Lessons from Developing Audio HTML Interfaces

Frankie James
Cordura Hall Room 128
CSLI, Stanford University
Stanford, CA 94305–4115
(650) 725–2312
fjames@cs.stanford.edu

ABSTRACT

In this paper, we discuss our previous research on the establishment of guidelines and principles for choosing sounds to use in an audio interface to HTML, called the AHA framework. These principles, along with issues related to the target audience such as user tasks, goals, and interests are factors that can help us to choose specific sounds for the interface. We conclude by describing scenarios of two potential users and the interfaces that would seem to be appropriate for them.

Keywords

audio interfaces, WWW, blind, human-computer interaction, HTML

INTRODUCTION

The WWW provides a vast quantity of on-line information. For blind people, who cannot access printed materials, the WWW may be the only source that they can use to access information for their daily lives, such as bus schedules and movie listings. Previous solutions to the problem of accessing the WWW, such as screen readers, have relied on the visual representations of web pages as the basis for audio renderings. Our research focuses on the idea that usable audio renderings can be produced in the same way that usable visual renderings are, that is, directly from the HTML markup of pages. This markup provides explicit information about the document structures the author intends to represent. By formatting the marked-up structures using audio cues presented along with synthesized speech, we can provide an interface to HTML that is at least as well formatted as the visual HTML renderings of browsers such as Microsoft Internet Explorer and Netscape Navigator.

The idea of providing audio formatting to speech in order to represent structures has a long history, including techniques used in radio broadcasting and children's story books on cassette. In addition, previous work has been done on the audio formatting of text for blind computer users in the area of access to mathematics and TeX documents. [16] We have used HTML as a starting point for analyzing audio formatting in general. User studies [7][8][9][10] have provided us with a set of principles and guidelines (called the AHA, or Audio HTML Access, framework) for choosing specific sounds to represent structures in interfaces for users with varying informational needs. Interface "variables," such as the user's background, the types of pages he or she typically uses on the WWW, and available audio marking techniques, also influence the design of interfaces based on AHA. In this paper, we present scenarios for two different users and interfaces that could be created to meet their individual needs.

AHA PRINCIPLES

We have tested various audio interfaces to HTML to compare different markings of particular elements and to determine how the markings affect users' perception of the document structures. The following principles (which are summarized in table 1) form the basis of the AHA framework for choosing sounds to use in a particular interface. The choice of specific sounds to be used will be affected by factors related to the expected users, which will be discussed later.

Table 1: AHA Framework

Vocal Source Identity	• number of voices • context switches
Recognizability	• sound identity • salient feature identity • identity of metaphor
Distraction	• number of sounds • signalling tendency • length of sounds • aesthetics

Vocal Source Identity

We have made extensive use of different speaking voices to mark HTML structures. This technique is borrowed from the use of multiple speakers in radio broadcasts such as sporting events, where each speaker has a certain role in the presentation of the game. Listeners who are used to this format quickly learn to expect different kinds of information from each speaker. There is also evidence from the psychological literature that speaker identity is remembered incidentally to what is said [5], which suggests that presenting certain types of information in a different voice will cause the information to be associated with that voice and, consequently, the structure being marked by it. Our empirical studies have explored the specific circumstances under which this powerful technique should be used.

Number of Voices

Reeves and Nass [17] have pointed out that when more than one voice is used in an interface, users attempt to establish and maintain relationships between the pairs of speakers (dominant-subordinate, etc.). Since the number of pairwise relations between *n* speakers is equal to *(n*(n-1)/2)*, Reeves and Nass maintain that the number of different speakers in an interface should be kept small.

Our studies confirmed the finding that the total number of speakers should be small, but for a different reason. By testing an interface that used speaker changes to mark several of the major document structures (headings, blockquotes, lists, etc.) against one that only used a speaker change in one place (to mark links), we found that, while speaker changes in almost all cases produced a slight improvement in subjects' ratings for appropriateness and liking, the single speaker change in the second interface produced a marked effect. The appropriateness of links in this interface was significantly higher than for any other tag.

This finding is consistent with the visual analogy where document structures are all marked by the same type of change, for example, a color change. In this case, the color can help the user to see that the text is marked, but if more than a few colors are used, the document begins to just look colorful and unmarked. Marking techniques (for the visual, color, for the audio, speaker change) lose salience when the gestalt of the document becomes a cacophony of similar markings.

