Navigating Information Spaces Through a Novel Immersive 3D Audio Interface

A Thesis
Presented in Partial Fulfillment of the Requirements for the
Degree of Master of Science
with a
Major in Experimental Psychology
in the
College of Graduate Studies
University of Idaho
by
Andrew Perry

Approved by:
Major Professor: Steffen Werner, Ph.D.
Committee Members: Ben Barton, Ph.D.; Todd Thorsteinson, Ph.D.
Department Administrator: Ben Barton, Ph.D.

August 2022

Abstract

Screen readers, braille displays, and voice-activated personal assistants (VAPAs) are the most common accessibility technologies aiding blind users in computer navigation tasks. However, several usability and performance issues have been identified with each method. Screen readers are constrained by high cognitive workloads (Theofanos & Redish, 2003), a loss of graphical information (Harper et al., 2006; Leuthold et al., 2008), and overall inefficiency (Lazar et al., 2007). Braille displays are often costly, and braille literacy has dropped drastically since the 1950s (National Federation of the Blind, 2009). VAPAs cannot handle complex tasks (Abdolrahmani et al., 2018) and force the user to spend a significant amount of time correcting misunderstood text (Azenkot & Lee, 2013). In the context of menu navigation, the current project analyzed the potential of a novel 3D audio interface to be a viable alternative to a conventional screen reader. Participants were tasked with navigating menu structures of varying depth and breadth to select a target item with three different interface styles (3D audio, screen reader, and visual). Results indicated the 3D audio interface was significantly slower, more error prone, and subjectively less-usable than the screen reader. However, the 3D audio interface showed larger performance improvements over the course of the experiment than did either the screen reader or visual interfaces, potentially indicating that more practice with this interface could eventually yield performance advantages over a screen reader.

Table of Contents

List of Figures

Section 1: Introduction

As of 2020, there are over 49 million people across the world who are blind (Bourne et al., 2020). For this population, interacting with websites and/or web-based applications is a difficult task that leaves them on the outside of digital information spaces (Brophy & Craven, 2007; Lazar et al., 2007). This difficulty is primarily rooted in the lack of usability and accessibility of current web technologies (Leuthold et al., 2008; Yoon et al., 2016; WebAIM, 2020). Accessibility can be thought of as the degree to which all users have access to system functionalities (Goodhue, 1986), while usability can be thought of as the degree to which a system conforms to the cognitive perceptions of a user as they are accomplishing a task within the system (Goodwin, 1987). Therefore, an accessible and usable Web for users who are blind should conform to their beliefs about performing online tasks. With online tasks becoming more and more intertwined with education, commerce, socialization, and work, blind users may be getting left behind in these areas.

In an effort to bridge the gaps in usability and accessibility for blind users, different technologies have been developed. The three main accessibility technologies available for these users are screen readers, haptic/tactile and braille displays, and voice-activated input. I will provide a short summary for how each of these technologies aid in web navigation and the limitations associated with each method.

Screen Readers

Most computers have built-in software that can analyze the layout and content of a screen and provide a text-to-speech translation that gets read to the user. This is done left to right, top to bottom in a sequential manner (Leuthold et al., 2008). However, through the use of keyboard controls, the user does have some control over what part of the screen is read at any given time. For example, the tab button and arrow keys are typically used to move across all the different interactive items in the space (links, menu bars, search fields, etc.). Additionally, screen reader users most commonly report using the headers of a web page to find information on a lengthy webpage (WebAIM, 2019). This effectively allows the user to scan the screen in a way that is similar to a sighted user sampling different areas of the screen, albeit with less graphical information available to them (Leuthold et al., 2008).

The two most popular screen readers are the Non-Visual Desktop Access (NVDA) and Job Access With Speech (JAWS) readers, accounting for 40.6% and 40.1% of users, according to a survey conducted by the Center for Persons with Disabilities at Utah State University (WebAIM, 2019). It is common for users to customize their readers to accommodate their interaction preferences. One common modification is the speech output rate, with many users increasing the speed to allow for quicker navigation times (Theofanos & Redish, 2003). In a study by Hochheiser & Lazar (2010) evaluating blind user performance within varying depth/breadth combinations in menu structures, the authors noted that about half of their 19 participants set their screen reader speech output to higher than 73, which is roughly equivalent to listening to a podcast at 1.5x regular speeds.

Screen Reader Limitations

Although the screen reader appears to be the most widely used and accepted accessibility technology for low-vision users, it still carries with it a number of downfalls that affect its practicality. Using a screen reader is often constrained by high cognitive workloads (Theofanos & Redish, 2003), a loss of graphical information (Harper et al., 2006; Leuthold et al., 2008), and overall inefficiency (Lazar et al., 2007). Additionally, web spaces often do not adhere to established requirements of the Web Content Accessibility Guidelines (WCAG). Research from WebAIM (2022) has shown that over 97% of website home pages contain WCAG failures.

By utilizing a screen reader as the primary tool for navigation, task completion becomes largely a listening task. While research that specifically addresses the online experience of blind users is very scarce (Leuthold et al., 2008), there are a number of constraints that have been identified when the task of computer navigation becomes a listening task:

- The cognitive resources of these users are split four ways: between the web browser, the website, the screen reader, and the interplay between them (Theofanos & Redish, 2003). This can create an overload in cognitive workload during these interactions (Thinus-Blanc & Gaunet, 1997).
- The innumerable number of keyboard shortcuts and the wide range of screen reader functionality puts strain on the user's cognitive resources (Theofanos & Redish, 2003).

- The user is always operating with less contextual information than a sighted user because of the sequential nature of the screen reader (Lazar et al., 2007).
- Lacking the ability to quickly scan the page leaves the user without access to much of their goal-relevant information (Di Blas et al., 2004).
- Many websites have complex layouts that cause the screen reader's feedback to become ambiguous (Lazar et al., 2007). Additionally, screen readers often mispronounce words (Theofanos & Redish, 2003), leaving the user struggling to understand the information being presented to them.
- Because the majority of screen information cannot be accessed simultaneously, the user loses nearly all spatial information.

Haptic/Tactile Displays

Haptic information has often been used as a method for communicating information when the other sensory systems are not free to process information. Tactile displays can generally be broken down into three different usage categories: orientation, navigation, or communication (Castle & Dobbins, 2004). For example, a deep sea diver may become spatially disoriented because of the lack of gravitational or visual cues one receives underwater. A haptic display vest may help orient the diver by vibrating in the direction of the water's surface. For communication, a common haptic alert is the vibration of a cell phone alerting when a phone call is being received. For navigation, tactile maps have been shown improve route and survey knowledge in blind users when compared to simply direct environmental experience (Espinosa et al., 1998). Tactile maps are displayed on special sheets of paper that produce raised elements to display spatial properties (points, lines, and regions), varying surface textures to show feature characteristics (dots/dashes, line height/thickness), along with braille labels for names and semantic information.

While haptic information has proven to be useful in these contexts, it can never provide the same richness of information as the visual system. This may be the largest problem a blind user experiences when using a braille display for computer navigation – they simply do not have access to the same amount of information as a visual user at any given time. This forces the global awareness of a blind user to be much lower.

Braille Displays

As an alternative device used in unison or in place of a screen reader is a refreshable braille display. These devices can be plugged directly into the computer and will convert the textual information that is on the screen into a braille display for the users to scan and read. Additionally, many braille displays contain buttons above the braille strip that can be used for both navigation and typing (see Fig. 1.1).

Figure 1.1 Image of the HumanWare Brailliant braille display connected to a laptop computer. Note the eight buttons located directly above the braille strip used for typing and navigation.

Braille Display Limitations. While the refreshable braille display does provide a better alternative than screen readers in terms of scanning screen information more quickly based on user discretion, there are a number of similar constraints that limit its usability:

- Refreshable braille displays are often very expensive, sometimes costing more than the computer they are plugged into.
- Another issue is their lack of display space. Most braille displays can only show a line of text with up to 40 characters in length. This again makes reduces access to contextual and goal-relevant information.
- Braille has fallen in literacy rates over the years, with only an estimated 10% of blind Americans able to read braille (National Federation of the Blind, 2009). In 1950, 50% of blind Americans could read braille. Furthermore, braille is difficult to learn for users who develop blindness later in life.
- While some progress has been made in presenting pictorial information as a “tactile image” (Cantoni et al., 2018), this technology has yet to be implemented into braille displays. This leaves the problem of graphical information being inaccessible to blind users unresolved.

Voice-Activated Inputs

Blind users also have the option of computer interactions via voice-activated personal assistants (VAPAs) and/or integrated computer software. VAPAs, such as the Amazon Echo or Alexa, are becoming smaller, cheaper, more accurate, and more prevalent around the world. A study by Pradhan et al. (2018) evaluating product reviews from Amazon Echo owners noted that a significant number of reviews mentioned the VAPAs utility for users who were visually impaired, suggesting many of these users are already aware of the potential VAPAs hold in providing accessibility to digital arenas they once thought inaccessible. However, these devices are currently used most often for simple retrieval tasks (e.g. local weather information) (Luger & Sellen, 2016) and leisure entertainment (e.g. playing music or controlling external devices) (Bentley et al., 2018; Lopatovska et al., 2019), leaving much of the fluid navigation process obtained with a keyboard and mouse still out of reach. This brings up the issue of a browsing vs. conversational interface. VAPAs operate via conversational commands (e.g. "Alexa, order more toilet paper from Amazon"), and do not have the functionality to allow browsing of information spaces.

Voice-Activated Input Limitations

VAPAs and/or voice-activated inputs have been shown to have some utility for visually impaired users in some specific contexts. In educational settings, Bouck et al. (2011) found that voice inputs helped visually impaired students operate their calculators in math classes. When inputting text to their mobile devices, blind users rated voice-input more favorably than normal sighted users, often citing its higher efficiency (Azenkot & Lee, 2013). While there is indeed some utility of these technologies for blind users, there are also a number of constraints limiting their overall effectiveness:

- Even with the significant increases in the accuracy of speech recognition software, blind users still spent a significant amount of time (about 80% of on-task time) correcting text misunderstood by the software (Azenkot & Lee, 2013).
- High cognitive loads when contemplating how to best phrase commands to the system (Cowan et al., 2017). This was particularly true with non-English speakers (Bogers et al., 2019). Additionally, the "low transparency" of the inner workings of the VAPA contributed to difficulties in properly choosing phrase interactions (Chen & Wang, 2018).

- Users have been shown to be more cautious or even embarrassed when using these technologies in public areas around third party observers (Cowan et al., 2017; Efthymiou & Halvey, 2016).
- The inability of VAPAs to change their speed of voice or handle complex, work-productive tasks (e.g. editing text or seeking information) limit their utility when compared to conventional screen readers (Abdolrahmani et al., 2018).
- Additionally, blind users felt VAPAs lacked more precise control and displayed less contextual information than conventional screen readers (Vtyurina et al., 2019). Screen readers are able to achieve better acceptance in these users by providing direct access to information like headings, links, lists, and tables.
- VAPAs are designed with the general consumer in mind, while screen readers are designed specifically for visually impaired users. While VAPAs are similar to screen readers in essence, this difference in design focus leaves VAPAs less accessible to visually impaired users than screen readers (Branham & Mukkath Roy, 2019).

