Sonifying HTML Tables for Audio-Spatially Enhanced Non-visual Navigation

Jonathan Cofino, Armando Barreto, Fatemeh Abyarjoo, Francisco R. Ortega
Digital Signal Processing Laboratory – Department of Electrical and Computer Engineering
Florida International University
Miami, FL 33174 USA
{jcofi001, barretoa, fabya001, forte007}@fiu.edu

Abstract— **Computing is now ubiquitous. Accessing the Internet is increasingly common from mobile devices through cellular networks. The Internet is now used for shopping, communicating, and for finding employment and housing. More information is readily available than at any time in human history; however, this information is often inaccessible to people with blindness or low-vision. In order to increase access to information for these individuals, information must be presentable in non-visual formats. Presently, screen readers are able to verbalize on-screen text using text-to-speech (TTS) synthesis; however, this vocalization is inadequate for browsing the Internet. Modern operating systems have generally relied on a graphical display and a mouse for interfacing. For individuals with vision loss, we propose to create and test an auditory interface that incorporates an auditory-spatial orientation with a standard keyboard for input. If information can be structured as a two-dimensional table, each link can be semantically grouped as a cell in a row within the auditory table. This provides a consistent structure for non-visual auditory navigation. After testing the auditory display prototype with 13 participants, it was shown that stereo panning was an effective technique for audio-spatially orienting non-visual navigation through a five-row, six-column HTML table as compared to a centered, stationary synthesized voice. Rather than be left out of ubiquitous computing, blind and low-vision individuals may benefit from increased efficiency and accuracy in non-visual navigation. This finding may also stimulate interest in non-visual computing while performing other tasks, such as walking, eating, or driving.**

Keywords—Auditory Displays; Screen Readers; Sound Spatialization; Stereo Panning; Non-visual Navigation

I. INTRODUCTION

Millions of Americans experience blindness and low-vision within their lifetimes [1]. According to the provisional report for the 2010 National Health Interview Survey, 21.5 million American adults aged 18 and older reported some vision loss. Of these people, 16.1 million Americans between the ages of 18 and 64 and 5.4 million Americans 65 years and older reported significant vision loss. At least 1.5 million Americans with vision loss use computers. Of these 1.5 million computer users, most will rely on some form of assistive technology. Those with low-vision may elect to use screen-magnifying technology that acts as a digital magnifying glass in order to enlarge select regions of the user's screen. Those with less vision may elect to use a screen reader.

Current screen-reading technologies (JAWS, WindowEyes) have enabled blind and low-vision (BLV) computer users to experience the Internet [2]. Using text-to-speech synthesis (TTS), a screen reader can verbalize the prose of a typical website. While this technology has been a boon for BLV users, it is limited by the sequential, one-track nature of audio. Where a sighted user could visually scan through a website while considering where to focus or navigate, a non-visual user must experience the website in a serial fashion. Where sighted users have a persistent display, BLV users must hear and rehear information temporally. Considering that most contemporary webpages must now be navigated two-dimensionally, screen readers are increasingly challenged by the navigation requirements. Without added verbiage, screen-reader users are unaware of the presentation and structure of the content they are browsing. Many BLV users must develop sophisticated navigation strategies involving the sorting of links, headings, tables, and other HTML features. The BLV user must rely that a given HTML webpage has been both constructed properly and well organized.

When the context of the content is integrated into the presentation, the stream of audio can become bogged down by extraneous cues, which distract from the actual content of the website and become irritating to the user. It is the goal of this research to determine if the synthesized speech can be sonically enhanced in order to provide a needed sense of spatial orientation without overburdening the listener's cognitive capacity. We will consider using such audio-centric techniques as pitch shifting, voice changing, and auditory spatialization (stereo panning).

It is the aim of this research to show that auditory tables can be successfully navigated and would be an appropriate reorganization technique for screen-reader facilitation. Amorphously structured websites could benefit from tabular reformatting [3] where website content would be both structured and separated semantically from other content. Since many websites are cluttered with sidebar advertising, this organization would serve as a crucial navigation aid, where desired content could be targeted efficiently. Now, many screen-reader users switch to a "print-friendly" or a "mobile" version of a website for a more accessible solution in order to decrease clutter and distraction. Our proposed technique would accomplish this in a semantically meaningful way.