Context Switches

Speaker changes do not seem to be appropriate for marking items when it is expected that the items to be marked will be found within a coherent text flow. For example, subjects became confused when emphasized text that occurred in the middle of a sentence was marked using a speaker change. This is because people do not expect to hear a different person say one or more of the words in a sentence being spoken by someone else. When users hear examples of this in an interface, their attention is drawn away from the content text and towards the speakers themselves, again trying to understand the relationship between the speakers that would allow them to work together to present a single thought.

The application of this guideline to a particular HTML structure depends greatly on the expected usage of that structure in the interface. For example, link points may be expected to fall within text flows if we assume that the user is interested in report-style documents where links are often used on words within a sentence to create a glossary of definitions. However, if we assume the user to be interested in index documents of the Yahoo! style [22], we would expect most link points to occur as list items that are *not* within other text flows. Additionally, returning to the first example, if we assume that the user is listening to report documents only to find the key concepts in the reports and not for thorough reading, we can ignore the fact that the links are within text flows.

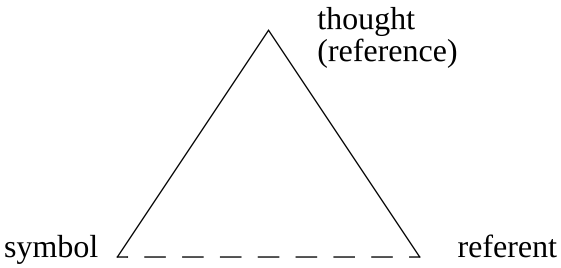
Recognizability

Several types of recognizability are valuable for the selection of sounds to be used in an audio HTML interface. These include the recognition of the identity of the sound and the recognition of the salient property of the sound, which can guide interpretation.

Sound Identity

Recognition of the identity of sounds in an interface by the target audience facilitates the mapping of those sounds onto the marked document structures. People often think of the sounds in an interface as being icons, or signs, of the things that they represent. Ogden and Richards [15] describe *linguistic* signs or words as being made up of a symbol, a thought (or reference), and a referent (see figure 1). In their explanation, words can only map to things in the world by first mapping the words to ideas and then mapping the ideas to things. The analogy for our situation is that sounds are used in an interface to map to ideas, and these ideas then map to document structures. By making sure that the identity of a sound is available to the user before she even hears the interface for the first time, we have a known symbol. At this point, all that remains for the user to learn is the reference that will allow her to map from the symbol (sound) to the referent (document structure). If, instead, the sound in the interface was new to the user (e.g., an unfamiliar melody), she would have to in some sense "learn" the symbol (sound) itself so that it would be familiar to her on the next hearing, and then after this internalization, she would again have to learn the reference.

Figure 1: The Parts of a Sign



Salient Feature Identity

The salient feature of sounds should also be readily identifiable. For example, certain sounds that might not be thought of as recognizable in the classical sense could have features that can be recognized as standing for a particular thing. Our second study showed that the use of tonal sequences whose pitch contours represented different levels were appropriate and fairly well liked by subjects. In this case, although the subjects had probably never heard the particular sequences used in the study before, they were able to identify the pitch contours as going up (to some degree) or down (to some degree) and then interpret that these degrees mapped onto heading levels in the interface. The recognition in this case has to do with the basic musical knowledge that some sequences of pitches go up and others go down. Musical knowledge of scales and their basis for interpreting higher and lower in tonal sequences form the reference, which is then mapped onto the referent, in this case, a heading of a particular level.

Salient feature identity is also important in the case where sounds are overlaid to mark document structures such as links. Since the number of words in any given link point can vary, the length of the accompanying sound will also vary in order to play for the duration of the link point and no longer. Therefore, in the strict sense, no two given links will be marked by identical sounds since the two sounds will be played for differing lengths of time. However, if the user knows that length is not the salient feature of the marking, he will be able to call all of the overlaid sounds that sound the same but have different lengths "identical."

Identity of Metaphor

The final type of recognizability explored in our interfaces is the identity of the sound's metaphor. Visual interfaces [18] use concrete icons such as folders that have obvious properties in the real world. The icons are used to represent interface objects that then have some of the same properties as the concrete items. This helps the user to interpret the sign not only by making the symbol (the folder) a known quantity, but also by suggesting the reference (something that holds other things) that will lead to an appropriate understanding of the referent (file system directory). Similarly in audio, when we use a sound with an obvious metaphor (e.g., the sound of a typewriter to mark typewritten text), the user firstly can recognize the sound for what it is (a typewriter) and then apply this to interpret what is being marked (typewritten text).

