Distinguishability vs. Distraction in 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 present the findings and conclusions from a user study on audio interfaces. In the experiment we discuss, we studied a framework for choosing sounds for audio interfaces proposed in [6] by comparing a prototype interface against two existing audio browsers. Our findings indicate that our initial framework, which was described as a separation between recognizable and non-recognizable sounds, could be better interpreted in the context of the distinguishability and distraction level of various types of sounds. We propose a definition of how a sound can be called distracting and how to avoid this when creating audio interfaces.

Keywords
audio interfaces, HTML, WWW

INTRODUCTION
Effective visual document reading and scanning depends on the viewer's ability to understand the document's structure, which is presented through the use of typographical conventions to distinguish between headings of various types, titles, paragraphs, lists, and figures. Effective visual HTML browsing works in much the same way, but with the additional feature of hyperlinks that allow the user to travel within and between documents to scan even faster or to find out about related material.

Typographical and layout conventions have been developed over centuries in such a way that the various marked structures are distinguishable without grabbing more of the user's attention than they deserve. For example, in this document, headings are distinguished from plain text by the use of bold, which is then broken down into all capital letters for heading level 1 and mixed case for heading level 2. The title is marked in bold, as well as in a larger font size. This gradation expresses to the user the relative importance of the title and the various heading levels in a way that is distinguishable while at the same time does not interfere with the reading of the text in the paragraphs.

People who are accessing HTML documents non-visually, such as blind computer users or people who are using handheld devices, need to be able to effectively read, scan, and navigate among documents. Audio presents its own set of challenges in accomplishing this. Because audio is essentially linear in its presentation, the need to hear and recognize important document structures quickly is increased in order to allow the user to rapidly obtain the information in the document. However, as Banks et al. [1] point out, "auditory attention must operate almost entirely by internal processing mechanisms [since] auditory attentional selection is not aided by any analog of visual fixation, which foveates attended items and diminishes interference from nonattended items by putting them into visual areas of lower acuity than the target." This means that if the user is presented with too much structural information, he or she must consciously shift attention away from it to understand and attend to the content text.

The experiment that is discussed in this paper was designed to test different marking techniques for producing audio renderings of HTML. We took the ideas from the AHA framework developed in [6] and applied them to create an audio interface that marked most of the main HTML structures[1] using multiple voices and non-speech audio cues. Our assumption was that by comparing this to two other audio interfaces that were already in use, we would learn which of the markings were appropriate and which weren't, as well as why.

However, what we have learned from this experiment is more far-reaching than just what sounds or voice changes are most appropriate for marking a particular tag. We have found out which of the various HTML tags are thought by users to be the most important to have clearly marked, which greatly influences the choice of sounds in the interface. We have also been confronted with the tension between distinguishability and distraction when marking document structures in audio.

INTERFACE DESIGN
This experiment was designed to test sounds chosen based on the AHA framework [6] against two publicly available audio web browsers (Emacspeak[2] by T.V. Raman [8] and

1. Certain HTML structures were left unmarked because their common uses are problematic. For example, table tags are generally used to create visual layouts rather than actual tables, so they were not marked.

pwWebSpeak[3] by Productivity Works, Inc. [7]). To eliminate differences between the interfaces that were not based directly on the auditory cues used to mark HTML elements, the sound palettes for all three interfaces were implemented in an audio browser called marcopolo))), designed by SONICON Development, Inc.[4] [9] In this way, no effects arose in the experiment from the fact that the Emacspeak and pwWebSpeak browsers (in their native forms), and the standard marcopolo))) browser, differ in their ability to handle user interaction with structures such as forms and tables.

Interface Sounds

The sound palettes of the three interfaces differed substantially in the types of sounds that were used as well as in their overall design philosophy. This section describes the sounds used and the general idea behind the different sound palettes.

AHA

The sounds for the AHA browser were selected based on information gathered in our pilot experiment [4][5] and through research in psychoacoustics. The main goal in AHA was to make as many different kinds of document structures distinguishable in the interface as possible, in much the same way that many different structures are distinguishable in a printed document. The AHA browser made use of multiple voices in the interface to distinguish between various large document structures, such as headings, lists, blockquotes, etc., as discussed in [6].

In order to mark smaller document structures, such as the difference between heading levels, AHA made use of various non-speech audio cues. These cues included both musical and non-musical sounds, as well as abstract tonal sequences. For instance, headings were marked by being read in a particular voice but were also preceded by a three-tone cue that differed depending upon the level of the heading. The contour of the tonal cue corresponded to the number of the heading level (see appendix).

