Presenting HTML Structure in Audio: User Satisfaction with Audio Hypertext

Frankie James
fjames@cs.stanford.edu
Computer Science Department, Stanford University

1.0 Introduction

Audio interfaces are becoming more prevalent in the world of computers. This is due to several factors. First, computers are being used to control many specialized devices and applications which are intuitively grounded in audio, such as voice messaging systems. Systems such as Phone Slave [17] take advantage of computer technology to provide a more intelligent way to leave and retrieve messages. The answering machine functions map well into conversational audio. Another factor leading to the need for more audio interfaces is the availability of PDAs. The screen on a typical PDA cannot accommodate a real GUI without causing eye strain. Adding audio output can make using a PDA as an internet terminal, word processor, or spreadsheet easier.

There is another reason why audio interfaces should be more widely available. 11 million Americans have some form of visual impairment, and about 1.5 million are totally blind. [10] The condition of blindness can be seen as an extreme case of the PDA problem mentioned above. While PDAs have limited screen real estate, to a blind user, every computer has *zero* screen real estate. Audio interfaces are needed to make them usable at all.

Current research in the area of interfaces for blind users focuses mainly on screen reader technology, which confronts the problem of GUI access by creating an auditory interface from the visual representation of the applications on the computer's screen. Audio is treated as a second-class interface modality, subordinate to the visual interface. To create more viable access solutions for blind users, designers must begin to treat audio as a first-class interface technology (see AsTeR [13], Emacspeak [14], and WebSpeak [18]).

2.0 Audio on the WWW

Every day, more information is being made available online in the form of electronic documents. Since the advent of the World Wide Web (WWW), hypertext (in particular, HTML) has become the medium of choice for the presentation of these documents. This is because HTML allows for the design of rich document structure, including tables, images, and hyperlinks, via a relatively simple command language.

Traditionally, blind computer users have accessed electronic documents through ASCII text files. This method is able to preserve the *textual content* of the document, but has problems when dealing with the *visual content* such as tables and figures. The visual content, which can be further subdivided into those visual elements which are used to indicate structure (like tables and the use of type face or style to denote headings) and those which are purely visual (such as pictures), is a fundamental part of any document. By using markup tags, HTML explicitly represents the structural visual content of a document as well as the textual content. This content is needed to get a sense of the document's overall structure, and also for navigation between documents.

Accessing the WWW is neither as wide-reaching a goal as GUI access (via a screen reader or Mercator [4]), nor is it as narrow as accessing a specific application such as a phone answering system. The WWW is comprised of different document types which range from structured reports to fill-out forms, and there are many ways to provide access to it.

2.1 Audio HTML

Figure 1 gives a representation of the creation 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 seems like the best strategy, it means that authors must create two documents for everything that they write: one in audio and one in the visual domain.

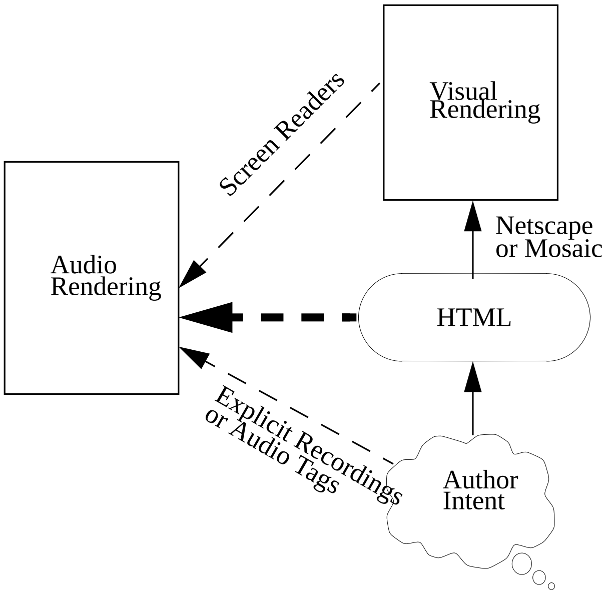

FIGURE 1. When to create the audio representation

Another way to create audio documents is by working directly from the visual representation, which is what screen readers do. [7] This is the current solution to Web accessibility for blind users. However, by the time the document has been presented visually, the explicit structural information in it has been made implicit. Recovering this structure is difficult, if not impossible. Screen readers also force blind users to interact with the documents spatially with documents since they are based on visual representations. Unfortunately, many blind users lack grounding in spatial and visual metaphors, and interactive screens do not map well to speech or Braille output. [16]

Finally, an audio rendering can be designed from the HTML representation of the document. Although the author's intent is not always truly represented in HTML[1], most of the visual elements important to navigation and structure are determined directly from the markup tags. This means that audio renderings can also be designed from the markup instead of by trying to determine the significance of visual elements. For example, headings can be identified by the tags <H1>...<H5>, rather than by guessing based on type size.