It is in examples when the identity of metaphor is absent that we see more clearly its importance. Our second experiment used an interface where various form elements were marked with familiar musical melodies, such as "Pop Goes the Weasel" and "Mary Had a Little Lamb." Everyone in the study seemed to be able to identify the tunes by name, but this interface was rated the lowest in the study for both appropriateness and likability of forms. Clearly, although subjects could identify the symbol (the tune name), they had no possibility of guessing from any feature of the tune its reference or referent. This is illustrated even more clearly by noting that the one marking that did have an obvious reference in this case stood out in subjects' responses. Text entry fields were marked with the tune from the game show "Jeopardy", which is used to indicate that the contestant needs to respond. Subjects were able to pick out this (cultural) feature of the tune to produce a reference and map to the referent, a text-entry field.

Distraction

Although our main goal is to provide the user with as much information as possible about the structures in a document, providing too much structural information can cause the user to be distracted from the document's textual content. This is a problem for our task, but not necessarily for all audio interfaces. For example, the goals in other interfaces, such as those for video games or process control [4] may be specifically to *cause* distraction from the primary task, such as to signal a problem in a factory simulation or to increase the cognitive load to add to the enjoyment of a game. However, in marking document structures, the primary task is reading, and the marked items do not signal problems requiring immediate attention.

Number of Sounds

Keeping the total number of sounds in an audio HTML interface small reduces distraction. Our second experiment showed that the interface using many sounds was rated the lowest for both general appropriateness and general likability, whereas the interface that used only a few sounds was rated the highest for general likability. By using many sounds in an interface, the cognitive load placed on the user is high, since there are many signs that have to be remembered to understand the document structure. Also, sound is played frequently in the interface since so many different structures are marked. A good visual analogy of this case is programs such as Microsoft Word that use button bars containing buttons for tens of functions at a time. In this case, users often complain that the interface is too "busy" and that it is difficult to pick out the important functions because of all of the extra clutter.

We can contrast this with the case of video games, where both the visual and auditory parts of the interface are quite cluttered. This profusion of information adds to the general excitement of the game and forces the user to expend more cognitive energy to reach the goal. This in turn increases the satisfaction in winning the game. Since the primary goal in video games is to maintain interest over time, adding distraction is a good use of sound in this case.

Signalling Tendency of Sounds

Distraction can also be caused by a failure to correlate the signalling tendency of sounds with various aspects of the tags to be marked, such as their importance to the user and their frequency. For example, it is not appropriate to draw a lot of attention to an insignificant structure such as a footnote. However, if headings are an important document element, a sound can be used that will attract more attention to them than to other structures.

Our second experiment demonstrated that subjects consider the address tag to be relatively unimportant, either because it is often misused or because it is apparent that the text is an address from the content. However, in one interface, addresses were marked with a prominent non-speech cue and a speaker change, which attracted a great deal of attention to the tag. This marking was not considered to be any more or less appropriate than that in another interface, which used a subtle overloaded voice quality cue. Therefore, there was a trade-off between the distinguishability of the cue in the first case and its obvious distraction from the text, and the relative indistinguishability of the second cue and its lack of distraction.

It is also important that structures that occur frequently in documents are marked with cues that do not draw as much attention as other sounds. For example, if a user commonly accesses index pages that present long lists, using a highly attention-getting cue (such as a bell) for list items will be a constant source of distraction. If a more subtle cue is used such as a quiet chime or tone, the sound can be shifted out of the user's attention more easily, thereby minimizing distraction.

There are clearly other types of interfaces where sounds can strongly signal the user away from her primary task. For instance, in a process simulation such as ARKola [4], using the sound of bottles crashing off of a machine effectively signals the user that there is a problem with the current state of the simulation and that an action needs to be taken immediately. In addition, few people would argue that the auditory signal for a nuclear meltdown should be something that did not attract a great deal of attention. However, when dealing with documents, there are few cases where the user needs to take an immediate action, and certainly the consequences of missing the occasional cue are less dire.

Length of Sounds

Distraction is also related to the length of the sounds used for marking document structures. Users listening to documents are generally trying to get as much information as they can as quickly as possible. If the structural cues in the interface take too much time to present, they will slow the user down. In both of our studies, there were comments that certain of the sounds used were too long, such as a repeated bell used in the first experiment and some musical cues used in the second. In all cases, the subjects were distracted from the text by having to wait for the sound to finish playing to start hearing the text again.