Other tags were marked by real-world sounds that corresponded metaphorically to the tag, such as images, which were marked by a camera sound. These sounds could be played before the textual content of the tag was read (as in the case of the image tag) or overlaid while the content text was read (as in the case of links, where the sound was played while the anchor text was being read). Finally, tags which did not have a strong correspondence to a real-world sound were marked by familiar musical themes.

ES

The Emacspeak philosophy is quite different from that for AHA. Basically, ES emphasizes the notion that HTML documents are generally used for browsing and finding *other* HTML documents, therefore, the ES interface stresses the marking of those elements that are most important for navigation *between* documents. All other structures are marked as unobtrusively as possible to keep them from conflicting with the goal of navigation.

The sounds in the ES interface, then, are a speaker change to mark links and voice inflection changes (changes in speech synthesizer parameters such as pitch range, base pitch, smoothness, etc.) to mark everything else. For example, headings are read with a certain amount of stress in the voice, which varies according to the importance (level) of the heading. Address sections and blockquotes are also read using a more inflected voice.

WS

The pwWebspeak interface uses linguistic cues to mark HTML structures. All of the HTML tags are preceded by short phrases such as “a section heading” or “list item.” In this way, WS users (who are English speakers) can use and understand the interface immediately upon first hearing. The cues, however, tend to take longer to play than non-speech audio cues and, since they are rendered in the same voice as the rest of the text, can be difficult to distinguish from content text.

EXPERIMENT DESIGN

The interfaces in this experiment were tested by about twelve blind subjects[5] with varying degrees of expertise in accessing the WWW via audio, all but one of whom said they used the WWW at least three hours per week. The experiment was designed to be run over the course of three weeks. Each week, participants received one of the three interfaces (AHA, ES, or WS) and were asked to use this interface as their primary web browser for at least three hours during the course of the week. At the end of the week, the participants were sent a user satisfaction questionnaire containing both free response and Likert scale questions regarding the appropriateness, likability, distinguishability, and importance of various HTML elements that had been marked in the interface.

After the participants used all three interfaces, they were sent an end questionnaire that asked them to choose, for each HTML element, which interface had marked this element most appropriately and which interface they liked the most regarding it. Participants were also prompted to give reasons for their choices.

RESULTS AND DISCUSSION

There were many interesting results produced by this experiment. In particular, an assumption we made in the pilot experiment that non-musicians would be able to distinguish between tonal cues that differed in their musical contour appears to have been proven. Also, the use of speaker changes to distinguish between major document structures appears to have been somewhat effective, and at least did not hinder the distinguishability or usefulness of the interface. We also made some new discoveries that can be con-

2. This interface is also referred to as “ES” in this paper.
3. This interface is also referred to as “WS” in this paper.
4. marcopolo))) runs as a Netscape plug-in for Windows 95 and uses the DECtalk Express speech synthesizer to render text and the SoundBlaster sound card to play wav files.
5. Only ten participants completed the experiment by working with all three interfaces. Two participants did not finish with the last interface.