2.2 Creating an Audio Browser

A commercial audio WWW browser (on the order of Netscape) would need to provide the following features:

- Speed Control
- Search
- Access to fill-out forms, email, and news
- Audio presentation of tagged HTML text

Some of these topics have already been addressed in previous research, such as Barry Arons' work on SpeechSkimmer [1] to control audio presentation rate. Other topics are important in a good commercial system, but are not being addressed in research, such as fill-out forms (see Henter-Joyce's Jaws for Windows [9] for a good example of forms access in a screen reader) and search.

This research focuses on the final item, the audio presentation of the tagged text. We are studying the design of ways to represent HTML structures in audio so that a browser which is dedicated to producing audio output will present most web pages in a way that is usable by blind people. The commercial product pwWebSpeak [18] and Emacspeak [14] are also in this vein, but neither has been formally studied. We plan to use our studies to create a framework for understanding how to represent document structure in audio and a set of guidelines for designing audio interfaces to HTML (called the Auditory HTML Access system, or "AHA").

3.0 The User Study

3.1 Interface Design and Setup

In designing the interfaces for this experiment, we explored the use of both non-speech sound effects and speaker changes to mark structure. 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)

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 2 by 4 mixed design. All subjects had at least a working knowledge of the WWW and web browsing.

For each interface, subjects had twelve minutes to try to perform the tasks on the task sheet, then they filled out a user-satisfaction questionnaire and moved on to the next

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

interface. Subjects were videotaped so that the author could later review the tapes and procedures.[2]

The sound effects in the interfaces were selected with the idea of auditory icons in mind. [5] An effort was made to choose sounds that seemed intuitively related to the structural element they were meant to represent. If there was no obvious sound, a short abstract effect was used. The choice of where to change speaker in the MS/V and MS/MS protocols was inspired by Geiselman and Crawley's Voice Connotation Hypothesis. [6] We used the analogy of a sports broadcast in which there is more than one announcer, each of whom has a specific role and presents only certain information. For example, in MS/MS, there is a "heading commentator" who only presents heading text, much like a color commentator in a hockey game only presents player statistics and analysis rather than play-by-play action. The utility and user reactions to the sound effects and voice changes are presented in the "Results and Discussion" section on page 3.

3.2 General Design

We designed the experiment using a "Wizard of Oz" format so we could test different interfaces without having to implement an HTML parser. The interface consisted of recorded[3] speech[4] and sound effects[5] in Hypercard running on a Macintosh.

The HTML pages used in the experiment were eight pages which are related to the Archimedes Project and CSLI at Stanford University.[6] The pages were chosen because they are substantially interlinked and represent a variety of page types that can be found on the WWW, and because they are related to this project.

The interfaces were designed to be used non-visually, and thus completely controllable via the keyboard. We assumed that users would need to have many ways to move around in a document to mimic visual skimming, so controls were made for jumping between headings, lists, etc. The decision was made not to *not* include controls for jumping between link points; this decision and its implications will be addressed in the "Results and Discussion" section on page 3.

3.3 Tasks

The tasks on the task sheets can be divided into three categories:

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

Answering specific questions
Content questions were used to focus subjects on the pages and to see if they could retrieve information by following links, finding appropriate sections, etc.

Describing document structure
These questions asked the subjects to reproduce or describe the structural elements on a page to see if the structuring techniques in the interfaces were usable.

Task sheets contained four tasks selected from at least two categories[7]. The task sheets were always given in the same order so they could be tested with each interface.

3.4 User Satisfaction

The user satisfaction questionnaires consisted of fifteen to seventeen questions about the usefulness or appropriateness of the various marking techniques. The question formats were Likert scales (e.g., a five-point scale ranging from very good to poor) and free-response. Subjects were given unlimited response time. Recorded responses included written and spoken comments gathered from the questionnaires and videotapes.

