Presenting HTML Structure in Audio: User Satisfaction with Audio Hypertext

Frankie James, *Stanford University*

Audio on the WWW

Every day, more information becomes available online as electronic documents. Since the advent of the World Wide Web (WWW), the medium of choice for electronic publishing has become hypertext—in particular, HTML. HTML allows the design of rich document structure, including tables, images, and hyperlinks, via a relatively simple command language.

To access electronic documents, blind computer users traditionally use ASCII text files. This method preserves the textual content of the document but cannot handle the visual content. Visual content is a fundamental part of any document. While some visual elements, such as pictures, are purely visual, other visual elements, such as tables or changes in type face that denote headings, indicate structure. By using markup tags, HTML explicitly represents both the the textual and the structural-visual content of a document. This representation is essential to understanding a document's overall structure and navigating between documents. Accessing the WWW, therefore, is neither as broad a goal as accessing general GUI, which disregards the particular application, nor as specific as accessing a particular application such as a phone answering system. The WWW is comprised of different document types that range from structured reports to fill-out forms, and there are many ways to provide access to it.

Audio HTML

Figure 1 represents the creation and use of a hypertext document in both the visual and auditory realms. If an audio document is designed straight from the author's intentions, it may correspond to the author making an explicit recording of the document or pieces of the document. While this is the most effective strategy for audio communication, it means that authors must create two versions, both audio and print, of each document.

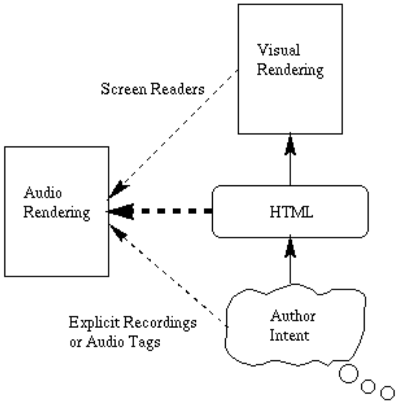

Figure 1: When to Create the Audio Representation

Another way to create audio documents is by working directly with visual representations, as screen readers do (Berkeley 1996; Edwards and Mynatt 1994). Although currently the solution to Web accessibility for blind users, this method has several weaknesses. By the time the document is presented visually, its explicit structural information has been made implicit. Recovering this structure is difficult, if not impossible. Screen readers, moreover, force blind users to interact spatially with documents. Many blind users lack grounding in spatial and visual metaphors, and interactive screens do not map well to speech or Braille output (Scadden 1996).

A final method of rendering an audio document is using the document's HTML representation. Although author intent is not always truly represented in HTML, [1] most of the visual elements important to navigation and structure are determined by markup tags. Audio renderings, consequently, can be designed from the markup tags instead of the visual representations. For example, headings can be identified with certainty by the tags, rather than guessed based on type size.

This research focuses on the audio presentation of HTML-tagged text originally designed for only visual use. We are studying ways to represent HTML structures in audio so that a browser dedicated to producing audio output will present most web pages so that blind people can use them. Emacspeak (Raman 1996) and the commercial product pwWebSpeak (Productivity Works) are also in this vein, and we have collaborated with their authors and analyzed their design choices. Our studies create a framework for understanding how to represent document structure in audio. Based on our studies, we are developing a set of guidelines (the Auditory HTML Access system, or "AHA") for designing audio interfaces to HTML.

The User Study

The pilot study compared several different audio presentation styles. The experiment was designed so that each of twenty-four paid subjects (twelve blind and twelve sighted) used the four interfaces in a random order, creating a two-by-four mixed design. All subjects had at least a working knowledge of the WWW and web browsing but not of any other audio-HTML access system.