While all three of the accessibility technologies discussed here have their place in aiding users during human-computer interactions, there are still extensive shortfalls that need to be addressed. Screen readers are often constrained by high cognitive loads, losses in graphical information, and low efficiency. Braille displays are very expensive, lack display space and subsequently the amount of information they can present, and are difficult to learn for users developing blindness later in life. VAPAs require much of the user's time be spent on making corrections of incorrect inputs, cannot be used for browsing, and carry with them privacy concerns such that users often will not use them in public.

Architecture of Information Spaces

The term "information architecture" was first coined by Richard Wurman back in 1975, and while his focus was mainly on the *presentation* of information, his reframing from information *design* to information *architecture* was useful to draw more attention to the structure and function of information. To be concerned with the architecture of an information space means to be concerned with its structure and foundation, and not simply its aesthetics. Any digital information space can be broken down into three integral parts: 1) information artefacts, 2) users, and 3) devices.

Information artefacts can be thought of as any object that aims to provide information about something. They are static and have the ability to be interacted with. Users refer to the individuals interacting with the information space at any given time. All information spaces are created with the intention of having some eventual user interaction. Devices are what deal with the syntax of interactions between the user and the information artefacts. The physical buttons, hardware/software combinations, and communication channels all fall within this category.

When constructing an information architecture, arguably the most important principle will be one of organization. There are three main steps in this process:

1. **Ontology:** This step consists of deciding on the entirety of information that will be needed in the architecture. The listing and naming of objects and pieces of information will take place here. The size of the architecture is directly related to this step.
2. **Taxonomy:** This step entails the classification of all information objects and showing how they relate to one another. This is closely related to the depth vs. breadth arguments of menu design.
3. **Choreography:** This step deals with the layout of information objects within the architecture. Understanding how the intended users will be interacting with the information space will largely determine how objects are choreographed. The most common structure seen in digital information spaces are networks of hierarchies for users to navigate through. Generally, the larger the network of objects the more difficult it is to navigate through.

Blind Users Navigating Information Spaces

Out of the three steps of the organizational process of information spaces, the most care needs to be given to the third step of choreography to ensure these spaces are as accessible as possible for blind users. With these users consistently working with higher cognitive loads, less graphical information, and less efficiency, having information spaces that are choreographed logically is integral to their usability for these populations. Naturally, larger information spaces become more difficult for blind users to navigate, as the likelihood of losing their place in the overall structure hierarchy increases. Additionally, spaces that have a poor taxonomy of items because of ambiguous classifications also decreases the ability of blind users to function within these spaces.

Web-Based Navigation vs. Menu Navigation

A common similarity between website structure and computer menu structures is the prevalence of hierarchical designs. Menu hierarchies are one of the most popular forms of menu structures, and websites often follow this same general structure with a home page (parent page) displaying the most relevant information and then a number of child pages with more specific information branching off from the home page. A common example of this are clothing websites (see Fig. 1.2), where articles of clothing act as their own separate nodes on the hierarchy and get subdivided into even more specific categories. A similar structure is found in most computer applications in the form of the application's menu bar (see Fig. 1.3). The menu bar serves as the initial access point, with different categories branching into more specific subcategories and functions.

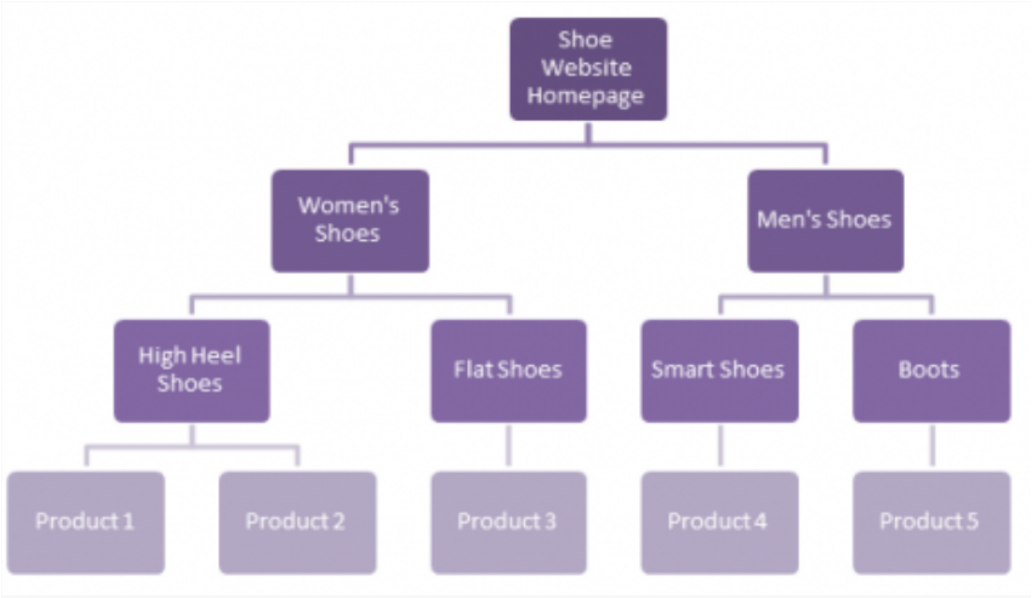

Figure 1.2 A hierarchical structure of a shoe website.

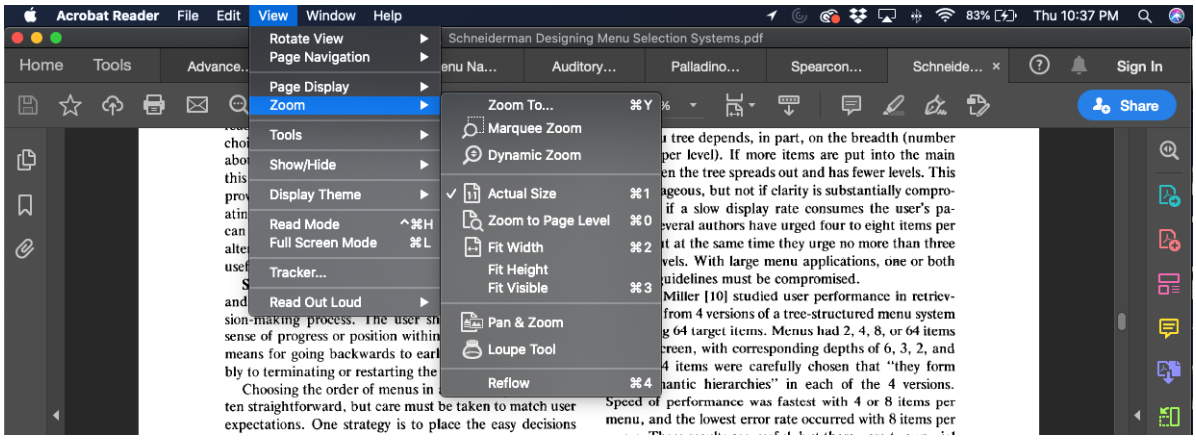

Figure 1.3 Example of a hierarchical menu displayed in the Adobe Acrobat application.

The structure of the information space is, of course, only half of the problem when discussing human-computer interactions. The human must operate within this space when they are tasked with achieving specific goals. Thus, it will be helpful to discuss the psychology underlying item selections in human-computer interactions.

Cognitive Processes in Item Selection

Visual Search

Before making a selection of a menu item, users must visually search for the desired item. Of course, for blind users, this is done through a different sensory system (with auditory or tactile information). Depending on the context in which the search is taking place, there are a few different schools of thought for how this search is performed. Kent Norman (1992) generally supports a serial processing theory, where the user moves through menu items in sequential order until finding the desired match they think will achieve their goal. Hornof and Kieras (1997), on the other hand, believe that people both randomly and systematically search menu structures and can process multiple menu items simultaneously instead of moving through them sequentially. However, this is more difficult for blind users, as their access to information at any given time is much more limited than normal sighted users.

Serial Processing. There are two contexts that Norman (1992) feels most affect the type of serial search a user will perform. First, if the user has a clear target in mind, they will run through each menu item sequentially until an adequate match is found, in which case the selection is then made (Fig. 1.4). However, if the user only has a partial idea of the target they are looking for, they are more likely to search through all of the menu items and then

decide which item most closely matches their target idea. Norman claims that although the second option takes more time, it is likely the more accurate of the two methods. Therefore, if a user is more interested in speed than accuracy, they are likely to adopt the first type of serial searching. Lending some support to Norman's serial processing theory, research by Byrne et al. utilizing eye tracking data has shown that users interacting with pull-down menus perform searches that occur top-to-bottom and are rarely random (Byrne et al., 1999).

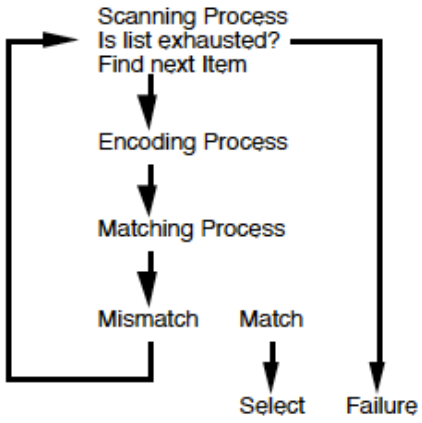

Figure 1.4 Norman's information processing model for search of a known target.

Random Visual Search. Hornof and Kieras (1997) attempted to validate Norman's theory that people generally move through menu items in a serial fashion. They did this by having participants perform a menu selection task and then compared the subject data to predicted data generated by a cognitive modeling tactic called EPIC (Executive Process Interactive Control). EPIC was developed my Kieras and Meyer (1997) and is similar to the GOMS (Goals, Operators, Methods, Selections) model in that it is used to predict completion times for specific processing and motor tasks based upon human performance literature. When comparing their subject data to their predicted data generated from the EPIC task analysis, Hornof and Kieras found that people do not stop and decide on items individually but instead process many items in parallel. They also found that people use both systematic top-to-bottom and random search strategies when making menu selections. Whether or not a person processes items in parallel largely depends on the field of view of the items being processed. If multiple items are within the foveal visual angle, which is where fine detail is perceived by the retina, then it is thought they can and will be processed in parallel. Again, this is a problem for blind users, as they have less information available to them at any given time.

Despite many theories being developed for how people process and move through menu structures, it is likely that it ultimately depends on the user's experience with the device/software and the context that they are using it in. For example, if someone has never used a computer before, they would probably move through a menu structure serially as they need to figure out what each menu item represents. However, if the user is very experienced and has used the same computer and computer program for years, they are likely to move through menu structures very quickly as they are already familiar with where the items are located within the program.