Figure 1: Link Ratings

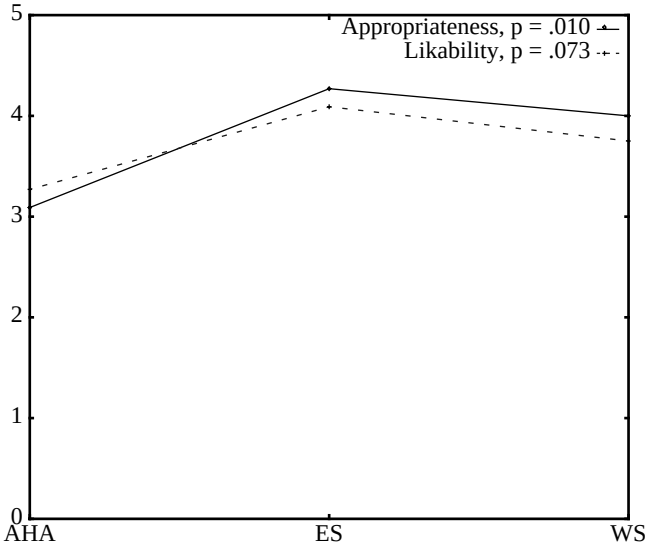
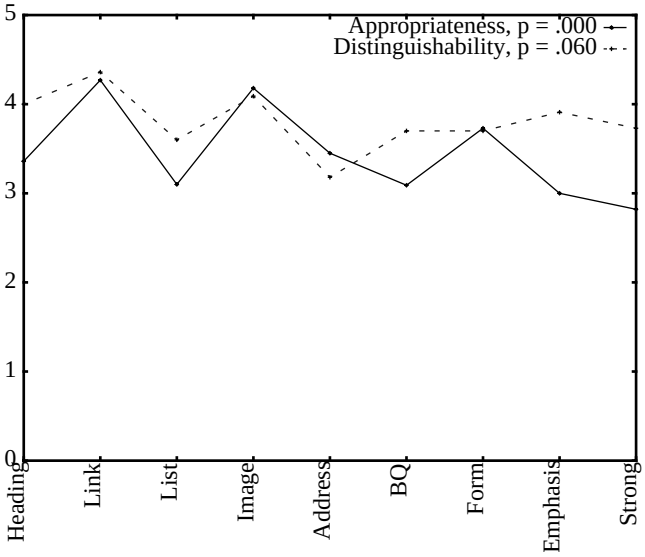


Figure 2: ES Ratings



solidated under the heading of distinguishability vs. distraction.

Speaker Changes

Although there is no direct evidence to prove that speaker changes are better than other marking techniques, the data does show a trend in that all of the structures that were marked in AHA by using a different speaker (except for headings) were rated as being somewhat more distinguishable from the surrounding text than the other interfaces. This suggests at least that the use of speaker change to mark large structures such as headings, lists, and forms does not hinder the usefulness or distinguishability of HTML structures.

In addition, the use of a speaker change in the ES interface to mark links was rated as being significantly more appropriate than the markings for the other two interfaces (see Figure 1), suggesting that, depending on the intended use of the document, speaker changes may be appropriate in more instances than we originally thought. [4][6] From data in the pilot experiment as well as psychology [3], we had maintained that speaker changes would only be appropriate if the text to be marked was not within another stream of text, such as a link on a word within a sentence. Although we do not deny our original assumption that speaker changes within a stream of text are inappropriate, we have been led to drop our assumption that links are generally found on words within text streams. Instead, we believe that there are clearly many more cases where link points are already separated from surrounding text by other structures such as lists and paragraph boundaries (see Yahoo! [10] for a good example).

For this reason, the high appropriateness of ES may not be as surprising as we first thought. If we assume that most of the links that subjects encountered were not within text streams, then high distinguishability between links and non-links would be a critical factor in appropriateness. In fact, in looking at the data for distinguishability (see Figure 2), we find that, in ES, the link marking was rated higher than any other tag. Since the link marking was the only one that did not use a voice inflection change, the link points were clearly the most salient element in the interface. Therefore, their high distinctiveness caused them to be rated the most appropriate in terms of both the ES appropriateness ratings as well as the ratings of link appropriateness across interface.

Tonal Sequences

In the pilot experiment, it was suggested that abstract tones were not very useful for presenting information to users because of the musical expertise required to distinguish between them. We conjectured in [6] that the use of tone sequences with different contours may in fact be useful. Therefore, as we discussed above, the AHA interface distinguished between heading levels by means of a three-tone sequence whose contour represented the heading level. Our results showed that users ranked the distinguishability between heading levels as significantly higher for AHA than for interface ES (see Figure 3), which used voice inflection to differentiate between them. Also, the ranking itself was between “good” and “very good”, suggesting not only that this interface was an improvement over ES, but also that it was in general a good way to mark heading levels.

This result is consistent with a study by Dowling and Fujitani [2] that found contour to be a very important feature in the recognition of melodies. In that study, listeners were presented with a target melody and were then played either an exact transposition of the melody (A), a melody with the same contour which was not an exact transposition (B), or a melody with a different contour (C). The study found that listeners were able to identify a melody as being of type A or B (and not C) very accurately. Additionally, the listeners could *not* distinguish *between* melodies of types A and B[6], suggesting that contour is much stronger than interval in memory for melodies. Therefore, by using a marking for heading levels where all of the information the user needed was in the contour seems to have been appropriate for general users.