Interface Design and Setup

In designing the interfaces for this experiment, we explored marking structures with both non-speech sound effects and speaker changes. The interfaces used in the experiment were based on four general formats:

- one speaker, few sound effects (OS/V)
- one speaker, many sound effects (OS/MS)
- multiple speakers, few sound effects (MS/V)
- multiple speakers, many sound effects (MS/MS)

OS/V used explicit linguistic cues (e.g., "level-one heading"), with the only sound effect being a tone indicating link points, which differed in pitch for followed and unfollowed links. OS/MS used various sound effects, including overlaid natural sounds (e.g., footsteps for within-document links), bracketed pure tones (to mark heading levels), and short natural sounds. MS/V used speaker changes to mark structures, plus a short beep following links to mark anchor points. MS/MS, finally, used a mixture of voice changes and sounds such as those described in OS/MS. For more details, see "Presenting HTML Structure in Audio: User Satisfaction with Audio Hypertext" (James 1996).

We designed the experiment using a "Wizard of Oz" format to test different interfaces without implementing an HTML parser. The interface consisted of recorded[2] speech[3] and sounds[4] in Hypercard running on a Macintosh. All eight HTML pages used relate to Project Archimedes and CSLI at Stanford University[5] and, thus, relate to this project. More importantly, these pages contain intricate interlinking and represent a variety of page types found on the WWW.

The interfaces were designed for non-visual use and, consequently, are keyboard controlled. Assuming that users would need many ways to navigate through a document--to mimic visual skimming--we made controls for jumping between headings, lists, etc. We selected sound effects based on auditory icons (Gaver 1986), choosing sounds that intuitively related to the structural elements that they would represent. If there was no obvious choice, we used a short, abstract sound.

To listen to sample sounds used in the study, visit *http://www-pcd.stanford.edu/frankie/pilot/*

Tasks

The tasks on the task sheets given to subjects fall under three categories:

Locating Information: To eliminate the effects of memorization or prior knowledge, subjects were asked to perform tasks such as finding the mention of the CSLI webmaster.

Answering specific questions: To focus subjects on the pages and to test whether they could retrieve information by following links, finding appropriate sections, etc., subjects answered content questions.

Describing document structure: To test the clarity of the structuring techniques in the interfaces, subjects were asked to reproduce or describe the structural elements on a page

The task sheets featured four tasks from at least two categories.[6] To be tested with each interface, the task sheets were given consistently in the same order.

User Satisfaction

User satisfaction was tested using questionnaires concerning the usefulness or appropriateness of the marking techniques. The question formats were Likert scales (e.g., a five-point scale ranging from very good to poor) and free-response. Subjects had unlimited response time. Recorded responses included written and spoken comments gathered from the questionnaires and videotapes.

Results

The results from this experiment are presented best according to the main structure types found in HTML documents. This paper notes our main findings; for a more detailed discussion, refer to "Presenting HTML in Audio: User Satisfaction with Audio Hypertext" (James 1996).[7]

Headings

The explicit markers in OS/V eased the identification of headings and their types for the users in the study, who were basically novices to audio web browsing. These results may not extend to more experienced users, however. In fact, seven subjects said that the explicit tag was too long or cluttered the presentation. Several added that, if more experienced, they would prefer a non-verbal tag. Subjects had trouble distinguishing heading levels differentiated by pitch in OS/MS and MS/MS, even when two headings sounded in succession. This result is supported by studies showing that non-musicians have difficulty distinguishing between tones that differ by pitch alone (Pitt and Edwards 1991; Portigal 1994).

Link Points (Anchors)

When subjects rated the usefulness of meta-information associated with links, they rated the interfaces that marked links with natural sounds (OS/MS and MS/MS) as significantly more effective than the other interfaces that used simple tones (OS/V and MS/V). Subjects had difficulty discerning the extent of the link text in the verbose protocols because the tone in the verbose protocols sounded either at the beginning or the end of the link text. Subjects also had trouble reacting in time to follow the link. OS/V also indicated whether links had already been followed by a pitch change in the link tone, but this slight change was difficult for subjects to perceive and not rated highly.

Lists

The list-recognition task produced different results for blind and sighted users. Blind OS/MS users had significantly more correct responses than did other blind users, even though OS/MS used a list bell that was rated as both too loud and too slow. Presumably, this overbearing cue forced users to focus on the list structure more than the cues in the other interfaces did. Within the sighted population, MS/V users generated significantly more correct responses than OS/V users. MS/V both separated list levels by speaker and marked list items with a short audio bullet, whereas the distinction between list levels was not clear in OS/V.