Sound length problems also encompass other issues, such as redundancy and the presentation of relevant information in the sound. For example, even when a sound is relatively short, if the salient feature of the sound is near the end, the user has to attend to almost the entire sound before interpreting it. In our second experiment, different tonal sequences were used to mark heading levels, where the pitch contours mapped to the different levels. However, each cue started on the same note. This meant that the first part of each of the sounds was identical, so that subjects had to wait to hear almost the entire sound to decide which level was being presented. In contrast, if each sequence had started on a different note, subjects may have been able to use this information to identify the heading levels more quickly.

The example of heading level markings also addresses the issue of sound redundancy. The initial tone in the sequences is in fact redundant, since it is basically used as a base for the pitch contour. It was apparently recognized as marking headings more quickly than it took to finish playing, which meant that the subjects had to wait to get more information. Another situation where the same problem arises comes in the case of overlaid sounds, where the user may recognize a sound well before it finishes playing, but is still forced to listen to it for the duration of the marked text. This occurred in the case of link markings in one of the interfaces of the second experiment. The sounds used obvious metaphors to relate them to the type of links being marked so that they were quickly recognized, but the sounds continued to play for the entire reading of the link text. This interface was consequently ranked lowest for both appropriateness and likability in terms of link points. Subjects also said that the sounds were "too long," even though the actual lengths of the sounds were constrained by the length of the link text, meaning that subjects never had to wait for the sounds to finish to continue hearing text.

The problem of overlaid sounds relates to evidence in the literature on shadowing [2] showing that even when people are told to ignore an audio stimulus and claim that they are doing so, they are actually still receiving information from this channel at a sub- or semi-conscious level. Therefore, if a sound is being played simultaneously while text is read, users will continue to process the sound the entire time, thus distracting them from the text. Again looking at process control, we can see that continuous overlaid sounds can be useful, such as when we monitor the sound of a copy machine in the background of performing another task to be sure that our copying job is completing successfully. However, the sounds in this situation are not marking a novel circumstance. Instead, the novel circumstance is when the sound stops.

Aesthetics

The discussion of distraction must also address the aesthetics of particular sounds in the interface. Some sounds are inherently more strident and abrasive than others, such as those commonly used in smoke alarms. In the case of alarms, however, the abrasiveness of the sound is important for drawing attention to the impending danger of the fire. In the case of marking document structures, there is little danger in ignoring or missing a cue; therefore, abrasive and annoying cues will cause unnecessary distraction.

Additionally, we can talk about the *general aesthetics* of our interface, which refers to the overall sound quality and beauty. For example, the addition of high quality speakers to a computer system adds beauty to an audio interface without changing the identity of the sounds themselves. This is analogous to switching from an 8-bit to a 16-bit color display and noticing that the 16-bit display is more appealing and beautiful. Also, in the same way that having an artist or graphic designer produce the final versions of icon designs improves the overall quality of an interface by maintaining a balance of color, line thickness, shadowing, etc., the use of a good sound designer can reduce end-user fatigue by ensuring that the sounds are not generally distracting because of poor timbral or harmonic quality.

INTERFACE VARIABLES

The sounds we choose in an interface are dependent on more than just the AHA principles. They also depend on factors related to particular users and their tasks. In this section, we will first outline the elements of users that constrain which sounds to use, and then address other variables related to the technologies and resources available to the interface designer that will further affect our choices.

User Variables

Several interdependent factors need to be considered in interface design: basic audience characteristics, expected page types to be encountered, and the relative importance of different tags in the interface. Varying one of these factors may have repercussions for the others. For example, if our audience is children, we may expect that they will visit certain types of pages, such as “fun” pages as opposed to corporate home pages. We can further assume that certain tags will be more important than others, for instance, links may be far more important than address sections. If our audience is not children, we will have different expectations.

Audience

One aspect of the intended audience is the age range. Just as there is a difference between a graphical drawing program designed for children, such as Kid Pix [11], and one designed for adults (MacDraw [13]), there should also be a difference in audio interfaces for children or adults.