4.0 Results and Discussion

The results from this experiment can be divided in terms of the main structure types that are available for marking in an HTML document. 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 Scheffé.

4.1 Headings

The four methods of presenting headings were analyzed by two Likert scale questions (question 6 and question 7) and one free-response question (question 8). The ANOVA model revealed high significance between blind and

[2] In order to prevent skewing of the results, the author did not conduct the experiment herself. The experimenter was Andrew Einaudi.

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

[4] Although it will be important to understand what effect less natural sounding voices will have on the users of an audio browser [15], the current study is focused on differentiating *between* voices. A less natural sounding set of voices could confound any related results.

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

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

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

sighted users (with blind users reacting more favorably to the headings than sighted) and also across interfaces. Post-hoc analyses revealed that in the question about distinguishing between heading levels, OS/V was rated significantly higher than the other three interfaces. The two Likert scale questions were then combined additively[8] to produce an overall heading rating. Again, there was significance between blind and sighted users and across interface. Post-hoc analysis showed that OS/V was rated significantly higher than the other three interfaces.

These results suggest that the explicit nature of the heading markers in OS/V made it easier for the subjects to distinguish headings and their types, but since the users were basically novices (they only used the interface for twelve minutes), results may not reflect the preferences of more experienced users. In fact, seven subjects commented that the explicit tag took too long or made the presentation seem cluttered, and several said that if they were more experienced, they would prefer a non-verbal marking.

The OS/MS and MS/MS protocols, which scored the lowest for the four interfaces, were apparently disliked because of the use of a relative tone as a marker for the heading level. Subjects commented that it was hard to distinguish the relative tones and decide what heading level was being presented, even when two headings were heard in succession. This is supported by studies which show that non-musicians have limited ability to distinguish between tones which differ only by pitch.[11][12]

As far as the use of multiple speakers to present heading levels is concerned, users rated MS/V, which used a separate speaker for each heading level, the lowest. This may be due to the fact that although the voices were distinguishable, it was not apparent which voice would stand for which heading level. In the MS/MS protocol which used one speaker for headings plus a tone which varied in pitch to indicate level, five subjects commented that the change of speaker to indicate headings was helpful. Therefore, the use of more than one speaker for heading levels seems to have caused confusion rather than clearly indicating heading text.

4.2 Link Points

Link points were rated in questions 1-4. There was no significance across interface in the subjects' ability to distinguish links from other text, but again, blind users rated the distinguishability of links significantly higher than did sighted users. In question 3 concerning the difference between link types, there was basically no difference at all between OS/MS and MS/MS. This was expected since the two interfaces presented links in the same way.

In question 2, subjects were asked to rate the usefulness of the meta-information associated with the links. The sound effect interfaces (OS/MS and MS/MS) were perceived as significantly better than the verbose interfaces (OS/V and MS/V). This indicates that the meta-information about link type found in the sound effect protocols was useful for the subjects, which is also confirmed by their written comments. Users stated that having several different natural sounds made it easy to distinguish between the link types, even though the mappings between the chosen sounds and their meanings was not always intuitive.

The other available meta-information was that in OS/V which indicated whether or not the link had already been followed. One of the blind subjects said that he liked having this information since sighted people have it in Netscape, but it was not rated very highly on question 2. This may again be due to the fact that the difference between followed and unfollowed links was indicated by a relative pitch change in the marking tone which was difficult for subjects to pick out. Another explanation is that the subjects in this test, although they had experience with web browsing, may not have been experienced enough to know that the link color in Netscape indicates whether or not the link has been followed.

Users also commented on interfaces OS/V and MS/V, which used a tone before and after the link point respectively. Many said that it was difficult to tell the extent of the link text and to react in time to follow the link. Since the beep in both of these cases came at one end of the link text or the other, the anchor text was not as clearly delimited as it was when accompanied by a background sound (as in OS/MS and MS/MS). A few users also said that the beeps sounded like the error tone.[9] Additionally, the use of a following beep in MS/V evoked responses such as "it sounds like the announcer says dirty words that get bleeped out." Clearly, there is a social impact to using beeps of different types (and in different positions) which should be addressed by interface designers using audio.

As we mentioned before, the keyboard interface allowed users to jump between certain kinds of structures, such as headings, lists, and graphical images. However, we did not provide the functionality to jump between anchor points. Several blind users found this to be annoying, since the screen readers and WWW access methods they were used to using allowed this functionality. In a non-

[8] Factor analysis of this combination yielded a Cronbach's alpha of .7297

[9] A non-sine wave tone was used in this experiment to indicate errors.

experimental system, such functionality should probably be added.