Figure 3: Heading Ratings

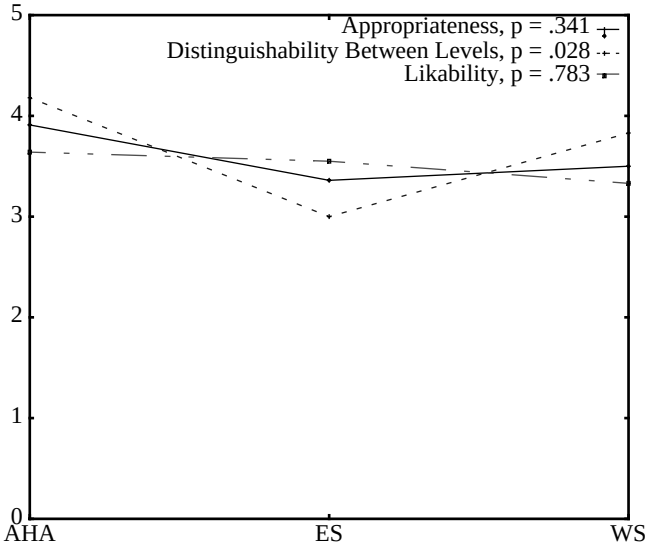
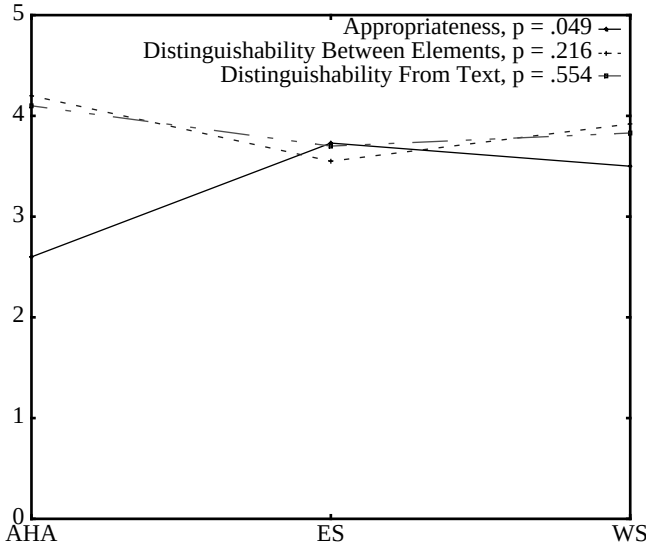


Figure 4: Form Ratings



One drawback to the use of the tonal contours in AHA was that subjects commented that the sounds took too long to present the relevant information. That is, the whole sound had to be heard to hear the contour and interpret the heading level. A possible improvement would be to start the heading cues on different pitch levels to see if this would aid users in making the heading level distinction more quickly.

Distinguishability vs. Distraction

Our major findings in this study concerned the idea of distraction in audio interfaces. Audio is inherently distracting; whereas in the visual domain, one can turn away from a stimulus' source and cease to be distracted by it, in audio, a stimulus in the environment cannot be turned away from. [1] Instead, conscious (and semi-conscious) processes are called upon to ignore the stimulus and attend to something else. We found many cases where there was a clear trade-off between making an HTML element distinguishable and causing it to be distracting.

The obvious question is, what makes a sound distracting? We will use this section to try to offer a definition of what is distracting in an audio interface to HTML while presenting the evidence that supports our theory that distraction is the most important criterion to keep in mind when selecting sounds to use in an audio interface.

Number of Different Sounds

Distraction can be caused by giving the user a large number of sounds from which to select when attempting to recognize a particular sound's meaning. A noisy environment increases the distraction factor of each of the sounds within it. The AHA interface, which tried to mark all of the document structures with distinct cues, was rated the lowest among the three interfaces in terms of both general effectiveness and general likability. On the other hand, the ES interface (which used few distinctive cues) was marked highest for likability. Although these general figures did not prove to be significantly different, there were other smaller cases within the data that were.

In the case of forms (see appendix), AHA marked the various form elements with familiar melodies. ES marked form elements in a more conversational style by using spoken cues to prompt the user to give input or make a selection. The appropriateness data showed a highly significant difference between AHA and ES (see Figure 4). However, the distinguishability between form elements (and the distinguishability of forms from text), while not significant, clearly produced a higher rating for AHA than for ES. This suggests that the musical cues did allow users to differentiate between form elements and between forms and text, but that the cues themselves did little to help the user remember what exactly was being marked. In addition, the appropriateness and likability ratings given to forms in the ES interface correlated to the choice of ES as being the most appropriate and most likable interface in terms of forms in the end data, as well as to the choice of ES as being in general the most appropriate and likable interface. The sheer number of different sounds available in the context of a form for AHA seems to have contributed to its low ratings, while ES' use of more subtle spoken cues contributed to it being rated higher.