Choice of sound should also be influenced by the skills of the audience. For example, for an audience of professional musicians, we can take advantage of their intricate knowledge of pitch or rhythm to provide document markings. Conversely, the average user generally cannot recognize the absolute pitches of tones, so the use of tones with differing only in pitch would be inappropriate for this audience.

The culture of the expected audience must also be taken into account. It is useful, for example, to select sounds for the interface that have obvious metaphorical relationships with the marked structures. However, these metaphors are generally not universal. In our “Jeopardy” example, for instance, there is a dependence on the fact that the target audience watches Jeopardy. The choice to use a siren or horn of a particular type may lose its metaphorical value for users in whose culture horns and sirens sound different. [12]

Other characteristics of the intended audience that can influence sound choice include the interests of users, which will directly affect the page types we expect users to encounter, and the users’ knowledge of HTML structures, which can affect their reactions to certain tags. For example, if users have no prior experience with the address tag as conveying contact information, they may be confused by a marking with a metaphor related to the mail or telephone.

Page Types

The types of pages that we expect users to visit need to be taken into consideration when designing an HTML interface. For example, do we expect users to access many report-style documents with a few links scattered in the text, or do we expect them to visit index pages that are mostly links with little other text? Questions such as this may also relate to the relative importance that will be placed on certain tags, but there are larger questions of page type that deal more with the general “sound-and-feel” of the interface we will design.

There are obvious differences between pages like the IBM corporate web site [6] and the web site for Metallica at Elektra records. [14] The visual styles are quite different, even though both pages make use of the common elements of links, headings, and lists. By varying things like the background color and font styles, the sites give completely different feelings that correspond to the types of information being presented. Of course, even when the look-and-feel is varied for different page types, there are some constant affordances that allow users to navigate across the whole web, which is comprised of many different page types. On both of these types of pages, link points are marked by an underline and a color change for the link text. Such common affordances need to be maintained in an audio interface.

Tag Importance

The importance of tags is related to the user’s task, such as whether he intends to read a report or just skim it to learn about key concepts. Link points will be more important in the latter case than in the former. On the other hand, blockquotes may be more important for the conscientious reader than for the skimmer. Choosing sounds that reflect the varying importances of tags for the target user group will help to create the most usable interface.

Tag importance is also dependent on tag usage, in that certain pages use tags more or less as the authors of HTML intended while others use tags to produce specific visual effects. This usage will affect how users react to the marked text (“this sound signals an address” or “this sound signals emphasized text”) and therefore what sound should be used. Clearly, if a tag such as that for addresses is never (or hardly ever) used to mark an actual address, a sound that brings to mind addresses will be inappropriate. It may be better in cases where the tag is not consistently used in any one way that it be left unmarked than to have it marked in a way corresponding solely to one usage.

Technological Variables

Technological constraints can also limit sound choices. Until recently, for example, the idea of using multiple voices in an interface has not been feasible because the text-to-speech synthesis technology did not allow it. However, it is now common for commercial synthesizers to have several voices available to users. The current largest constraint imposed by TTS technology is the use of paralinguistic cues, or voice quality changes, in interfaces. Certain high-end synthesizers such as the DECtalk allow this kind of fine-grained voice control, but many others do not.

Other resources available to the interface designer can also constrain her choices. If she has no access to a synthesizer, the designer will unable to use synthesized musical sounds in the interface and may be forced to look for pre-sampled sounds, perhaps in freeware libraries on the WWW. Furthermore, if the designer is constrained to using pre-sampled sounds, some sounds may not be readily accessed or may be of too poor quality to use in an interface. The use of genre sounds can also be limited or expanded by their availability on the web. While there is no problem with finding samples of sounds related to Star Trek, for example, sounds related to other television programs may be unrepresented on the WWW or on sound effect collections available on CD.

A further possibility for audio interfaces is spatialized or so-called 3-D sound. However, the technology required to implement a decent 3-D sound system, including speakers[1] and the software to control spatial positioning of sounds, is not readily available to the general public. A design using 3-D audio would have to assume that the target audience had access to the appropriate equipment.

SCENARIOS

This section describes two scenarios of typical users of an audio interface to the WWW. The scenarios demonstrate how the choice of audience affects what page types the browser needs to be prepared to handle, how important certain tags will be, and thus what sounds should be used in the interface.

Scenario 1: Gaining General Knowledge and Entertainment

Our first scenario focuses on using the WWW to gain general knowledge about topics. The sounds are chosen to facilitate skimming and navigation between documents.