4.3 Lists

Although there were no specific questions regarding list presentation in the questionnaires, many users commented on the presentation of lists in each interface. The list bell used in the OS/MS and MS/MS protocols was marked as being too loud by most users (see discussion in Section 4.4, “Volume”). Users also said that the bell sound in OS/MS was too slow, since you had to wait for it to ring three times at the top and twice at the bottom of each list. However, blind OS/MS users had significantly more correct responses than other blind users in the task dealing with list recognition, so this presentation style cannot be completely invalid.

Sighted MS/V users had significantly more correct responses than sighted OS/V users. This could perhaps be due to the dual cue of a list bell plus a different speaker to indicate list nesting. This makes sense when we think about how lists are structured in the visual domain. Lists are cued to the visual user by both a bullet or number of some type and an indentation that indicates the list level. Even if the bullet is unchanged between list levels (or left out completely), the indentation is enough for users to understand that this is a nested list. In MS/V, the list bullet remains the same but the other list cue (in this case speaker change instead of indentation) is enough for the users to understand the nested list structure.

On the other hand, the use of alternating speakers in list presentation in MS/MS was commented on by subjects as being “annoying” and “useless”. This is because it does not provide any structural information besides a separation of adjacent list items, which is redundant since this separation is also indicated by an audio bullet. The analogous situation to this in the visual would be to have list items represented in two different (alternating) colors. This method distracts from the actual list structure, instead calling attention to the separation between list items which is already marked by a bullet or number.

4.3.1 Pauses

The most significant result in the area of pauses was between blind and sighted subjects. Blind subjects, who are more used to using audio to get information, found the pauses too long in general and wanted to speed up the presentation. On the other hand, sighted subjects, who have little experience with audio computer interfaces, found the pauses to be too short and the presentation too fast.

In question 11, all subjects found it difficult to distinguish between different pause types. This has a couple of explanations. First, Hypercard is inherently asynchronous, so the pause lengths may be inconsistent from one usage to the next. Therefore, it is difficult to control pause length well enough to produce a noticeable difference. Secondly, the pause change is again a relative change between two markings, which may be hard for novice users to notice.

4.4 Volume

The overall volume rating (in question 13) revealed a significant difference between OS/MS and the other three interfaces. Further analysis showed that this significance can be accounted for by the sound effect used to mark lists in the interface. The bell used in OS/MS was much too loud in comparison to other sounds in the interface and proved to be distracting. Analysis of question 14 also revealed that MS/MS was significantly louder from OS/V and MS/V. This may also be due to the unusually loud bell used to mark lists.

The free-response questions on volume also revealed that users had problems with the use of relative volume, such as the volume difference between speakers (because of indication of bold text or simply because of general loudness of voice) and the volume difference of the list bell in MS/MS to indicate list level.

4.5 Overall Rating

The question corresponding to overall rating (question 16) produced inconclusive results, due to its vague nature. However, in the general comments question, many users picked a favorite interface. Although the totals were not significant, ten subjects chose OS/V, probably due to the explicit tags discussed in 4.1. MS/V received five “best” votes and comments like “this interface seemed more friendly.” OS/MS and MS/MS each got three “best” votes.

Task analysis yielded across-interface significance[10] (with MS/V rated highest), but post-hoc tests found no pairwise significance. We are encouraged by the apparent usefulness of multiple voices to convey structure in audio HTML.

5.0 Conclusions

There are many specific conclusions that can be made regarding this experiment, such as that certain of the sound effects should have been softer, etc. However, there were

[10] Significance was found across interfaces (p = .044) when using squared results.

also several general concepts that we learned as a result of this experiment, which are described below.

5.1 Novice Users

Although it may not be a novel concept, this study revealed that novice users of a system, in this case a system for accessing HTML using audio, like to have information presented very explicitly within the interface. For each interface, users were given a sheet describing the sound effects and voices used to present the structures, but even this was not enough for many of them to feel comfortable with the interface and to remember and understand all of the sounds and voices. What they preferred, at least in this stage of their experience with the system, was the interface that explicitly said what structures were what.

