5-22-2014

Using Sonic Enhancement to Augment Non-Visual Tabular Navigation

Jonathan M. Cofino
Florida International University, jcofi001@fiu.edu

Armando Barreto
Florida International University, barretoa@fiu.edu

DOI: 10.25148/etd.FI14071117

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

USING SONIC ENHANCEMENT TO AUGMENT NON-VISUAL TABULAR NAVIGATION

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ELECTRICAL ENGINEERING

by

Jonathan Cofino

2014

To: Dean Amir Mirmiran
College of Engineering and Computing

This dissertation, written by Jonathan Cofino, and entitled Using Sonic Enhancement to Augment Non-Visual Tabular Navigation, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Malek Adjouadi

Jean Andrian

Naphtali Rishe

Armando Barreto, Major Professor

Date of Defense: May 22, 2014

The dissertation of Jonathan Cofino is approved.

Dean Amir Mirmiran
College of Engineering and Computing

Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2014

DEDICATION

To my parents and grandparents, who always instilled in me the value of education through their own examples. The role that their guidance, patience, understanding, and unwavering support played in my life can not be overstated.

ACKNOWLEDGMENTS

I would like to express the most sincere gratitude for my major Professor, Dr. Armando Barreto, for his guidance and support in navigating the research process. The National Science Foundation supported the research described in this dissertation through grants HRD-0833093 and CNS-0959985. Finally, I would like to recognize my committee members Drs. Malek Adjouadi, Jean Andrian and Naphtali Rishe for their expertise and support.

Over the last few years, many individuals have provided me support, friendship and inspiration. Scott Schiller, the creator of the SoundManager2 JS API, provided the framework for implementation of dynamic sound playback within a browser, without which this project would never have existed. Eric Barrette and Elly du Pré, of the Lighthouse of Broward, R. David New of the Miami Beach Council of the Blind, as well as the staff of the Lighthouse of the Palm Beaches and the FIU Disability Resource Center provided the invaluable participants, without whom this research would have been meaningless. I am particularly grateful to Fatemeh Abyarjoo, Jian Huang, Francisco Ortega, and other graduate students at the FIU DSP Lab, with whom I have experienced and shared all of the trials and tribulations that a graduate student encounters in his career. Finally, and most importantly, I would like to thank my parents and grandparents, without whose love and encouragement the completion of this endeavor would never have been possible.

ABSTRACT OF THE DISSERTATION

USING SONIC ENHANCEMENT TO AUGMENT NON-VISUAL TABULAR NAVIGATION

by

Jonathan Cofino

Florida International University, 2014

Miami, Florida

Professor Armando Barreto, Major Professor

More information is now readily available to computer users than at any time in human history; however, much of this information is often inaccessible to people with blindness or low-vision, for whom information must be presented non-visually. Currently, screen readers are able to verbalize on-screen text using text-to-speech (TTS) synthesis; however, much of this vocalization is inadequate for browsing the Internet. An auditory interface that incorporates auditory-spatial orientation was created and tested. For information that can be structured as a two-dimensional table, links can be semantically grouped as cells in a row within an auditory table, which provides a consistent structure for auditory navigation. An auditory display prototype was tested.

Sixteen legally blind subjects participated in this research study. Results demonstrated that stereo panning was an effective technique for audio-spatially orienting non-visual navigation in a five-row, six-column HTML table as compared to a centered, stationary synthesized voice. These results were based on measuring the time-to-target (TTT), or the amount of time elapsed from the first prompting to the selection of each tabular link. Preliminary analysis of the TTT values recorded during the experiment showed that the populations did not conform to the ANOVA requirements of normality and equality of variances. Therefore, the data were transformed

using the natural logarithm. The repeated-measures two-factor ANOVA results show that the logarithmically-transformed TTTs were significantly affected by the tonal variation method, $F(1,15) = 6.194$, $p = 0.025$. Similarly, the results show that the logarithmically transformed TTTs were marginally affected by the stereo spatialization method, $F(1,15) = 4.240$, $p = 0.057$. The results show that the logarithmically transformed TTTs were not significantly affected by the interaction of both methods, $F(1,15) = 1.381$, $p = 0.258$. These results suggest that some confusion may be caused in the subject when employing both of these methods simultaneously. The significant effect of tonal variation indicates that the effect is actually increasing the average TTT. In other words, the presence of preceding tones increases task completion time on average. The marginally-significant effect of stereo spatialization decreases the average log(TTT) from 2.405 to 2.264.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

Computing has become ubiquitous as commercial, educational, and recreational activities have all shifted to the digital world. Much of this computing involves a graphical user interface (GUI), and this interface typically employs a windows, icons, menus, pointer (WIMP) style of interaction. GUIs necessitate a feedback loop for successful interaction: the sighted person manipulates a pointer (realized visually as a cursor), which the sighted user tracks visually across a screen, often to click on icons (which are inherently visual), which further launches applications or initializes processes. Throughout these interactions, graphical (and to a lesser extent, sonic) output is used to convey the status and progress of the operating system at any given time. For individuals who are blind, these visual cues are inaccessible.

Screen readers have become the go-to solution for visually impaired users using computers. While they serve as an invaluable tool for these users, the sometimes excessive verbosity of the synthesized speech can overwhelm the listener's cognitive memory and patience. Extraneous messages about status and format can crowd out the intended content. It is desirable to seek a non-verbal method for conveying status information that is both short in duration and unobtrusive to the listener.

1.2 Objective of Research

This research aims to determine whether sonic enhancements can augment the non-visual browsing experience of a blind user navigating a webpage. Two specific enhancements, stereo spatialization and tonal variation, will be employed to enhance

the synthesized speech that dominates a typical screen-reading system. Stereo spatialization will serve as an acoustical cue suggesting an auditory-spatial relationship that exists inherently in the visual-spatial presentation. The tonal variation technique is another method of conveying progression and differentiation through a continuum of space.

1.3 Significance of this Research

Auditory interfaces allow for non-visual computer usage. They would make computing more accessible to the blind community. Also, individuals who interact with the Web while traveling may wish to navigate non-visually; for example, a jogger may want to focus on the path ahead as opposed to menus and submenus.

1.4 Structure of the Dissertation

The rest of this dissertation is organized according to the following structure.

Chapter Two describes the background research, concepts, and terminology pertinent to this dissertation. The various forms of screen-reading technology are introduced. In addition, the synthesis of speech and the theory of tables are discussed. Previous works studying the efficacies of earcons and auditory icons are explored. Later, tonal variation and stereo spatialization are introduced as the two main factors in this research study. Earlier studies examining non-visual tabular navigation are reviewed. Finally, transcoding, the process of reformatting web content, is examined.

Chapter Three details the methodology used to conduct the research experiment. Prosodic modification using the PSOLA algorithm is described in detail. The auditory cues used in the experiment are listed and detailed. The two main sonic augmenta-

tion techniques, stereo spatialization and tonal variation, are explained. Finally, the experimental design strategy is presented.

Chapter Four describes the experiment design. It details the implementation of the software packages as well as the hardware components used in the experiment. The tabular content presented to the subject participants is listed. The training and testing phases of the experiment are detailed. Finally, the demographic information of the subject participants is presented.

Chapter Five presents and discusses the results of the research experiment. Insights gleaned from the personal interviews with the subject participants are shown. The quantitative analysis of the experimental results is elaborated in three sections: tonal variation, stereo spatialization, and the interaction of these two factors. Qualitative feedback is reported. Finally, the discussion of the results is presented.

Chapter Six reports the conclusion of this research study and discusses how this research may potentially be furthered in the future.

CHAPTER 2

BACKGROUND

This chapter provides the background information for this dissertation. It begins with an overview of screen readers, followed by a description of various technologies and techniques employed in the sonification of interfaces. In the final section, the basis for constructing tables and navigating tabular data is explored, and methods for transcoding raw visual content into highly organized and annotated content are covered.

2.1 Screen Readers

Screen readers are software applications that are capable of re-interpreting a computer's standard output from a visual presentation into a narrative audio presentation. Typically, this audio representation is synthesized speech (text-to-speech, or TTS) in conjunction with sound icons (earcons, auditory icons described in Sections 2.4.2 and 2.4.1, respectively). This assistive technology (AT) approach can be contrasted with Braille outputs and screen magnifiers, which of course require at least some visual capability.

2.1.1 Screen Reading Applications

Most of the more popular operating system (OS) platforms provide a screen reader. The Microsoft Windows environment has included Microsoft Narrator since the Windows 2000 OS. The Macintosh OS X platform provides VoiceOver, which is also present in the iOS mobile platform used in the ubiquitous iPhone and iPad devices. BlackBerry and Android mobile devices provide their own mobile screen reading applications as well. Commercially-developed screen readers include JAWS from Freedom

Scientific, Window-Eyes from GW Micro, and the ZoomText Magnifier/Reader from AiSquared, while the NonVisual Desktop Access (NVDA) provides an open-source alternative.

2.1.2 Types of Screen Reader

Command-Line Interface (CLI) Screen Readers

CLIs utilize text as the primary input/output methodology. Commands can be typed as text using a keyboard, while output is typically displayed as text on a visual monitor. The one-dimensional nature of this text allows a screen-reader to represent an auditory narration of typed commands and textually-displayed output.

Graphical User Interface (GUI) Screen Readers

GUIs use graphics and characters arrayed two-dimensionally on a pixellated screen. Typically, a user will operate a point-and-click device like a mouse to launch applications and processes by visually guiding a cursor to an on-screen object and selecting it with a click. This visual arrangement of icons resists a simple narrative overview: visual icons persist over time, while auditory narration is transient in nature and lacks spatial relevance.

Self-Voicing Applications (SVA)

SVAs provide an auditory interface without the presence of an external screen reader. A significant group of SVAs include talking web browsers. These include Connect Outloud from Freedom Scientific and WebAnywhere from the University of Washington. Recently, mainstream browsers have included self-voicing capabilities, including the Fire Vox extension for the Mozilla Firefox web browser.

Mobile Web-based Applications

With the growing ubiquity of mobile devices, both sighted and visually-impaired users are increasingly taking advantage of spoken-interactive features of their devices. Examples of this technology include Siri for Apple's iOS platform and Google Now for its Android platform. By using the user's speech input and the device's synthesized-speech output, the need for a visual display is obviated. Presently, the capabilities for this technology are somewhat constrained by the limited nature of speech, typically the intelligent personal assistant answers short questions or launches a task.

2.2 Synthesized Speech

Synthesized speech is the artificial generation of human speech. Both hardware and software have been employed for speech synthesis. A typical application of this technology is text-to-speech (TTS). Effective TTS systems are able to convert text into human-like, intelligible speech. A typical TTS system includes three phases: textual analysis, linguistic analysis, and wave form generation, as shown in Figure 2.1. The text analysis phase serves as a pre-processing step which turns symbols like abbreviations and numbers into written-out words. Once the raw text has been fully tokenized, each word is phonetically transcribed into prosodic units such as phrases, clauses, and sentences. Prosody refers to rhythm, stress, and intonation of speech. Together, phonetic transcription and prosodic information create the symbolic representation used by the synthesizer to create the sound of the speech.

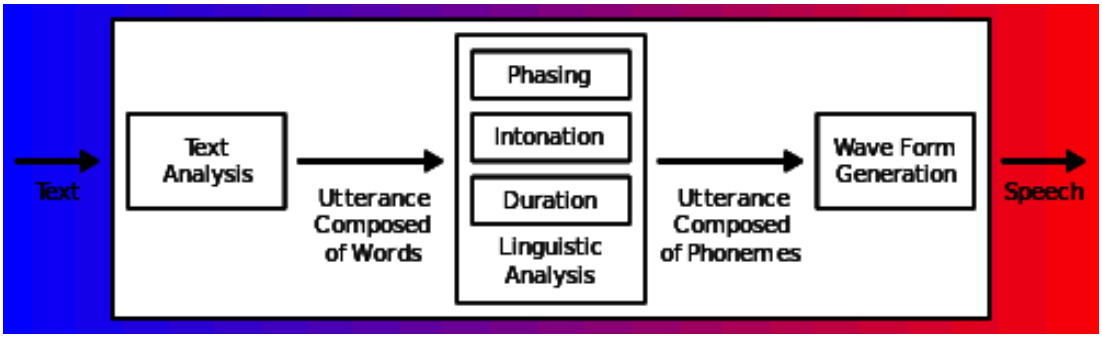


Figure 2.1: Overview of a typical TTS system.

Table 2.1: Tabular Calendar Example - January 2014

Sun.	**Mon.**	**Tue.**	**Wed.**	**Thu.**	**Fri.**	**Sat.**
x	x	x	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	x

2.3 Tables

Tables are composed of simple components, which can be organized into increasingly more complex structures. The fundamental building block of any table is the cell. Cells may contain data or labels, which describe the content of other cells. Cells can be grouped as arrays together in rows or columns. These arrays of cells are often associated with a label cell known as a header cell. These arrays can be visually grouped using boundary lines, which also frequently serve to frame the entire table and separate header cells from data cells. Table 2.1 is a monthly calendar, using column headers that clearly delineate which calendar days are particular days of the week.

Table 2.2 depicts the sales revenue of three different fast-food chains. This table contains both row and column headers, where each cell is the intersection of a row and column. Practically, this means that each data cell is the sales revenue resulting from a singular food item at a particular chain.

Table 2.2: Numerical Table Example - Sales in Millions ($)

x	Hamburgers	Fries	Broccoli Poppers
McDoogle's	4.3	3.2	0.0
Sausage Factory Co.	0.0	7.8	0.0
Salad Shack	0.0	2.1	9.4

2.4 Mapping Sound to Data and Events

While the visual modality tends to dominate the display of information, interfaces may be enhanced by the utilization of the auditory channel to supplement or replace the visual channel. Gaver suggests three types of sonic representations as extensions to the use of graphical icons to represent objects visually [Gav86]:

- **Symbolic-arbitrary mappings**: This representation relies on learned social conventions. The sound of a telephone ringing may signal an attempt to connect to a communications channel, while the sound of a slamming door may signal the closing of a program or process.
- **Nomic mappings**: This representation relies on the physical properties of sound. For example, "dragging" a computer file across a desktop may be aurally represented with frictional, abrasive sound.
- **Metaphorical mappings**: This representation uses similarities between the object and its representation. For example, as a user descends a textual document using a scrollbar, he or she may simultaneously hear a descending glissando note that connotes the sensation of falling.

The mappings listed above serve as a basis for auditory icons, which are further described in Section 2.4.1. While these mappings can create meaningful sonic relationships, many computing processes and events resist sonic symbolization. Earcons, described in Section 2.4.2, are more flexible in that they are not reliant on analogies and can be arbitrarily composed.

2.4.1 Auditory Icons

Auditory icons serve as sonic metaphors posing for visual icons representing computing events and tasks. Seeking to justify auditory icons as intuitive, Gaver stated that "natural sounds are related to events in a principled, systematic way (described by physics), and people learn this mapping from early childhood in their interactions with the world [Gav86]." Implemented in his auditory interface, SonicFinder, Gaver uses the sonic qualities of materials to relate the type and size of files [Gav89]. For example, text files may sound wooden, and larger files may generate louder sounds. Opening a file would be conveyed with a tapping sound, as one may tap an object in the physical world. Dragging a file object would be conveyed as a scraping sound, while copying would be analogized to a pouring sound.

2.4.2 Earcons

Blattner *et al.* defined earcons as "nonverbal audio messages used in the user-computer interface to provide information about some computer object, operation, or interaction [BSG89]." The earcon technique involves using composed melodies/rhythms to indicate information [Sum85, BRK96]. Typically, earcons are made up of short, rhythmic motives that can be combined to produce more complex earcons. A motive is defined as "... a brief succession of pitches arranged to produce a rhythmic and tonal pattern sufficiently distinct to allow it to function as an individual, recognizable entity [Ber05]." Blattner *et al.* define rhythm and pitch to be the *fixed parameters* of motives, while timbre (tone color), register (range of pitches), and dynamics (the relative volume of a sound or note) are the *variable parameters* of motives [BSG89]. This may be accomplished through concatenation of the earcons, known as compound earcons.

Earcon Studies

In a study evaluating sonified tables and earcons, the response time of listeners answering questions about data was reduced by 10 percent when the auditory enhancement was used [SSK09].

Vargas and Anderson investigated a sonified automobile interface for controlling an automobile's accessory functions [VA03]. Their interface implemented a hierarchical menu structure using earcons as an enhancement of the speech describing the menu items aloud. Navigating with directional arrow keys, test users were exposed to the interface in two variations, with and without earcons. The earcons (when present) preceded the speech, as the researchers found that simultaneous playback of earcons and speech cause the overall playback to heard unclearly. At the top-level of their auditory menu, each item received an earcon representation as a different instrument (timbre) and motif (chord, or grouping of differently pitched notes). For example, the "lights" family timbre is sonified as a piano; the "windshield wipers" family timbre is sonified as a chorus; the "ventilation" family uses bells for sonification; the "radio" family sounds like horns. As users descended the hierarchy, menu items on lower levels inherited the timbre and motif from their parent node and were differentiated using melody and rhythm.

Thirty-six participants were randomly placed into either the Speech-Only or Speech-and-Earcons Group. Participants were trained using a graphical version of the interface by performing five practice tasks in addition to self-guided exploration. On average, task completion under the Speech-and-Earcons condition took an extra 18% of time (statistically significant) as compared to the Speech-Only condition. The researchers estimated that earcons extend the duration of speech alone by 90%. However, the number of keystrokes needed to complete the tasks increased by 15% (statistically significant) in the Speech-Only condition, as compared to the Speech-and-

Note	A	A#	B	C	C#	...	F	F#	G
MIDI pitch	0	1	2	3	4	...	125	126	127

Table 2.3: MIDI pitch parameter to musical note mapping.

Earcon condition. To look at the efficiency of either method, the keystroke rates for the Speech-Only condition was 1.0 keys/sec, while the keystroke rate for the Speech-and-Earcon condition was 0.75 keys/sec. This implies that the addition of earcons actually is more efficient, as less keystrokes are necessary per unit time.