Another case that suggests too many sounds were present in the AHA interface was that of link markings (see appendix). AHA allowed the user to distinguish between link types by marking each link with a sound that indicated the type of file that was being linked to[7], while ES and WS did not. AHA users rated their ability to distinguish between link

6. There were some differences found between cases where the target melody was tonal or atonal, and additionally when the B melody was tonal or atonal. However, the main effect, that A and B melodies were harder to differentiate than either A and C or B and C, was still true in all cases.

7. File type identification was approximated by looking at the file extension.

types as being good (4 out of 5) and the importance of knowing a link's type was rated as being important (4.12 out of 5) in general, but the AHA ratings for links in terms of both appropriateness and likability were both significantly lower than for the other two interfaces (see Figure 1). In addition, there is a strong correlation between the distinguishability between the link types in AHA and both the appropriateness and likability of AHA's link markings. This means that people who found it hard to distinguish between link types (presumably due to the confusion of hearing more than one sound to mark the same type of structure) tended to rate the appropriateness and likability of the link types low, whereas people who did not find it hard thought the markings were both appropriate and likable. Apparently, then, the low ranking of AHA for link markings is not based on the fact that marking link types is unnecessary, but rather on the distraction factor of having more than one sound to mark the same type of structure.

Importance of Marked Tag

Sounds can also be distracting by calling more attention to a tag than the tag itself warrants. If a tag is of low importance, marking it with a highly distinguishable sound can cause it to sound as if it is more important than it really is. As a visual example, think about the marking of a footnote within some text. Generally, this is marked by a small superscripted character, indicating that, while it contains more information about the text being footnoted, it is not of central importance to the task of reading the document. Think now about what would happen if a footnote was marked by a colored character in a font size that was significantly larger than the rest of the text. This marking would be distracting and would seem to suggest that the footnote text was terribly important, perhaps even more so than the rest of the text. Figure 5 shows the relative importances of the various HTML elements, as rated by the study participants.[8]

We can see from this figure that the address tag was rated to be of the lowest importance of all of the tags in the interface. If we look at the appropriateness rating for the address tag across the three interfaces, we find no difference between the ratings ($F = .60$, $p = .559$). This is striking because the marking used in AHA was very distinctive, whereas the marking in ES was a voice inflection change that was also used to mark blockquotes. If the participants' answers to this question were reflective of the address marking itself, we would expect to find some difference between the ratings for the two interfaces, presumably that AHA was preferred because the address tag was very distinguishable or that ES was preferred because it was not. However, there was no correlation in the data between the distinguishability of address markings in AHA and either appropriateness or likability, indicating that the distinguishability of address tags did not affect the other ratings, reaffirming its unimportance.[9] Since the rating for an overloaded marking such as that found in ES is the same as for AHA and the tag is not very important, we can conclude that there was a trade-off between the distinguishability of the AHA marking versus its distraction from the rest of the interface. In the free response data, about half of the subjects stated that marking the address tag was unnecessary, therefore, any marking of it could be labelled as being distracting.

Aesthetic Annoyance of Sound

The aesthetic annoyance level of sounds can also cause distraction. In our pilot experiment (see [4] and [5]), we found that there was one sound (a bell used to mark lists) in the OS/ME interface that was too loud and strident compared to the rest of the sounds. This influenced the fact that this interface received a low general rating. In the current experiment, we were able to provide separate controls for the volume of the sounds and the volume of the speaker, thus eliminating the volume problem. We also tried to select sounds that were not as jarring as some of the sounds used in the pilot experiment.

Length of Sound

The length of a sound is also a contributing factor in its tendency to distract. Long sounds, or those whose salient features occur towards the end of the sound, are more distracting than short sounds. Short sounds, on the other hand, can be used in a wider variety of situations without causing distraction.