Pauses

The most significant result in the area of pauses is the comparison of blind and sighted subjects. Blind subjects, who are accustomed to using audio to retrieve information, found the pauses too long and the presentation too slow. Sighted subjects, who have little experience with audio computer interfaces, on the other hand, found the pauses too short and the presentation too fast.

Volume

The overall volume rating showed that OS/MS was significantly louder than the other interfaces, and that MS/MS was significantly louder than OS/V and MS/V. Because the users had full control over the main volume, they most likely found certain sound effects too loud in comparison with the rest of the interface and, thus, distracting.

Overall Rating

Many users chose a favorite interface in their general comments. Although the totals were not significant, ten subjects chose one speaker with minimal sound effects (OS/V), probably due to the explicit tags discussed in the headings section. Multiple speakers with minimal sound effects (MS/V) received five "best" votes and comments like "this interface seemed more friendly". The two interfaces using multiple sound effects each got three "best" votes. Task analysis yielded across-interface significance (with MS/V rated highest), but post-hoc tests found no pairwise significance.

General Conclusions

Several general concepts for using audio HTML resulted from this study and are discussed briefly below.

Novice Users

This study confirmed our intuition that novice users of a system, in this case a system for accessing HTML using audio, prefer information presented very explicitly within an interface. Even though users were given a sheet describing the audio markings used in each interface, many still did not remember or understand all the sounds and voices. They preferred, at least in this stage of their experience, the interface that explicitly stated the identity of the various structures. Subjects commented that if they had more experience with the system, they might prefer having sound effects or voice changes because such effects would cut the presentation time. To address this issue, further tests will study users who work with the system for the period of a few days.

Relative Sound Changes

Relative sound changes are difficult for users to distinguish. This effect was evident for almost all the major HTML structure types. For example, users disliked the use of relative pitches to indicate heading level or followed versus unfollowed links, longer pauses to indicate paragraph boundaries, and volume changes to indicate bold text and list level nesting.

Users found natural sounds more distinguishable and easier to remember than sounds that differed relatively. Presumably, even sounds which are less natural such as earcons (Blattner, et al., 1990) but differ non-relatively (e.g., by melody rather than by pitch or volume) would be more effective than relative sound changes.

Recognizable Sounds

Although our findings indicated that some of the natural sounds used did not suggest their meanings effectively, subjects generally reacted to them more favorably than to artificial sounds such as beeps. Even poorly chosen natural sounds seemed easier to use than simple tones, perhaps because of their distinguishability. Gibson (1966) points out that "[m]eaningful sounds vary in much more elaborate ways than merely in pitch, loudness and duration." Such elaborate differences enhance distinguishability and memorability in auditory interfaces.

Distinguishability also extends to other sounds not classified as "natural," such as musical themes or sounds associated with popular culture.[8] In *Auditory Scene Analysis: The Perceptual Organization of Sound* (1994), Bregman discusses the fact that a familiar melody is heard more easily out of a sound mixture than an unfamiliar one. He suggests that people listen for familiar sounds using a "schema-driven attentional process." Therefore, if the sounds in an audio interface are chosen because they are familiar and distinguishable, users should find them easier to hear, recognize, and ultlimately associate with HTML-document structures.

Speaker Changes

Another significant finding is that speaker changes can effectively indicate structure. Speaker change is used in radio, for example, to present structure, but little research explores the use of similar changes in computer interfaces to present structure. This study showed that a speaker change effectively signals a macro-level structure such as a heading (in MS/MS) or a level of list nesting (in MS/V). The study also demonstrates that voice changes as indications of micro-level structures tend to be distracting. The use of three different speakers to indicate heading levels in MS/V, the use of two alternating speakers to separate list items in MS/MS, and the use of a separate speaker to present bold text in MS/V evoked unfavorable comments from users.