2.5 Tonal Variation

Ramloll *et al.* explored a tonal variation approach based on a study involving numerical tables and pitches [RBYR01]. In their study involving both blind and sighted individuals, they explored a relationship between pitch and numerical trends. Although the relationship between frequency and pitch (the psychoacoustic perception of frequency) is non-linear and user-specific [SVN37], relationships between pitches are recognizable and can be exploited. By mapping an arbitrary range of numbers to a MIDI scale, melodic sequences were created in order to give a sense of numerical trend to a non-visual user. This mapping is expressed in Equation 2.1:

$$y = \left(\frac{x - S_{min}}{S_{max} - S_{min}} \right) \times 127 \tag{2.1}$$

where y is rounded to the nearest integer and is the value of the MIDI pitch parameter representing an integer x in an arbitrary set of integers ranging from S_{min} to S_{max}. The result of this mapping can be seen in Table 2.3.

2.6 Spatial Audio

2.6.1 Lateralization

Human beings are able to localize a sound source's angular position. This is accomplished by comparing the wavefront at the two ears in the horizontal plane [B+94]. The relative differences between these two wavefronts are quantified in terms of interaural time differences (ITDs) and interaural intensity differences (IIDs). Looking at Figure 2.2, the sound source at position **A** is azimuthally located at 0°, which is directly in front of the listener. Since the path lengths from the sound source to each of the ears are equidistant, the wavefront arrives at both ears with equal intensity and at the same time. The sound source located at position **B** is shifted 60°clockwise from position **A**. The paths, no longer being equal, cause the sound source waveform to arrive delayed at the left ear relative to the time that it arrives at the right ear.

The aural cues of ITD and IID are frequency dependent. Below around 1 kHz, ITDs are useful for detecting path length differences. For frequencies above about 1.5 kHz (smaller wavelengths), the human head acts as an obstacle, where an ear is "shadowed" from a source emanating from the opposite side of the head. This "shadowing" is the result of attenuation of the source at the opposite ear and results in IID. As frequency increases, wavelength decreases and the "shadowing" effect becomes more pronounced for a head of fixed size.

2.6.2 360°Audio

Walker and Brewster *et al.* explored the possibility of sonifying a daily agenda from the DateBook application of a personal digital assistant [WBMN01]. Hypothesizing that the DateBook's vertically linear array of daily appointments could be circularly

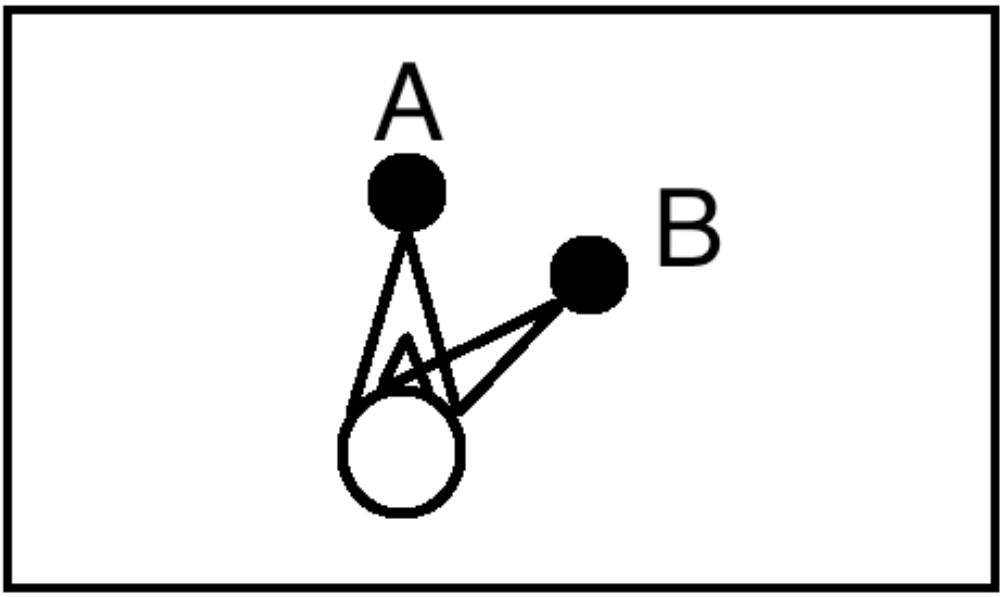


Figure 2.2: A listener in an anechoic chamber, with a sound source oriented directly ahead on the median plane (A) and displaced to 60˚azimuth (B).

positioned along a virtual auditory clock face (see Figure 2.3), the researchers had subjects interact with the sonic calendar display.

They found that when comparing two modalities (visual linear display and circular sonic display) for displaying the agenda, the test subjects were able to correctly recall details about the agenda 88.3% versus 70.2% of the time for the auditory and visual presentations, respectively. They also noted that their subjects reported less physical demand, time pressure, and a higher sense of performance when using the audio-centric modality.

A study of localized hearing was conducted in Japan in 2006 [OISM06]. As shown in Figure 2.4, blind or blindfolded listener-subjects were seated in the center of a circular array of speakers. Ohuchi *et al.* found that blind individuals could outperform sighted individuals in attempting to locate the perceived origin and intensity/distance of sound. The average azimuthal localization error for sighted users was

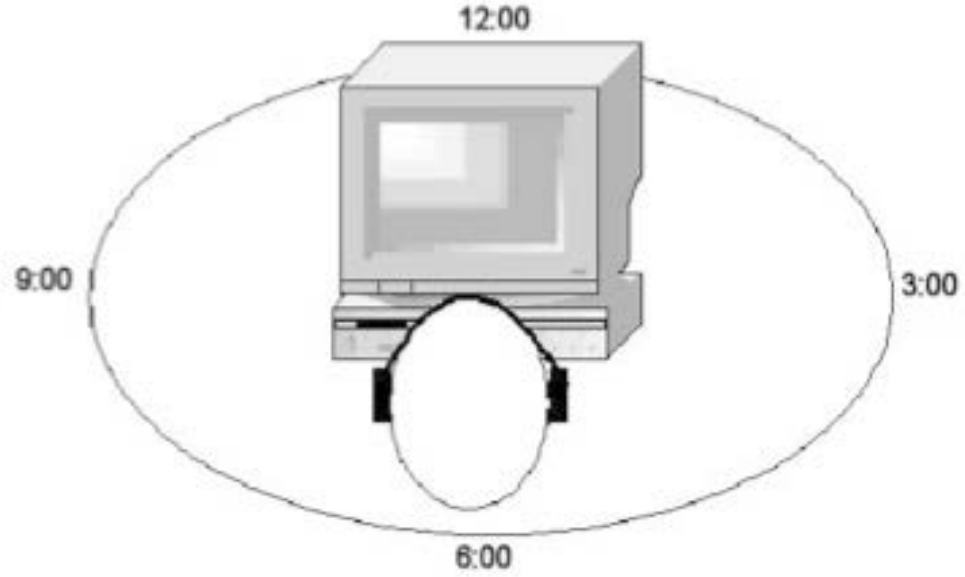


Figure 2.3: Time-space mapping of the auditory DateBook display.

approximately four°larger than that for blind users.

2.6.3 Stereo Panning

In the study conducted by Ramloll *et al.* [RBYR01], the audio signal was panned, or shifted in auditory perceptual space: when starting at the leftmost cell in a row, the audio would be heard in only the left ear, gradually moving along a sonic continuum towards the right ear, until the focus reached the rightmost cell and could only be heard in the right ear. Similarly, Ramloll *et al.* elected to use this horizontal stereo panning method when vertically traversing a column (column-motion from top to bottom is auditory motion from left to right). The mapping of focal location to stereo encoding in the MIDI framework was accomplished according to the following formula:

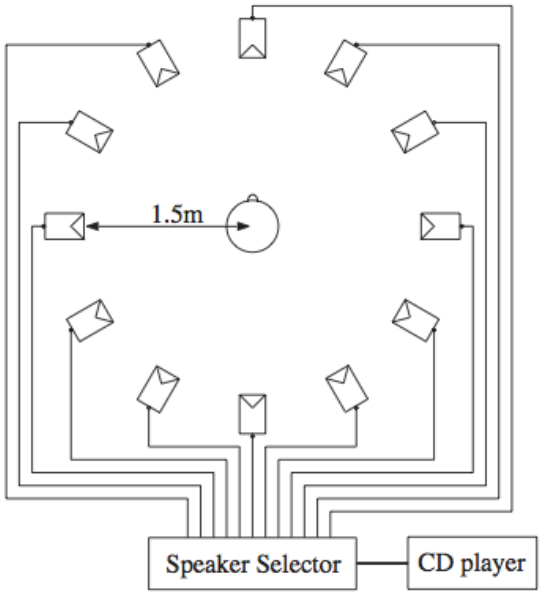


Figure 2.4: Equipment for measuring sound localization acuity.

$$P = \left(\frac{R}{N}\right) \times 127 \tag{2.2}$$

where P (rounded to the nearest integer) is the panning number, R the rank (starting with 0) of the cell in the current column or row and N is the number of cells in the current column or row.

Simultaneous panning and tonal variation reinforce the sense of movement as the focus moves throughout the numerical table. When prompted to answer numerical questions based on the tables presented, they found that their visually impaired subjects took roughly three times as long to answer, on average, when using only speech as compared to when they answered these questions with the pitch enhancements.

In a preliminary study to this research, Cofino *et al.* investigated the relative efficacies of the Vector-Based Amplitude Panning (VBAP) and Linear Panning (LP) methods. In the VBAP technique for two loudspeakers, an active arc of perceived sound is created (see Figure 2.5).

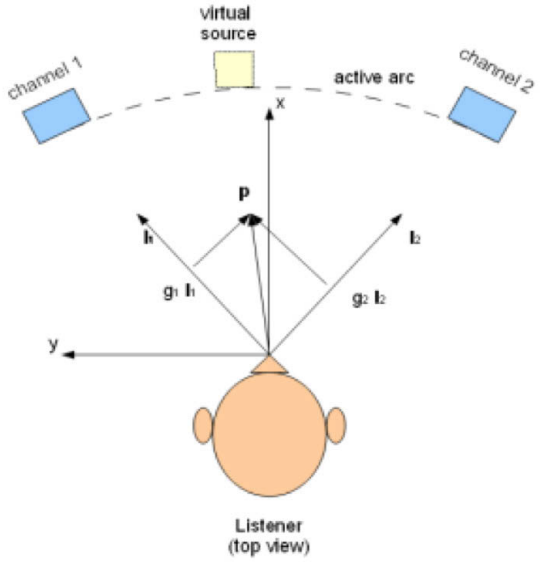


Figure 2.5: Stereo configuration formulated with vectors.

The vector to the virtual source, $\boldsymbol{p}$, can be expressed as a linear combination of the speaker vectors:

$$\mathbf{p} = g_1 l_1 + g_2 l_2. \tag{2.3}$$

where g_1 and g_2 are the gain factors for channels one and two, respectively.

The previous equation can be expressed in a matrix form:

$$\mathbf{p}^T = \mathbf{g}\mathbf{L}_{12} \quad \text{where} \quad \mathbf{g} = \begin{bmatrix} g_1 & g_2 \end{bmatrix} \quad \text{and} \quad \mathbf{L}_{12} = \begin{bmatrix} \mathbf{l}_1 & \mathbf{l}_2 \end{bmatrix}^T. \tag{2.4}$$

The inverse of $\mathbf{L}_{12}$ exists as long as the speakers are not placed collinearly. Then, the gain vector is:

$$\mathbf{g} = \mathbf{p}^T \mathbf{L}_{12}^{-1} = \begin{bmatrix} p_1 & p_2 \end{bmatrix} \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}^{-1}. \tag{2.5}$$

To perceive all virtual sources as emanating from equidistant points (i.e., within an active arc of constant radius), $\mathbf{g}$ is normalized to maintain a constant overall power level:

$$\mathbf{g}^{\mathbf{scaled}} = \frac{\mathbf{g}}{\sqrt{{g_1}^2 + {g_2}^2}}. \tag{2.6}$$

To simplify the implementation of the VBAP method in two dimensions, an angular approach is desirable. As shown in Figure 2.6, the virtual source is at an angle ϕ, between the two channel speakers at equiangular positions $(-\phi_0, \phi_0)$ on either side of the x-axis.

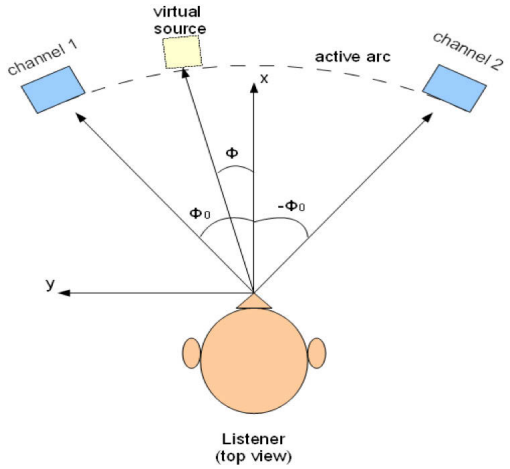


Figure 2.6: Stereophonic panning using the angular approach.

It can be shown that the gain vector $\mathbf{g}$ satisfies the tangent law. The two-channel stereophonic loudspeaker configuration matrix $\mathbf{L}$ components and virtual source position components are derived from Figure 2.5 and Figure 2.6:

$$l_{11} = l_{21} = \cos(\phi_0) \tag{2.7}$$

$$l_{12} = -l_{22} = \sin(\phi_0) \tag{2.8}$$

$$\mathbf{p}^T = \begin{bmatrix} p_1 & p_2 \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi \end{bmatrix} \tag{2.9}$$

The inverse of matrix $\mathbf{L}_{12}$ can be determined:

$$\mathbf{L}_{12}{}^{-1} = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}^{-1} = \frac{\begin{bmatrix} l_{22} & -l_{12} \\ -l_{21} & l_{11} \end{bmatrix}}{l_{11}l_{22} - l_{21}l_{12}} \tag{2.10}$$

The equation can now be reformulated:

$$\mathbf{g} = \frac{\begin{bmatrix} p_1 l_{22} - p_2 l_{21} & p_2 l_{11} - p_1 l_{12} \end{bmatrix}}{l_{11}l_{22} - l_{21}l_{12}} \tag{2.11}$$

$$g_1 = \frac{\cos\Phi \sin\Phi_0 + \sin\Phi \cos\Phi_0}{2\cos\Phi_0 \sin\Phi_0} \tag{2.12}$$

$$g_2 = \frac{\cos\Phi \sin\Phi_0 - \sin\Phi \cos\Phi_0}{2\cos\Phi_0 \sin\Phi_0} \tag{2.13}$$

$$\frac{g_1 - g_2}{g_1 + g_2} = \frac{2sin\Phi \cos\Phi_0}{2\cos\Phi \sin\Phi_0} = \frac{\tan\Phi}{\tan\Phi_0} \tag{2.14}$$

It is now apparent that VBAP satisfies the stereophonic tangent law.

In the preliminary experiment conducted by Cofino *et al.*, sonic source locations were created at five angles: Φ_1 = 40°, Φ_2 = 20°, Φ_3 = 0°, Φ_4 = -20°, Φ_5 = -40°. Φ_0, the angular position of the speakers, is a constant 40°.

The angular VBAP was compared to simple linear panning (LP). LP is described by the following equations:

$$\text{LG} = \frac{1}{2}(1 - \text{pan}) \tag{2.15}$$

$$\text{RG} = \frac{1}{2}(1 + \text{pan}) \tag{2.16}$$

where "LG" and "RG" are the left/right gains (respectively), and the parameter "pan" corresponds to the desired horizontal panning position, i.e., "pan" = -1 at the

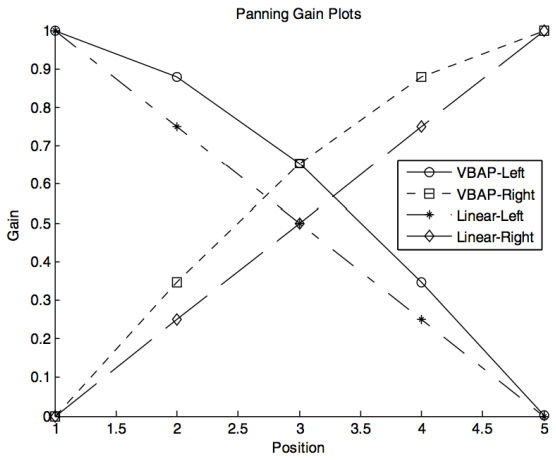


Figure 2.7: Gain Plot Comparisons - LP vs. VBAP.

left speaker position and "pan" = +1 at the right speaker position.

VBAP vs. LP Experiment Design Methods

The experiment sought to evaluate the accuracy of virtual sound localization (from five possible locations) by a computer user, when the VBAP and LP methods are used for audio spatialization.

Twenty sighted participants were recruited to determine the apparent localization of virtual sound source, consisting of white noise lasting two seconds. Each of the two panning methods was used twice for each of the five virtual positions, yielding 20 sound playback trials that each subject needed to identify as emerging from one of the five pre-set virtual source locations. To avoid ordering effects, both the order of these virtual source locations and the spatialization method used were randomized for each participant.

In each trial, the subject was asked to identify and key-in the spatial location position (1, 2, 3, 4, or 5, with position 3 located directly in front of the subject) perceived as the origin of the sound. Each trial was scored as "correct" (if the subjectṣ numerical answer exactly matched the numeric identifier of the virtual sound

placement) or "incorrect" (otherwise). Accordingly, percentages of accuracy could be derived for each subject: when the spatialization was performed with VBAP, when the spatialization was performed with LP, and an overall accuracy percentage (considering both spatialization methods).

As can be seen in Figures 2.5 and 2.6, the subject and the two speakers form an isosceles triangle. Each speaker was placed 30 inches from the subjectȁs chest with a lateral between-speakers distance of 38.57 in. This implies that using the linear panning technique the listening subject should perceive that the five virtual sound sources were placed at 9.64-in. intervals. While Pulkki states that in two-dimensional amplitude panning the speaker angle is typically chosen as +/-30°, we sought to establish a compromise between the several factors affecting the layout: angles, distances, etc., within the practical need to fit the setup on a typical desk. We decided to increase the horizontal distances to obtain a better differentiation between virtual locations; thus, the speaker angle was set to +/-40°.