As mentioned before, subjects did comment that if they were more experienced with the system, they might prefer having sound effects or voice changes since it would cut down on the presentation time. To deal with these changes over time, further testing in the form of a more longitudinal study needs to be performed. The next round of experimentation will be more of this nature so that we can analyze feedback for users who have worked with the system for the period of a few days.

5.2 Relative Changes

A major finding of this study is that relative sound changes are difficult for users to distinguish in a useful way. This factor cut across almost all of the major HTML structure types. For example, the use of a relative pitch change to indicate heading level or followed versus unfollowed links, the use of longer pauses to indicate paragraph boundaries, and the use of volume changes to indicate bold text and list level nesting were all commented on in a negative way.

To design an audio HTML formatting system which is intuitive to users, relative changes such as these will need to be avoided. Users found natural sounds more distinguishable (even if their mappings were unintuitive) and easier to learn and remember than sounds which differed relatively. Presumably, even the use of less natural sounds such as earcons [2] which differed in non-relative ways (e.g., by melody rather than by pitch, volume, or instrument) would be more effective than the relative changes.

5.3 Recognizable Sounds

Although our findings indicated that some of the natural sounds used in the interfaces did not readily suggest their meanings, in general, subjects reacted to the natural sounds more favorably than to artificial sounds such as beeps. That is, even poorly chosen natural sounds are easier to use than simple tones because of their distinguishability. Gibson [8] points out that “[m]eaningful sounds vary in much more elaborate ways than merely in pitch, loudness and duration”. These elaborate differences enhance their distinguishability and memorability in auditory interfaces.

Distinguishability can also extend to other sounds which are generally not classified as “natural,” such as musical themes or sounds associated with popular culture.[11] In [3], Bregman discusses the fact that a learned (or familiar) melody can be more easily heard out of a sound mixture than an unfamiliar one. He suggests that the subjects listen for the familiar melody using a “schema-driven attentional process.”[12] It is the familiarity of the sound and the subject’s prior experience that allows her to segregate a sound mixture and hear the target melody.

These main ideas of distinguishability and prior experience of the user to particular sounds are important when choosing sound effects for use in an audio interface. If sounds are chosen because they are familiar and distinguishable, users should find it easier to hear and recognize the sounds, which will in turn make it easier to associate these sounds with HTML document structures.

5.4 Speaker Changes

Another significant finding is that the use of speaker changes to indicate structure can be effective in certain circumstances. Speaker change is used (for example) in radio broadcasts to present structure, but little research has been done on the use of similar changes in computer interfaces to do the same. This study showed that when a speaker change is used to indicate a macro-level structure such as headings as opposed to text (in MS/MS) and a level of list nesting (in MS/V), it can be an effective tool.

The study also showed that voice changes which are made to indicate micro-level structures tend to be ineffective and 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 was ineffective and evoked unfavorable comments from users. These cases were extreme uses of voice change and proved to be more of a hindrance in the presentation of HTML.

Clearly, macro-structures such as a nested list or address text create sections which are separable from the rest of

[11] For example, most people would not consider the “communicator sound” in Star Trek to be a natural sound, but it is easily recognizable by any Trekker.

[12] See discussion in [3] on page 411.

the document. Using a voice change to mark these is intuitive because of our human expectations when we hear someone new begin to speak. We expect the new speaker to add to the discussion, but to be expressing a thought separate from the previous speaker. This is in direct opposition to marking micro-structures such as bold text using a new speaker, since intuitively we do not merge the words of two people into a single sentence or thought.

6.0 Restructuring HTML

In addition to learning more about the specific sounds which are effective in audio HTML interfaces, this study has also caused us to focus our attention on HTML itself and to think about what inferences we can make about the HTML author's intentions given her use of markup tags. For example, the <HR> tag has no document semantics, but makes a horizontal rule. In general practice, it is used to separate two document sections from each other. Therefore, we can infer that whenever a horizontal rule is used, it is a signal for a section change. Similarly, a list which contains no list items or a description list containing no titles is almost always used to simply indent text in a document. Inferences that we make such as these allow us to create a richer presentation of the HTML document's structure in audio.

While the tags mentioned above are *generally* used in these ways, they are sometimes used differently to produce unexpected results. For example, a horizontal rule of a specified (short) width may be used to underline some text or outline an image. However, if we tag these structures using a non-speech effect which may *suggest* a semantics related to the tag without *explicitly saying* what those semantics are, the user should not have too many cognitive mapping problems. That is, if we mark each horizontal rule with the words "section break," we may confuse the user in the cases where the horizontal rule is used for a different purpose. But if horizontal rules are marked by a sound effect or voice change, the breaking semantics are suggested but not explicitly stated. In this case, the user can recognize a creative use of the tag without being given a false interpretation by the interface mechanism, i.e., that this is a section break.