Depth vs. Breadth Considerations

When analyzing the depth and breadth of a menu structure, we are looking at the number of levels a menu has (its depth), and the number of menu items present at each level (its breadth). There is no one size fits all recommendation for how deep or broad a menu structure should be, and often designers need to make compromises with their selection of levels and menu items in order to find the design that best suits the needs of the population their designing for, as well as the tasks that said population will be performing most often. Too many items within one level may lead to clutter or information overload, subsequently increasing the time a user needs to spend on a task and potentially increasing error rates. However, too few items on a screen and an excessive amount of levels a user must navigate through can also be problematic by degrading the user experience and increasing drop-off rates. A thorough user testing process and multiple iterations of designs may be necessary to eventually find the correct balance of depth vs. breadth for any one menu structure.

However, there has been past research done already that indicates when certain depth/breadth ratios may be best utilized. Generally, it seems that broader menus have outperformed deeper menus in most of the past research. Snowberry et al. (1982) conducted a study comparing four different verbal hierarchies, each containing a total of 64 words and asking participants to search for target words within the list of items. One hierarchy contained binary choices at six levels (2^6), another with eight choices at two levels (8^2), a third with four choices at 3 levels (4^3), and finally a the broadest level containing all 64 words. They found that after conducting their trend analyses that search speed and accuracy improved as a function of menu breadth (Snowberry et al., 1982). Other researchers have found similar performance advantages when comparing broader menus to deeper menus

(Dray et al., 1981; Kiger, 1984; Miller, 1981; Shneiderman, 1986). However, there has been other research stating the menu structure should be developed depending on the complexity of the task, as well as the screen size it will be performed on (Chae & Kim, 2004). In the example of mobile phones, Geven et al. (2006) found that users preferred deeper menus with fewer items, presumably because of the smaller screen real estate to work with on these devices.

Depth vs. Breadth for Blind Users. In a study by Hochheiser & Lazar (2010), the authors attempted to replicate past research done by Larson & Czerwinski (1998) by analyzing performance of blind users as they navigated hierarchical structures of varying depth/breadth combinations. Additionally, they evaluated a subjective measure of lostness to determine the extent that blind users would feel lost within certain menu structures. They had 19 blind participants complete selection tasks from three different menu hierarchies using a common screen reader software, JAWS 8.0. The experiment was performed on a single laptop running Windows XP with Internet Explorer 7.0 to access the pages, and the menus consisted of an 8 x 8 x 8, 16 x 32, and 32 x 16 menu structure. Also, the speech output rate was set to a moderately fast speed to mimic the output speeds most commonly used by screen-reader users (Theofanos & Redish, 2003).

The authors hypothesized that working memory constraints would cause users to perform best within the 8 x 8 x 8 menu hierarchy, as there would be fewer items at each level for the users to hold in their working memory. However, their results showed that subjects had the shortest completion times in the 16 x 32 condition and the longest times in the 8 x 8 x 8 condition. Additionally, the subjective ratings of lostness were most favorable for the 16 x 32 condition and least favorable for the 8 x 8 x 8 condition. These results indicate that depth vs. breadth considerations for blind users may actually be relatively similar to normal sighted users.

Choice Response

Once the user has finished their visual search and has decided upon an item to be selected, they must provide an input to the device to make their choice selection. This can often be done by a number of different means. A person may use command line language, a pointing device like a mouse, or a finger to a touchscreen. For blind users, this is most often accomplished with the assistance of a screen reader to ensure they are selecting their intended

target. Because pointing with our fingers is something that we all perform since infancy, Norman (1992) claims that this method of point selection is the most intuitive, followed by mouse, joystick, and finally cursor keys. Depending on the method of item selection, there are some models that can be used to predict human performance and response times in specific scenarios. Fitts' Law, for example, can predict the amount of time it takes to move a pointer device to its desired target by taking the distance of the target from the pointer device and dividing it by the size of the target (Fitts, 1954). There are also cognitive modeling techniques that can be used to perform task analyses and predict the amount of time it will take a person to complete specific tasks. For example, GOMS can predict the performance of an expert user by accounting for the times it takes for the person's sensory organs to perceive a certain stimulus, process the information, and then perform the desired physical response (Card, Moran, & Newell, 1983).

Another concept often cited when describing response times in menu structures is the Hick-Hyman Law. In separate but parallel studies, William Hick (1952) and Ray Hyman (1953) found that choice reaction time increased linearly with the amount of stimulus information, or bits, that were present. The increase of choice RT was constant with each doubling of alternatives available for selection, and describes the added processing time needed when making selections from multiple items.

This finding was replicated by Landauer and Nachbar (1985) in the context of touchscreen item selection from alphabetic and numeric menu trees. Subjects were tasked to select a "goal" item of either a specific number or word from a menu tree of varying depths and breadths. Their results indicated that selection times were a logarithmic function of the number of selectable items on each level (e.g. reaction times increased linearly as the number of items on screen doubled) (see Fig. 1.5).

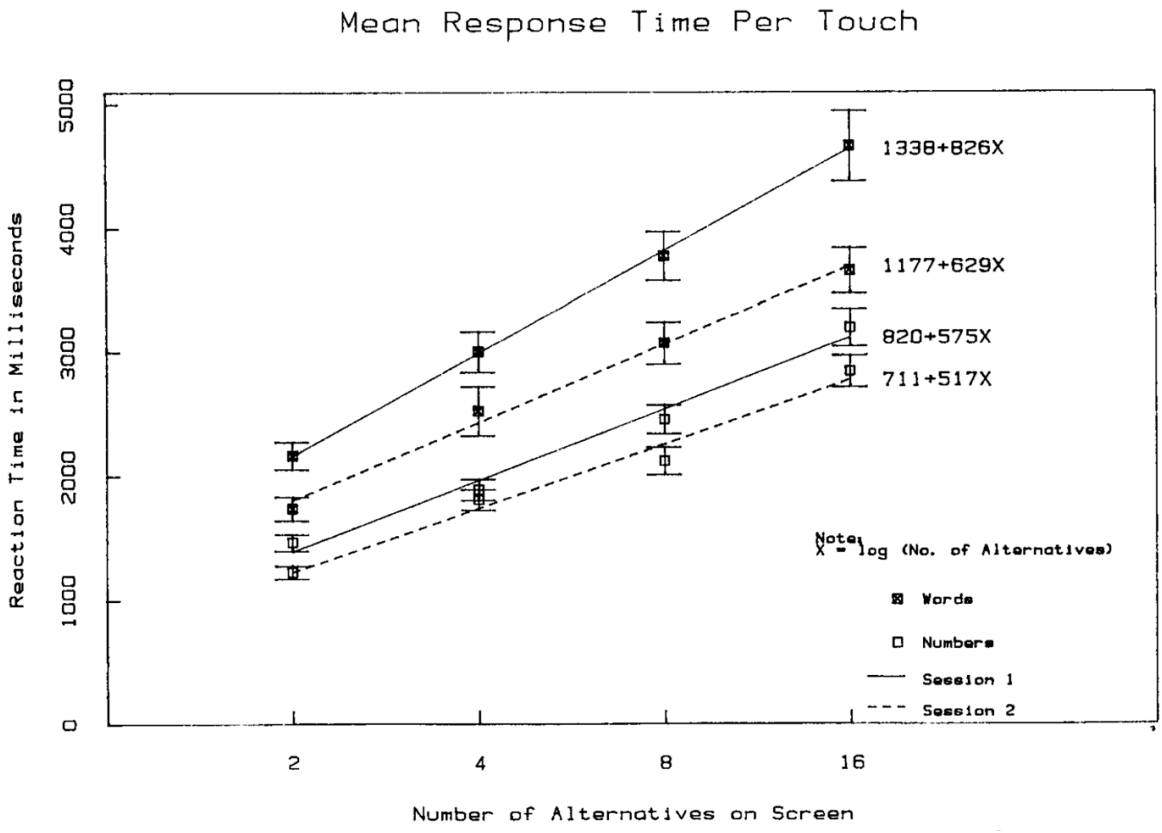

Figure 1.5 Response times and their relation to breadth of item levels.

While their results indicated longer response times for broader levels of items, they noted that the overall response time needed to find a target item was lower for broader menu trees than in deeper ones (see Fig. 1.6). This is because navigating a broader menu tree will result in fewer selection steps, and each selection step adds a constant amount of human response time.

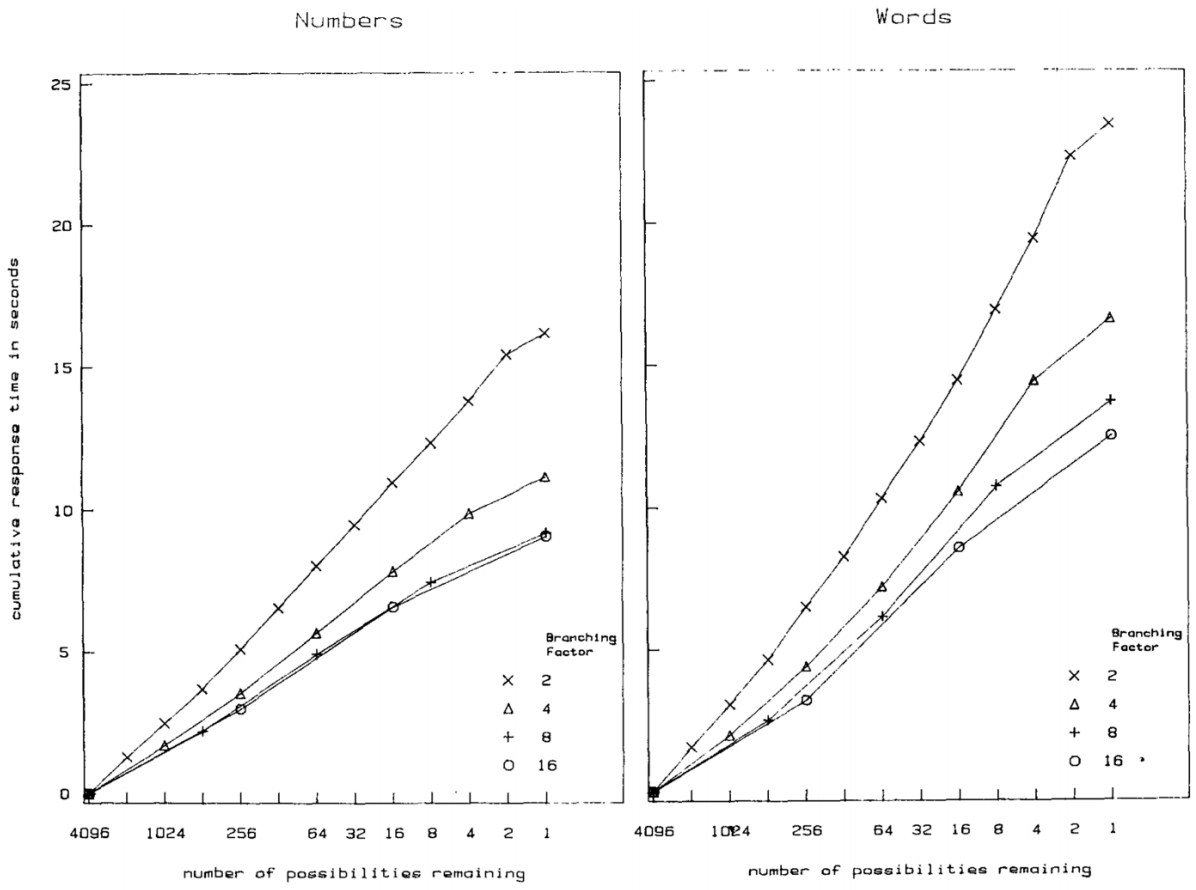

Figure 1.6 Total response time and its relation to breadth of item levels (branching factor).