The results of the virtual sound source localization experiment were analyzed. For the the VBAP method, the percentage of correct answers was 85% (std. dev. = 18.209%) while the percentage of correct answers for the linear panning (LP) method was 87% (std. dev. = 13.416%). As a result of the small sample size of this experiment and the non-normality of the data, a simple statistical t-test was inappropriate. Rather, the Wilcoxon signed-rank test was used as it is a non-parametric test. The (VBAP-Linear) difference resulted in six negative ranks, nine positive ranks, and five ties. Of the negative ranks, the mean rank was 10.50 and the sum was 63.00. Of the positive ranks, the mean rank was 6.33 and the sum of ranks was 57.00. These results of the Wilcoxon test yielded a Z-statistic (Z=-0.175, p=0.861). Noting that the significance value p was much greater than 0.05, the perceptual difference between both methods of stereo spatialization was deemed statistically insignificant. Therefore, this

research study will employ linear panning (LP) for stereo spatialization.

2.7 Non-visual tabular navigation

Tables are a product of visually-oriented written language. The inherent structure of the table dictates the context of tabular content, as opposed to prose. Where prose may be read aloud by a screen-reader for a passive listener, tabular data requires an active navigator to browse through the non-linear data. In short, tables have no obvious aural equivalent [Wri81].

Researchers in Japan developed an early prototype system for non-visually accessing HTML tables [OA98]. In order to conform to the limitations of Screen Reader/2, each cell in the HTML table was converted into its own HTML file. These HTML files could then be interlinked in a grid-like fashion (see Figure 2.8). A HTML table index, listing all of the tabular cells as hyperlinks, was created for navigation through the grid. Users could access the surrounding cells (if available) by pressing a two-key sequence and could reach the extremes of the table (top, bottom, left/right edge) by pressing chords (simultaneous keystrokes). The researchers found through observation that the visually-impaired users were able to understand simple tables using both the index and grid arrangements of links.

2.7.1 EVITA

Yesilada *et al.* devised the EVITA table browsing system for visually-impaired people [Yes00]. The EVITA system is based on the notions of browsing and navigation. "Navigation suggests an opportunity of movement within the local environment [GHS00]," while browsing is the process by which a user selects units of information deemed valuable according to his or her needs and interests [CR93]. Together, naviga-

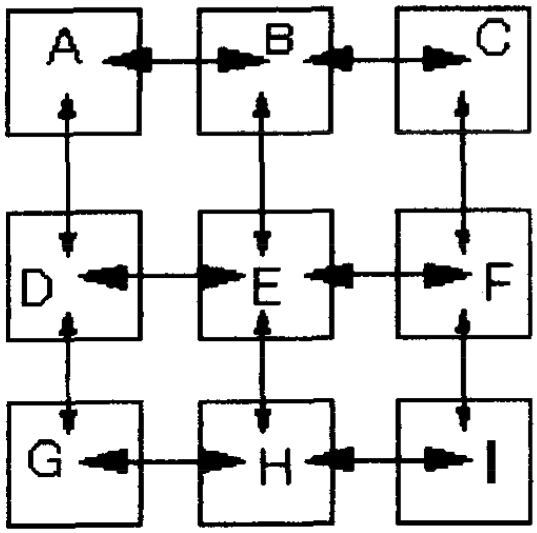


Figure 2.8: Tabular navigation links

tion and browsing allow the user to actively control the flow of information [YSGH04]. Navigation in the EVITA system is implemented in three levels. Low-level navigation functions (Level-1) allow the user to move cell-by-cell, specifically to adjacent cells and cells at either extreme of the current row/column. These functions are based on a "current" cell concept. Considering the transient nature of an auditory display, the notion of a current focus of attention must be emphasized systemically where visual persistence is absent [RBYR01]. Higher-level navigation is realized in Level-2, where a user may take an "action" towards a "target." Actions include moving to an adjacent or first/last "target," namely a row or column. The highest level tasks (Level-3) are performed by identifying the row-column intersection and comparing two rows/columns.

2.7.2 Non-visual News Table Navigation

In a preliminary experiment performed as a precursor to the simulated online purchasing exercise described in Chapter 4, a study was conducted to determine the

Business	Sports	Lifestyle	Technology	Health
1.) In Cautious Times, Banks Flooded with Cash	1.) NBA Plans to Cancel Two More Weeks of the Season	1.) The Future of Mangoes at the Fairchild Farm	1.) Microsoft to Partner With China's Leading Search Engine	1.) Gastric-bypass patients more prone to alcoholism
2.) Home Prices Up in Half of Major U.S. Cities	2.) Lakers Announce Ham Will Join Coaching Staff	2.) Miami Culinary Institute Grand Opening	2.) Homework Help Site Has a Social Networking Twist	2.) Study Links Smoking Drug to Cardiovascular Problems
3.) BP Earnings Slip 3.7% on Lower Production	3.) Redskins Lose Top Running Back for 2011	3.) Ralph Lauren Tells the Secrets of His Success	3.) The Collision Over Traffic Cameras	3.) New Study Implicates Environmental Factors in Autism
4.) Consumer Confidence Dips to Recession Levels	4.) Crowd Noise Causes Cardinals' Bullpen Confusion	4.) Wynwood Art Fair a Fundraiser for Homeless	4.) Hackers Claim Small Breach on Apple Site	4.) Don't just sit there: More health risks found
5.) Help for Underwater Homeowners	5.) FIFA Proposal: Fewer Penalty Area Red Cards	5.) Some Arts Groups Struggle, Some Succeed in South Florida	5.) Verizon dropping unlimited smartphone data plans	5.) Gum disease delays pregnancy
6.) Netflix Stock Hammered	6.) Panthers Send Canadiens to Sixth Straight Loss	6.) Broadway Debut After a Life of Opera	6.) Next generation video games let players control the story	6.) Overweight men have poorer sperm count

Figure 2.9: An auditory news table, five column layout.

efficacy of tonally adjusting the fundamental pitch of the synthesized speech. Within the context of a tabularly arrayed online news website, the user would perceive the voice descend in pitch as he/she analogously descended the focus of his/her navigation through a tabular news column. This approach to dynamic speech parameter modification seemed to either confuse or entertain the listener-navigators. The novelty of the modified voice, in conjunction with current event information that would vary in user familiarity, lead to very long completion times, large numbers of keystrokes, multiple task reminders, and requests for help to the study administrator. While this study did not yield objective results, it did reinforce the importance of maintaining a high-quality, consistent voice for auditory table narration.

As a pilot study, 13 sighted individuals were recruited to test the efficacy of spatialized/enhanced audio for an auditory purchasing table. The results of this preliminary study can be observed in Table 2.7.2.

TABLE	TTT_Mean	TPM_Mean	Moves_Mean
A	13.513	2.046	7.169
B	10.112	1.583	6.415
C	11.251	1.723	6.492
D	11.539	1.703	6.600

With a relatively small sample size of sighted individuals, a statistically significant treatment effect was not expected to be achieved. However, these results motivated

the researchers to persevere in their study of blind tabular navigation.

2.8 Transcoding

Transcoding for web accessibility is the process of converting visual web content into screen-reader accessible content. Using an intermediate proxy server, content is re-formatted (often serially) as it passes from web server to client browser [AT08]. To make visual webpages accessible, content is serialized by removing layout tables and extraneous content such as navigational bars and advertising. Ideally, the intermediate transcoding server would be able to serialize webpage content automatically and on-the-fly.

Annotation, a technique where the webpage developer marks which content is essential, must be authored manually and be standardized. Annotations describe both the meaning of website content as well as its role within the context of webpage navigation. Collaborative solutions such as crowd-sourcing may expedite the manual labeling of content needed for effective transcoding [TKK+08]. Leveraging the social interconnectedness that pervades the present Internet experience, social accessibility awards points to anyone who describes website content. These descriptions are then incorporated into the transcoded webpages as, for example, alt-tags, and further serve as annotations for transcoding.

Listing 2.1: Sweetwater University's Homepage

```
1 <!-- Advertisement for Joe's -->
  <a href="ad.html">
3   <img src="ad.gif">
  </a>
5 <a href="ad.html">Eat at Joe's!</a>
```

```
<br>

<!-- Sweetwater University Homepage -->
<p><b>Sweetwater University Homepage</b></p>
SU is Florida's newest public research university.
```

Asakawa and Takagi proposed a system for structural annotation [AT00]. Using HTML-style tags, they extended the existing markup with a annotative markup file. In the original HTML file (see Listing 2.1), there are two HTML groups: a banner advertisement and the content for a university homepage. The banner advertisement is coded first, with no alt-text to indicate that the image is actually a clickable link for Joeś. The screen reader user is forced to navigate through the advertisement before coming to the desired content, the university homepage.

Listing 2.2: Annotations to make SU's Homepage Accessible

```
<annotation url="http://www.sweetu.edu/" author="
    VisionaryCitizen@sweetu.edu">
  <group title= Banner members= /html/body/a[1] | /
      html/body/a[2] >
    <role type= advertisement />
    <importance value= -1 />
    <description>
      This is a banner advertisement for Joe's restaurant.
        Both the clickable image and clickable link lead to
        Joe's homepage.
    </description>
```

```
  </group>
9
  <group title= Homepage  members=  /html/body/p[1] | /
     html/body/text()[1]  >
11    <role type= main content />
    <importance value="1"/>
13    <description>
       This is the homepage for Sweetwater University. You can
           find information about classes, research, and
          athletics here.
15    </description>
  </group>
17 </annotation>
```

The annotation (see Listing 2.2), in an HTML-like format, gives semantic meaning to the original HTML. The group tags (and their corresponding title attributes) clearly delineate the banner advertisement from the homepage. The author of the annotation can be reached through e-mail. Most importantly, the purpose of the group is defined by the role tag. The importance tag is used for displaying the relative priority for a piece of content, where -1 represents most unimportant and +1 represents most important.

The final transcoded HTML (see Listing 2.3) reflects both the prioritization of content over extraneous advertising. Also, the transcoded page includes images preceding each annotated group. These images serve as a placeholder for alt-text that helps define the semantic purpose of the content as well as it's starting and ending.

Listing 2.3: Sweetwater University's Transcoded Homepage

```
<!-- Sweetwater University Homepage -->
<img width="16" height="16" src="http://www.sweetu.edu/access
    /homepage.gif" alt="This is the homepage for Sweetwater
    University. You can find information about classes,
    research, and athletics here.">
<p><b>SU Homepage</b></p>
SU is Florida's newest public research university.
<img width="0" height="0" src="" alt="end of group">

<hr>

<!-- Advertisement for Joe's -->
<img width="16" height="16" src="http://www.sweetu.edu/access
    /banner_ad.gif" alt="This is a banner advertisement for
    Joe's restaurant. Both links lead to Joe's homepage.">
<a href="ad.html">
  <img src="ad.gif">
</a>
<a href="ad.html">Eat at Joe's!</a>
<img width="0" height="0" src="" alt="end of group">
```

An annotated webpage can be broken down into smaller components. As a token example, the *New York Times* website for January 17, 2014 is depicted in Figure 2.10. The highlighted "A" section is the "title bar," having the title as well as the date of issue, weather report, and links to social media. The "B" sections represent extraneous advertising not linked to news content. The "C" sections represent a typical

title-author-summary trio representing a link to a particular article or other content. The "D" sections represent a list of links to articles and other content, lacking summaries or authors. The "E" section is the navigation bar. The "F" section represents the membership and portal functionality of the website relating to having a *New York Times* account or subscription.

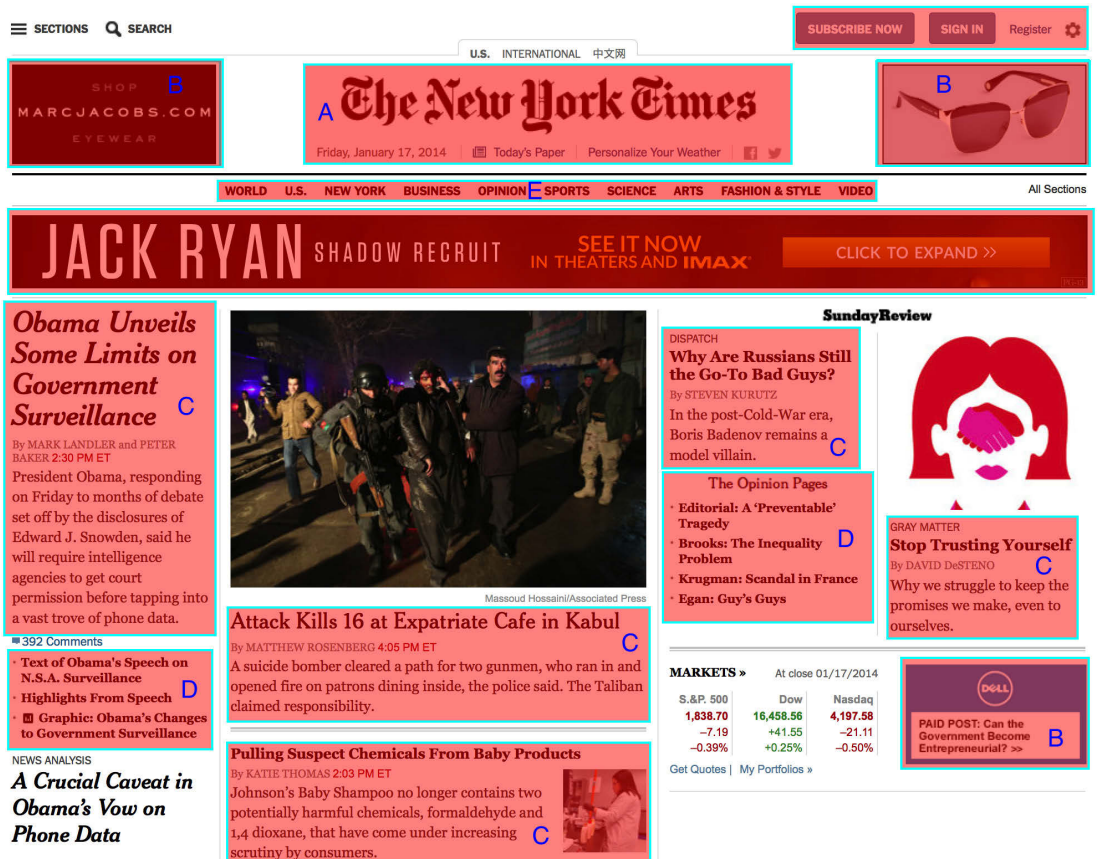


Figure 2.10: Possible annotation scheme for a newspaper website.

Several methods exist for transcoding pages; these methods can be grouped into five approaches [YSHG07]: creating a text-only version, incorporating user preferences, employing heuristics, leveraging semantics, and optimizing for the mobile web. Text-only webpages are inherently amenable to screen-reading technology. As the Web has grown more interactive and dynamic, the text-only approach often leaves out content that is not easily converted to text and therefore creates a separate experience

for screen-reader users. Allowing users to set their own preferences involves serializing content on the server side, which is redundant with client-side screen-reading technology. The heuristic approach involves using experience-based algorithms to transcode a webpage. While some success has been achieved for screen-magnifiers using this technique, it has not yet been successful for screen-readers. Semantic transcoding depends on external annotations based on controlled vocabularies. These vocabularies can be used to define structural properties of webpages as well as to define specifications for a mobile web device.

This brief review of prevailing approaches to transcoding is meant to illustrate and document the feasibility or reorganizing, on-the-fly, the contents of ordinary web pages, which is not always arrayed in a regular, grid-like format, to a configuration that could be modeled as a table. This is highly relevant to the work reported here, as the focus of this research is on navigation through collections of items that are already set as aggregates of rows and columns, i.e., navigation through tables.

2.9 Summary

In this chapter, the background work supporting this research was presented. A brief overview of screen-reader development and applications was shown. Synthesized speech and the theory of tables were explained. The concept of mapping sound to data, specifically with regard to tonal variation and stereo spatialization, was shown to have been explored through preliminary experiments. A previous study investigating non-visual tabular navigation was presented. Finally, attempts to reformat webpage content for screen readers were shown to be developing and increasingly feasible.

CHAPTER 3

METHODOLOGY

This chapter outlines the various alternative approaches considered towards the goal of enhancing the non-visual navigation of a webpage formatted as a table. This chapter also summarizes the rationale for selecting two specific sonic manipulations for the final implementation of the experimental setup used by the blind volunteers in the main experimental phase of this research.

3.1 PROSODIC MODIFICATION - A Preliminary Study

Prosody refers to the stress (the relative emphasis that may be given to certain syllables in a word), rhythm, and intonation (the variation of spoken pitch) of speech. Whether the speaker is asking a question, issuing a command, or declaring a statement is determined by prosodic features of speech. Certain non-literal devices of language (irony, sarcasm) rely on prosody for their communication as opposed to grammar or vocabulary.

In a pilot study preceding the main experimental phase of this research, it was sought to determine whether the pitch of the synthesized speech could be varied in order to signal the tabular focal position. The pilot study featured a table of news headlines, where each column represented a section (Business, Technology, Sports, Weather, etc.) of a conventional newspaper. Listeners could ascertain the vertical focus (or depth) by perceiving an increasingly lower-pitched narrative voice as the focus shifted downward. The prevailing motivation behind this was that variation of the synthesized speech itself would obviate the need for additional sound effects.

Speech (synthesized or natural) is a complex process that can be modeled in terms of a speaker's source and filter. The larynx, or voice box, serves as the generator of

sound with the ability to modify both pitch (a particular fundamental frequency) and volume. This source is then modulated through the vocal tract (laryngeal cavity, the pharynx, the oral cavity, and the nasal cavity). Vowels and consonants are formed based on the positions of the tongue, lips, mouth, and pharynx.

3.1.1 Typical PSOLA-based Pitch Shifting System

A basic PSOLA-based pitch shifting system features three sequential processes, as shown in Figure 3.1.