Audience

Jerry is a blind Stanford undergraduate who often uses the WWW for recreation, browsing through topics of interest and trying to find cool new things. He has just been assigned to do a term paper on machine translation, a subject that he knows little about. He wants to use the WWW to learn about the key issues and related ideas in order to get started on his paper. Since Jerry often uses the WWW for fun, he likes the sounds in his interface to be entertaining. He is *not* a musician, so he cannot easily interpret musical distinctions, such as pitch differences, when they are used to present tags.

Page Types

Jerry typically browses Yahoo! to find interesting things, so he often encounters index pages and other pages with many links. When he does find report-style documents for his term paper, he generally skims to find key ideas or interesting points rather than devoting a long time to reading each document thoroughly.

Tag Importance

Jerry is interested in knowing about the links on a page and typically encounters many of them during a browsing session. He also uses document headings to decide what parts of the page to read thoroughly. He finds the other tags, such as lists, to be unimportant because they are used throughout the document and their structuring effects are redundant with those for links in the index documents. Other structures are relatively unimportant because he is not thoroughly reading most of the documents that he encounters.

Sounds

Links in this interface are preceded by a short sound indicating the link type (such as a typewriter for links to text documents). The link text is read in a different voice than the rest of the text. The reason for this is that the different voice will draw Jerry's attention, which is appropriate since links are very important to him. Our guideline of avoiding speaker changes within a text flow is not broken often, since most of the pages Jerry visits are index pages where the links are not in the middle of the text. The reason for using sounds to indicate link type is that, since Jerry encounters many links, he needs to be able to decide quickly whether or not any one is worthy of following. By knowing the link's type, he has more information to help him choose.

Headings are quite important to Jerry for determining key concepts in the document, but gaining an exact understanding of the document framework is not necessary. Therefore, headings are preceded by a horn sound of varying type (car horn, truck horn, bicycle horn, etc.). The horn type indicates the level of the heading, using a hierarchy based on the size of the corresponding vehicle to map to the levels. The high signalling potential of the horns is effective for drawing Jerry's attention, while the possible confusion among middle-level headings ("was that a big car or a small car?") is not a big problem since a general idea of the structure is sufficient for his needs. We should also note that while it seems that the difference between the sounds is a musical parameter change, in fact, the changes are related to different objects in the real world and can be interpreted as such. So, instead of Jerry having to interpret the musical quality of the horns, he can think about the kind of vehicle that has a horn *like* that.

Images in the interface are preceded by a camera sound, and the ALT text is read following the sound. The camera sound is short and not very distracting, giving a direct metaphorical mapping from the sound to the tag.

Jerry often encounters lists on index pages. In these cases, the whole page is usually one big list or a series of lists. Therefore, a strong marking for lists would both get in the way of Jerry's attention to the links, and would be fatiguing since the sound would be repeated quite often. Also, each list item usually consists of one link, and the link sound is already an effective means of separating the links. Therefore, lists are marked by a pause before the items.

Blockquotes are also marked by a slight pause in this interface, since, on the pages Jerry visits, they are typically used to indent or center portions of text. In reports where the blockquotes are actually used for quotations, Jerry does not spend much time reading them. In both cases, there is really no need for an explicit cue, so the pause is sufficient.

In the documents Jerry encounters, text changes are sometimes used to mark key concepts. Therefore, he is somewhat interested in paying attention to the text changes. However, we do not want to use too strong a marking for italics or bold since they are also often used just to provide emphasis. For this reason, the interface uses voice inflection changes to mark text changes, in much the same way that a human reader would read such markings with more voice stress.

Finally, address sections are also marked by a voice inflection change. This is because the address content is not especially important to Jerry since he does not intend to contact authors of documents, but the visual representation of the address tag is italicized text. By using a voice change, the cases in which the address tag is misused will not become confusing.

1. Although headphones may be used in place of speakers, the current trend in 3-D audio design is to move away from the use of headphones due to their tendency to cut users off from the rest of their sonic environment. [19]

Scenario 2: Gaining In-Depth Knowledge

In our second scenario, we look at someone who is interested in gaining in-depth knowledge about a particular field. There is some overlap with the first scenario in terms of both the types of pages visited and the relative importance of some of the tags, but the difference in interests affects the choice of sounds in this interface.