While it is useful to understand the cognitive process of item selection by reducing the process down to simply visual (or aural) search and choice response, this should be incorporated into a more global view of goal selection and action. For this purpose, Norman's Seven Stages of Action will be briefly discussed.

Norman's Seven Stages of Action

Don Norman (2013) proposed a seven-stage model of action to explain the relationship between users and systems. Additionally, it provides a framework to help identify exactly where a breakdown in the interaction of user and system takes place. As this model approximates for both cognitive and physical activities (Zhang et al., 1999), it is a useful model for mapping the interactions of blind users navigating computer menus. The seven stages of action are as follows:

1. **Goal** (form the goal). The interaction begins once the user identifies the overall goal to be achieved (e.g. selecting a specific item from a menu hierarchy).

2. **Plan** (the action). After the goal has been established, the user must identify how they can move forward to achieve the goal. In this example, how will they move through the hierarchy and select the item?
3. **Specify** (an action sequence). Once the plan of action has been established, the user must identify the specific steps needed to work through the action plan. If using a screen reader, the user will use keyboard shortcuts to move through the menu items and then ultimately select the desired item.
4. **Perform** (the action sequence). The user will the interact with the system to accomplish the desired goal. Blind users will rely on auditory information to provide context to their cursor location, while using keyboard shortcuts to move about and select items.
5. **Perceive** (the state of the world). After performing an action, the user is then left to perceive the changed state of the system. For blind users, this means listening to the output of the screen reader for context as they move about and make their selection.
6. **Interpret** (the perception). Once the user has perceived the changing system state, they must interpret the changes and determine if their desired goal has been met. A blind user must interpret the feedback of the screen reader to know if their task has been completed.
7. **Compare** (the outcome with the goal). Based on their interpretation of the screen reader feedback, the user will compare the current state of the system with the user's goal state to see if they match.

Additionally, Norman identifies two different "gulfs" that the user must navigate to achieve their goal: The Gulf of Execution and the Gulf of Evaluation. The Gulf of Execution is crossed by the user as they attempt to understand how the system operates, while the Gulf of Evaluation is crossed as the user tries to understand what happened during their interaction. This framework is useful to keep in mind as it provides a model to diagnose pain points for blind users during their interactions with computers.

For blind users, they are most often utilizing auditory information to aid them in their navigation and item selections. Before describing the goal of the proposed research, it will be useful to give a brief overview of the different ways auditory information has been used to carry information, even outside the context of accessibility for blind users.

Auditory Techniques Used for Contextual Information

Multiple Resource Theory (Wickens, 1991), states that there is a limited amount of mental resources available to a person at any given time. This pool of resources is utilized for a number of different mental operations, ranging from sensory-level processing to meaning-level processing. Allocation of these resources accounts for the performance within different tasks, modalities, processing. This explains why dual-task performance is more likely to be negatively affected by performing similar tasks than would be by performing dissimilar tasks. Designers have taken advantage of this by providing multiple different modalities for information presentation during complex task completion. For example, auditory alerts are an important part of cockpit design. Pilots have an abundance of visual information to be processed while flying a plane, which may cause them to miss valuable information being displayed in their cockpit. Auditory alerts are often used to help drive the pilot's attention to an area of need or provide information without the pilot needing to divert their gaze from more important navigation tasks.

Another example is the use of auditory information in support of anesthesiologists. Anesthesiologists are often tasked with inducing specific levels of unconsciousness that allow patients to undergo invasive surgeries. As part of this task, anesthesiologists must monitor the patient's state during unconsciousness while also performing subsequent tasks. Sonification of patient respiratory or heart rate information is used to allow the anesthesiologist to properly monitor the patient while performing their separate tasks.

In the context of human-computer interactions for blind users, sonification techniques fall into three different categories: earcons, audio icons, and spearcons. I will provide a brief overview of each technique and the limitations associated with each one.

Earcons

Earcons are brief, nonverbal audio cues that convey information and/or feedback to a user to alert them that a specific interaction or event has taken place. This is achieved by systematically manipulating pitch, register, timbre, rhythm, intensity, and other sound qualities to represent the specific interaction, object, or operation. The term earcon was first seen in 1985 in a technical report by Denise Ann Sumikawa titled "Guidelines for the integration of audio cues into computer user interfaces" (Sumikawa, 1985). However, the first tests evaluating the effectiveness of earcons were not conducted until 1992 when

Brewster, Write, and Edwards concluded that earcons were more recognizable than unstructured bursts of sounds as long as timbre, pitch, and rhythm were used appropriately (Brewster, Wright, & Edwards, 1992). This also led to Brewster and his colleagues developing guidelines for how to implement earcons in regards to these specific sound qualities.

Additionally, earcons have been shown to provide navigational cues in hierarchical menu structures (Brewster, Raty, & Kortekangas, 1996). In this study, a hierarchy of 27 nodes and four levels were constructed, with each node having their own distinct earcon attached to it. Participants were then asked to identify their position within the hierarchy based upon listening to an earcon. The results showed that participants were able to identify their hierarchical position with 80% accuracy.

Medical device technology is a field that relies heavily on earcons to display information while healthcare workers on performing subsequent tasks. Earcons are often used in these scenarios so that clinicians can continually monitor a patient's state without having to rely on constant visual information to do so. This frees them up to perform other tasks for their patients. Watson and Sanderson demonstrated that anesthetists can accurately monitor a patient's respiratory vitals via sonification (Watson & Sanderson, 2001). Also, by relying on auditory stimuli instead of visual, the anesthetists were freed up to perform other concurrent tasks more effectively while still monitoring their patient's state. In this experiment, participants were asked to perform a primary task of making true or false judgements about basic arithmetic problems. They were also asked to perform a secondary task of monitoring patient status. To monitor patient states, participants were provided with either sonification only, sonification plus visual display, or visual display only. The results indicated that anesthetists were able to accurately monitor patient status with all three of the monitoring conditions. However, the sonification condition allowed the anesthetists to most accurately perform their primary task of arithmetic judgements, answering 96% of those questions correctly vs. 91% and 92% accuracy for the visual display and the sonification plus visual display conditions.

A medical device that is commonly used with an auditory display is a pulse oximeter. Pulse oximeters are noninvasive devices that measure how well your heart is pumping oxygenated blood throughout the body. Standard pulse oximeter displays use varying pitch

and alarms to communicate hemoglobin oxygen saturation levels (SpO_2). However, research from Paterson et al. has shown that by adding slightly more complexity to sound qualities (varying pitch plus tremolo and acoustic brightness), anesthetist were able to be faster and more accurate in their detections of transitions to and from target SpO_2 ranges (Paterson et al., 2020). For this study, 20 experienced anesthetists supervised a junior colleague (played by a confederated) during two airway surgery scenarios: one scenario using the advanced pulse oximeter display and one scenario using the standard display. During the experiment, participants were distracted with other tasks such as paperwork and operating room commotion to simulate the distractions typically encountered in these scenarios. They were then asked to identify when SpO_2 transitioned between preset ranges (target, low, and critical) as well as when other vitals transitioned out of target ranges. While visual displays were displayed for other vital signs, the numerical value for SpO_2 was always excluded. In addition to being faster and more accurate in their detection of SpO_2 transitions, their accuracy also increased when identifying the precise range SpO_2 fell in after a transition had occurred when compared to a standard pulse oximeter display.

Earcon Limitations. Although earcons have been shown to aid navigation within hierarchical menu systems, they are somewhat limited in the type of systems they can be used for. For earcons to be used confidently, the system will need to be one that is not expected to change over time and will need to be simple in the types of interactions it handles. This is because earcons cannot effectively handle systems where new items are added regularly and/or get reorganized depending on how the user is interacting with it. For example, if a menu hierarchy needs a new item added to its system, it may not be a problem if the item is added to the bottom of the hierarchy because a new earcon can be generated to represent its position on the bottom of the hierarchy. However, if the new item that is generated needs to be added in the middle of the hierarchy, a problem arises where all items below this newly added item will now need their earcon altered because their place in the hierarchy has been slightly modified. This problem becomes even more evident in more complex systems that display items algorithmically based on the items the user most commonly uses. Earcons simply cannot keep up with that amount of flexibility within a system.

Another issue with earcons is that the sounds used to convey specific items are arbitrary and unrelated to items they are representing. While this is beneficial for allowing earcons to be adapted to a wide range of menu systems, it also means that training will be required for any user interested in learning the system. Additionally, there is no standardization for how earcons are used, so there is likely to be little carryover from one system to the next, forcing users to need training for each new earcon system they come across. Dingler et al. compared the learnability of the sonification techniques of earcons, audio icons, spearcons, and speech, and demonstrated that earcons were much more difficult to learn than were spearcons or speech (Dingler, Lindsay, & Walker, 2008). For this study, 39 undergraduate students were presented with a grid of commonly seen items around their campus (e.g. bench, fountain, garbage can, stairs) and asked to identify them using only specific sonification techniques. Participants were randomly assigned to each condition (earcons, audio icons, earcons + audio icons, spearcons, or speech) and given a round of training to understand how each of the 20 items would be displayed sonically. Then, they were tasked with identifying each item and were asked to participate until they were able to identify every item without any errors. Results showed that both spearcons and speech conditions performed the best, with mean training cycles both equaling 1.14, and aggregate percentage accuracies of 99.64%. Spearcons, on the other hand, performed the worst out of the conditions, with an average number of cycles equaling 8.5, and an aggregate percentage accuracy of about 74%.

Audio Icons

Audio icons are similar to earcons in that sound is still being used to provide feedback about an action, interaction, or event that has or is currently taking place. However, unlike earcons, audio icons are typically non-musical and are made to resemble the event that is taking place. For example, when an email is sent, an audio icon that often represents that action is the sound of a jet taking flight to represent the email flying quickly through the air to its destination. The first interface to use strictly audio icons for its sonification techniques was seen in 1989 when William Gaver developed the SonicFinder for Apple Computers (Gaver, 1989).

A benefit that audio icons hold is that because their sounds are analogous to the actions that are taking place, very little, if any, training is required to understand them. This would

also support the use of similar audio icons being used across cultures and languages. Additionally, making the computer interface consistent both in its auditory and visual dimensions should increase the *direct engagement* of the user, which Hutchins et al. describe as working within the world of the task instead of the computer (Hutchins, Holland, & Norman, 1986). This increases the feeling of having a transparent interface.

Audio Icon Limitations. While audio icons have the benefit of needing less training because their sounds are analogous to the action that is taking place, they are limited in their usage because many interactions within computer technologies are not easily represented by sound (e.g. “connecting to server” or “save file”). Also, assuming a sound can be created for a specific action, this sound needs to be created manually, which becomes more time consuming than auto-generated earcons and can become problematic in dynamic systems.