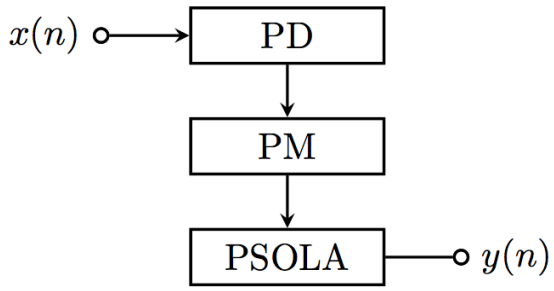


Figure 3.1: Common PSOLA-based pitch shifting system overview with pitch detector (PD), pitch marker (PM) and pitch shifter (PSOLA).

Pitch Detection

Pitch detection (PD) computes the current fundamental frequency of the input signal. The PSOLA algorithm critically depends on the accurate determination of the instantaneous pitch of the signal. In [vdKZZ10], pitch determination was computed using the YIN algorithm [DCK02]. Periodicity of a signal can be evaluated using a difference function; essentially, by examining how a signal correlates with itself over a series of lagging intervals, the greatest values of an autocorrelation function indicate likely periodicity of lag l. Expressed as a difference function (See Equation 3.1) which sums squared differences at lag l:

$$d_t(l) = \sum_{n=1}^{N-1} (x(n) - x(n+l))^2 \tag{3.1}$$

In the difference function, low values of $d_t(l)$ indicate similarity at a particular lag l.

$$d'_t(l) = \begin{cases} 1, & l = 0 \\ \frac{d_t(l)}{\frac{1}{l}\sum_{n=1}^{l} d_t(n)} & \text{else.} \end{cases} \tag{3.2}$$

The normalized mean difference function (NMDF, see Equation 3.2) starts off with a value of 1 for lag l=0 and then only drops below one where the current value is below all of the previous lags. A threshold can be defined such that when the NMDF drops below this threshold a local minimum can be found. This local minimum can be further refined through parabolic interpolation. **MATLAB** code for the parabolic-interpolated YIN algorithm can be seen in Listing 3.1.

Listing 3.1: YIN Pitch Detection Algorithm

```
function [pitch, x_out] = yinDAFX_new(x_in, fs, f0min, f0max, fhop
    )
% function pitch = yinDAFX(x, fs, f0min, f0max, fhop)
% Author: Adrian v.d. Knesebeck
% determines the pitches of the input signal x at a given hop
    size
%
% input:
% x         input signal
% fs        sampling frequency
% f0min     minimum detectable pitch
```

```
% hop         hop size
%
% output:
% pitch       pitch frequencies in Hz at the given hop size
%
% This source code is provided without any warranties as
% published in DAFX book 2nd edition,
% copyright Wiley & Sons 2011, available at
% http://www.dafx.de. It may be used for educational
% purposes and not for commercial applications without
% further permission.

% initialization
disp('Frame Processing:'),
taumax = round(1/f0min*fs);
yinLen = 1024;
flen = yinLen + taumax + 1;
[fx,fpad] = linframe(x_in,fhop,flen,'pad');

[~,frames] = size(fx);
yinTolerance = 0.22;
k = 0;
pitch = zeros(1,frames);

for f = 1:frames % frame processing
    k=k+1;
```

```
    xframe = fx(:,f);

    % Computing autocorrelation
    a_t = xcorr(xframe(1:yinLen));
    r_t = zeros(taumax+1,1); r_t(1) = a_t(yinLen);
    a_t = a_t(yinLen:end);
    c_t = xcorr(xframe(yinLen+1:yinLen+taumax+1),xframe(1:
        yinLen));

    for tau = 1:taumax
        r_t(tau+1) = r_t(tau) - xframe(tau)^2 + xframe(yinLen
            +tau)^2;
    end
    yinTemp = r_t(1) + r_t(1:taumax) - 2*(a_t(1:taumax)+c_t
        (1:taumax));

    % calculate cumulated normalization
    yin = yinTemp .* (1:taumax)' ./ cumsum(yinTemp);

    % determine lowest pitch
    tau=1;
    while(tau<taumax)
        if(yin(tau) < yinTolerance)
            % search turning point
            while (yin(tau+1) < yin(tau))
                tau = tau+1;
```

```
            end
            % interpolating to find corrected pitch
            x = tau-1:tau+1; y = yinTemp(x); % nearest
                neighbor coordinates
            [p,~,mu] = polyfit(x',y,2);      % parabolic
                interpolation
            x_new = tau-1:1/100:tau+1;
            y_new = polyval(p,x_new,[],mu);
            [~,i] = min(y_new);
            pitch(k) = fs/x_new(i);
            break
        else
            tau = tau+1;
        end
        % if no pitch detected
        pitch(k) = 0;
    end
end

pitch(pitch>f0max)=0;
x_out = linunframe(fx,fhop,fpad);
end
```

The lag l corresponds to the fundamental period in samples, which can be converted to time (sec) using Equation 3.3:

$$period = \frac{l}{f_s} \tag{3.3}$$

where f_s is the sampling frequency. The fundamental pitch can be found by inverting the period.

MATLAB code for finding the fundamental pitch within a grain of the signal may be found in Listing 3.2.

Listing 3.2: Basic PSOLA Pitch Shifting Algorithm

```
function [ out ] = psola( in,m,alpha,beta )
% in      input signal
% m       pitch marks
% alpha   time stretching factor
% beta    pitch shifting factor

P = diff(m); % compute pitch periods
in = in(:);
% remove first pitch mark
if m(1) <= P(1)
    m = m(2:length(m));
    P = P(2:length(P));
end

% remove last pitch mark
if m(length(m)) + P(length(P)) > length(in)
    m = m(1:length(m)-1);
else
```

```
    P = [P P(length(P))];
end


Lout = ceil(length(in) * alpha);
out = zeros(1,Lout); % initialize output signal
tk = P(1) + 1; % output pitch mark
while round(tk)<Lout
    [~, i ] = min(abs(alpha*m-tk)); %find analysis segment
    pit = P(i);
    if ( m(i)+pit > length(in) )
        break;
    end
    gr = in(m(i)-pit:m(i)+pit) .* hanning(2*pit+1);
    iniGr = round(tk)-pit;
    endGr = round(tk)+pit;
    if endGr>Lout
        break;
    end
    out(iniGr:endGr) = out(iniGr:endGr) + gr'; % overlap new
       segment
    tk = tk + pit/beta;
end


% Normalizing 'out'
out = out/max(abs(out)); out = out(:);
```

Pitch Marking

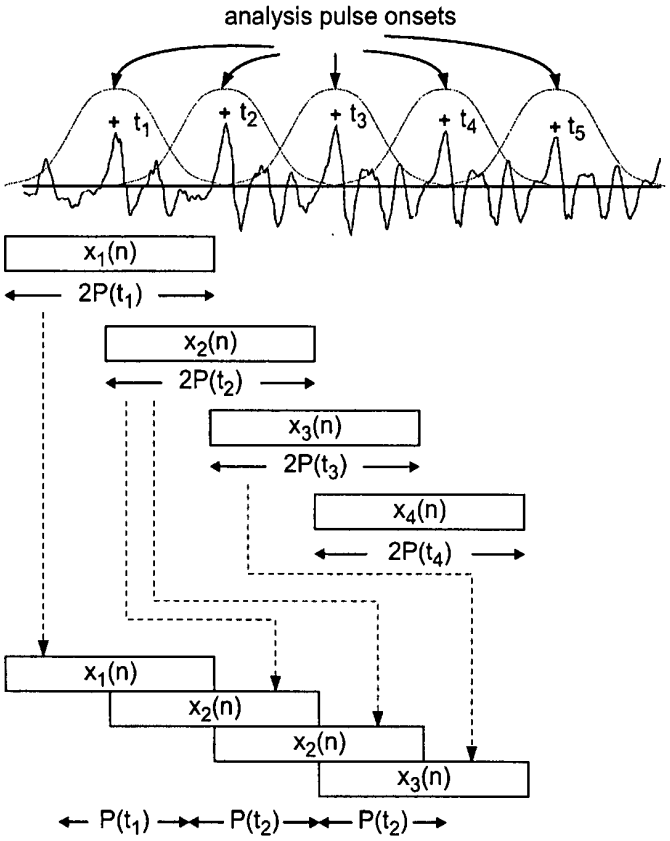


Figure 3.2: Analysis Phase of PSOLA

The analysis phase of PSOLA involves setting pitch marks at intervals corresponding to the pitch periods found in the pitch detection phase. In Figure 3.2 (image from [San09]), five analysis pulse onsets correspond to five pitch marks $t_{1:5}$. The first four of these pulse onsets can be extracted as grains ($x_{1:4}(n)$), where each grain has length $2P(t_i)$ corresponding to the pitch period of pitch mark t_i. As the detected pitch varies, the pitch marks are set at intervals of inverse size with respect to pitch. The length of a pitch period in samples can be determined according to Equation 3.4:

$$P(m_i) = m_{i+1} - m_i \tag{3.4}$$

Determining the position of pitch marks is not a trivial problem. The pitch marks provide the center of the segmentation windows of the PSOLA algorithm, which ultimately affects the quality of the sound quality of the pitch-shifted audio. In the case of speech, sections are considered to be voiced or unvoiced depending on the whether the glottis is engaged and resonating. Voiced sections generally exhibit semi-periodic qualities, while unvoiced sections generally resemble transient noise. During unvoiced sections, the "pitch period" may be kept at the constant rate of the previous voiced section of speech.

A method for finding the pitch marks of a complete voiced frame has been described [LJ04, MVV06]. The global maximum of the waveform in the voiced frame is identified and becomes the first pitch mark, t_i. From this initial pitch mark, first approximations of pitch marks may be made in the left- (t_{i-1}) and right- (t_{i+1}) ward directions using the pitch period, which are further refined:

$$t_{i+1} = \max([t_i + \delta P_0, t_i + (2 - \delta)P_0]) \tag{3.5}$$

$$t_{i-1} = \max([t_i - \delta P_0, t_i - (2 - \delta)P_0]) \tag{3.6}$$

where P_0 is the pitch period and δ is a factor in the range of 0.5 to 0.9. This is effectively using a peak-search approach to finding pitch marks.

A more refined approach to finding pitch marks is described by von dem Knesebeck *et al.* [vdKZZ10]. The pitch period P is determined by the interval between consecutive pitch marks.

$$P(m_i) = m_{i+1} - m_i \tag{3.7}$$

A Hanning window of length L_i is applied to each grain according Equation 3.8:

$$L_i = 2P(m_i) \tag{3.8}$$

where a Hanning window can be described:

$$w(n) = 0.5(1 - \cos(\frac{2\pi n}{L_i - 1})) \tag{3.9}$$

where m_i represents the center of the window. Since the quality of the pitch shifting relies critically on the positioning of the pitch marks, care must be taken in their placement. The window should ideally contain the grain's maximum energy. This necessitates a strategy for effectively placing pitch marks without relying on amplitude or glottal pulses. First, a center of energy $\hat{m}$ of short signal interval x_a is found:

$$\hat{m} = \frac{\sum_{n=1}^{N} |x_a(n)|^2 n}{\varepsilon_x} \tag{3.10}$$

where ε_x is the energy of x_a:

$$\varepsilon_x = \sum_{n=1}^{N} |x_a(n)|^2 \tag{3.11}$$

The short time signal x_a is a small region of the signal block and represents an analysis region. The size of x_a corresponds to the pitch period P. The upper(l_u)/lower(l_l) limits of the region x_a are defined:

$$l_l(m_i) = l_u(m_{i-1}) + 1 \tag{3.12}$$

$$l_u(m_i) = l_u(m_{i-1}) + P(m_i). \tag{3.13}$$

These pitch marks $\hat{m}_i$ based on the center of energy serve as a first approximation to a more refined placement of the pitch mark.

von dem Knesebeck provides MATLAB code for finding pitch marks in Listing 3.3.

Listing 3.3: PSOLA Code for Finding Pitch Marks

```
function [m] = findpitchmarks(x,Fs,F0,hop,frameLen)
% Author: A. von dem Knesebeck
% x         input signal
% Fs        sampling frequency
% F0        fundamental frequencies
% hop       hop size of F0 detection
% frameLen  length of frame
%
% This source code is provided without any warranties as
% published in DAFX book 2nd edition,
% copyright Wiley & Sons 2011, available at
% http://www.dafx.de. It may be used for educational
% purposes and not for commercial applications without
% further permission.

% Initialization
m           = 0;    % vector of pitch mark positions
```

```
P0          = zeros (1,length(F0));
index       = 1;
local_m     = [];   % local pitch marker position

% processing frames i
for i = 1:length(F0);
    % set pitch periods of unvoiced frames
    if (i==1 && F0(i)==0); F0(i) = 120;     % 120Hz in case
       no preceding pitch
    elseif (F0(i)==0);     F0(i) = F0(i-1);
    end
    P0(i) = round(Fs/F0(i));                % fundamental
       period of frame i
    frameRange  = (1:frameLen) + (i-1)*hop; % hopping frame
    last_m = index;                         % last found
       pitch mark

    % beginning periods of 1st frame
    j = 1;  %period number
    if i==1
        % define limits of searchFrame
        searchUpLim = 1 * P0(i);
        searchRange = (1 : searchUpLim);
        [~,loc]  = max(x(searchRange));
        local_m(j)  = round(loc);
```

```
        % beginning periods of 2nd - end frame
    else
        searchUpLim = searchUpLim + P0(i);
        local_m(j)  = last_m + P0(i);
    end % beginning periods of 1st - end frame

    % remaining periods of 1st - end frame
    index = local_m(1);
    j = 2;  % grain/period number
    while( searchUpLim + P0(i) <= frameRange(end))
        % define range in which a marker is to be found
        searchUpLim = searchUpLim + P0(i);
        local_m(j)  = local_m(j-1) + P0(i);
        index = local_m(j);
        j = j+1;
    end %while frame end not reached
    m = [m local_m];
end % processing frames i

% finishing calculated pitch marks
m = sort(m);
m = unique(m);
m = m(2:end);
```

PSOLA Pitch Shifting

Originally proposed by Moulines *et al.* [MC90], pitch synchronous overlap and add (PSOLA) is a digital signal processing (DSP) technique which is useful for modifying the pitch and duration of a speech signal. The technique divides a speech signal into small overlapping segments. The segments can be moved farther apart or closer together to lower or raise the pitch of the speech, respectively. Similarly, the segments can be repeated or removed in order to slow down or speed up the speech, respectively. After the segments are manipulated, the overlap-add (OLA) method is used to recombine them into the modified signal.

Valbret *et al.* describe the PSOLA process in three steps [VMT92]:

Analysis "the speech waveform is decomposed into two components: a flattened source signal containing much of the prosodic information, and a global envelope component which accounts for the resonant characteristics of the vocal tract transfer function together with the spectral characteristics of the glottal excitation."

TD-PSOLA The Time-Domain PSOLA algorithms are applied to the source signal to alter the prosodic parameters (pitch, duration).

Synthesis "The synthesis signal is obtained from the modified excitation source and the modified envelope."

The analysis and synthesis algorithms for PSOLA can be described according to [DDPZ02]: Analysis algorithm:

1. Determination of the pitch period $P(t)$ of the input signal and of time instants (pitch marks) t_i. These pitch marks are in correspondence with the maximum amplitude or glottal pulses at a pitch-synchronous rate during the periodic part of the sound and at a constant rate during the unvoiced portion. In practice $P(t)$ is considered constant $P(t) = P(t_i) = t_{i+1} - t_i$ on the time interval (t_i, t_{i+1})

2. Extraction of a segment centered at every pitch mark t_i by using a Hanning window with length $L_i = 2P(t_i)$ (two pitch periods) to ensure fade-in and fade-out.

Synthesis algorithm for every synthesis pitch mark $\tilde{t}_k$:

1. Choice of the corresponding analysis segment i (identified by the time mark t_i) minimizing the time distance $|\alpha t_i - \tilde{t}_k|$.
2. Overlap and add the selected segment. Notice that some input segments will be repeated for $\alpha > 1$ (time expansion) or discarded when $\alpha < 1$ (time compression).
3. Determination of the time instant $\tilde{t}_{k+1}$ where the next synthesis segment will be centered in order to preserve the local pitch, by the relation

$$\tilde{t}_{k+1} = \tilde{t}_k + \tilde{P}(\tilde{t}_k) = \tilde{t}_k + P(t_i).$$

3.1.2 Enhanced PSOLA Pitch Shifting System

The "Enhanced PSOLA Pitch Shifting System" extends the basic system previously discussed by adding preprocessing for transient detection (TD) and extrapolation (EXT).

Transient Detection

Because PSOLA is intended to shift the *periodic* portions of a signal, it would be an improvement to first remove the transient and non-periodic parts of a signal (see Equation 3.14) before applying PSOLA.

$$r(n) = x(n) - t(n) \tag{3.14}$$

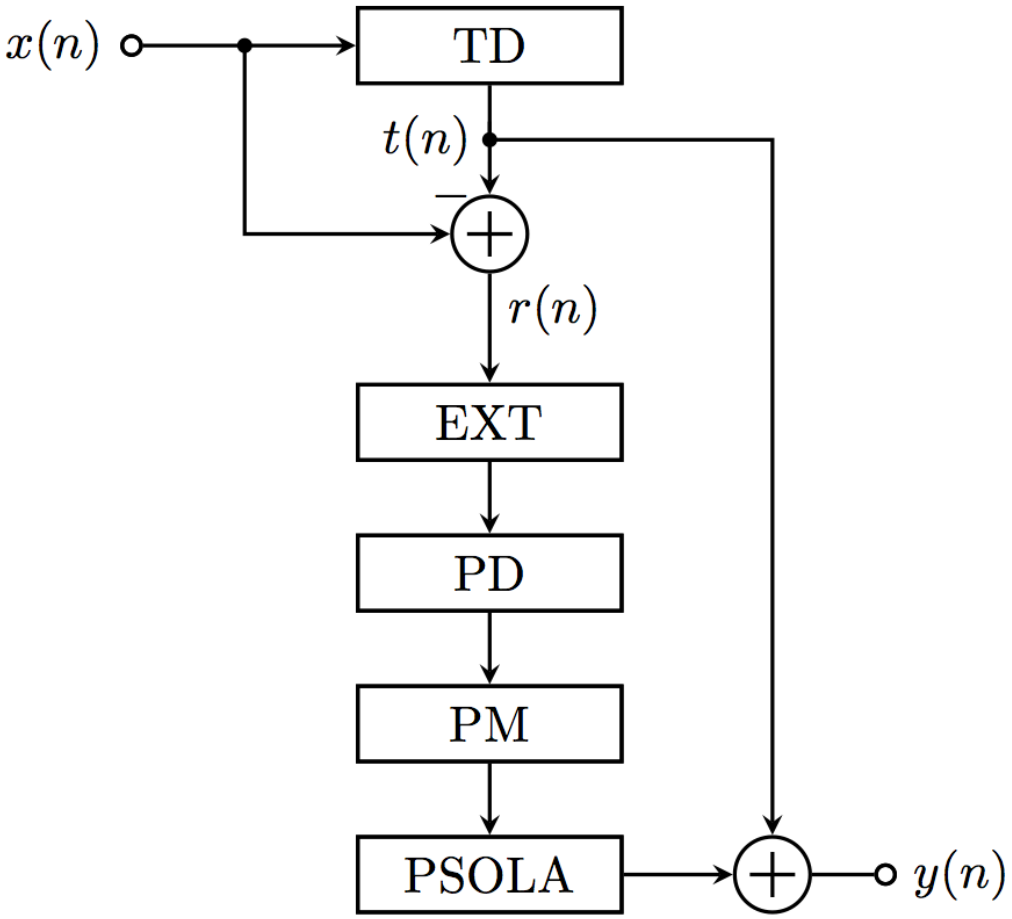


Figure 3.3: Enhanced PSOLA system overview. The common PSOLA system extended by the transient detection (TD) and the extrapolation (EXT).

where $x(n)$ is the signal, $t(n)$ is the transient part of the signal, and $r(n)$ is the residual part of the signal.