Audience

Amy is a blind woman and a researcher in particle physics. Her use of the WWW is generally for learning the latest-breaking news and reading articles of interest on-line. She also uses the WWW to keep up with public transportation and movie schedules, since she goes out often but does not drive. She does not like to be distracted by a lot of sounds in an interface, but still wants to have all of the relevant structures marked clearly and efficiently. Amy is also not a musician and does not want to be distracted from her reading by having to interpret musical differences between sounds.

Page Types

The kinds of pages that Amy uses for her research are mainly structured reports containing only a few links. She generally reads the pages carefully to understand the content and only follows links if they seem especially important. When she looks at schedule pages, they are typically ones with which she is somewhat familiar, so again she concentrates more on the new content than on links away from the document.

Tag Importance

Headings, lists, and other document structures help Amy to build a framework of the ideas in the document into which her knowledge of the content can be fit. She does not want to be distracted by links in the text while she is reading, but an indication that links are present is acceptable.

Sounds

Headings are very important for establishing a framework of the documents that Amy usually reads, but the sounds should be short and understated to avoid distracting her from the document content. Therefore, the levels are marked using a three-tone sequence (as was used in our second experiment [10]), where the pitch contour maps onto the heading level. Each sequence begins on a different tone to allow a possible way to recognize the marked level more quickly. [21][2] The heading text in this interface is read using a different speaker, whom we can call the "heading reader." This speaker change attracts Amy's attention to the headings and allows her to more rapidly group the headings together in her mind.

Lists in this interface are somewhat important, in that they are sometimes used in reports to present a set of key points. Amy wants to be able to rapidly distinguish lists, but they are clearly not as important as the headings and should be marked less saliently. The marking we choose to use here is a short bell sound preceding the link items, which can act as an audio bullet, plus a speaker change (using a voice we will call the "special reader"). The bell clearly separates the list items, while the "special reader" gives an indication of the scope of the various list items and that the list is a structured section to which Amy should pay some attention.

Images are again marked using the "special reader," who in this case presents the spoken cue "image caption" and then reads the ALT text for the image. The alternative voice indicates that this is a special section (apart from the regular text) and the spoken cue is easily interpretable.

Amy does not generally follow links within the documents she reads until she has finished reading and is deciding what topic she wishes to explore next. Therefore, links need to be marked in a subtle way to indicate their presence without distracting Amy from her reading. In this interface, links are marked by a short beep preceding the link text. This sound is short and sufficiently different from the list bell to avoid confusion.

Blockquotes are marked in this interface by the use again of the "special reader" to read the text, and are preceded by a short cue that is somewhat "quote-like."[3] Blockquotes can be somewhat important in the documents that Amy uses because they are often used to mark actual quotations or equations that need to be offset from the rest of the text. By marking these sections with the speaker change, there is a signal that this section is special without causing a major problem if the tag is simply being used for formatting. In addition, the short preceding sound allows Amy to differentiate between this and other uses of the "special reader" in the interface.

Text changes and address sections are both marked in this interface by voice inflection changes. As we mentioned in the previous scenario, address tags are often misused simply to create segments of italicized text. In these cases, we would not want to use a marking that suggested the other usage. In addition, if the address tag is in fact being used correctly, Amy can usually tell this from the content text, which will contain a street or e-mail address.

CONCLUSIONS AND FUTURE WORK

These two scenarios illustrate how sound choices in audio interfaces can take general principles into account while still allowing the flexibility to accommodate different kinds of users. The scenarios overlap in that some of the same types of pages are used, but in different ways. There may also be overlaps between the particular sounds used to mark certain tags even though the motivations may be slightly different, as in the use of voice inflection changes in both cases to mark text changes.

[2] It should be noted that the cited experiment was in fact testing congruency effects between pitch and pitch change, but the final results can be interpreted to suggest that a correct congruency matching can cause a positive reinforcement of level, when level is primarily represented by contour (i.e., pitch change) and secondarily marked by initial pitch.

[3] The sound used is that which was used in the second experiment for marking blockquotes in the AHA interface. This sound is based on the sound found in Victor Borge's *Phonetic Punctuation*. [3]

No one interface is appropriate for all users in all situations. However, we are not so pessimistic as to think that a different interface needs to be created for each individual user, or, even more extreme, for each individual user's current mood. Although our two example users differ in their preferences, leading to different interface designs, Jerry and Amy can represent whole classes of users with similar preferences who, as a group, can benefit from the same interface. We can envision the design of alternative sound representations as parallel to the approach taken by the designers of a variety of standard applications such as drawing programs. Competing programs share core functionality while differing in many aspects of the detailed functions and the look and feel of the interface. (See, for example, Kid Pix [11], MacDraw [13], and AutoCAD [1]). In the same vein, Microsoft Windows 95's notion of sound schemes has led to the creation of a few interfaces suitable for different points in the user space. The structure allows for further individual customization, but we do not see the emergence of separate interfaces for each individual.