This research study proposes to exploit the sound localization abilities of blind persons in order to enhance their sonic browsing. Ohuchi et al [4] demonstrated that blind persons localize sound with greater acuity, on average, when compared with sighted persons. To obtain these experimental

results, 12 speakers were spaced evenly at 30° intervals along a circular array in the azimuthal plane. A listening test subject was placed in the center of this array. Each sound was played from exactly one of the twelve different physical locations (speakers) at a time. While the localization results from these experiments are impressive, it would be impractical to expect a typical user to set up such a large array of speakers in a home or office. Since most users typically have a stereo audio playback setup, our spatial-sonic browser will not utilize multichannel audio to ensure practicality and promote widespread adoption.

Auditory spatialization can be implemented through stereo panning, where each of the stereo speakers emits a coherent signal with a variation of amplitude. This amplitude difference creates the perception of a single auditory source emanating from a virtual position located between the actual speakers. By varying the relative amplitudes, the virtual source can be "positioned" anywhere on a sonic continuum between the two loudspeakers.

II. PREVIOUS WORK

A. Other Researchers

Goose and Möller [5] implemented an audio browser that could be accessed non-visually, possibly by a telephone or while driving an automobile. Their paper describes how an HTML document could be sonically traversed: the column of text would be represented as a "stage-arc" where the beginning (screen upper-left) is represented as the left-most end of the arc, and the right-most end of the arc represents the end of the text (screen lower-right). Intra-document linking is sonified by an aircraft metaphor: a lift-off sound connotes leaving the current position, then the audio is panned in the direction of the destination content, then a landing sound effect connotes reaching the targeted link location. BLV users are given a sense of the length of a document through auditory-spatial cuing much the same way that a sighted person is given this sense through a scrollbar.

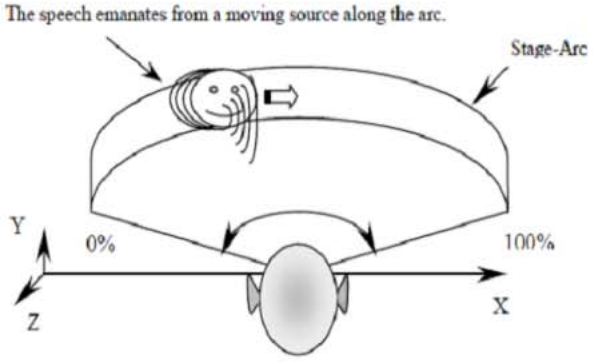

Figure 1. Goose and Möller's stage-arc sonification.

B. Previous Studies

In a previous (unpublished) study, an "audio-column" implementation of the sonic tabular browser was attempted. The content model was that of a newspaper, where each tabular column would represent a traditional newspaper section (Sports, Business, Lifestyle, etc.) and occupy a distinct vertical slot in auditory space. It was reasoned that vertical spatialization could be represented through such techniques as pitch-shifting the voice or using different voices. Research [6] has shown that individuals commonly associate variations of pitch with the corresponding sensation of vertically rising and falling, which would suggest a natural sonic-navigation approach with pitch variation (granted this is with pitched tones, not voices). When this approach was attempted with BLV subjects, many found the pitch-shifted synthesized speech to be interesting, yet their average times-to-target (TTT) did not seem to decrease. It seems that changing the quality of the synthesized speech is disturbing and may confuse the listener when she is navigating.

Through experimentation and user feedback, we have found that the most distinguishing characteristic of sound that is not also distracting is horizontal spatialization, or stereo panning. Pitch shifting or using differing voices serves as a distraction to the listener-navigator. Any inconsistency in the voice quality seems to serve as an impediment to enhanced navigation. In addition, finding a consistent pitch interval that is both sufficiently large without making the first/last pitches too extreme is subjective for each listener. Finding a characteristic of sound that could be intuitively perceived without too much extra cognitive loading proved to be nontrivial.

One misstep in previous testing was to use current events as a means for generating a data table of news headlines. Participants were asked to find a correct headline based on a question relevant to that headline. Cultural differences as well as varied familiarities with local institutions and current events hampered the navigation in this testing. It was decided to use simplified tables, such as those that may occur in a typical (online) shopping task instead of newsgathering. It is our aim to have the content used to populate the tables to be as universally familiar as possible.