A method for transient detection and extraction for audio coding is described in [EN06]. Their first step is to employ a transform for frequency decomposition. Since the signal being extracted is transient, the time resolution of the transform must be high enough to detect steep temporal offsets. To accomplish these aims, Edler and Niemeyer elected to use the Modified Discrete Cosine Transform (MDCT, Eq. 3.15), and its complementary Modified Discrete Sine Transform (MDST, Eq. 3.16):

$$X_{MDCT}[m] = \frac{2}{\sqrt{L}} \sum_{n=0}^{L-1} w[n]x[n] \cos(\frac{2\pi}{L}(n + \frac{L}{4} + \frac{1}{2})(m + \frac{1}{2})) \tag{3.15}$$

$$X_{MDST}[m] = \frac{2}{\sqrt{L}} \sum_{n=0}^{L-1} w[n]x[n] \sin(\frac{2\pi}{L}(n + \frac{L}{4} + \frac{1}{2})(m + \frac{1}{2})) \tag{3.16}$$

with m=0, ..., $\frac{L}{2} - 1$

The MDCT and MDST represent the real and imaginary parts of the signal spectrum, respectively. For both the MDCT and MDST, a windowing function ($w[n]$) of length L was implemented as a half-sine wave with a window length of 256 samples. Put together, the full complex spectrum is given by Equation 3.17:

$$X[m] = X_{MDCT}[m] + jX_{MDST}[m] \tag{3.17}$$

An envelope function (see Equation 3.18) is then established to detect sharp onsets while ignoring steep decays:

$$X_{env}[b][m] = max\left\{\sqrt{X^2_{MDCT}[b][m] + X^2_{MDST}[b][m]}, \frac{1}{2}X_{env}[b-1][m]\right\} \tag{3.18}$$

where b is the temporal index of the subsequent blocks and m is the index of the spectral bands. The calculation of the parameters based on the spectral envelope of band m is illustrated in Figure 3.4.

For a given transform block b_0, a "viewport" can be determined by including

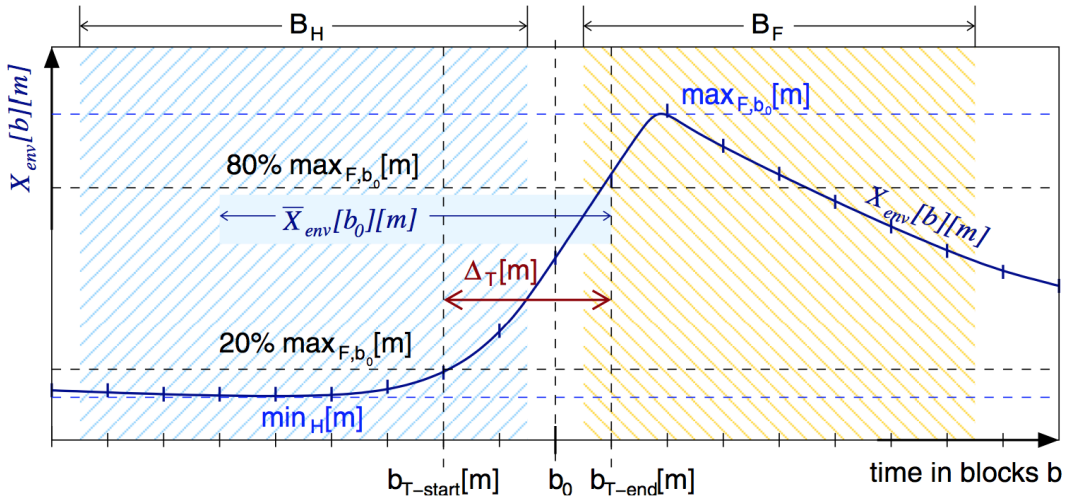


Figure 3.4: Calculation of parameters for the transient detection for one frequency band m based on the temporal envelope $X_{env}[b][m]$.

$B_F = 7$ transform blocks into the future and $B_H = 8$ transform blocks into the past (history). At this point, the block containing the end of the onset ($b_{T-end}[m]$) may be determined. Edler and Niemeyer define this block as where the envelope function $X_{env}[b][m]$ for the first time exceeds 80% of $max_{F,b_0}[m]$ in the viewport before the maximum. Next, the onset block ($b_{T-start}[m]$) is determined. It is defined as the block b, whose envelope value $X_{env}[b][m]$ for the first time falls below 20% of $max_{F,b_0}[m]$ going backwards in time starting from $b_{T-end}[m]$. The onset duration can be determined (see Equation 3.19) by comparing the starting and ending blocks:

$$\Delta_T[m] = b_{T-end}[m] - b_{T-start}[m]. \tag{3.19}$$

From here, the minimum value $min_{F,b_0}[m]$ in the viewport before $max_{F,b_0}[m]$ can be determined. Finally two means are calculated. First, the mean envelope including the onset (see Equation 3.20), and second, the mean envelope value before the onset (see Equation 3.21):

$$\bar{X}_{env}[b_0][m] = \frac{1}{B_H} \sum_{b=b_{T-end}[m]-B_H+1}^{b_{T-end}[m]} X_{env}[b][m] \tag{3.20}$$

$$\bar{X}_{env,H}[b_0][m] = \frac{1}{B_H} \sum_{b=b_{T-start}[m]-B_H}^{b_{T-start-1}[m]} X_{env}[b][m] \tag{3.21}$$

Considering the the relative increase of the envelope at the onset for the frequency band m, a weighting factor $r[b_0][m]$ can be calculated according to Equation 3.22:

$$r[b_0][m] = \frac{max_{F,b_0}[m] - min_{F,b_0}[m]}{\bar{X}_{env}[b_0][m]} \tag{3.22}$$

Since transients typically occur in more than one frequency band, neighboring frequency bands are taken into account. In this case, $M = 7$ neighboring frequency bands are considered on each side. The relative onset duration $\Delta_{T,M}[m]$ (Equation 3.23) is used as a transient detection criterion:

$$\Delta_{T,M}[m_0] = \frac{1}{R_{m0}} \sum_{m=m_0-M}^{m_0+M} \frac{M+1-|m-m_0|}{M+1} r[b_0][m]\Delta_T[m] \tag{3.23}$$

with

$$R_{m0} = \sum_{m=m_0-M}^{m_0+M} \frac{M+1-|m-m_0|}{M+1} r[b_0][m] \tag{3.24}$$

Finally, the relative onset duration $\Delta_{T,M}[m]$ is compared to a threshold. Edler and Niemeyer have determined that a threshold of three has delivered desirable results,

so the final criteria for transient detection can be stated (see Equation 3.25):

$$\begin{gathered}\forall(b,m) \text{ between } b_{T-start}[m] \text{ and } b_{T-end}[m] \\ \text{(b,m) is a transient, if} \quad \Delta_{T,M}[m_0] \leq 3. \end{gathered} \tag{3.25}$$

Now that a criterion for detecting transients has been established, the signal can be decomposed into its complementary transient (see Eq. 3.26) and stationary components (see Eq. 3.27):

$$X_T[b][m] = cf[b][m] \cdot X_{MDCT}[b][m] \tag{3.26}$$

$$X_S[b][m] = (1 - cf[b][m]) \cdot X_{MDCT}[b][m] \tag{3.27}$$

where $cf[b][m]$ is the cross fade factor. The cross fading of the transient and stationary components allows for a slower onset. The cross fade factor can be calculated according to Equations 3.28 and 3.29:

$$\text{if (b,m) is a transient:}$$

$$cf[b][m] = max\left\{ d_T \cdot cf[b-1][m], \frac{X_{env}[b][m] - c_x \bar{X}_{env,H}[b][m]}{X_{env}[b][m]} \right\} \tag{3.28}$$

$$\text{else}$$

$$cf[b][m] = d_T \cdot cf[b-1][m] \tag{3.29}$$

Edler and Niemeyer chose $c_x = \frac{1}{8}$ and $d_T = \frac{9}{10}$.

A MATLAB implementation of this transient detection approach, implemented according to the methods described above, is presented in Listing 3.4.

Listing 3.4: Transient Detection

```
function [y_orig,y_synth,y_trans,y_stat,fs] = ...
    trans_det_marios(fname,cx,fhop,flen)
%TRANS_DET_MDCT Separates a signal into transient and stationary components
% As described in
%    "Detection and Extraction of Transients for Audio Coding"
%   by  Niemeyer AND Edler
%   ABSTRACT
% An algorithm for the detection and extraction of transient signal
% components is presented. It is based on the detection
% of sharp onsets of the signal power in time direction of the complex time
% frequency domain. Afterwards the detected
% transients are extracted in the corresponding MDCT spectrum. The audio
% signal containing only the extracted transients
% is synthesized using the inverse MDCT. In an audio coding application
% this transient signal and a resulting residual
% signal can be coded separately using specifically optimized coders.

clc,
%% Truncate signal to non-zero and normalize
```

```
[in,fs] = wavread(fname);
in = in(:,1);
ii = find(in~=0);
sig_start = ii(1); sig_end = ii(end);
in = in(sig_start:sig_end);
in = in/max(abs(in));

%% Segmenting into window-frames
if nargin < 4
    flen = 256;
end
if nargin < 3
    fhop = flen/2;
end

[fx,fpad] = linframe(in,fhop,flen,'sym');
fx_win = winit(fx,'sinewin'); % windowing signal block
L = fpad(5); blocks = fpad(6);

%% Computing the Temporal Envelope
[X_MDCT, X_MDST] = mdctv(fx_win);

X_env = sqrt( X_MDCT.^2 + X_MDST.^2 );
X_env(:,2:end) = max( X_env(:,2:end), 1/2 * X_env(:,1:end-1)
    );
```

```
%% Detecting Transients (Main Loop)
% viewport (8 past blocks + current block + 7 future blocks)
BF = 7; % Seven blocks in the future, from b0 reference block
BH = 8; % Eight blocks in the past,   from b0 reference block

max_F = zeros(L/2,blocks); idx = zeros(L/2,blocks);
X_env_bar = zeros(L/2,blocks); X_env_H_bar = zeros(L/2,blocks
    );
r = zeros(L/2,blocks); min_F = zeros(L/2,blocks);
trans = zeros(L/2,blocks); start = zeros(L/2,blocks);
for b0=BH+1:blocks-BF
    [max_F(:,b0),idx(:,b0)] = max(X_env(:,b0:b0+BF),[],2);
    idx(:,b0) = idx(:,b0) + (b0-1);

    % Finding the index of start and end of the onset
    delta_T=zeros(L/2,1); b_T = zeros(L/2,2);
    for m = 1:L/2
        %80 of max is block T_end
        j = b0; % block index
        while X_env(m,j) < 0.8 * max_F(m,b0)
            if j == idx(m,b0)
                break;
            end
            j = j+1;
        end
        b_T(m,2) = j;
```

```
        %20 of max is block T_start
72      k = b_T(m,2);
        while X_env(m,k) > 0.2 * max_F(m,b0)
            if k == b0-BH
                break;
            end
77          k = k-1;
        end
        b_T(m,1) = k;

        % blocks of onset
82      delta_T(m) = b_T(m,2) - b_T(m,1);
    end

    % Mean envelope calculation
    X_env_bar(:,b0) = sum(X_env(:,b_T(:,2)-BH+1:b_T(:,2)),2)/
        BH;
87  start(:,b0) = max(b_T(:,1)-BH,1);
    X_env_H_bar(:,b0) = sum(X_env(:,start(:,b0):b_T(:,1)-1)
        ,2)/BH;

    % Weighting factor 'r'
    min_F(:,b0) = min(X_env(:,b0-BH:b_T(:,2)),[],2);
92  r(:,b0) = ( max_F(:,b0) - min_F(:,b0) ) ./ X_env_bar(:,b0
        );
```

```
    % Transient detection in neighborhood of 15 frequency
        bands
    M = 7; % neighboring frequency bands
    delta_T_M = zeros(L/2,1);
    for m0 = M+1:L/2-M
        num = 0; den = 0;
        for m = m0-M:m0+M
            num = num + (M+1-abs(m-m0))/(M+1) * r(m,b0) *
                delta_T(m);
            den = den + (M+1-abs(m-m0))/(M+1) * r(m,b0);
            delta_T_M(m0) = num/den;
        end
        if ((b0>=b_T(m0,1))&&(b0<=b_T(m0,2))&&delta_T_M(m0)
            <=3)
            trans(m0,b0) = 1;
        end
    end
end

%% Crossfading
if nargin < 2
    cx=1/8;
end
dT = 9/10;
cf = zeros(L/2,blocks);
```

```
cf(:,1) = 1 - cx*X_env_H_bar(:,1) ./ X_env(:,1);
for b=2:blocks
    cf(:,b) = dT*cf(:,b-1); %disp(size(cf(:,b)))
    ii = find(trans(:,b));
    if numel(ii)==0, continue,  end
    cf(ii,b) = max( cf(ii,b), 1 - cx*X_env_H_bar(ii,b) ./
       X_env(ii,b));
end
% MDCT coefficients for transient and stationary parts of
   signal
X_T = cf .* X_MDCT; X_S = (1-cf) .* X_MDCT;

%% Extracting Transient/Stationary Components
% Inverse MDCT
fy = imdctv(X_MDCT); fy_trans = imdctv(X_T); fy_stat = imdctv
   (X_S);

% Rewindowing
fy = winit(fy,'sinewin');
fy_trans = winit(fy_trans,'sinewin'); fy_stat = winit(fy_stat
   ,'sinewin');

% Overlap-adding (OLA)
y_orig  = linunframe(fy,fhop,fpad);
y_trans = linunframe(fy_trans,fhop,fpad);
y_stat  = linunframe(fy_stat,fhop,fpad);
```

```
y_res    = y_orig-y_trans;
```

Extrapolation

Extrapolation of a signal is a predictive procedure that is performed in both the forward and backward directions. For a given vector $\mathbf{x} = [x_1, x_2, x_3, \ldots, ..., x_N]$, the following $[x_{N+1}, x_{N+2}, \ldots]$ and preceding $[\ldots, x_{-2}, x_{-1}, x_0]$ samples are calculated by forward and backward extrapolation, respectively. Sample prediction is performed according to Equation 3.30:

$$\tilde{x}_n = \sum_{i=1}^{M} h_i x_{n-i} \tag{3.30}$$

using the impulse response coefficients $\mathbf{h}$. These coefficients are obtained from the prediction error coefficients $\mathbf{a} = [1, a_1, a_2, \ldots, a_p]$ and modifying them according to Equation 3.31:

$$\mathbf{h} = [h_1, h_2, \ldots, h_M] = [0, -a_2, -a_3, \ldots, -a_p]. \tag{3.31}$$

The Burg method has been proven to provide good results in predicting error coefficients with regards to audio signals [Bur68]. If $\mathbf{x}$ contains at least M known samples ($N \geq M$), a forward prediction of the succeeding sample x_{N+1} is possible, resulting in $[x_1, x_2, x_3, \ldots, x_N, x_{N+1}]$. The next sample x_{N+2} is computed using the new last M samples. When this procedure is iteratively applied, an arbitrary amount of new samples may be generated. In order to extrapolate W samples using an IIR filter, the following procedure may be used:

1. Calculate impulse response coefficients $[h_1, h_2, \ldots, h_M]$.

2. Initialize the filter with M past known samples ahead of the extrapolation region.

3. Feed a zero vector of length W into the filter.

After the extrapolated signal is processed by the PSOLA algorithm, the extracted transients must be correctly replaced. Since the transients are surrounded by pitch-shifted periodic grains of the original signal, the transient part of the signal must be crossfaded into the periodic parts in order to avoid a cracking sound. The fading function suggested is a half Hanning window:

$$x_{fadein}(n) = 0.5(1 - \cos(\frac{2\pi n}{N-1})), \tag{3.32}$$

$$x_{fadeout}(n) = 1 - x_{fadein}(n) \tag{3.33}$$

$$\text{with } n \in [0, \frac{N-1}{2}].$$

The width N of the window is defined as twice the length of the fade's transition region.

MATLAB Code for the Enhanced PSOLA Pitch Shifting System

The overall enhanced PSOLA system coded to include transient extraction and extrapolation is shown in Listing 3.5.