Spearcons

Spearcons, which stands for “speech-based earcons”, were created in 2006 by Walker et al. in an attempt to improve performance and usability of menu-based interfaces (Walker, Nance, & Lindsay, 2006). Spearcons are similar to both earcons and audio icons in how they are used. However, unlike both earcons and audio icons, spearcons use speech as the defining audio characteristic for each item or action. Spearcons have the benefit over audio icons in that they can be generated automatically by taking the text of a specific action (e.g. “save file”), and converting that to audio with text-to-speech (TTS) software. Then, without changing the pitch of the audio, the speech is sped up to the point where it is no longer comprehensible as speech. Spearcons also inherit the same quality of audio icons in that they become unique to the specific icon or action they are representing. However, this uniqueness is phonetic instead of metaphorical. A spearcon of an object could be thought of as that object’s auditory fingerprint.

Additionally, there is some evidence that spearcons outperform both earcons and audio icons in search time and accuracy in menu navigation. Walker et al. (2006) had participants navigate a 5x5 menu structure for specific target items using earcons, audio icons, and spearcons, and found that spearcons resulted in both faster and more accurate navigation to the target items. The mean time to completion using spearcons was 3.28 seconds, compared to 4.12 for audio icons and 10.52 for hierarchical earcons. The mean percentage accuracy for spearcons was 98.1%, compared to 94.7% for audio icons and 94.2% for hierarchical earcons.

Spearcon Limitations. If the system being represented is simple and unlikely to ever change, and where navigation is particularly important, earcons may outperform spearcons in their ability to aid hierarchical navigation. Also, spearcons are dependent on language, whereas earcons are not. This has the potential to create problems if an interface is translated from one language to another. Additionally, if spearcons are being utilized in a space that includes concurrent verbal tasks such as listening or speaking to another person, a performance decrement of spearcon identification may result. Davidson et al. conducted two experiments evaluating the ability of non-clinicians to identify multiple patient spearcons while performing concurrent tasks such as reading, listening, or speaking tasks. Their results indicated similar performance decrements during the saying and listening tasks, but no accuracy reductions in the reading or "no task" conditions (Davidson et al., (2019). This reduction in identification accuracy may be explained by Wickens' Multiple Resource Theory, as the concurrent tasks of listening/speaking are competing for the same resources within the subject's verbal processing systems (Wickens et al., 2013). Furthermore, spearcons must be learned and paired with their corresponding word items in order to be utilized by the user. This means they can only be used in limited contexts where the user has an opportunity to be trained on the spearcons beforehand. This excludes them from being used in rapidly changing environments often found in computer navigation.

The sonification and text-to-speech techniques presented thus far describe linear, serial outputs. However, the human auditory system has the ability to spatially separate incoming information, and this ability can potentially allow additional contextual information to be recognized by the user. I will provide a summary of relevant literature evaluating spatial audio in human-computer interfaces.

Spatial Audio

In 1953, Edward Colin Cherry attempted to understand more about how humans recognize speech by conducting a set of experiments where subjects were exposed to multiple concurrent messages and were asked to identify different properties of the played messages (Cherry, 1953). Specifically, Cherry was interested in the "cocktail party effect", where we can seemingly tune out other voices in a room in order to process the speech by a specific person of interest. Part of his experiment involved a task where subjects were presented with concurrent, yet different, spoken messages with each message being played to

separate ears. What Cherry found was that subjects were able to pay attention to one message and accurately repeat back what the message said. However, when asked about the message being played to the other ear, they were unable to repeat back what was said, and often times were unable to even identify what language was being spoken. While subjects could not process exactly what was being spoken into the ear that was being "tuned out", they were able to accurately identify if the voice was male or female. This indicates two things: 1) that the processing of information can be accurately moved from each ear at will, and 2) when the focus of attention is moved towards one ear, there are some properties of sound from the other ear that can still be identified.

This may have implications for designing auditory interfaces by allowing some amount of parallel processing to occur. Traditionally, auditory interfaces display sounds serially in an attempt to mitigate confusion or interference between items. However, if designed correctly, there may be a way to display information simultaneously by utilizing this ability to process certain sound properties spatially. Lorho et al. (2001) demonstrated this as a possibility by spatially separating auditory items (by head related transfer functions or stereo panning) and found that subjects were able to identify multiple items simultaneously (Lorho et al., 2001). These same authors also found that spatial audio can be utilized for demonstrating the depth and breadth in hierarchical menu structures (Lorho et al., 2002). Other interesting research conducted by A. Walker & Brewster (2000) indicates that spatial audio can be utilized in the same way a progress bar shows location within menu levels. In this study, participants were able to perform background monitoring tasks more effectively while using the auditory progress bar compared to the visual one, lending support to using spatial audio in multi-modal designs to aid performance in certain tasks.

Spatial Audio and Human-Computer Interfaces

Spatial audio has also been studied to understand its utility for devices aiding blind or visually impaired users within human-computer interactions. With the majority of blind users using a screen reader for computer navigation, spatial audio has been evaluated as a potential addition to this technology. In a study by Sodnik et al. (2012), the authors evaluated a custom interface design utilizing spatial audio and its performance against a conventional screen reader paired with a braille display. Sodnik et al. designed an auditory interface that could provide contextual information about text alignment, text style, and table dimensions. The

system architecture (see Fig. 1.7) used to create this interface is fairly simple, and could be utilized rather easily with current computer technologies. For text alignment and table dimension information, the authors designed a spatial positioning module based on a Cartesian coordinate system that would provide a spatialized output into the user's headphones depending on certain textual information. For example, any text that was centrally aligned was played at a coordinate position of (0˚, 0˚, 0º), with the first coordinate representing the x-axis, the second coordinate representing the y-axis, and the third coordinate the z-axis of three-dimensional space. Left aligned text was positioned at (-20º, 0º, 0º) and right aligned text was positioned at (20º, 0º, 0º) (see Fig. 1.8). Additionally, table dimension information was provided by spatially representing each cell of the table, with the center cell of the table acting as the anchor point where no spatialization is perceived. Finally, text styles of bold, italic, and underlined text were represented by changes in pitch and output rate. Italic text was represented by an increase in pitch and speed by 20%. Bold text was represented by a decrease in pitch and speed by 20%. Underlined text was represented by an increase in speed by 40%.

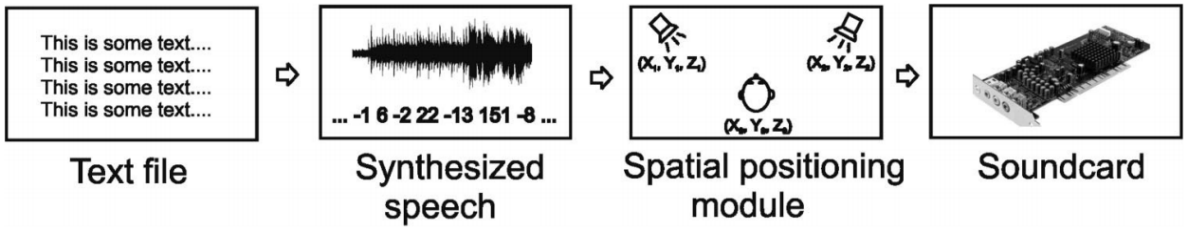

Figure 1.7 Sodnik et al. system architecture, consisting of a speech synthesizer, spatial positioning module, and soundcard.

POS (−20°, 20°, 0°)	POS (−10°, 20°, 0°)	POS (0°, 20°, 0°)	POS (10°, 20°, 0°)	POS (20°, 20°, 0°)
POS (−20°, 0°, 0°)	POS (−10°, 0°, 0°)	POS (0°, 0°, 0°)	POS (10°, 0°, 0°)	POS (20°, 0°, 0°)
POS (−20°, −20°, 0°)	POS (−10°, −20°, 0°)	POS (0°, −20°, 0°)	POS (10°, −20°, 0°)	POS (20°, −20°, 0°)

Figure 1.8 Example of coordinate systems used for spatializing audio of a 5 x 3 table. (0º, 0º, 0º) represents no spatialized output. Imagine each cell of this table consists of a single speaker, with the speakers stacked in a wall-like fashion. The sound coming from each speaker would represent a single item within the table and allow spatialization of items to be perceived by the listener. The size of the tables ranged from 3 x 2 to 5 x 5.

To evaluate the potential performance benefits of their constructed interface, the authors created six different reading tasks across two conditions (spatial audio interface vs. conventional screen reader + braille). In each task, the subject was instructed to read a passage consisting of roughly 10 paragraphs with around 30 words per paragraph. Each passage consisted of a table with varying dimensions, specific text alignments, and with at least one portion of bold, italic, or underlined text. Four outcome measures were collected: (1) task completion times, (2) correctness of the perceived information of text alignment and styles, (3) the correctness of perceived table structure information, and (4) a subjective evaluation of the constructed interface. The authors found that the spatial audio interface was more than twice as fast than the conventional screen reader, with an average completion time of 3 minutes and 12 seconds vs. 8 minutes and 38 seconds for the screen reader. Additionally, correctness in identification of text alignment, style, and table information was nearly identical, with no significant differences between the two conditions. Also, the subjective evaluations of the spatial audio interface were positive, with the most common comment stating the high intuitiveness of the design.

While Sodnik et al. elected to not use any space behind the user to display auditory information, earlier designs have used the back hemisphere of the listener to provide specific types of information. Crispien et al. (1994) developed a spatial audio representation of screen information by taking 2D screen items and transforming them into audio items that were displayed as three-component vectors in space. The items were all displayed in the frontal hemisphere of the listener, leaving the back hemisphere open for displaying contextual information like warning messages or help information (Crispien et al., 1994).

Frauenberger et al. (2004) also developed a spatial audio interface to be used to be more efficiently interact with GUI displays. In their interface, users were able to navigate a structured menu, perform text inputs, select items from a list, and confirm various messages and alerts. The metaphor they used in their design was that of a user being in a virtual room with up to six items displayed in a semi-circle in front of them. In order to interact with an item, the user would move their head and select items with keyboard interactions. By using head tracking software, an individual item was played when gazed upon by the user, which was represented as an audio icon playing on a loop. A study with 10 participants (6 with normal vision and 4 with visual impairments) was conducted to analyze the usability of the

design. Users were asked to perform different selection tasks, as well as entering information into the system with the keyboard. Interestingly, no performance differences were found between normal sighted and visually impaired users. However, the authors did find that complex interactions and applications could be modeled with the use of spatial audio interfaces (Frauenberger et al., 2004).

Goose and Moller (1999) proposed a 3D audio-only representation of an HTML web based page. The purpose of this construction was to aid the user in their recognition of the physical structure of the document, since when text is the only information of the HTML file displayed, much of its context gets lost. On top of displaying the structure of the document spatially, other sonification techniques were used to help prevent the user from becoming disoriented when taken to a new document upon clicking on a link. The use of audio icons and earcons aided in this aspect. Specific earcons were used to alert the user when they were interacting with a link that would take them to a new document. If selected, an audio icon of a spacecraft departing one location and landing on another was used to represent the arrival to a new location (Goose & Moller, 1999).