Our future goals include implementing the two proposed scenario interfaces using the marcopolo))) plug-in for Netscape designed by SONICON Development, Inc. [20] In addition, further work remains to be done in the area of extending the types of HTML tags that can be rendered using AHA principles. These extensions include the appropriate renderings of more "interactive" HTML constructs, such as forms and tables, as well as the rendering of higher-level structures, such as frames.

A somewhat different type of extension of the AHA framework is in the development of non-visual interfaces for sighted users, for example, in situations where the eyes are busy as when driving a car. In this case, there are different problems that need to be addressed related to the creation of audio environments where the user is not expected to focus his attention on the interface, but rather to attend to it in the background of another task.

REFERENCES

1. *AutoCAD*. Autodesk, Inc., 1997

2. Banks, W. P., et al. Negative Priming in Auditory Attention. *Journal of Experimental Psychology: Human Perception & Performance*, 21(6): 1354–1361, December 1995.

3. Borge, V. (speaker). Phonetic Punctuation. *Caught in the Act* [Analog Sound Disc Recording]. Columbia Recording, 1950.

4. Gaver, W., R. Smith, and T. O'Shea. Effective Sounds in Complex Systems: The ARKola Simulation. *CHI '91 Conference Proceedings*, New Orleans, 1991, pp. 85–90.

5. Geiselman, R. E. and J. M. Crawley. Incidental Processing of Speaker Characteristics: Voice as Connotative Information. *Journal of Verbal Learning and Verbal Behavior,* 22(2):15–23, 1983.

6. IBM Corporation. See http://www.ibm.com/

7. James, F. Presenting HTML Structure in Audio: User Satisfaction with Audio Hypertext. *ICAD '96 Proceedings,* Xerox PARC, 4–6 November 1996, pp. 97–103.

8. James, F. Presenting HTML Structure in Audio: User Satisfaction with Audio Hypertext. *CSLI Technical Report 97-201*. CSLI, Stanford University, January 1997.

9. James, F. AHA: Audio HTML Access. *The Sixth International World Wide Web Conference.* Ed. by Michael R. Genesereth and Anna Patterson, Santa Clara, CA, 7–11 April 1997. IW3C2, pp. 129–139.

10. James, F. Distinguishability vs. Distraction in Audio HTML Interfaces. *SIDL-WP-1997-0077, Submitted to CHI '98.*

11. *Kid Pix*. Broderbund Software, 1996.

12. Lodding, K. N. Iconics: A Visual Man-Machine Interface. *Proceedings of the Third Annual Conference and Exhibition of the National Computer Graphics Association, Inc.*, Vol. 1, Washington, D.C., June 1982, pp. 221–233.

13. *MacDraw.* Claris Software, 1996.

14. Metallica. See http://www.elektra.com/metal_club/metallica/menu.html

15. Ogden, C.K. and I.A. Richards. *The Meaning of Meaning.* Routledge & Kegan Paul, Ltd., London, 1923.

16. Raman, T.V. *Audio System for Technical Readings.* Ph.D thesis, Cornell University, 1993.

17. Reeves, B. and C. Nass. *The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places.* Cambridge University Press, New York, 1996.

18. Rogers, Y. Evaluating the Meaningfulness of Icon Sets to Represent Command Operations. *People and Computers: Designing for Usability.* Cambridge University Press, Cambridge and New York, 1986, pp. 586–603.

19. Sawhney, N. and C. Schmandt. Design of Spatialized Audio in Nomadic Environments. *ICAD '97 Proceedings,* Xerox PARC, 3–5 November 1997, pp. 109–113.

20. SONICON Development, Inc. marcopolo))), 1996. See http://www.webpresence.com/sonicon/marcopolo.

21. Walker, B. and A. Ehrenstein. Congruency Effects with Dynamic Auditory Stimuli: Design Implications. *ICAD '97 Proceedings,* Xerox PARC, 3–5 November 1997, pp. 7–11.

22. Yahoo! See http://www.yahoo.com/