There are some tags in HTML which are very ambiguous and are used in many different ways. The address tag is a good example of this. Address tags are used for everything from setting apart the webmaster's address on an HTML page to creating an acknowledgments section. The only real consistency in its usage seems to be that address text is usually meant to distinguish meta-information about the document (or document section) from the rest of the text. The table tag is also used in many different ways which are clearly in opposition to its original intention, such as providing a 2-dimensional framework for HTML authors to use for formatting text.

Tags such as these are good examples of how HTML tags were created to serve one purpose, but were later used to fulfill another because of how they are represented in the popular browsers such as Netscape. If the address tag, for example, was formatted by Netscape to produce an outlined box with an icon of a stamp in the upper right corner and the text in the middle, it would likely not have come to be used for anything but contact information for the HTML document.

There are also a number of tags and tag attributes which are under consideration in the HTML working group as additions to HTML which would be useful for audio interfaces. The adoption of the "REL" and "REV" attributes on link and anchor tags, which has been proposed for inclusion in HTML 3.0, would be helpful in determining link type and link semantics. Currently, the link type can be approximated by examining the target URL (as in OS/MS and MS/MS) and determining whether the link remains within the current document or targets a different document. The REL and REV attributes allow for the specification of things like the author of a document and the next or previous documents in a series. A browser could make use of this information to allow users to send email to a page's author or to browse a multi-document collection in an author-specified order. While some of these functions, like navigation order, can now be approximated by adding a row of icons at the bottom of a document, these icons often have no ALT text. The use of REL and REV attributes would therefore benefit both blind and sighted users, since the semantics are clear and the presentation mode can be chosen by the individual browser.

Another HTML 3.0 proposal is the use of style sheets to format documents. With style sheets, documents can be formatted according to the whims of the author ("I want all heading 3 text to be blue on a red background, in Times-Roman font") while maintaining the explicit structure of the HTML. Today, if the author wants all of the headings to be red, he simply tags the text as large and red rather than as a heading. Style sheets would obviate the need for authors to use tags which change the text style directly and would therefore make it easier to present HTML in non-visual media.

HTML as it exists today contains many examples of tags that are commonly used in ways which are different from the way in which they were originally intended, especially in the "informal" web pages such as individual home pages, sites for fan clubs, etc. Causing widespread changes in HTML authoring practice would be impossible in this group of web users, since their main goal is to design creative and visually interesting web sites which take little advantage of the ability to explicitly represent structures in HTML. (Notice how many of these sites contain the line “Enhanced for Netscape,” suggesting that the tags used are specifically designed for the visual formatting provided by Netscape.)

On the other end of the spectrum of web authors, however, are those who create web sites for businesses (based on existing intranets) or universities. These sites can house huge collections of documents which have been converted to web formats from legacy formats based either electronically or in paper. These collections typically use the same “look-and-feel” for all of the documents which they contain and are highly structured into hierarchies and other orderings. Convincing institutions or universities to use HTML tags more consistently could help them to maintain the structure and look-and-feel of their collections on the web, and would also be enforceable by the employment structure of the institution.

It is authors closer to this second group who design web pages that are most easily representable in alternative media such as sound. Since structure and consistency is so important, the idea of using logical tags such as <H2> and then deciding later how such headings should be formatted seems more natural and better fits their goals than it does for authors whose main goal is to be creative or shocking. Therefore, the idea of allowing blind users to browse the web in audio is not far-fetched, because much of the “important” information on the web (such as government and educational pages) is of this highly-structured, “clean” genre. The non-visual user will admittedly still run into problems when accessing the “fun” pages, but hopefully the audio techniques recommended by AHA will be robust enough to provide at least some level of access.

7.0 Future Work

This study clearly indicates that certain sounds and certain types of sound changes are more effective in presenting HTML structures than others, and also that speaker change can provide an effective way to mark certain kinds of document structures. Our future plans include a more focused study to understand the usefulness of sound markings and voice changes for more experienced users of audio interfaces. We also hope to provide a list of suggestions for the improvement of HTML so that authors whose main goal is to provide consistent and structured documents can do so in a manner that also makes it easier to access these documents in audio.