Listing 3.5: Enhanced PSOLA Code

```
function [out, pm, y_stat, y_trans, fs] = ...
    psola_jmc_frames(fname,f_range,hop,frameLen)
%PSOLA_JMC Enhanced PSOLA system, including transient
   extraction
```

```
%% Transient Detection
%(method from "Detection and Extraction of Transients for
    Audio Coding," by
% Niemeyer and Edler, using MATLAB functions from Marios
    Athineos)
[~,~,y_trans,y_stat,fs] = trans_det_marios(fname,frameLen,hop
    );
% y_stat is the "stationary" signal, a.k.a the residue signal
    from
% transient removal, y_trans is the transient signal

%% Pitch Detection (YIN algorithm, as in DAFX 2nd. Ed., 2011,
    U. Zolzer)
f0min = f_range(1); f0max = f_range(2);
[F0] = yinDAFX(y_stat,fs,f0min,f0max,hop);

%% Pitch Marking (basic method in DAFX 2nd. Ed., 2011, U.
    Zolzer)
[pm] = findpitchmarks(y_stat,fs,F0,hop,frameLen);

%% PSOLA (Pitch-Synchronous Overlap-Add, DAFX 2nd. Ed., 2011,
    U. Zolzer)
% alpha = 1 => constant time-scale,
% beta => pitch scale factor
[out] = psola( y_stat,pm,1,beta ); out = out + y_trans; out =
    out/max(abs(out));
```

```
end
```

3.2 Framework for this Research

The presentation of outlines describing several types of signal processing manipulations that can be applied to audio and speech signals makes it possible to summarize the framework for this research.

An inherent limitation in screen-reader presentation is verbosity. Fundamentally, a user will want exactly as much information as he or she desires and nothing more. In his famous and widely cited work, Miller showed that a typical person has a working short-term memory of seven chunks of information, plus or minus two chunks [Mil56]. When a user receives less than he or she desires, he or she may become confused, disoriented, or frustrated by the lack of information. Conversely, if he or she hears too much information for a prolonged period, he/she will become annoyed at the excess speech directed at him/her.

The aim of this research is the development of a non-visual, auditory system for blind and low-vision individuals to access tabular data. This study seeks to determine the efficacy of organizing information into tabular structures with auditory cues for guidance. Sighted consumers of tabular data are able to access the inherent semantic meaning of the table from its structure. They use the visual intersection of rows and columns to glean meta-information, intuit trends by scanning across rows and columns, and can immediately sense the size and shape of the data. It is of interest to determine if digital audio manipulations could afford similar capabilities to blind navigators of web-based tabular data.

3.2.1 Critical Questions to be Answered by the Research

The questions posed by this study aim to determine whether sonifications, stereo spatialization, and tonal variation have an impact on non-visual tabular navigation, as measured by time-to-target (TTT) and navigational moves.

1. Do screen-reader users have significantly different times- to-target (TTT) when the synthesized speech is stereo-spatialized?
2. Do screen-reader users have significantly different times-to- target (TTT) when the synthesized speech is preceded by a tone varying in pitch?
3. Do screen-reader users have significantly different times-to-target (TTT) when the synthesized speech is both stereo-spatialized and preceded by a tone varying in pitch?

3.3 Auditory Cues

3.3.1 Non-verbal Cues

This study investigates the usefulness of non-verbal cues to efficiently guide a non-visual user through a tabular data structure. When confronted with the task of aurally presenting non-linear content, screen-readers must make a trade-off between added verbiage and the potential for a lack of focus. This study replaces verbosity with non-verbal cues and sonic parameter variation. According to [BBR+03], "the inclusion of non-speech audio lead[s] to significantly faster, and more accurate, completion of tasks. In addition, the workload [involved in non-visually navigating sonified tables], ... measured using NASA TLX scales [HS88], [is] significantly lowered when non-speech sounds [are] available." These conclusions are based on the findings of by Ramloll *et al.* [RBYR01].

Besides the unique issues associated with conveying two-dimensional information aurally, there must be a sonic mechanism to keep non-visual navigators within the confines of the table. When boundaries cannot be perceived visually, Brown *et al.* suggest the following [BBR+03]:

> Users should be informed when they reach the boundary of a table or a graph. Upon reaching the boundary a non-speech sound (such as a percussive sound), which is distinct from the sounds heard in the table, should be played to indicate this. In addition they should not be allowed to cross these boundaries. If they attempt to do so the boundary sound cue should be repeated to indicate that they are still at the boundary and cannot cross it.

This study implemented an "electrified fence" sound as a sonic metaphor to reinforce the concept of a tabular boundary. Participants were not allowed to navigate outside of the table, as that would be confusing to the users. Audible buzzes were sounded to indicate the pressing of an incorrect key. These sound effects are critical to non-verbally indicate navigational status, errors, and real-time events.

3.3.2 Verbal Cues

While non-verbal cues have been found to be helpful in increasing the efficiency of non-visual tabular browsing, " users can easily become lost in a large table presented in audio [BBR+03]." Brown *et al.* further instructed:

> . . . users can easily become lost in a large table presented in audio so they should be able to access information about their location (row and column numbers) at any time. However, unnecessary information is distracting and overloads short-term memory so users should be allowed to choose

> when they want to access their location information rather than being forced to hear it at all times.

Since this study intends to replace verbal cues with non-verbal cues, the tables are all small in dimensions (five rows, six columns).

3.3.3 Hybrid Approach

Theoretically, the best approach would be to directly manipulate the qualities of the speech output. No tones or sound effects would need to be appended; this would mean that the overall duration of presentation would be reduced. The speech could be represented by different synthesized voices, where variables like perceived gender, age, pitch, stress/emphasis, and timbral qualities could be varied. In addition, speech could be accelerated or decelerated to suit the listener.

While the above-mentioned manipulations may seem ideal, speech quality and consistency is paramount. In a series of six studies by Brown *et al.*, the researchers found that "changing the pitch of speech output voice ... does not improve memory performance [BBR+03]." They continue to say:

> "memory performance is not improved when different speech output voices are used to organize verbal information into larger units [aurally]. One voice should be used to present tabular data verbally. This should not be manipulated by imposing auditory structure on the information with a view to improving memory performance."

According to these considerations by Brown *et al.*, as well as a quality assessment of the speech shifted in pitch by the custom implementation of the extended PSOLA system, it was decided that the main experimental phase of this research would faithfully employ a singular male voice at a fixed register and pace.

3.4 Shortcuts

Brown *et al.* [BBR+03] that "if a user presses another navigation key while any speech information is being read out, the speech should be stopped immediately. Ramloll *et al.* found that it was frustrating for users to have to wait for speech to finish before moving on [RBYR01]." This study has faithfully implemented this suggestion. Navigational keystrokes interrupt any previously triggered speech or sound effects. In general, no speech or sound effect outputs are played simultaneously, which aids the overall clarity of the sonified table interface. In addition, they inform that "navigation shortcuts are extremely helpful, so [that] users should be able to jump to the start or end points of rows or columns at any time [BBR+03]." This study, emphasizing the semantic grouping of the row arrays, implemented shortcuts to the beginning and end of the row using the 's'-key to reach the leftmost cell and the 'e'-key to reach the rightmost cell of a given row. The (perhaps) more obvious choices of the 'l'- and 'r'-keys were avoided, as the 'l'-key would require the user to either move the left hand over from the other keys ('c','e', and 'r'), or to have the user shift the right hand from the directional arrows, which are critically important for navigation.

3.5 Spatialization Techniques

By replacing verbally explicit descriptions of cellular focus with dynamic relative spatialization, some verbosity can be eliminated. Sonic spatialization can be accomplished by leveraging two fundamental properties of human psychoacoustics, namely sonic localization and pitch perception.

3.5.1 Stereo Panning

Human hearing is binaural in nature. By comparing the volume (human perception of sonic amplitude) sensed by each ear, a human perceives sonic localization. The psychoacoustic phenomenon of sonic localization can be exploited to artificially localize sound sources. In his pioneering work, Blumlein discovered that by simultaneously playing a sound from two loudspeakers at a differential in amplitude, a person will perceive the sound as originating from somewhere along the continuum of space between the two loudspeakers [Ale99]. The apparent location of origination is relative to the ratio of amplitudes, where the speaker with greater amplitude will drag the virtual sound source closer to it from the middle. The technique of varying the perceived sound location is called stereo panning. The synthesized speech audio output from the screen reader can be panned (in real-time) to indicate the location of a text item along the horizontal dimension of the table. Ramloll *et al.* studied the use of "stereo panning through headphones . . . to indicate the location of a cell within a table [RBYR01]." Although the benefits of this technique were not formally evaluated, pilot studies suggested that "the majority of participants were pleased with the panning, and none found the mapping confusing [BBR+03]." In this research study, stereo panning will only be horizontally representative. This is to say, listener-subjects will perceive stereo spatialization (if activated) while traversing across a row, but never while traversing vertically through a column. It has been recommended not to place headphones on blind and low-vision individuals, as their immersive auditory characteristic may serve to alienate them from the real-world acoustic environment.

3.5.2 Tonal Variation

Humans can perceive changes in sonic frequency as distinct pitches on a one-dimensional continuum [Rit67]. When the frequency of a tone is increased or decreased by a factor of two, humans typically perceive a psychoacoustic relationship known commonly as an octave. In Western music, octaves are typically split into scales of eight tones. These scales can be exploited to correspond to sonic locations. A tone, prepended to the TTS representation of each item, can be manipulated to achieve graduated relative auditory orientation. This prepended tone also serves to alert the user-navigator that he or she is focused on a link (gleaned from Googles ChromeVox accessibility extension).

3.6 Main Experimental Design Strategy

Upon review of the literature and preliminary experiments outlined through the previous sections in this chapter the strategy for the main experimental phase of this research was decided. This phase would explore the potential benefits of stereo spatialization and/or variations of the prepended tone in the acoustic rendering of each cell in a table.

In particular, it was decided that the linear panning (LP) technique should be used, as the preliminary experimentation that compared it to VBAP panning, revealed no significant advantage in using the more complex approach. Further, linear panning is frequently used in many available implementations of this effect and would, therefore, not require highly specialized custom implementations. This is in keeping with the intent expressed at the beginning of the project to strive for as much portability as possible.

Similarly, it should be observed that there are advantages in applying tonal variation manipulations to the tone that is prepended to the informational speech describing each cell in a table:

1. The critical quality of the speech description is not compromised.
2. The tone-to-location mapping can still be implemented.
3. As the multiple variations of the prepended tone are completely independent of the contents of the cell they can be prerecorded and simply selected according to cell location, in real time.

3.7 Summary

This chapter described the methodology behind this research study. A preliminary study investigating the feasibility of prosodic modification for non-visual tabular navigation was described. The framework of this research study was enumerated in the form of research questions. The hybrid approach (using both verbal and non-verbal cues) was discussed. Finally, the spatialization techniques (stereo and tonal variation) were explained as they pertain to this research study.

CHAPTER 4

EXPERIMENT DESIGN

In order to address the critical questions identified previously for this research, an experimental system was created that would selectively provide auditory guidance by means of stereo spatialization, variation of a prepended tone, or both. Different levels of auditory guidance were made available in different sonified tables that several blind volunteers were asked to navigate, after a brief training phase.

4.1 Software Design

Considering that the Web is currently the most common way of sharing information, the sonified tables were implemented as an accessible website. The sonified table architecture was implemented with the three most common Web development frameworks:

HTML HyperText Markup Language

JS JavaScript

CSS Cascading Style Sheets

The content of the tables was generated in HTML (specifically using the $< table >$ ... $< /table >$ tags, see Listing 4.4). JavaScript, a client-side dynamic scripting language, was used to facilitate the interactions between the participant and the tables. SoundManager2 was the core JavaScript API that made calls to Adobe Flash, enabling real-time dynamic audio playback within the Google Chrome browser, which acted as a virtual machine. A custom JavaScript object was written to handle the keyboarding inputs and dynamic sound playback. All of the speech was synthesized with the online service iSpeech (http://www.ispeech.org/). Pre-synthesized audio files

(synthesized speech, tones, and sound effects) were used to prevent variable latency in speech playback during navigation.

To give an impression of the custom JavaScript, Listing 4.1 describes how the interface handled keystrokes in real-time to move the cellular focus throughout the sonified HTML table.

```
function leftArrowPressed() {
  if(eval(curr_focus.id[1])!==1){
    curr_focus = document.getElementById( 'D' + eval(--
       curr_focus.id[1]) );
    curr_focus.focus();
  } else{
    sm.stopAll(); sm.play('warning_edge');
  }
}

function rightArrowPressed() {
  if(eval(curr_focus.id[1])!==5){
    curr_focus = document.getElementById( 'D' + eval(++
       curr_focus.id[1]) );
    curr_focus.focus();
  } else{
    sm.stopAll(); sm.play('warning_edge');
  }
}

function downArrowPressed() {
  var headline_index = eval(curr_focus.id[3]);
  if (headline_index != rows){
```

```
    headline_index++;
    var new_id = curr_focus.id.slice(0,3) + headline_index;
    curr_focus = document.getElementById(new_id);
    curr_focus.focus();
    if (display) { $('#focus').html(curr_focus.innerHTML); }
    return false;
  }
  else{
    sm.stopAll(); sm.play('warning_edge',{pan: 0});
    return true;
  }
}

function upArrowPressed() {
  var headline_index = eval(curr_focus.id[3]);
  if ( (headline_index==1 && transposed) || (headline_index==0
      && !transposed) ){
    sm.stopAll(); sm.play('warning_edge',{pan: 0});
    return true;
  } else{
    headline_index--;
    var new_id = curr_focus.id.slice(0,3) + headline_index;
    curr_focus = document.getElementById(new_id);
    curr_focus.focus();
    if (display) { $('#focus').html(curr_focus.innerHTML); }
    return false;
  }
}
```

Listing 4.1: Directional Arrow Keystroke Handler Functions

Listing 4.2 shows how the stereo spatialization was performed. The range (left: -100, right: +100) represents the panning spectrum with the "0" value representing a source audio-spatially centered between the speakers.

```
function setPanning(column){
  if (pan_option){
    switch (eval(column[1])){
      case 0:
        panning = 0;
        break;
      case 1:
        panning = -100;
        break;
      case 2:
        panning = -50;
        break;
      case 3:
        panning = 0;
        break;
      case 4:
        panning = 50;
        break;
      case 5:
        panning = 100;
        break;
      default:
        console.log('Spatialization FAIL');
```

```
    }
  }
  else{panning = 0;}
  return panning;
};
```

Listing 4.2: Stereo Panning Handler Function

Listing 4.3 shows have the keystroke events were handled.

```
// Capturing the keystroke code
if (!e) var e = window.event;
var code; //alert(code);
if (e.keyCode) code = e.keyCode;
else if (e.which) code = e.which;

curr_focus = document.activeElement; // Setting the focus

// Binary navigation variables
self.border_reached = false, self.wrong_key = false, self.
   target_key = false, self.record_position = true;
self.space_bar = false, self.instruct = false, self.locate =
   false;

switch (code) {
  case 67: // 'c' key to repeat the categories
    sm.stopAll();
    if (train_sequence[train_index]!='c_key'){
      sm.play('wrong_key');
      self.wrong_key = true;
```

```
    }
    else{
      autoCatReader(1,true);
      self.record_position = false;
      return;
    }
    break;

  case 82: // 'r' key to read instructions
  case 74: // 'j' key to read instructions
    sm.stopAll();
    if(train_sequence[train_index]=='search'){
      self.instruct = true; heard_instruction[link] = true;
         instruct_repeats[link]++;
      sm.play('find', { onfinish: function(){self.instructions
         [link].play();} });
      console.log('Instruction #' + link + ': playing...');
      self.record_position = false;
    }
    else{ sm.play('wrong_key'); self.wrong_key = true;}
    break;

  case 32: // spacebar
  case 101: // numpad 5
    self.targetKeyPressed();
    self.target_key = true;
    self.record_position = false;
    break;
```

```
case 69: // 'e' (shortcut) key in transposed view to go to
    the right-most cell in the row
  if (eval(curr_focus.id[1])==cols) {self.record_position =
     false;}
  else{
    curr_focus = $('#C' + cols + 'R' + curr_focus.id[3]);
       curr_focus.focus();
    if(display){ $('#focus').html(curr_focus.html()); }
  }

case 83: // 's' (shortcut) key in transposed view to go to
    category description
  curr_focus = document.activeElement;
  sm.stopAll(); must_press_s = false;
  curr_focus = $('#COR' + curr_focus.id[3]);
  curr_focus.focus();
  self.target_key = false;
  self.handleKeyboard(e, false);
  break;

case 37: // left
case 100: // numpad 4
  self.border_reached = self.leftArrowPressed();
  break;

case 39: // right
case 102:// numpad 6
```

```
      self.border_reached = self.rightArrowPressed();
      break;

  case 38: // up
  case 104: // numpad 8
      self.border_reached = self.upArrowPressed();
      break;

  case 40: // down
  case 98: // numpad 2
      self.border_reached = self.downArrowPressed();
      break;

  default:
      sm.stopAll(); sm.play('wrong_key');
      self.wrong_key = true;
      num_wrong_keys[link]++;
      break;
}
```

Listing 4.3: Custom JavaScript Keystroke Handler

All of the sonified tables were constructed in HTML, with each cell consisting of a hyperlink. These hyperlinks, although not linking to another page or HTML element, served as a vehicle for sonified speech using the SoundManager2 JavaScript API (http://www.schillmania.com/projects/soundmanager2/). The sound playback, with respect to narrative speech feedback, would be interrupted if the user chose to change focus mid-audio-stream. This was an intentional decision so that audio files would not play concurrently. An example of one row of an HMTL sonified table is in

Listing 4.4. All of the audio files were stored in the MP3 format to take advantage of the format's compression scheme. Each table cell had a unique ID identifying its unique location, where the second character denoted the column location and the fourth character denoted the row location. The "sm2_button" CSS class is used to signal to the SoundManager2 JS API that the link is a playable MP3 file.

```
<tr>
  <!-- Row 1 -->
  <th>
    <a class="sm2_button" id="C0R1" href="./Audio/Purchasing/
       categories/vegetables.mp3" name="C0R1">Vegetables</a>
  </th>
  <td>
    <a class="sm2_button" id="C1R1" href="./Audio/Purchasing/
       vegetables/carrot.mp3" name="C1R1">Carrot</a>
  </td>
  <td>
    <a class="sm2_button" id="C2R1" href="./Audio/Purchasing/
       vegetables/potato.mp3" name="C2R1">Potato</a>
  </td>
  <td>
    <a class="sm2_button" id="C3R1" href="./Audio/Purchasing/
       vegetables/cucumber.mp3" name="C3R1">Cucumber</a>
  </td>
  <td>
    <a class="sm2_button" id="C4R1" href="./Audio/Purchasing/
       vegetables/onion.mp3" name="C4R1">Onion</a>
  </td>
  <td>
```

```
    <a class="sm2_button" id="C5R1" href="./Audio/Purchasing/
       vegetables/lettuce.mp3" name="C5R1">Lettuce</a>
  </td>
</tr>
```

Listing 4.4: HTML Sonified Table (Row)

4.2 Hardware Utilization

A generic, ergonomic Microsoft Windows-oriented USB keyboard was provided for user control. This keyboard had keys that were raised and distinct from one another to accommodate touch-typing. An ordinary pair of portable, amplified loudspeakers was utilized for stereo playback of the synthesized speech and sound effects. The speakers were placed approximately 50 cm. in front of each user, separated by about one meter laterally. Many participants opted to use their own keyboards, as they were equipped with raised landmarks that helped orient them with respect to the keyboard layout. All navigations were performed and all data were recorded with an Apple MacBook Pro laptop.