As an example, blockquotes were rated lower than the average in terms of their importance in the interface (see Figure 5), which suggests (as was the case with the address tag) that a distinct cue would be distracting and should be avoided. However, the likability data did not follow this trend. The AHA interface, which used a short sound cue to mark blockquotes[10], was rated higher in both likability and appropriateness than ES, which used the same overloaded voice change used to mark address sections (see Figure 6). Again differing from the case of addresses, there was a strong correlation in AHA and a somewhat weaker correlation in ES between the distinguishability of blockquotes from text and both appropriateness and likability. The distinguishability (or lack thereof) of the blockquote from text therefore influenced the other ratings, even though the importance of the tag was low. Apparently, a sound that is short enough does not cause the same level of distraction as does a longer sound.

The form rating for AHA is also reflective of the distraction caused by long sounds. About 70% of the subjects commented that the form sounds in AHA were too long, and were far longer than the spoken cues used in ES and WS. Also, there was a correlation in AHA between the distinguishability of forms (both form elements and from text) and the likability of the form marking, but not between distinguishability and appropriateness. The other two interfaces showed correlations for both factors. Since

8. Figure 5 uses the average importance ratings across interfaces since there is no significance along this dimension, except for the strong tag.

9. Although there was no significance between interfaces, AHA was rated the highest in terms of distinguishability of address tags from text.

10. The cue was an abstract cue that was meant to sound like the sound used in Victor Borge's "Phonetic Punctuation."

Figure 5: Tag Importance

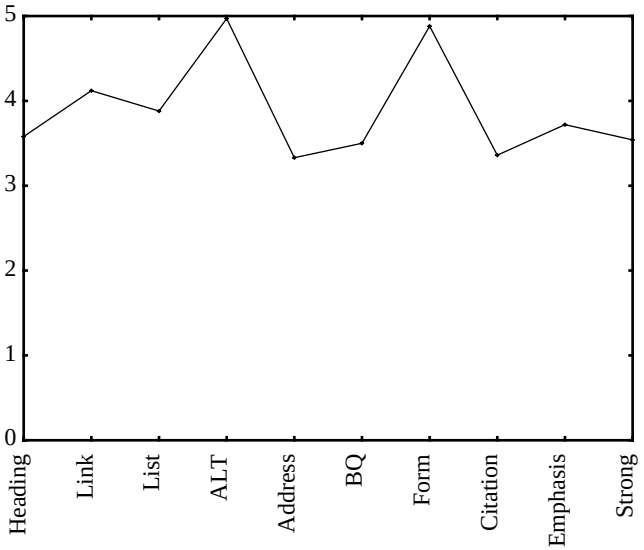
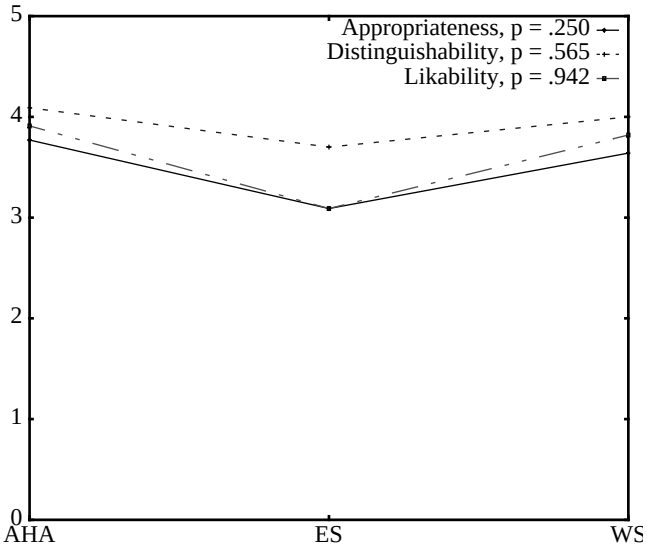


Figure 6: Blockquote Ratings



distinguishability was rated lower for ES and WS (see Figure 4), we again see that the appropriateness of the form marking is not based on distinguishability and that the high distinguishability (or distraction, in this case) in AHA contributed to its low likability rating for forms.

Subjects also said that the heading sounds used in AHA were too long and that the information in the cue that was needed to determine what heading level was being presented came too late in the sound. The appropriateness and likability of headings was basically the same across interfaces (see Figure 3). However, the distinguishability between heading levels in AHA correlates strongly to the appropriateness of the marking but not its likability, which is reversed in ES. Therefore, while users thought it was appropriate to clearly mark heading levels in this way, they did not particularly like the markings, presumably because they were distracted from the content text by having to wait for the cue to finish (or almost finish) before knowing what heading level was presented. This is opposed to the case of ES, where the heading level marking is made continuous via the voice change and the interpretation time is dependent solely on the user's ability to hear the inflection change and assign it to a heading level.