8.0 Appendix: Structural Elements in the Interfaces

8.1 Headings

OS/V
“level x heading”

OS/MS
tone preceding and following heading text, varying from high to low pitch for heading levels H1 down to H3

MS/V
different speakers read each level of headings

MS/MS
headings read by one speaker, tones also used as in OS/MS

8.2 Link Points

8.2.1 Previously followed links

OS/V
Preceded by a short, low-pitched beep

OS/MS
See Table 1 for various link sounds

MS/V
Short beep presented after link text

MS/MS
See Table 1 for various link sounds

8.2.2 Not previously followed links

All of the interfaces except for OS/V used the same marking as for links that had already been followed. OS/V in this case preceded the link text with a short, high-pitched beep.

Link Type	Sound Effect
different location within document	footsteps
another HTML file	telephone ringing
mailto link	doorbell

TABLE 1. Link Sounds for OS/MS and MS/MS Protocols

8.3 Lists

8.3.1 Unordered Lists

OS/V
"the following is an unordered list"..."end of list". Each list item preceded by "first item" or "next item".

OS/MS
List preceded by an audio bullet (bell) played three time, followed by bell played twice. Each list item preceded by one bell.

MS/V
Audio bullets (ding) precede each list item. Each level of list nesting read by a different speaker.

MS/MS
Two speakers alternate reading list items. List bullets are used as in OS/MS, except the volume of the bell decreases for each level of list nesting.

8.3.2 Ordered Lists

OS/V
"the following is an ordered list"..."end of list". Items preceded by "item number x".

OS/MS
Same as unordered list, except each item also preceded by "item number x".

MS/V
Same as unordered list, except each item also preceded by "item number x".

MS/MS
Same as unordered list, except each item also preceded by "item number x".

8.4 Paragraph Boundaries

Each interface marked paragraph boundaries by a pause which was slightly longer than the normal pause between sentences.

8.5 Address

OS/V
"address"..."end address"

OS/MS
sound of a door knock

MS/V
A different speaker reads the address text

MS/MS
Sound of a door knock, and the text is read by a different speaker.

8.6 In-Line Images

8.6.1 Plain Images

OS/V
"there is an inlined image", ALT text read

OS/MS
sound of a camera precedes any ALT text

MS/V
different speaker says "there is an inlined image", reads ALT text

MS/MS
sound of a camera, different speaker reads ALT text

8.6.2 Image Maps

OS/V
"there is an image map"

OS/MS
sound of a camera, "there is an image map"

MS/V
different speaker says "there is an image map"

MS/MS
sound of a camera, different speaker says "there is an image map"

8.7 Bold Text

OS/V
Bold text spoken in a louder voice

OS/MS
Bold text spoken in a louder voice

MS/V
Bold text read by a different speaker with a louder voice.

MS/MS
Bold text read by current speaker in a louder voice

9.0 Appendix: Tasks

Task Set 1

- What topic does John Perry see as being central in the study of logic?
- Find the webmaster for CSLI.
- Describe (or reproduce) the list structure for the list at the top of the Archimedes Project Description page.
- Name at least three of the topics being pursued by CSLI researchers.

Task Set 2

- Find the text that mentions how to get more information about CSLI.
- Name three tasks involved in language technology.
- Find John Perry’s email address.
- Name two of the dimensions of the problem of human-computer interactions.

Task Set 3

- Find the Associate Director of CSLI.
- How many mailto links are on the CSLI Home Page?
- What kinds of people are involved in CSLI’s human-computer interaction projects?
- Find where Northeastern University is mentioned.

Task Set 4

- List all of the headings on the CSLI Home Page.
- Reproduce or describe the structure of John Perry’s home page.
- What does Greg Edwards see as being the future of interface design?
- Find the two main prototypes currently being designed by the Archimedes Project.

