b.boulder.ibm.com/wsdd/library/techarticles/0201_hopson/0201_hopson.html

4. Hori, M., Kondoh, G., Ono, K., Hirose, S. and Singhal, S. (2000). Annotation-Based Web Content Transcoding. In *Proceedings of the 9th International World Wide Web Conference*. Amsterdam. May 15-19, 2000.

5. James, F. (1998). Lessons from Developing Audio HTML Interfaces. *ACM Conference on Assistive Technologies*. April 15-17, 1998, Marina del Rey, CA USA, pages 27-34.

6. Lucas, B. (2000) VoiceXML for Web-Based Distributed Conversational Applications, *Communications of the ACM 43*(9): 53-57.

7. Raman, T.V. (1996) Emacspeak—a speech interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* pp 66-71. Vancouver, British Columbia, Canada.

8. Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization. *Proceedings of the IEEE Symposium on Visual Languages*, September 3-6, 1996, pp. 336-343.

9. Wang, Q.Y., Shen, M.W., Shi, R.D. and Su, H. (2001). Detectability and Comprehensibility Study on Audio Hyperlinking Methods. *Proceedings of the Human-Computer Interaction - Interact '01*, pp. 310-317.

10. Williamson, C. and Shneiderman, B. (1992). The Dynamic HomeFinder: Evaluating Dynamic Queries in a Real-Estate Information Exploration System. *In Proc of ACM SIGIR 92 (June 1992),* 338 – 346.

11. Yankelovich, N. (1996). How do Users Know What To Say? *ACM Interactions*, V3, N6, Nov/Dec 1996.