Relevancy of Sound

The relevancy of a particular sound can also influence how distracting it will be. Sounds that are completely unrelated to the tag being marked are more likely to cause distraction than those that are not.

If we look at the data regarding image markings in the three interfaces (Figure 7), we see that AHA was rated significantly higher than the others both in terms of appropriateness and likability. AHA used a camera sound to indicate images, where ES and WS both used spoken cues. The camera sound was short and quite distinctive, and subjects commented on how much they liked the sound, saying it was both short and very related to the marked tag.

In contrast, forms, which we have mentioned before, were rated lowest in AHA for appropriateness. Anecdotal data from study participants suggested, however, that not all of the musical sounds in the form markings were equal. In particular, the sound used to mark text fields, the "Jeopardy" theme, was commented on favorably. Presumably, the subjects who were familiar with Jeopardy recognized this sound as the signal to give a verbal response. The cue for a selection list was "Old McDonald", which was intended to be relevant (albeit a bit "cute") for hearing about a list of things, was not noticed as being relevant, as was also the case for the other musical cues that were not intended to be related to anything in particular.

Number of Occurrences of Sound

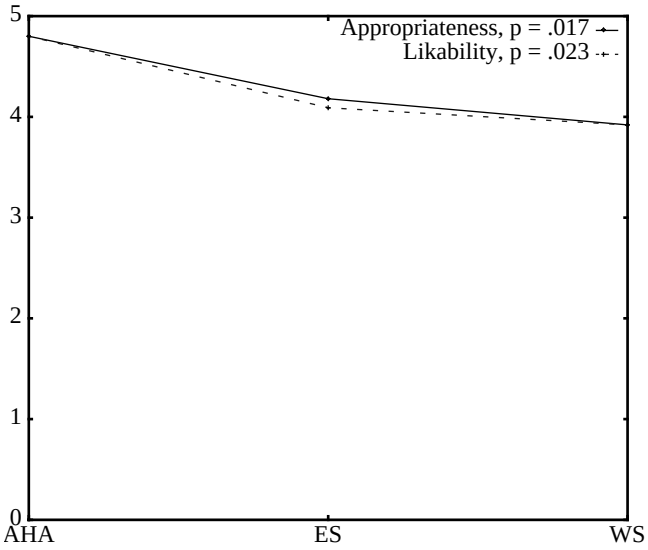
Some sounds in an auditory interface may just be played so frequently that they cause distraction. For example, for index pages (such as Yahoo!) where the text is mostly links, the link marking is played almost constantly. Several subjects complained about this, saying that the sound (in AHA, at least) was getting annoying because it was being played over and over again. It is clear that if a particular tag occurs frequently in a document, it should be marked with a cue that will cause the least amount of distraction.

Overlaid Sounds

Overlaid sounds in general would appear to be inherently distracting. There is evidence in the literature on shadowing [1] 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.

In this experiment, the link sounds for AHA, which overlaid the link text, were rated significantly lower for both appropriateness and likability than the other two interfaces. Subjects commented by saying that the link sound was "too long", which seems strange given that the sound was overlaid and therefore depended entirely on the length of the hyperlink text. This suggests that what people may have

Figure 7: Image Ratings



been reacting to was not the sound's length per se, but to the fact that the sound continued to be played even after the user had already processed and interpreted it to be marking a link. The additional playing time was redundant and caused distraction for the users.

CONCLUSIONS

In the AHA framework we proposed in [6], we maintained that the conventional split in audio interfaces between those using musical sounds and those using non-musical sounds was not the most important dimension for selecting sounds. We hypothesized instead that it was more important to look at sounds as being either recognizable or unrecognizable, and that this dimension was the most relevant. We were still left, however, with troublesome cases such as the usefulness of so-called "iconic" sounds and speaker changes, which are hard to define as "recognizable", and also recognizable sounds that are unrelated to the HTML tags they signify. In this experiment, we set out to determine where these border cases would be appropriate in audio interfaces by creating an interface using them and evaluating it against other marking techniques, such as overlaid natural sounds, spoken cues, and short musical themes.