4.3 Tabular Content

Each auditory table consisted of commonly purchased items in physical stores, including a virtual grocery, pharmacy, home goods/furniture store, and gift store. These stores were chosen for their simplicity and universality. To maximize the chance of familiarity of tabular content with the subjects, brand names and proper nouns were completely avoided.

Figure 4.1: Ergonomic Microsoft Keyboard

Table 4.1: Grocery Auditory Table

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

4.4 Experimental Procedure

4.4.1 Training

Each subject was asked to navigate an "Audio Training Table" which would acquaint him or her with the various functions and keystrokes required by the interface design (see Figure 4.2). Subjects were prompted to find links within the table. After successfully employing the directional arrows to focus on the intended target, subjects would then press the "spacebar" to confirm their selection. The training module then required the subject to use the "s" key, emphasizing the use of the shortcut to

Table 4.2: Personal Care Auditory Table

Cosmetics	Lipstick	Eyeliner	Mascara	Skin Cream	Eye Shadow
Hair Care	Hairspray	Hair Brush	Hair Dye	Hair Gel	Shampoo
Oral Care	Toothpaste	Mouth Wash	Toothbrush	Dental Floss	Toothpick
Baby Care	Pacifiers	Bottles	Formula	Diapers	Baby Wipes
First Aid	Splint	Brace	Bandaid	Crutch	Antiseptic

Table 4.3: Gifts Auditory Table

Jewelry	Bracelet	Anklet	Earrings	Diamonds	Necklace
Clothing	Jacket	Pants	Shirts	Shoes	Skirts
Toys	Dolls	Board Games	Card Games	Building Blocks	Puzzles
Electronics	Video Games	Laptops	Music Players	Cameras	Tablet Computers
Tickets	Concerts	Amusement Parks	Movies	Sporting Events	Comedy Shows

the categorical column. In this way, the categorical column could be instilled non-visually. After successfully finding three links in the above-described manner, the subject would then be prompted to press the "c" key to recall the category headers in a successive top-to-bottom sequence. After training the basics of tabular navigation, subjects were exposed to an "Audio Spatialization Demonstration," which consisted of a single five-cell row (see Figure 4.3).

This row was stereo spatialized such that each cell had been panned at a proportional angle in auditory space (see Listing 4.2). After moving the audio cursor from left-to-right and then from right-to-left, the subject had the opportunity to hear all five sonic source locations at least once. The next exercise consisted of tonal variation training. As in the audio spatialization exercise, a five-cell row (see Figure 4.3) would

Audio Table Training

Clothing	Underwear	Pants	Hats	Belts	Shirts
Doctors	Podiatrist	Dentist	Cardiologist	Pediatrician	Neurologist
Food	Yogurt	Hamburger	Pizza	Chicken	Taco
Home Care	Pool Care	Plumber	Maid	Electrician	Gardener
Transportation	Car	Train	Airplane	Bus	Taxi

Figure 4.2: Training Table for Augmented Auditory Table Navigation

Table 4.4: Home Care Auditory Table

Furniture	Chairs	Tables	Couches	Desks	Beds
Home Office	Scissors	Paper	Staples	Pencils	Pens
Kitchenware	Fork	Knife	Spoon	Plates	Glasses
Cleaning Supplies	Mop	Broom	Soap	Duster	Vacuum Cleaner
Lawn Care	Lawnmower	Shovel	Rake	Hedge Clipper	Fertilizer

Audio Spatialization Demonstration

Column 1	Column 2	Column 3	Column 4	Column 5
Left	Middle Left	Middle	Middle Right	Right

Tonal Variation Demonstration

Column 1	Column 2	Column 3	Column 4	Column 5
Left	Middle Left	Middle	Middle Right	Right

Figure 4.3: Training Rows for Stereo Spatialization and Tonal Variation

be navigated horizontally. However, the user would hear the preceding tone rise in pitch as he or she navigated from left-to-right, and descend in pitch from right-to-left.

4.4.2 Testing

After training, each participant was redirected to one of the testing tables. After a 90-second free-exploration phase, the subject would be prompted to navigate to a link. This navigation would be repeated, for a total of ten navigations being performed for each table. These navigations, unbeknownst to the navigator, were identical in location for all four sonified tables (see Listing 4.5).

```
function BasicMP3Player() {
  \\...
  this.articles = ["C2R1","C5R3","C3R2","C1R5","C4R4","C1R2","
     C2R1","C4R3","C3R5","C5R4"];
  \\...
}
```

Listing 4.5: Cell Locations for Navigations

4.5 Pool of Experimental Subjects

Sixteen individuals volunteered to participate in this study. They ranged in age from 20 to 67 years old, with an average of 44.1 years (std. dev. = 13.7 years). Adult subjects who met the following criteria were considered for the study:

- legal blindness
- experience with touch-typing
- fluency in English

No prior experience with screen-reading technology was required; however, many of the subjects were well acquainted with various screen-readers. Subjects were recruited from disability resource centers throughout the greater South Florida area.

4.6 Summary

In this chapter, the design of the experiment for this research study was described. The design of the various software packages, MAMP stack, and APIs used and bundled were described. The hardware and peripherals used in this study were listed. The tabular content presented in the form of routine online shopping exercises was presented. The various stages of the experimental procedure were enumerated. The experimental subject participants were described, with their demographic data listed and summarized.

CHAPTER 5

RESULTS & DISCUSSION

Each participant's navigational data were recorded in real-time using custom JavaScript (the jQuery and SoundManager2 JavaScript libraries were used extensively). Each movement and relevant keystroke were recorded such that the specific path taken (and number of moves) is known. These JavaScript variables were then permanently stored into a MySQL database using MAMP as a local server environment, consisting of:

Mac OS X as the operating system,

Apache as the web server,

MySQL as the database management system, and

PHP as server-side scripting language.

To assess user confusion and error, problematic events were recorded. Such events included the pressing of an irrelevant key to the defined navigational system, a wrong link selected, or a border bounce (when the user attempted to navigate beyond the boundaries of the table).

5.1 Personal Interviews

Information about the efficacy of the study and personal screen reader use was obtained based on oral interviews conducted with 14 of the subjects. By far, Freedom Scientific's JAWS was the most commonly used screen reader. The length of time of screen reader use varied from less than one year to up to 17 years. Seven of the subjects reported that they use Microsoft's Excel spreadsheet software with a screen

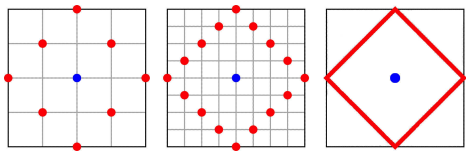
reader. Only two subjects stated that they presently take advantage of online shopping (eBay, Amazon.com, etc.), with many citing inaccessibility for screen readers as a primary hindrance.

5.2 Quantitative Analysis

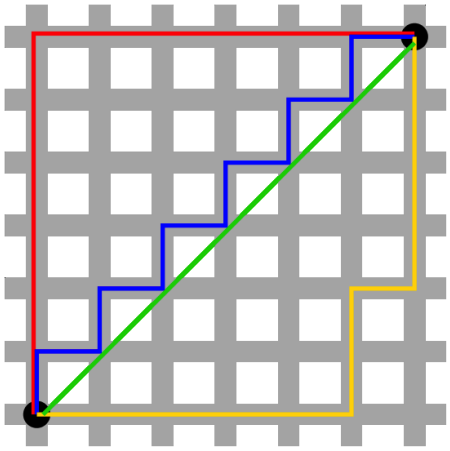
As each participant performed his or her various navigations in each navigational table, his or her navigational data were recorded, such as the TTT and the number of moves needed to reach the stated target. While these two aspects of navigation are readily recorded and understood, it is less clear how to compare navigational data from one user to another as well as from one navigation task to another. The users mental model heavily influences non-visual screen reader usage [KSB03]. Each navigation is unique in that the user must use his or her mental model of the table to map out a course toward the intended target. For a sighted navigator, this would merely be the Manhattan distance (also known as taxicab geometry), or the shortest path of horizontal and vertical moves needed to reach the target. The taxicab distance, d_1, between two vectors $\mathbf{p}$, $\mathbf{q}$ in an n-dimensional real vector space with fixed Cartesian coordinate system, is the sum of the two lengths of the projections of the line segments between the points onto the coordinate axes. More formally,

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i| \tag{5.1}$$

where $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ are vectors.



(a) "Circles" in Taxicab Geometry



(b) Manhattan Distance Paths

Figure 5.1: Manhattan Distances in Taxicab Geometry

For a coordinate plane scenario mimicking the tabular format of this study,

$$d_1(\mathbf{p}, \mathbf{q}) = |p_1 - q_1| + |p_2 - q_2| \tag{5.2}$$

where $\mathbf{p} = (p_1, p_2)$ and $\mathbf{q} = (q_1, q_2)$ are vectors.

Manhattan distances (a result of taxicab geometry) are exemplified in Figure 5.1. In Figure 5.1a, the three sets of "circles" ($radius = \{2, 4, n\}$, (left-to-right, respectively)) represent sets of equidistant points in taxicab geometry. Figure 5.1b depicts four arbitrary paths from the bottom-left point to the upper-right point. Notice that the straight diagonal path represents the Euclidean distance, while the other three adhere to taxicab geometry.

For the non-visual navigator, the path followed in the experimental tables may be much less straightforward. Many blind navigators will choose to return to the

category-column, move vertically to the appropriate category-row, and then navigate across to the intended cell. In this scenario, there is no ideal path or number of movements. The existence of the shortcut to the category-column complicates the notion of shortest path. However, this capability was considered necessary to emulate a real-life navigational environment.

Initially, a time-per-move (TPM) metric to equitably compare across different tables and participants was considered. The TPM was computed by dividing the TTT by the number of moves in each navigation. In theory, this metric considers both aspects of navigational efficiency:

- reducing movement(s)
- decreasing task-completion time

A larger number of movements involves more listening and imposes a larger mental burden. Unfortunately, a low TPM could also be achieved by simply moving from cell-to-cell in a random or arbitrary fashion and then clicking the appropriate link upon eventual happenstance. This behavior is clearly undesirable, and it may represent frustration or lack of mental exertion on the part of the non-visual navigator. Noting this, TTT was elected as the metric for evaluating non-visual navigation.

The Statistical Package for the Social Sciences (SPSS) software package was used to analyze the results of the experiment. The general linear model was used for our two-factor repeated-measures ANOVA analysis. The two factors considered were labeled "Tonal" and "Stereo," to indicate the presence (2) or absence (1) of variation in the prepended tone and the linear panning in rendering the synthetic speech, respectively. The SPSS syntax is shown in Listing 5.1.

```
GLM TTT_A_AVG_log TTT_B_AVG_log TTT_C_AVG_log TTT_D_AVG_log
  /WSFACTOR=Tonal 2 Simple Stereo 2 Simple
  /METHOD=SSTYPE(3)
```

```
  /PLOT=PROFILE(Stereo*Tonal)
  /EMMEANS=TABLES(Tonal) COMPARE ADJ(BONFERRONI)
  /EMMEANS=TABLES(Stereo) COMPARE ADJ(BONFERRONI)
  /EMMEANS=TABLES(Tonal*Stereo)
  /PRINT=DESCRIPTIVE TEST(MMATRIX)
  /CRITERIA=ALPHA(.05)
  /WSDESIGN=Tonal Stereo Tonal*Stereo.
```

Listing 5.1: SPSS Syntax - Repeated Measures: log(TTT)

Preliminary analysis of the TTT values recorded during the experiment showed that the populations did not conform to the ANOVA requirements of normality and equality of variances. Therefore, in order to meet the normality and equality of variances assumptions of ANOVA, the data were transformed using the natural logarithm [Fie09].

The means and standard deviations of the logarithmically-transformed TTTs for each navigational table are listed in Table 5.1. This table also shows the assignment of auditory manipulations to each of the four tables navigated by each subject. So, for example, the "B" navigation table included speech panning ("Stereo"=2), but did not include tonal variation ("Tonal"=1). Conversely, the "C" navigation table included tonal variation ("Tonal"=2) but lacked panning of the speech ("Stereo"=1).

The ANOVA results (see Table 5.2) show that the logarithmically-transformed TTTs were significantly affected by the tonal variation method, V=0.292, F(1,15)

Table 5.1: Within-Subjects Factors - log(TTT_AVG)

Tonal	**Stereo**	**Table Letter**	**Mean**	**Std. Deviation**
1	1	A	2.3698	.55103
1	2	B	2.1366	.35275
2	1	C	2.4394	.43523
2	2	D	2.3920	.47332

Table 5.2: Tests of Within-Subjects Effects - log(TTT_AVG)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Tonal	.423	1	.423	6.194	.025
Error(Tonal)	1.024	15	.068		
Stereo	.315	1	.315	4.240	.057
Error(Stereo)	1.114	15	.074		
Tonal * Stereo	.138	1	.138	1.381	.258
Error(Tonal*Stereo)	1.501	15	.100		

Table 5.3: Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	87.196	1	87.196	582.327	.000
Error	2.246	15	.150		

= 6.194, p= 0.025. Similarly, the results show that the logarithmically transformed TTTs were marginally affected by the stereo spatialization method, V=0.220, F(1,15) = 4.240, p=0.057. The results show that the logarithmically transformed TTTs were not significantly affected by the interaction of both methods, V=0.084, F(1,15) = 1.381, p=0.258. This result suggests that some confusion may be caused in the subject when employing both of these methods simultaneously. While statistical significance was observed for some of the factors, the practical meaning of this significance will be explored in the following sections.

A statistical test of between-subjects effects is stated in Table 5.3. Noting the results of the test, $F(1, 15) = 582.327$, $p < 0.001$, the extreme significance indicates a great degree of variability between the subjects, which should not be trivialized. Each participant had a different level of attention, capacity for recall, relationship with semantic information, hearing ability, exposure to synthesized speech, familiarity with screen readers, etc.

Table 5.4: Estimated Marginal Means: Tonal Variation

Tonal	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	2.253	.100	2.040	2.467
2	2.416	.104	2.194	2.637

Table 5.5: Pairwise Comparison: Tonal Variation

(I)	(J)	Mean Difference (I-J)	Std. Error	Sig. b	95% Confidence Interval for Difference b	
					Lower Bound	Upper Bound
1	2	-.163*	.065	.025	-.302	-.023

Based on estimated marginal means
* The mean difference is significant at the alpha = 0.05 level.
b. Adjustment for multiple comparisons: Bonferroni.

5.2.1 Tonal Variation

The significant effect of tonal variation was stated earlier; however, Tables 5.4 and 5.5 indicate that the effect is actually *increasing* the average TTT. In other words, the presence of preceding tones increase task completion time on average.

5.2.2 Stereo Spatialization

As can be seen in Tables 5.6 and 5.7, the marginally-significant effect of stereo spatialization decreases the average log(TTT) from 2.405 to 2.264.

Table 5.6: Estimated Marginal Means: Stereo Spatialization

Stereo	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	2.405	.114	2.161	2.648
2	2.264	.089	2.074	2.455

Table 5.7: Pairwise Comparison: Stereo Spatialization

(I)	(J)	Mean Difference (I-J)	Std. Error	Sig. b	95% Confidence Interval for Difference b	
					Lower Bound	Upper Bound
1	2	0.140	.068	.057	-.005	.286

Based on estimated marginal means
a. Adjustment for multiple comparisons: Bonferroni.

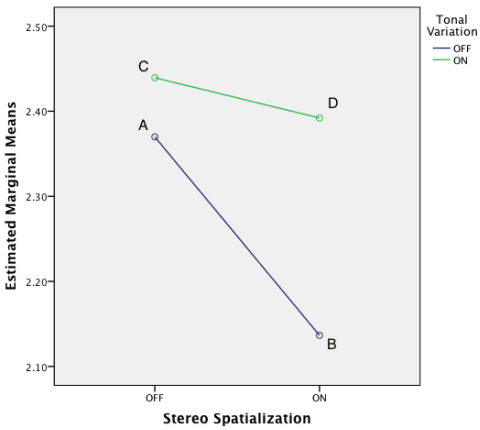


Figure 5.2: Estimated Marginal Means - Interaction of Tonal Variation and Stereo Spatialization

Table 5.8: Estimated Marginal Means

Tonal	Stereo	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	2.370	.138	2.076	2.663
1	2	2.137	.088	1.949	2.325
2	1	2.439	.109	2.208	2.671
2	2	2.392	.118	2.140	2.644

5.2.3 Interaction of Tonal Variation and Stereo Spatialization

The marginal means of the tabular navigation experiment can be seen in Figure 5.2 and Table 5.8. This figure and the corresponding table may be the best summary for the experimental results and ensuing analysis in this research.