In previous trials, it was wrongly assumed that training could be accomplished through simple verbal explanation and that the spatialized/pitch-shifted voice synthesis would be self-explanatory. This assumption proved troublesome and a source of confusion. It became clear that extensive training of all relevant table navigation functions would be necessary. For ease of testing, all navigation between testing modules was made seamless and automatic. The recording of data needed to be accomplished automatically and unobtrusively. To accomplish these two objectives, we have designed a PHP/MySQL database to automatically collect all navigation data and seamlessly transition the subjects from one auditory table to another.

III. METHODS

A. Testing

In this testing implementation, sighted subjects were tested. To simulate blindness, an opaque folder covered the laptop's screen. It was determined that the sighted subjects should not have been blindfolded, as that would hamper their ability to operate the standard keyboard of the laptop. The subjects were trained to use the non-visual table browser and it was explained that they would be simulating an online purchasing task.

B. Description

TABLE I. GROCERY SHOPPING TABLE

Vegetables	Carrot	Potato	Tomato	Onion	Lettuce
Fruits	Banana	Apple	Lemon	Orange	Coconut
Bakery	Bread	Cake	Pie	Cookie	Muffin
Meat	Beef	Chicken	Pork	Turkey	Duck
Drinks	Beer	Juice	Milk	Soda	Tea

A short tone was prepended to the synthesized speech. This tone will serve to simultaneously alert the listener-navigator to the existence of a hyperlink (as implemented in ChromeVox, an extension to Google's popular Chrome browser) as well as to provide a mechanism for varying pitch without altering speech quality. This additional tone may add duration to the temporal stream but should enhance the spatial mapping of the table for the blind user, leading to greater navigational efficiency and orientation.

As a sample, a table of groceries is shown in Table I. To address the challenge of vertical audio spatialization, the categorical tables were inverted: the tables are categorically grouped by rows rather than by columns. This has one distinct advantage in the vertical dimension: the category/rows are now self-differentiating through semantics. Horizontal spatialization can now help orient the listener-navigator where there is a lack of natural semantic or alphabetic ordering. No voice modification or vertical spatialization is necessary.

To evaluate the efficacy of the system, the subjects were evaluated on their navigational performance: their time(s)-to-target (TTT) and number of moves-to-target were recorded. In addition, subject error and confusion were recorded. On a per link basis, each navigation path was tracked, noting how many moves were necessary to find the target cell/link. Along each path, the numbers of wrong links and inappropriate key presses were recorded. In addition, the number of times that a subject attempts to move outside the boundary of the table was noted.

To help guide the listener-navigator, an "electrified-fence" sound effect will be heard whenever the subject attempts to move beyond the table's boundaries. This sound effect acts as an auditory metaphor and helps to reinforce the containment of the tabular navigation. Constant auditory feedback was used to keep the listener-navigator aware of her position within the table and to alert her to the status of her search.

C. Training

An extensive training module has been designed. The three major keyboard functions are trained independently. In order to return to the category column that precedes the data cells, the subject is instructed to move into the data cells and press the 's' key to return herself to the category cell of her presently focused row. This is repeated for each row in order to ingrain the importance of this crucial shortcut. Next, a brief search exercise is conducted. The subject must press the 'r' key to read/recall the instruction for table navigation. After (re)hearing the instruction, she must proceed to navigate the table using the directional arrow keys to find the proper cell, confirming her choice by pressing the 'space bar.' Through this exercise, it should be clear how to proceed in the testing phase. In order to recall the categories within the table, the subject must press the 'c' key. The categories will be read in order proceeding from top to bottom.

D. Audio Spatialization

Stereo panning and tonal variation will be implemented to horizontally spatialize the columns of the data table. Stereo panning is a technique exploiting the psychoacoustic phenomenon that causes a listener to perceive a 'sound source' emanating from between two loudspeakers. An auditory arc can be conceptualized as a continuum through which an auditory source can be spatialized. Tonal variation involves using a pitched tone to represent progress through a continuum. The listener perceives frequencies relative to one another, causing a sense of a musical scale that can be used to represent a distinct series of pitches to represent movement.

While most listeners cannot identify pitches absolutely by their frequency (perfect pitch), most listeners have a sense of relative pitch, i.e. sensing when one frequency is higher or lower than another one. The two variables, panning and pitch variation, will be toggled as shown in Table II.

TABLE II. TABLE SPATIALIZATION METHODS

Table Letter		**Panning**	
		Off	**On**
Pitch Variation	**Off**	A	B
	On	C	D

Table II shows that Table A serves as a global control where panning and tonal variation are turned off. Tables A and B serve as controls to Tables C and D, respectively, with regard to pitch variation. Tables A and C serve as controls to Tables B and D, respectively, with regard to stereo panning.