What we have found in this experiment is that our initial assumption about recognizable sounds as well as our intuitions about iconic sounds, speaker changes, and musical themes, can all be better fit into a framework that starts with the foundation that first and foremost, a sound to be used in an audio interface should not cause undue distraction. Because audio has the inherent tendency to alert us and draw our attention, we must be careful when choosing sounds to avoid drawing attention when it is not necessary, and to only draw as much attention as the marked structure warrants. To accomplish this, we enumerate the following guidelines for selecting sounds in an audio interface:

- Keep the total number of sounds in the interface small.
- Choose sounds based on their potential to attract attention and correlate this to the amount of attention required by a particular document structure.
- Choose sounds that are aesthetically pleasing.
- Choose sounds that are short and that present relevant information early.
- Choose sounds that are clearly related to the structure being marked, either structurally or metaphorically.
- For commonly occurring document structures, choose unobtrusive sounds.
- Avoid overlaid sounds.

Clearly, these guidelines are both material- and individual-dependent. There may be people who find overlaid sounds quite useful, for example, or situations in which even document structures that occur frequently should be marked strongly to attract the user's attention. But, by following these guidelines, the sounds in an audio interface can be chosen in a way that reduces distraction in most cases. This will allow the user to focus firstly on the content text and then to gain structural information through their own attentional efforts, rather than having all of the structural information in a document forced upon them at once.

APPENDIX

The following section is a listing of some of the interface markings mentioned in this paper.

Headings

AHA marked heading levels using sequences of three tones:

- Heading 1: C2 E2 G2[11]
- Heading 2: C2 E2 E2
- Heading 3: C2 C2 C2
- Heading 4: C2 C2 G1
- Heading 5: C2 G1 G1
- Heading 6: C2 G1 C1

ES used voice inflection changes to mark the heading levels, and WS used the spoken cues "A heading entitled" (level 1), "A subheading" (level 2), and "A section heading" (all other levels).

Forms

AHA marked form elements using musical themes:

- text field: "Jeopardy" theme
- password field: "Twinkle Twinkle Little Star"
- checkbox: "Pop Goes the Weasel"
- radio button: "Mary Had a Little Lamb"
- select list: "Old McDonald"
- button: "Star Spangled Banner"

ES and WS both used spoken cues to mark form elements.

Links

AHA marked links to different file types using overlaid sounds:

- HTML file: phone ringing

11. C1 = 130.8 Hz, G1 = 196 Hz, C2 = 261.6 Hz, E2 = 329.6 Hz, G2 = 392 Hz

- text file: typewriter
- mailto: phone dialing
- sound file: old phonograph

ES used a speaker change to mark links, and WS preceded link points with the spoken cue “link”.

Addresses

AHA marked addresses by playing the sound of a stamp machine before and after address sections. ES used a voice inflection change, and WS used the spoken cue “address”.

ACKNOWLEDGMENTS

We would like to thank SONICON Development, Inc. for their help in setting up and running this experiment, as well as Cliff Nass, Hank Strub, and Terry Winograd for their help in analyzing and interpreting the data. Also, special thanks to all of the experiment participants for their time and effort.

REFERENCES

1. Banks, William P., et al. Negative Priming in Auditory Attention. *Journal of Experimental Psychology: Human Perception & Performance*, 21(6): 1354–1361, December 1995.
2. Dowling, W.J. and D.S. Fujitani. Contour, Interval, and Pitch Recognition in Memory for Melodies. *Journal of the Acoustical Society of America,* 49(2):524–531, 1971.
3. Geiselman, Ralph E. 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.
4. James, Frankie. Presenting HTML Structure in Audio: User Satisfaction with Audio Hypertext. *ICAD ‘96 Proceedings,* Xerox PARC, 4–6 November 1996, pp. 97–103.
5. James, Frankie. Presenting HTML Structure in Audio: User Satisfaction with Audio Hypertext. *CSLI Technical Report 97-201*. CSLI, Stanford University, January 1997.
6. James, Frankie. 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.
7. Productivity Works. pwWebSpeak, 1996. See http://www.prodworks.com/pwwebspk.htm.
8. Raman, T.V. Emacspeak—Direct Speech Access. *ASSETS ‘96: The Second Annual ACM Conference on Assistive Technologies,* New York, April 1996. ACM SIGCAPH, Association for Computing Machinery, Inc., pp. 32–36.
9. SONICON Development, Inc. marcopolo))), 1996. See http://www.webpresence.com/sonicon/marcopolo.
10. Yahoo! See http://www.yahoo.com.