In Figure 5.2, lines connect pairs of means, with each line-pair representing either tonal variation engaged or disengaged. The line for tonal variation engaged (ON), includes the results for Tables C and D, and lies above the line for tonal variation engaged (OFF), which includes results from Tables A and B. Additionally, both lines connecting the data points have negative slope. This indicates that the effect of tonal variation causes users to have longer task completion times (on average) regardless of stereo spatialization, while the effect of stereo spatialization causes users to have shorter task completion times (on average), regardless of the engagement of tonal variation.

5.3 Qualitative Feedback

Based on oral interviews conducted as a follow-up to each battery of tabular navigations, many users remarked that they would be more likely to engage in spreadsheet editing (e.g. Microsoft Excel) and online purchasing (e.g. eBay, Amazon, etc.) if tabular navigation were streamlined and presented in an audio-spatial modality. They remarked that the synthesized speech exceeded the quality that they were accustomed to in their existing screen-reading software.

Several subjects reported that they reflexively moved their heads to follow the perceived origin of the moving voice under the stereo spatialization treatment. This behavior clearly shows that the psychoacoustic effect of sonic localization is readily

perceived and intuitive. There was a mixed reception to the pitched tones; some users noticed them and made use of them, while others simply ignored them. Some users noted that their hearing was weak in one ear and that, for them, tonal variation was more effective for auditory orientation.

5.4 Discussion

We opted to use an earcon-based approach as opposed to an auditory icon approach, noting the lack of well-correlated sounds for our purchasing exercise. Considering that time for training, testing, and interviewing would be at a premium, we opted not to pursue additional training necessary to associate auditory icons to specific text or rows/columns, as that would be ineffectual in a real-world browsing scenario. We believe that earcons, which can be arbitrarily composed and varied, serve as the best means for distinguishing rows and columns for arbitrary, dynamically changing text.

After performing an extensive literature review, it seems as if no other research has previously incorporated both earcons and stereo spatialization into auditory table navigation. While much has been made of numerical table sonification through tones [RBYR01], and spatialization has been considered for documents [GM99], taking advantage of these augmentations where textual content is concerned has not yet been addressed.

We have decided to use artificial tables in lieu of tables captured from the real world. Our rationale behind this decision considers the wide range of experiences that our subjects possess as a result of cultural, linguistic, educational, and cognitive differences. To be more specific, age is often a factor with familiar recognition of technology and current events, while alimentary preferences are informed by culture and experience. The diversity present in the South Florida metropolitan area ne-

cessitated using relatively universal objects encountered in the daily lives of many, such as staple foods, home furnishings, and the like. This philosophy was inspired by research into the role of aging with regard to fluid intelligence [TRH+12]. In their study, generic categorical links from eBay were harvested such that specifics of brand name and technological details were eluded by keeping the semantic relationships at a higher level. While it may be illuminating to have subjects traverse a live, real-world website, the aforementioned reasons confound attempts to measure sonification efficacy against simple semantic understanding.

Noting the significant effect of stereo spatialization on tabular navigation, we would strongly encourage web designers to consider a tabular layout as a blind-accessible alternative to conventional, visually-oriented webpages. While there has been some progress in automating the process of converting arbitrarily arranged content into semantically structured webpages, we believe that having a tabular format in mind (similar to the use of alt-tags) from the onset would best benefit the non-visual browser.

Our design philosophy was to keep everything as simple and intuitive as possible. Rather than opting for the use of head-related transfer functions (HRTFs), which are customized for each user relative to their unique pinnae (ear anatomy) and suffer from front-to-back confusion, we opted to use generic, off-the-shelf stereo speakers. Headphones, which vary in style and shape, were not considered, as they tend to create the psychoacoustic perception of hearing sound move within one‘s own head and restrict a blind person from remaining aware of his/her ambient surroundings. While a full 3-D spatialized sound environment may have been novel and allowed for far more possibilities of sonic localization, the availability of such systems is prohibitive to a user who must be able to access his or her computer on-the-go. Three-dimensional sonic spatialization would require a highly customized implementation of hardware and

software not available to the typical user. It has been further noted that HRTF sonic environments suffer from "front-to-back" confusion that may be listener dependent.

We opted to use what we believed to be intuitive features of auditory perception: sonic localization and relative pitch. It was originally expected that these intuitive features would not need extensive training and explanation. Gougoux *et al.* note that pitch discrimination in early-blind (0-2 years from birth) listeners is enhanced compared to both sighted people and those having acquired blindness later in childhood [GLL+04]. Early-blind infants take advantage of cerebral plasticity, which is considered at its peak in the first years of their lives. Through the study conducted by Gougoux *et al.*, they found that early-blind listeners were much more adept at discriminating pitches played sequentially in both the temporal and spectral domains. Since it is already challenging to find a representative sample of blind screen reader users in a given geographic area, we did not consider this factor (onset time of blindness). Through the course of the subject testing, a few of the most experienced screen-reader users remarked that they ignored the tonal guidance in order to better focus on the navigational task at hand. While this was an unexpected reaction in our participants, it does indicate that the treatment can be tolerated without great annoyance by those who wish to ignore it. Some other users stated that their hearing had deteriorated in one ear relative to the other and, consequently, they would prefer the tonal variation. For more effective implementation, users must be given the choice to enable one or both of the sonic enhancements to adjust for auditory limitations in pitch perception and sonic localization.

5.5 Summary

This chapter discussed the results of the experiment in this research study and detailed the statistical analyses performed to interpret those results. The feedback from personal interviews with the subject participants was described. Both qualitative and quantitative results seem to indicate that the panning of the speech describing the content of the table cells was easily perceived and beneficial for the navigation of the table. Similarly, quantitative and qualitative outcomes of the experiment seem to indicate that the spatial connotation of the variations in the prepended tones were not as easy to assimilate by the subjects and don not seem to be particularly beneficial to the users in their tabular navigation. Anecdotal comments from the subjects seem to confirm the notion that panning was perceived convincingly and effortlessly by the users (even causing some of them to turn their heads in the direction of the virtual sound source). Conversely, the assimilation of the spatial guidance meant to be provided by the variations of prepended tones seems to have been less intuitive for some users, a few of which actually opted to ignore this acoustic clue.

CHAPTER 6

CONCLUSIONS & FUTURE WORK

6.1 Conclusions

The marginally significant effect of stereo spatialization confirms the expectation that the presence of auditory enhancements could have an impact on the efficiency and comfort experienced by the experimental users in navigating the tables. This result highlights the potential benefits that may be possible by solely implementing a simple form of stereo panning that does not require any hardware changes (other than the inclusion of commonly available amplified stereo speakers). Fortunately, the minimal software modifications needed (such as utilization of the freely available SoundManager2 JavaScript utility) would be transparent to an actual end-user of this approach. The results obtained in the experiments and associated statistical analyses have revealed that the tonal variations that were applied to indicate spatial relations between cells were not as intuitive and easy to assimilate as was hypothesized. While it has been shown that tonal variations may be useful in other interaction contexts, such as the manipulation of auditory scrollbars by blind users [YW08], they were not as useful in our experiment. We speculate that this may be due to the mental effort necessary to map pitch change to spatial displacement during the browsing of our navigational tables. Perhaps, some level of musical training may also be necessary to fully appreciate tonal variation.

Noting the lack of help provided by the tonal variation method, the interaction of the combined enhancements did not mutually complement each other, as we would have hoped. Since the tonal variation method itself does not seem to be well assimilated, it does not enhance or complement the stereo spatialization method. Rather

than making navigation more efficient, the combination of the two methods seems to be adding to the listener-navigators cognitive burden.

6.2 Future Work

To better emulate a real purchasing exercise, items for purchase could be gleaned from websites such as Amazon or eBay [TRH+12]. In their work to assess the impact of fluid intelligence in information search, Trewin et al. required each of their participants to pass through multiple tables on the way to his or her targeted item for purchase. These tables were not necessarily square (NxN) tables, as real-world purchasing tasks may necessitate a limited number of wider rows. This hierarchical structure would better emulate auditory menus as commonly experienced in telephonic customer service [ENLS06]. Also, a shopping cart feature could be implemented for added realism.

The opinions from the experimental subjects captured after their participation in this research suggest that it may be beneficial if the two features of sound spatialization (prepended tonal variation and speech panning) could be enabled independently in a future prototype, to suit the preferences of each non-visual user, according to his/her auditory sensitivities and preferences.

To further evaluate a realistic scenario, noise and distractions could be incorporated into the study to account for suboptimal listening environments. Synthesized speech, in a realistic scenario, would be rendered on-the-fly by a dynamic TTS synthesizer. This synthesizer may cause unpredictable latencies and unnatural utterances, which would mimic an actual screen-reading experience.

BIBLIOGRAPHY

[Ale99] Robert Charles Alexander. *The inventor of stereo: the life and works of Alan Dower Blumlein*. Focal Press, 1999.

[AT00] Chieko Asakawa and Hironobu Takagi. Annotation-based transcoding for nonvisual web access. In *Proceedings of the fourth international ACM conference on Assistive technologies*, pages 172–179. ACM, 2000.

[AT08] Chieko Asakawa and Hironobu Takagi. Transcoding. In *Web Accessibility*, pages 231–260. Springer, 2008.

[B+94] Durand R Begault et al. *3-D sound for virtual reality and multimedia*, volume 955. AP professional Boston etc, 1994.

[BBR+03] Lorna M Brown, Stephen A Brewster, SA Ramloll, R Burton, and Beate Riedel. Design guidelines for audio presentation of graphs and tables. International Conference on Auditory Display, 2003.

[Ber05] Martin Bernstein. *An introduction to music*. Kessinger Publishing, LLC, 2005.

[BRK96] Stephen Brewster, Veli-Pekka Raty, and Atte Kortekangas. Earcons as a method of providing navigational cues in a menu hierarchy. In *People and Computers XI*, pages 169–183. Springer, 1996.

[BSG89] Meera M Blattner, Denise A Sumikawa, and Robert M Greenberg. Earcons and icons: Their structure and common design principles. *Human–Computer Interaction*, 4(1):11–44, 1989.

[Bur68] John Parker Burg. A new analysis technique for time series data. *NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics*, 1, 1968.

[CR93] Shan-Ju Chang and Ronald E Rice. Browsing: A multidimensional framework. *Annual review of information science and technology (ARIST)*, 28:231–76, 1993.

[DCK02] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[DDPZ02] P Dutilleux, G De Poli, and U Zölzer. Time-segment processing. *DAFX: Digital Audio Effects*, pages 201–236, 2002.

[EN06] Bernd Edler and Oliver Niemeyer. Detection and extraction of transients for audio coding. In *Audio Engineering Society Convention 120*. Audio Engineering Society, 2006.

[ENLS06] Elsa Eiriksdottir, Micheal Nees, Jeff Lindsay, and Raymond Stanley. User preferences for auditory device-driven menu navigation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 2076–2078. SAGE Publications, 2006.

[Fie09] Andy Field. *Discovering statistics using SPSS*. Sage publications, 2009.

[Gav86] William W Gaver. Auditory icons: Using sound in computer interfaces. *Human-computer interaction*, 2(2):167–177, 1986.

[Gav89] William W Gaver. The sonicfinder: An interface that uses auditory icons. *Human-Computer Interaction*, 4(1):67–94, 1989.

[GHS00] Carole Goble, Simon Harper, and Robert Stevens. The travails of visually impaired web travellers. In *Proceedings of the eleventh ACM on Hypertext and hypermedia*, pages 1–10. ACM, 2000.

[GLL+04] Frédéric Gougoux, Franco Lepore, Maryse Lassonde, Patrice Voss, Robert J Zatorre, and Pascal Belin. Neuropsychology: pitch discrimination in the early blind. *Nature*, 430(6997):309–309, 2004.

[GM99] Stuart Goose and Carsten Möller. 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)*, pages 363–371. ACM, 1999.

[HS88] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload*, 1(3):139–183, 1988.

[KSB03] Sri Hastuti Kurniawan, Alistair G Sutcliffe, and Paul Blenkhorn. How blind users' mental models affect their perceived usability of an unfamiliar screen reader. In *INTERACT*, volume 3, pages 631–638, 2003.

[LJ04] Cheng-Yuan Lin and Jyh-Shing Roger Jang. A two-phase pitch marking method for td-psola synthesis. In *INTERSPEECH*, 2004.

[MC90] Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5):453–467, 1990.

[Mil56] George A Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[MVV06] Wesley Mattheyses, Werner Verhelst, and Piet Verhoeve. Robust pitch marking for prosodic modification of speech using td-psola. In *Proceedings of the IEEE Benelux/DSP Valley Signal Processing Symposium, SPS-DARTS*, pages 43–46, 2006.

[OA98] Toshiya Oogane and Chieko Asakawa. An interactive method for accessing tables in html. In *Proceedings of the third international ACM conference on Assistive technologies*, pages 126–128. ACM, 1998.

[OISM06] Makoto Ohuchi, Yukio Iwaya, Yôiti Suzuki, and Tetsuya Munekata. A comparative study of sound localization acuity of congenital blind and sighted people. *Acoustical science and technology*, 27(5):290–293, 2006.

[RBYR01] Rameshsharma Ramloll, Stephen Brewster, Wai Yu, and Beate Riedel. Using non-speech sounds to improve access to 2d tabular numerical information for visually impaired users. In *People and Computers XVInteraction without Frontiers*, pages 515–529. Springer, 2001.

[Rit67] Roelof J Ritsma. Frequencies dominant in the perception of the pitch of complex sounds. *The Journal of the Acoustical Society of America*, 42:191, 1967.

[San09] J B Sanjaume. Audio signal transforming. Technical report, 2009.

[SSK09] Dimitris Spiliotopoulos, Panagiota Stavropoulou, and Georgios Kouroupetroglou. Acoustic rendering of data tables using earcons and prosody for document accessibility. In *Universal Access in Human-Computer Interaction. Applications and Services*, pages 587–596. Springer, 2009.

[Sum85] Denise Ann Sumikawa. Guidelines for the integration of audio cues into computer user interfaces. Technical report, Lawrence Livermore National Lab., CA (USA), 1985.

[SVN37] Stanley Smith Stevens, John Volkmann, and Edwin B Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8:185, 1937.

[TKK+08] Hironobu Takagi, Shinya Kawanaka, Masatomo Kobayashi, Takashi Itoh, and Chieko Asakawa. Social accessibility: achieving accessibility through collaborative metadata authoring. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, pages 193–200. ACM, 2008.

[TRH+12] Shari Trewin, John T Richards, Vicki L Hanson, David Sloan, Bonnie E John, Cal Swart, and John C Thomas. Understanding the role of age and fluid intelligence in information search. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 119–126. ACM, 2012.

[VA03] Maria LM Vargas and Sven Anderson. Combining speech and earcons to assist menu navigation. In *Proceedings of the 2003 International Conference on Auditory Display. Addison-Wesley*, 2003.

[vdKZZ10] Adrian von dem Knesebeck, Pooya Ziraksaz, and Udo Zölzer. High quality time-domain pitch shifting using psola and transient preservation. In *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[VMT92] Hélène Valbret, Eric Moulines, and Jean-Pierre Tubach. Voice transformation using psola technique. *Speech Communication*, 11(2):175–187, 1992.

[WBMN01] Ashley Walker, Stephen Brewster, David McGookin, and Adrian Ng. Diary in the sky: A spatial audio display for a mobile calendar. In *People and Computers XVInteraction without Frontiers*, pages 531–539. Springer, 2001.

[Wri81] Patricia Wright. Tables in text: the subskills needed for reading formatted information. *The Reader and The Text*, pages 60–69, 1981.

[Yes00] Yeliz Yesilada. Browsing tables when you cannot see them. *Unpublished Masters Thesis The University of Manchester*, 2000.

[YSGH04] Yeliz Yesilada, Robert Stevens, Carole Goble, and Shazad Hussein. Rendering tables in audio: the interaction of structure and reading styles. In *ACM SIGACCESS Accessibility and Computing*, number 77-78, pages 16–23. ACM, 2004.

[YSHG07] Yeliz Yesilada, Robert Stevens, Simon Harper, and Carole Goble. Evaluating dante: Semantic transcoding for visually disabled users. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 14(3):14, 2007.

[YW08] Pavani Yalla and Bruce N Walker. Advanced auditory menus: design and evaluation of auditory scroll bars. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, pages 105–112. ACM, 2008.

VITA

JONATHAN COFINO

EDUCATION

Jan. 2009 - Aug. 2014 (expected)	Ph.D., Electrical Engineering
Florida International University	Miami, FL
Jun. 2006 - Dec. 2008	M.S., Electrical Engineering
Florida International University	Miami, FL
Sep. 2002 - May. 2006	B.S., Music Engineering Technology
University of Miami	Coral Gables, FL

PUBLICATIONS AND PRESENTATIONS

Journal Publications

Cofino, J.; Barreto, A.; Abyarjoo, F.; Ortega, F.R. *"Blind Assistive Spatialized Screen-reading"* Behaviour & Information Technology (submitted)

Conference Proceedings

Cofino, J.; Barreto, A.; Abyarjoo, F.; Ortega, F.R. (2013). *"Sonifying HTML tables for audio-spatially enhanced non-visual navigation,"*
Southeastcon, 2013 Proceedings of IEEE
Jacksonville, FL

Cofino, J.; Barreto, A.; Adjouadi M. (2013). *"Comparing Two Methods of Sound Spatialization: Vector-Based Amplitude Panning (VBAP) Versus Linear Panning (LP),"*
Innovations and Advances in Computer, Information, Systems Sciences, and Engineering, 359–370

Cofino, J.; Barreto, A.; Adjouadi, M. (2012). *"Sonically spatialized screen reading: Aiming to restore spatial information for blind and low-vision users,"*
Southeastcon, 2012 Proceedings of IEEE
Orlando, FL

Presentations

"B.A.S.S. Blind Assistive Spatialized Screen-reading"
29th International Technology and Persons with Disabilities Conference
Podium Presentation - San Diego, CA (March 19, 2014)

"B.A.S.S. Blind Assistive Spatialized Screenreading"
ACM Richard Tapia Celebration of Diversity in Computing
Poster Presentation - Seattle, WA (Feb. 6, 2014)

"Audio-Spatialized Tables: Navigating Non-Visually Through Two-Dimensional Web-Based Data Structures"
Assistive Technology Industry Association
Podium Presentation - Orlando, FL (Jan. 30, 2014)