This leads me to the research question of the current study. Can an immersive 3D audio interface be utilized with virtual reality technology to aid blind users navigating information spaces? With the current technology allowed by headsets like the Oculus Rift, the amplitude of audio objects within the information space can be modulated based on the user's head position. I hypothesize that this sound modulation can aid the user by allowing them to quickly sample the structure of the information space and identify their target items more quickly and with less error than conventional screen readers.

Section 2: Navigating Information Spaces Through a Novel Immersive 3D Audio Interface

Despite advancements of accessibility standards, most users with visual impairments are still severely disadvantaged in their use of modern digital information and websites. While conversational interfaces are growing in importance, most browsing still depends on the visual display of information and GUIs. This leaves users who are visually impaired dependent on universal screen readers which results in mediocre experiences at best. Past research has shown that screen readers constrain user actions because of high cognitive workloads (Theofanos & Redish, 2003), a loss of graphical information (Harper et al., 2006; Leuthold et al., 2008), and overall inefficiency (Lazar et al., 2007). The purpose of the current study is to understand the potential of a new, immersive 3D auditory interface in navigating complex information spaces.

While much past research has been done on the tradeoffs between different amounts of depth and breadth within visual menu structures (Miller, 1981; Dray et al., 1981, Snowberry et al., 1982; Kiger, 1984; Shneiderman, 1986), there has been less research on auditory interfaces. New developments in spatial audio and modern VR headsets allows designers to create new and innovative, immersive 3D auditory interfaces. The following study compared two different types of auditory interfaces (screen reader & novel immersive 3D audio) and a standard visual menu to assess the potential usefulness of immersive 3D audio for navigation for low-vision users or in situations where the eyes of a user are already occupied. The 3D audio interface presented three menu options within a particular level simultaneously (described in more detail below).

A 3x2 factorial design was used, with three different interface styles (visual menu, screen reader, and immersive 3D audio) assessing two different depth/breadth combinations for menu structure (4^3 and 8^2 hierarchies). The dependent variables consisted of time to task completion (measured in hundredths of a second), number of errors (defined as incorrect item selection), and subjective usability scores (measured by System Usability Scale).

The research question being addressed is as follows: Will the immersive 3D auditory interface provide better performance and subjective ratings during menu navigation than the conventional screen reader? One of the main differences between the 3D auditory interface and the conventional screen reader lies in the potential to present the auditory information

simultaneously from multiple locations, whereas the traditional screen reader has to present the information sequentially for each item at a given level of the menu. Based on the human auditory system and its ability to spatially localize auditory stimuli, the listener should be able to pinpoint which direction a specific audio item is presented at and direct their attention towards it by moving their head in that direction. Additionally, to leverage past research from Cherry (1953), the 3D audio interface varied the gender of the speech that was vocalizing the audio items in an attempt to improve the discernability of each audio item.

The expectation was to find a main effect for interface type, with the visual menu recording fastest time to completion times and fewest errors, but with the 3D audio interface outperforming the screen reader. Additionally, based on past research by Hochheiser and Lazar (2010), we expected to find a main effect for menu type, with broader menus outperforming deeper menus in all conditions.

Method

Participants

Participants (n = 16) were obtained primarily through snowball sampling. The experimenter recruited subjects by contacting classmates, students, and friends to participate initially. After they completed the study, they were encouraged to ask anyone from their social circle if they'd be willing to also participate. All subjects were paid $20 for their participation in the study.

The average age was 25.9 (range 21-33), with no participants reporting any hearing and visual deficits. All but three participants were students of the University of Idaho, with four majoring in Human Factors, two in Virtual Technology and Design, two in Computer Science, and one each in Biology, Criminology, Communication, Advertising, and Public Health. Participants were asked about their level of experience with VR, with the majority (n = 13) selecting either "novice", or "none" for their experience levels. Two participants indicated they were frequent users of VR, with one additional participant selecting "expert" as their experience level. The primary purpose for participant use of VR was gaming (n = 11).

Project Description

The study was modeled after past research on navigation of menu structures (Hochheiser & Lazar, 2010). Users were tasked with navigating two different menu

structures with varying degrees of depth and breadth in search for specific target items. They navigated these menu structures with three different interface styles: A simulation of the NVDA screen reader, the novel immersive 3D audio, and a standard visual menu common to current graphical user interfaces. The dependent variables recorded were time to task completion, number of errors, and a subjective measure of usability called the System Usability Scale (SUS) (Brooke, 1996).

All conditions were tested through the Unity game engine, utilizing the Unity Experiment Framework (UXF) plugin for data collection. Depth vs. breadth of menu structures were varied similarly to previous research by Hochheiser and Lazar (2010), with participants navigating two separate menu structures for each condition. These menu structures were one 8^2 and one 4^3 menu.

Time-to-Completion and Error Measurements. Each trial measurement for time began immediately after the participant hit the space bar in the auditory conditions or the "start" button in the visual conditions. If the participant at any point requested to see the target instructions again, the timer was stopped as to not add any time that was not related purely to menu navigation. All times were recorded to one-hundredth of a second. Errors were defined as any incorrect item selection. For example, if the participant was tasked with selecting "walleye" from the fish subcategory of the animals category, but instead selected "clothes" and then "hats" before realizing they were moving down the wrong menu path, this would be counted as two errors as they selected two incorrect items before finally moving down the correct path.

Visual Menu Condition

The visual menu condition simulated a menu style common among most computer programs (see. Fig. 2.1). The first level of items in each menu condition were displayed similarly to a toolbar, with each item opening a drop-down menu of the items to be selected from the following level. For the 8^2 condition, the second level contained the final target item. However, for the 4^3 condition, the items in the second level branched off into a third level, which then contained the final target item. Refer to the appendix for a hierarchical representation of each of the menu structures.

The participants navigated this condition with a mouse, while making selections of items by left-clicking the mouse. All target items were found in the final level of each of the

menu structures. Participants were not considered done with their trial until they selected the correct target item. Time-to-completion and error numbers were recorded via UXF for each individual trial.

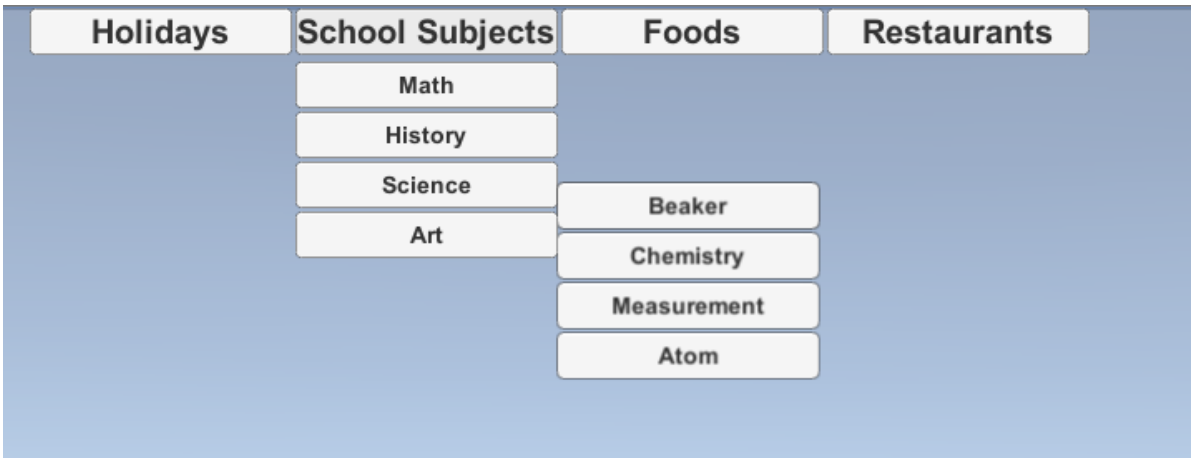

Figure 2.1 Example of what the participants saw for the 4^3 visual menu condition of their practice trial.

Screen Reader Condition

The screen reader condition simulated the sound and controls of the NVDA screen reader. In order to have the data collected within Unity, this condition had the screen reader functionality built into Unity instead of relying on the screen reader software itself. This was due to compatibility issues with screen reader software and Unity.

Since the conditions under which the screen reader was used for this experiment were rather limited, the full functionality was not needed. However, it is worth explaining exactly how the participants navigated the menu structures in the screen reader condition. For navigating and reading webpages via NVDA, users most commonly scan the headings of the page. For example, in Fig. 2.2 a user would navigate the different headings by pressing the "H" key on the keyboard. This would take them from the title of "Important foods in the world" to the headings of "Chocolate", "Dark chocolate", and "Milk chocolate". Additionally, they could move backwards through headings by hitting "Shift + H".

Figure 2.2 Headings and body text. NVDA users can jump from heading to heading by pressing the "H" key on the keyboard. To move backwards through headings, they would press "Shift + H". Jumping from line to line requires the up and down arrow keys, while moving back and forth between letters requires the left and right arrow keys.

While the experimental condition did not consist of headers with body text, it did consist of a hierarchical menu with what NVDA would refer to as a "tab panel". Tab panel items can be navigated by using the up and down or left and right arrow keys. To select an item and move to the next level of the tab panel hierarchy, a user would hit the down arrow on the keyboard. To move back to the previous level, they would hit back arrow. Once in the final level of a menu structure, the space bar can be used to select a particular item. These same controls were used for the experiment.

If you refer back to Fig. 2.1 above as an example, the participant would use the left and right arrow keys to move between the top level of items (Holidays, School Subjects, Foods, and Restaurants). To move into the sub-level of school subjects, the participant would press the down arrow once they hear "school subjects". In the sub-level, the up and down arrow keys would be used to move between items. To move into the sub-level of science (third level of overall menu), the participant would use the right arrow key and then the up and down arrow keys to move between the "science" items. The space bar would then be used to select an item from the final level.

There was no visual information available to the participants in either the screen reader nor the immersive 3D audio condition in an attempt to simulate blindness. Both of these conditions were completed while wearing the VR headset.

Immersive 3D Audio Condition

The 3D auditory interface for this study was developed in the Unity game development program. It featured a virtual scene that was accessed through the VR headset, with the headset itself serving as the tool for navigating the menu structures. For each menu structure, the menu items were placed in a half-circle envelope around the user in 3D space (see Fig. 2.3). Each menu item was represented by a slightly compressed spoken version, with three items played simultaneously on a loop as the user enters the scene. However, based on the user's head position, which was tracked via the Oculus headset, the sound amplitude of each menu item was modulated, with the gaze of the user dictating the item with the highest amplitude. Only three items were playing at any given time, with items surrounding the gazed upon item displaying decreased amplitudes. Additionally, the voices used to generate the audio of each item alternated between male and female in an effort to increase the discriminability of the menu items.

To select an item from the first level and move into its sub-level, the participant was required to gaze at the particular item and press the space bar. The same process was followed for any subsequent levels, with the trial ending once the participant selected the correct item in the final level of each menu structure. If the participant made a mistake and needed to navigate backwards to an earlier level, they would simply press the backspace button while gazing at any item in the current level.