Clearly, macro-structures, such as a nested list or address text, create sections that are separable from the rest of the document. Using a voice change to mark these sections fits our natural expectations of hearing a new speaker. We expect the new speaker to add to the discussion but to express a thought separate from the previous speaker. Contradicting this expectation is using a new speaker to mark microstructures; people do not naturally merge the words of two speakers into a single sentence or thought.

Future Work

This study indicates that 1) certain sounds and certain types of sound changes are more effective in presenting HTML structures than others and that 2) speaker change can mark certain kinds of document structures effectively. In addition to determining which specific sounds are effective in audio HTML interfaces, this study has focused our attention on the HTML itself and the inferences we can make about the HTML author's intentions based on markup tags. More research is needed to determine what tags could be added or changed in HTML to help web-page authors more explicitly express their design intentions.

Our future plans include a study of the usefulness of sound markings and voice changes for more experienced users of audio interfaces. We also intend to produce a proposal containing guidelines for augmenting HTML so that web authors who seek to provide consistent, structured documents can ensure that their documents are audio accessible.

References

Berkeley Systems, Inc. outSPOKEN. Available at *http://access.berksys.com/*

Blattner, Meera M. et al. (1990). Earcons and icons: Their structure and common design principles. In Ephraim P. Glinert (Ed.), *Visual Programming Environments: Applications and Issues* (pp. 582-606). Los Alamitos, CA: IEEE Computer Society Press.

Bregman, Albert S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Edwards, W. Keith & Mynatt, Elizabeth. (1994). An architecture for transforming graphical interfaces. In *Proceedings of the ACM:*

UIST '94 (pp. 39-47). New York: ACM Press.

Gaver, William W. (1986). Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction*, 2, 167-177.

Gibson, J.J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

James, Frankie. (1996). Presenting HTML structure in audio: User satisfaction with audio hypertext. Stanford University Digital Libraries Working Paper, SIDL-WP-1996-0046.

Pitt, Ian J. & Edwards, Alistair D.N. (1991) Navigating the interface by sound for blind users. In D. Diaper and N. Hammond (Eds.), *People and Computers VI: Proceedings of the HCI '91 Conference* (pp. 373-383). Cambridge, UK: Cambridge University Press.

Portigal, Steve. (1994). Auralization of document structure. Master's thesis, University of Guelph.

Raman, T.V. (1996). Emacspeak-direct speech access. In *ASSETS '96: Second Annual ACM Conference on Assistive Technologies* (pp. 32-36). New York: ACM SIGCAPH, ACM.

Reeves, Byron & Nass, Clifford. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York: Cambridge University Press.

Scadden, Lawrence A. (1996). Blindness in the information age: Equality or irony? *Journal of Visual Impairment and Blindness*, November 1984, 394-400.

Productivity Works. pwWebSpeak, 1996. Available at *http://www.prodworks.com/pwwebspk.htm*

Footnotes

[1] Because of its limitations, HTML's tags are often used "creatively" to produce visual effects desired by authors.

[2] Special thanks to Dave Barker-Plummer, Andrew Beers, Mark Greaves, Stephanie Hogue, Claire James, Connie James, and Dick James for providing the recorded speech.

[3] Although it is important to understand what effect less natural sounding voices have on the users of an audio browser [11], this study is focused on differentiating between voices. Less natural sounding voices could confound any related results.

[4] Sound effects were obtained from freeware libraries or were recorded via SoundEdit Pro using ordinary household objects.

[5] The pages used in this experiment can be found at http://www-pcd.stanford.edu/~fjames/testpages/

[6] Task Set 2 did not contain a document structure task.

[7] Results were obtained by analyzing raw scores of the scales using a repeated measures ANOVA model. Statistical significance of the pairwise comparisons was based on post-hoc tests, including Student-Newman-Keuls, Tukey hsd, and Scheffi.

[8] Most people would not call the Star Trek communicator sound "natural", but it is easily recognizable by any Trekker.

Author Information

Frankie James

Computer Science Department
Stanford University
fjames@cs.stanford.edu