10.0 Appendix: Questionnaire Questions

1. Rate your ability to distinguish anchor points from other text in the interface.

    ___ Very Good

    ___ Good

    ___ Adequate

    ___ Not Very Good

    ___ Poor

2. The meta-information associated with the links (the “Read URLs” and sound effects) was:

    ___ Very Helpful

    ___ Somewhat Helpful

    ___ Confusing

    ___ I Didn’t Use It

3. Rate your ability to distinguish between the different kinds of anchor points in the interface:[13]

    ___ Very Good

    ___ Good

    ___ Adequate

    ___ Not Very Good

    ___ Poor

4. What did you like the best about the way anchor points were presented in the interface? The least?

5. What did you like the best about the sound effects used in the interface? The least?[14]

6. Rate your ability to distinguish among the various heading levels using this interface:

    ___ Very Good

    ___ Good

    ___ Adequate

    ___ Not Very Good

    ___ Poor

7. Rate your ability to distinguish headings from other types of text using this interface:

    ___ Very Good

    ___ Good

    ___ Adequate

    ___ Not Very Good

    ___ Poor

8. What did you like the best about the way in which the headings were presented? The least?

9. The pauses between sentences were:

    ___ Way Too Long

    ___ Too Long

    ___ Appropriate

    ___ Too Short

    ___ Way Too Short

10. The pauses between paragraphs were:

    ___ Way Too Long

    ___ Too Long

    ___ Appropriate

    ___ Too Short

    ___ Way Too Short

11. Rate your ability to distinguish between the pauses between paragraphs and those between sentences:

    ___ Easy

[13] Only in the questionnaires for the OS/MS and MS/MS protocols.

[14] Only in the questionnaires for the OS/MS and MS/MS protocols.

___ Hard

___ Didn’t realize there was a difference

12. General comments on pauses?
13. The overall volume of the interface was:

___ Too Loud

___ Appropriate

___ Too Soft

14. The volume of the sound effects as compared to the speech in the interface was:

___ Too Loud

___ Appropriate

___ Too Soft

15. General Comments on volume?
16. Rate the effectiveness of this interface in presenting the overall structure of the documents:

___ Very Good

___ Good

___ Adequate

___ Not Very Good

___ Poor

17. General Comments?

11.0 References

[1] Barry Arons. *Interactively Skimming Recorded Speech.* PhD thesis, M.I.T., February 1994.

[2] Meera M. Blattner, et al. Earcons and Icons: Their Structure and Common Design Principles. In Ephraim P. Glinert, editor, *Visual Programming Environments: Applications and Issues.* IEEE Computer Society Press, Los Alamitos, CA, 1990.

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

[4] W. Keith Edwards and Elizabeth D. Mynatt. An architecture for transforming graphical interfaces. In *Proceedings of the ACM: UIST ‘94*, pages 39-47, New York, 1994. ACM Press.

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

[6] Ralph E. Geiselman and Joseph M. Crawley. Incidental processing of speaker characteristics: Voice as connotative information. *Journal of Verbal Learning and Verbal Behavior*, 22(2):15-23, 1983.

[7] Berkeley Systems, Inc. outSPOKEN. See http://access.berksys.com/

[8] J.J. Gibson. *The Senses Considered as Perceptual Systems.* Houghton Mifflin, Boston, 1966.

[9] Henter-Joyce, Inc. Jaws for Windows, 1996. See http://www.hj.com/jfw.htm

[10] John M. McNeil. Americans with disabilities: 1991-1992. Current population report, series p70-33, U.S. Bureau of the Census, Washington, D.C., 1993. Published by the U.S. Government Printing Office.

[11] Ian J. Pitt and Alistair D.N. Edwards. Navigating the interface by sound for blind users. In D. Diaper and N. Hammond, editors, *People and Computers VI: Proceedings of the HCI ‘91 Conference*, pp. 373-383, Cambridge, UK, August 1991. Cambridge University Press.

[12] Steve Portigal. *Auralization of Document Structure.* Master’s thesis, University of Guelph, 1994.

[13] T.V. Raman. *Audio System for Technical Readings.* PhD thesis, Cornell University, May 1994.

[14] T.V. Raman. Emacspeak-direct speech access. In *ASSETS ‘96: The Second Annual ACM Conference on Assistive Technologies*, pp. 32-36, New York, April 1996. ACM SIGCAPH, Association for Computing Machinery, Inc.

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

[16] Lawrence A. Scadden. Blindness in the information age: Equality or irony? *Journal of Visual Impairment and Blindness*, pages 394-400, November 1984.

[17] Chris Schmandt and Barry Arons. A conversational telephone messaging system. *IEEE Transactions on Consumer Electronics*, 30(3):xxi-xxiv, 1984.

[18] Productivity Works. pwWebSpeak, 1996. See http://www.prodworks.com/pwwebspk.htm