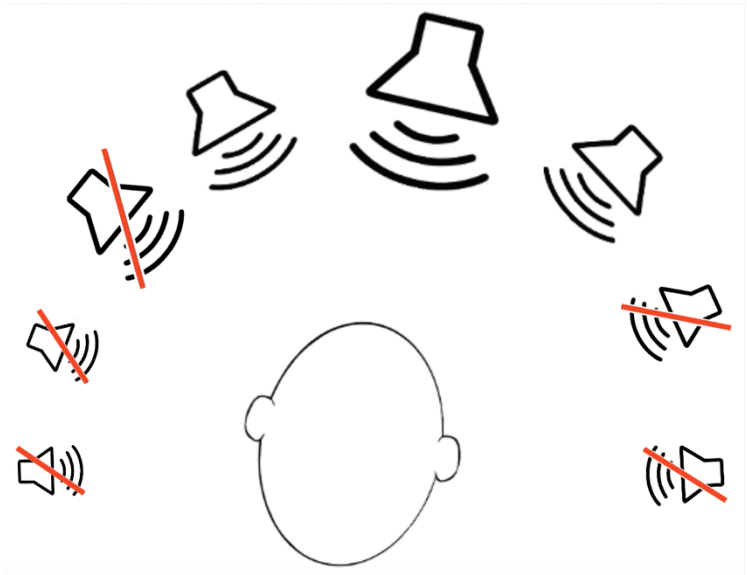

Figure 2.3 Example of 3D immersive audio condition with an 8-item level. Note that the user's gaze will dictate the amplitude of the items in the virtual reality scene.

Study Design

A 3x2 factorial design was used, with three different interface styles (visual menu, screen reader, and immersive 3D audio) assessing two different depth/breadth combinations for menu structure (4^3 and 8^2 hierarchies). Each of the menu structures contained the same amount of total items (64). Additionally, the targets were balanced across trials such that each category and sub-category of the menu structures occurred an equal amount of times for each participant's session. Also, no target item was repeated at any point for each participant.

Each participant completed two rounds of 16 blocks, with the first eight blocks of round one consisting of only the visual condition. In the second round, the final eight blocks contained the visual condition. Each of the eight blocks of the visual condition consisted of four trials (twice with 8^2 and twice with 4^3) where they would search the menu structure for a target item. For the eight auditory blocks, each block consisted of four trials for both the screen reader and 3D immersive audio conditions (again twice with 8^2 and twice with 4^3), resulting in eight trials per auditory block. These eight auditory blocks were also balanced such that the number of blocks beginning with a particular interface (i.e. 3D immersive audio vs. screen reader) were equal. For each block, all trials of one condition were completed before moving on to the next condition. Furthermore, the order that each participant saw the blocks were randomized.

Finally, the position of the target item was balanced such that it occurred the same amount of times across all blocks and conditions. Target position is defined as its particular place in the menu structure. Using Fig. 2.4 as an example, the item of "walnut" found in the "trees" category would have a final position of four, since that's where it lies on the final sub-level.

	1	2	3	4	5	6	7	8
	utensils	states	music	geography	presidents	automobiles	trees	technology
1	spatula	north dakota	rap	mountain	trump	volvo	cottonwood	computer
2	fork	vermont	blues	ocean	clinton	honda	spruce	monitor
3	knife	texas	funk	jungle	obama	cadillac	sycamore	printer
4	spoon	louisiana	disco	crater	biden	chevy	walnut	television
5	grater	wisconsin	jazz	forest	washington	dodge	mahogany	headphones
6	colander	nebraska	punk	volcano	jackson	porsche	chestnut	router
7	tongs	florida	pop	glacier	reagan	toyota	maple	mouse
8	whisk	ohio	soul	island	lincoln	ferrari	redwood	keyboard

Figure 2.4 Visualization of the position of items within the 8^2 menu structure.

The dependent variables consisted of time to task completion (measured in hundredths of a second), number of errors (defined as incorrect item selection), and subjective usability scores (measured by System Usability Scale).

Materials

Computer. All sessions were completed on a custom-built PC running Windows 10 as its operating system. The primary specs of the computer were as follows:

- Processor: Intel Core i7-5820k CPU @ 3.3GHz
- GPU: Nvidia GTX 980-Ti
- Samsung 850 pro 512GB drive
- Seagate Barracuda 7200.14 1TB drive
- RAM: 16GB total – G.SKILL F4 DDR4 3000 C15 2x8GB clocked at 2133MHz
- Motherboard: Asrock X99X Killer

Virtual Reality Headset. For both the screen reader and immersive 3D audio condition, participants completed each task while wearing an Oculus Rift S. However, instead of using the Oculus Touch controllers, participants used the space bar on the keyboard as their selection input. For example, in the 3D audio condition, if the user's gaze was on an item they desired for selection, they would simply press the space bar to "select" that item from the menu hierarchy. Also, for the screen reader condition, participants navigated the menus with the arrow keys. To simulate blindness, no visual information was given inside the headset for either of the auditory conditions.

Unity Game Engine. All experimental and practice conditions have been built with the Unity game development engine, version 2019.3.5f1.

Unity Experiment Framework (UXF) 2.0. Unity Experiment Framework (UXF) 2.0 was used for the data collection of this experiment (Brookes et al., 2020). UXF is a package that can be downloaded through the Unity asset store and it provides the functionality needed for data collection with human subjects. It allows the researcher to take advantage of the powerful tools Unity provides for task creation by providing numerous scripts that can be imported in the scene and used to collect data on the variables of interest. This takes much of the programming burden off the researcher. Additionally, it provides the functionality for

random assignment to conditions, creation of blocks and trials, and the ability to host the experiment on the web for remote data collection. All data collected was uploaded to a private GitHub repository (excluding any identifying information) so the primary researcher could access the session data remotely.

System Usability Scale. The System Usability Scale (SUS), which was developed by John Brooke (1996), was used as the subjective usability measure for the screen reader and 3D immersive audio conditions. Past research has indicated the SUS to be a highly robust and versatile tool in evaluating a product or system's usability (Bangor et al., 2008; Lewis, 2018). It is a 10 item questionnaire with likert-style questions regarding the user's perceived usability of a system or product. For example, one item would be the statement "I would imagine that most people would learn to use this system very quickly.", and the user would respond with a 1-5, with 1 indicating "strongly disagree" and 5 indicating "strongly agree". For the full list of items, please refer to the appendix section. The SUS was given at the end of the experimental session.

Pilot Testing

Pilot testing was performed across a number of iterations with multiple members of the Cognition and Usability Lab at the University of Idaho. This pilot testing was crucial in determining the amplitudes of the menu items in both auditory conditions. Additionally, a handful of iterations were tested to identify the best method of amplitude modulation in the 3D audio condition. The final result of this pilot testing was that no more than three items at any given time would be playing concurrently in the 3D audio condition, with the two items immediately next to the gazed upon item displaying amplitudes slightly lower than the primary item. In the first few iterations, all menu items were played concurrently in an effort to provide more information and support faster completion times. However, pilot testing revealed that this was too overwhelming and, as a result, became counterproductive.

Additionally, this pilot testing helped determine the size of the envelope containing the items in the 3D audio condition. Initially, the envelope was too wide, requiring a large degree of neck rotation to interact with every menu item. The final iteration reduced the amount of neck rotation needed to what was deemed a more comfortable range.

Site of Study

All data was collected in the Cognition and Usability Lab on the University of Idaho campus.

Procedure

Upon entry to the study, participants were given a one-page demographics questionnaire, as well as a debriefing about what the study will entail and its intended purpose. Once the questionnaire was completed, participants were introduced to their practice session. While we expected the subjects to need minimal guidance in the visual menu condition, there was some training that was expected before they were comfortable with both auditory interfaces. Within the practice condition was a set of instructions that explained how to interact with each of the three interface styles. As an addition to the visual instructions, the experimenter also explained verbally how they were to interact with each interface. Each participant was encouraged to take as much time as they needed to explore the three conditions and become comfortable with the controls.

At the beginning of each trial the participant was presented with a target to search for. Once ready, the participant would hit the space bar to begin their trial for the auditory conditions, or press the "start" button for the visual interface. For each trial, auditory feedback was given upon selection of correct or incorrect menu items, with a buzzer indicating incorrect and a dinging bell indicating correct. If at any point during a trial the participant forgot what their target was, they could ask the experimenter for help and he would bring up the help screen by pressing the escape key which would present their target again.

Once they had chosen the correct item for each interface, thus completing their practice session, they were then presented with their first eight blocks of visual menus for the experimental trials. They completed four separate trials in each visual block (twice with each of the 8^2 and 4^3 menus). Upon completion of the visual blocks, the participant was then given their eight auditory blocks, which each consisted of four trials for each of the screen reader and 3D immersive audio conditions. This completed their first round of trials. They were then given their second round of trials, this time with the auditory blocks completed first.

Upon finishing their second round of testing, the participants were given the SUS to evaluate the usability of each of the auditory interfaces. In addition to the SUS, participants

were given two likert-scale items asking, “how easy was it to hear/discriminate the different words?” and “how easy was it for you to navigate the auditory interfaces with the keyboard keys?”. These were scored on a scale from 1-5, with 1 indicating “very difficult” and 5 indicating “very easy”. Finally, they were also asked for any general comments about the study or interfaces they interacted with.

The decision was made to exclude the two likert-scale items from any analysis. There were two main reasons for this: 1) The wording of the first item was unclear in what it was asking. To hear vs. discriminate words are different processes altogether, so to include them both in the same question was a mistake that was not caught until the conclusion of the study. 2) The second items asking about the easiness of navigation with arrow keys was seen as a redundant measure since the SUS was already in use to measure ease of use. Thus, the decision was made to rely on the more well-established SUS as the measure of usability.

Section 3: Results

Time-to-Completion

To evaluate the time-to-completion data, a two-way repeated measures ANOVA was run in SPSS. First, to check for outliers, the studentized residuals were generated for each interface and menu type. Fig. 3.1 shows that participant 14 was the only one with residuals greater than ± three. This participant indicated before they began their trial that they were very tired. Their data was not excluded from analysis because while their 3D audio times were deemed extreme outliers by SPSS, their times across all conditions were slower than the average. Thus, it was felt their poor performance was spread fairly evenly across each condition.