In our previous work [7], we showed that there was little difference in the technique used for stereo panning. The playback medium is much more important. Listeners often report the sensation of hearing sound sources "move within their heads" when wearing headphones while stereo loudspeakers effectively externalize sound, which yields a wider range of spatialization.

IV. RESULTS

TABLE III. TIME-TO-TARGET (TTT)

TTT	Mean	Std. Deviation
A	13.513	3.633
B	10.112	2.022
C	11.251	3.643
D	11.539	3.825

TABLE IV. MAUCHLY'S TEST OF SPHERICITY

Measure: Spatialization_Method

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Table	0.431	9.034	5	0.109

TABLE V. TESTS OF WITHIN-SUBJECTS EFFECTS

Measure: Spatialization_Method

Source		F	Sig.
Table	Sphericity Assumed	3.6	0.023
	Greenhouse-Geisser	3.6	0.043
	Huynh-Feldt	3.6	0.033
	Lower-bound	3.6	0.082

TABLE VI. PAIRWISE COMPARISONS

(I) Table	(J) Table	Mean Difference (I-J)	Std. Error	Sig.[a]
A	B	3.401*	1.027	0.037
	C	2.262	1.114	0.391
	D	1.974	0.851	0.233
B	A	-3.401*	1.027	0.037
	C	-1.139	0.754	0.942
	D	-1.427	1.071	1.000
C	A	-2.262	1.114	0.391
	B	1.139	0.754	0.942
	D	-0.288	1.388	1.000
D	A	-1.974	0.851	0.233
	B	1.427	1.071	1.000
	C	0.288	1.388	1.000

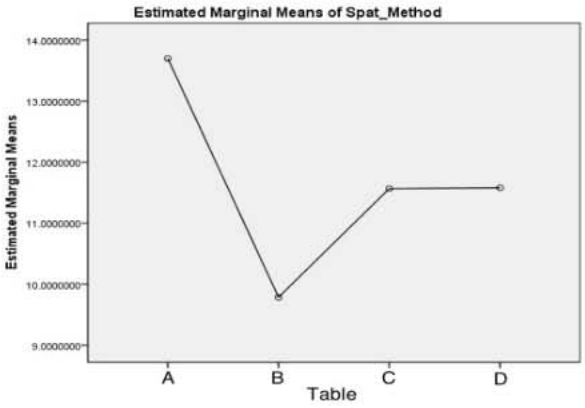

Figure 2. Table spatialization method means.

V. ANALYSIS AND DISCUSSION

As shown in Table V, the effect of spatialization method is significant at the p=0.05 level. Since Mauchly's test of sphericity is statistically non-significant (p=0.109), the within-subject effects of audio spatialization and frequency shifting can be treated without correction for sphericity. Looking at post-hoc pairwise comparisons with Bonferroni correction, Table A (no spatialization) had significantly higher TTT than did Table B (stereo panning only). Tables C and D (both with tonal variation) had very similar TTT on average, with Table B having the fastest TTT, slightly faster than Tables C and D.

As shown in Fig. 2, the control Table 'A' (no panning, no pitch variation) has the longest average TTT of the four audio spatialization methods. This method has the least amount of auditory guidance and serves as a control method to the two audio guidance methodologies. Effectively, Table A represents the current state of the art with unmodified TTS. Table B, the audio panning only approach had the shortest TTT of the four tables. Some test subjects noted that the sense of horizontal space let them know the boundaries of the table without necessarily hearing the boundary sound. It seems that stereo spatialization is more readily perceived than is pitch/tonal variation.

A particularly challenging aspect of this undertaking is the issue of speech synthesis clarity. Many participants remarked that certain words were either unclear or oddly uttered. Without semantic context, the synthesized speech can often sound bizarre with emphasis on an inappropriate syllable, and one-syllable words may sound completely distorted. In testing the sonified tables as an exercise in purchasing items, there is no escaping the fact that different individuals relate and categorize common objects differently. Many remarked that "a tomato is a fruit, not a vegetable" or that "a vacuum is a kitchen appliance, since that's where I keep mine." It is nearly impossible to account for everyone's cultural and experiential differences with regard to semantic meaning and categorization. Some speakers of other languages were tested, and unfamiliarity with vocabulary posed issues with them.

With regard to pitch variation, one subject remarked that he would have preferred to have heard a diatonic scale as opposed to the wider intervals implemented. Other researchers

have experimented with pitched intervals and have found that it may be somewhat arbitrary to determine a suitable width depending on the number of intervals desired and the total frequency range.

VI. CONCLUSION

We have determined that stereo panning is a potentially effective enhancement of tabular non-visual browsing as performed with a screen reader. Horizontal audio spatialization was perceived readily after minimal training while maintaining the quality of the synthesized speech. We suggest that stereo panning does not necessarily overburden the cognitive abilities of the listener-navigator and leads to increased non-visual navigational efficiency and accuracy.

VII. FUTURE WORK

It has been widely accepted that a typical person can only recall "seven plus or minus two" items [8]. As newer information is assimilated, she begins to forget older information, much like a buffer. In keeping with this axiom, we have limited our tables to five columns and five rows. Sophisticated algorithms will be needed in order to segregate website content into tables of similar dimensions. It may also be possible to "chunk" this information by having tables of tables; in other words, each cell in the primary table contains a link to yet another table, similar to hierarchical telephone menus.

It would be desirable to integrate the techniques used in this study with a broader screen-reading application in order to facilitate more practical web browsing. Using the screen reader available, a blind individual could navigate into a table and then begin to use these techniques.

As two speakers yield a linear continuum of spatialization, we have considered implementing a four-speaker rectangular array to make an actual plane or wall of sound, directly analogous to the two-dimensional table. The benefits of this implementation would need to be weighed against the costs of increased computational burden, specialized software, additional drivers, as well as the impracticality of a multi-channel array of speakers that must be placed precisely in a rectangle.

Headphone based approaches may be studied. Head-related transfer functions (HRTF) allow for audio processing such that a headphone-wearing listener may experience the psychoacoustic sensation of 360° audio swirling around her head. One limitation of this technique relates to the need for determining each listener's HRTF based on their unique pinnae. Some members of the BLV community have reservations about headphone use, as it may restrict awareness of their general environment and surroundings, although this does not seem to be a universally supported opinion. On the other hand, headphones provide privacy while loudspeakers alert others to the activities of the BLV user. Such decisions should ultimately be made by the BLV user herself.

ACKNOWLEDGMENTS

This work was sponsored by National Science Foundation grants HRD-0833093, and CNS-0959985.

The authors of this paper would like to extend their profound gratitude and appreciation to:

- Lighthouse of Broward in Fort Lauderdale, FL
  - http://www.lhob.org/staff-directory
- Disability Resource Center at Florida International University in Miami, FL
  - http://drc.fiu.edu/
- Miami-Dade County Medical Examiner Laboratory Staff
  - http://www.miamidade.gov/medicalexaminer

REFERENCES

[1] "Living with Vision Loss." *American Foundation for the Blind.* N.p., n.d. Web. 12 Jan. 2013. <http://www.afb.org/section.aspx?FolderID=2>.

[2] M.A. Hersh and M.A. Johnson. "Screen Readers." *Assistive Technology for Visually Impaired and Blind People.* London: Springer, 2008.

[3] Y. Hwang, J. Kim, and E. Seo. "Structure-aware Web Transcoding for Mobile Devices." IEEE Internet Computing 7.5 (2003): 14-21. Print.

[4] M. Ohuchi, Y. Iwaya, Y. Suzuki, and T. Munekata. "A Comparative Study of Sound Localization Acuity of Congenital Blind and Sighted People." Acoustical Science and Technology 27.5 (2006): 290-93.

[5] S. Goose and C. Möller. 1999. A 3D audio only interactive Web browser: using spatialization to convey hypermedia document structure. In Proceedings of the seventh ACM international conference on Multimedia (Part 1) (MULTIMEDIA '99). ACM, New York, NY, USA, 363-371.

[6] Roffler, Suzanne K., and Robert A. Butler. "Localization of Tonal Stimuli in the Vertical Plane." The Journal of the Acoustical Society of America 43.6 (1968): 1260-6. Print.

[7] Cofino, J.; Barreto, A.; Adjouadi, M.; "Sonically spatialized screen reading: Aiming to restore spatial information for blind and low-vision users," Southeastcon, 2012 Proceedings of IEEE , vol., no., pp.1-6, 15-18 March 2012

[8] Miller, George A. "The Magical Number Seven, plus or minus Two: Some Limits on Our Capacity for Processing Information." Psychological Review 63.2 (1956): 81-97. Print.