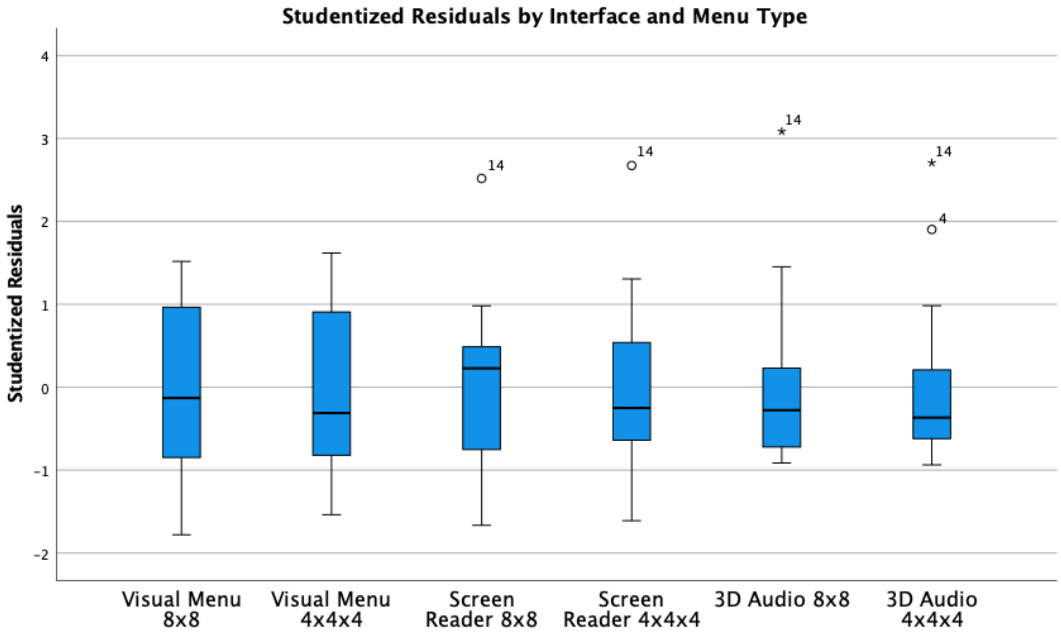

Figure 3.1 Studentized residuals by interface and menu type.

Times for the 3D audio condition in each menu type violated the assumption of normality ($p = .001$ for 8^2, $p = .002$ for 4^3), assessed by Shapiro-Wilk's test of normality on the studentized residuals. However, no other condition violated normality, and with ANOVAs being considered robust against violations of normality (Laerd Statistics, 2015), the analysis was continued without transforming any data.

Mauchly's test of sphericity indicated that the assumption of sphericity had been violated for the within-subjects comparisons between interfaces $\chi^2(2) = 29.799$, $p < .001$. The

Greenhouse-Geisser correction was applied to correct for this violation and any future violations of sphericity.

There was a significant two-way interaction between interface and menu type, $F(1.191, 17.863) = 10.810$, $p = .003$, $\eta_p^2 = .419$. To evaluate the simple main effects, a one-way repeated measures ANOVA was conducted. There was a statistically significant simple main effect in completion times between the visual menu, screen reader, and 3D audio interface styles across both the 8^2 and 4^3 menu structures. For the 8^2 menu type, the visual menu performed best ($M = 3.15$, $SE = 0.10$), screen reader second-best ($M = 7.16$, $SE = 0.27$), and 3D audio performing worst ($M = 15.66$, $SE = 1.44$), $F(1.028, 15.416) = 69.269$, $p < .001$. For the 4^3 menu type, the visual menu again performed best ($M = 3.77$, $SE = 0.17$), screen reader second-best ($M = 8.57$, $SE = 0.49$), and 3D audio again performing worst ($M = 13.93$, $SE = 1.25$), $F(1.132, 16.976) = 67.222$, $p < .001$.

Additionally, there was a significant simple main effect in completion times between the 8^2 and 4^3 menu types across all interface types. For the visual menu, the broader 8^2 menu ($M = 3.15$, $SE = 0.10$) outperformed the deeper 4^3 menu ($M = 3.77$, $SE = 0.17$), $F(1, 15) = 43.938$, $p < .001$. For the screen reader condition, the broader 8^2 menu ($M = 7.16$, $SE = 0.27$) again outperformed the deeper 4^3 menu ($M = 8.57$, $SE = 0.49$), $F(1, 15) = 22.267$, $p < .001$. The 3D audio condition, on the other hand, had the deeper 4^3 menu ($M = 13.93$, $SE = 1.25$) outperform the broader 8^2 menu ($M = 15.66$, $SE = 1.44$), $F(1, 15) = 4.644$, $p = .048$.

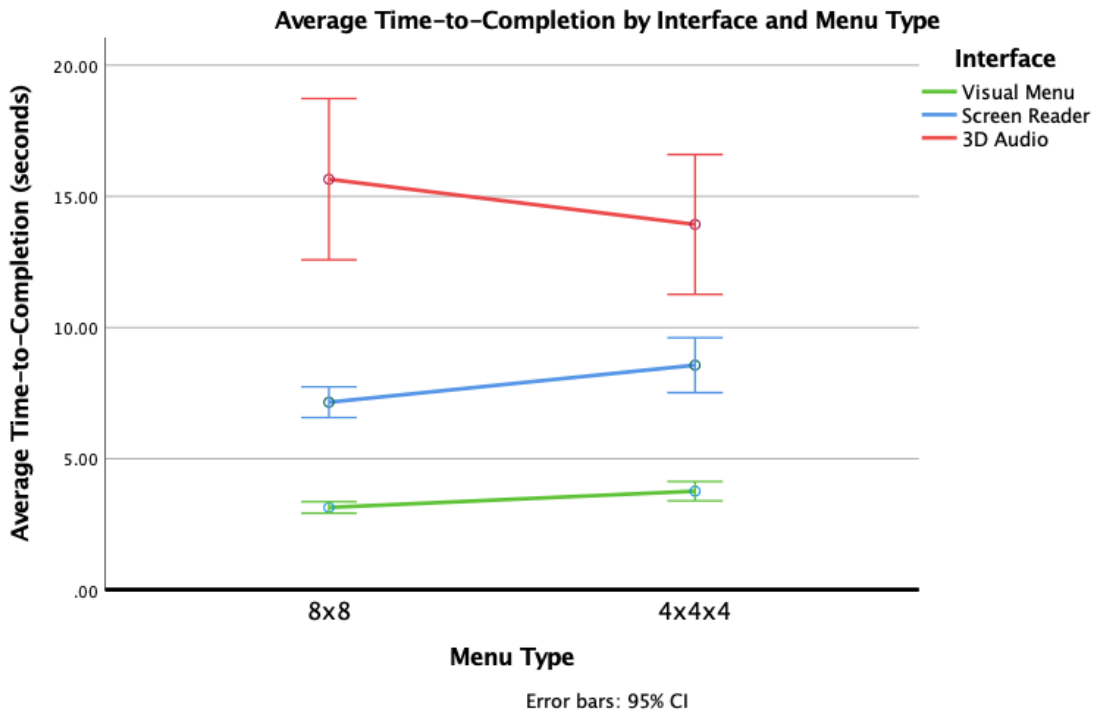

Figure 3.2 There was a statistically significant difference in completion times across all interface styles, regardless of menu type. For the visual menu and screen reader conditions, the broader 8^2 menus outperformed the deeper 4^3 menus. For the 3D audio condition, the deeper 4^3 menu outperformed the broader 8^2 menu.

Furthermore, while there did appear to be improvements in mean scores between rounds one and two across all participants, the improvements in time did not reach a level of significance to indicate a three-way interaction ($p = .276$). However, there was a two-way interaction between interface type and round $F(1.264, 18.960) = 15.329$, $p < .001$, $\eta_p^2 = .505$. A one-way repeated measures ANOVA was used to determine where these differences were found. There was a statistically significant difference between rounds one and two for the visual menu across both menu types. For the 8^2 menu, the times improved from round one ($M = 3.27$, $SE = 0.11$) to round two ($M = 2.93$, $SE = 0.11$), $F(1, 15) = 8.949$, $p = .009$. For the 4^3 menu, the times again improved from round one ($M = 3.83$, $SE = 0.20$) to round two ($M = 3.54$, $SE = 0.14$), $F(1, 15) = 7.266$, $p = .017$.

There was also a statistically significant difference between rounds one and two for the screen reader condition across both menu types. For the 8^2 menu, the times improved from round one ($M = 7.67$, $SE = 0.38$) to round two ($M = 6.33$, $SE = 0.22$), $F(1, 15) = 33.326$, $p < .001$. For the 4^3 menu, the times improved from round one ($M = 9.16$, $SE = 0.63$) to round two ($M = 7.40$, $SE = 0.54$), $F(1, 15) = 9.019$, $p = .009$.

Finally, there was another significant difference between rounds one and two for the 3D audio condition across both menu types. For the 8^2 menu, the times improved from round one ($M = 16.48$, $SE = 1.62$) to round two ($M = 14.08$, $SE = 1.37$), $F(1, 15) = 18.682$, $p < .001$. For the 4^3 menu, the times again improved from round one ($M = 15.68$, $SE = 1.62$) to round two ($M = 11.91$, $SE = 1.00$), $F(1, 15) = 13.960$, $p < .005$.

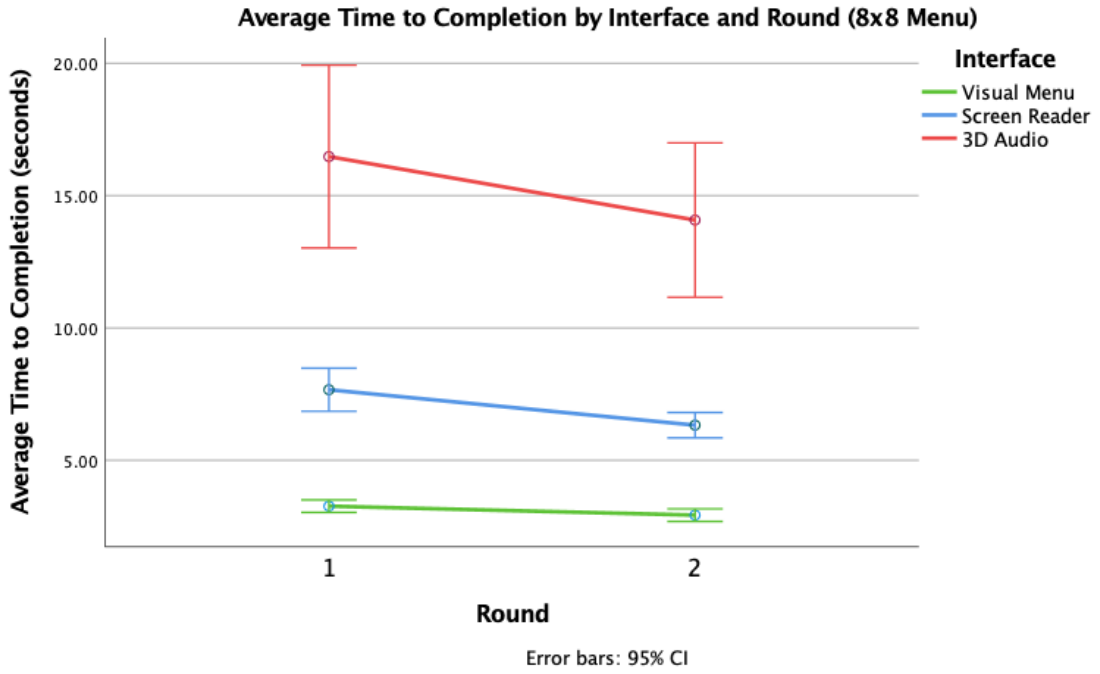

Figure 3.3 Statistically significant improvements in time across all interfaces in the 8^2 menu from round one to round two.

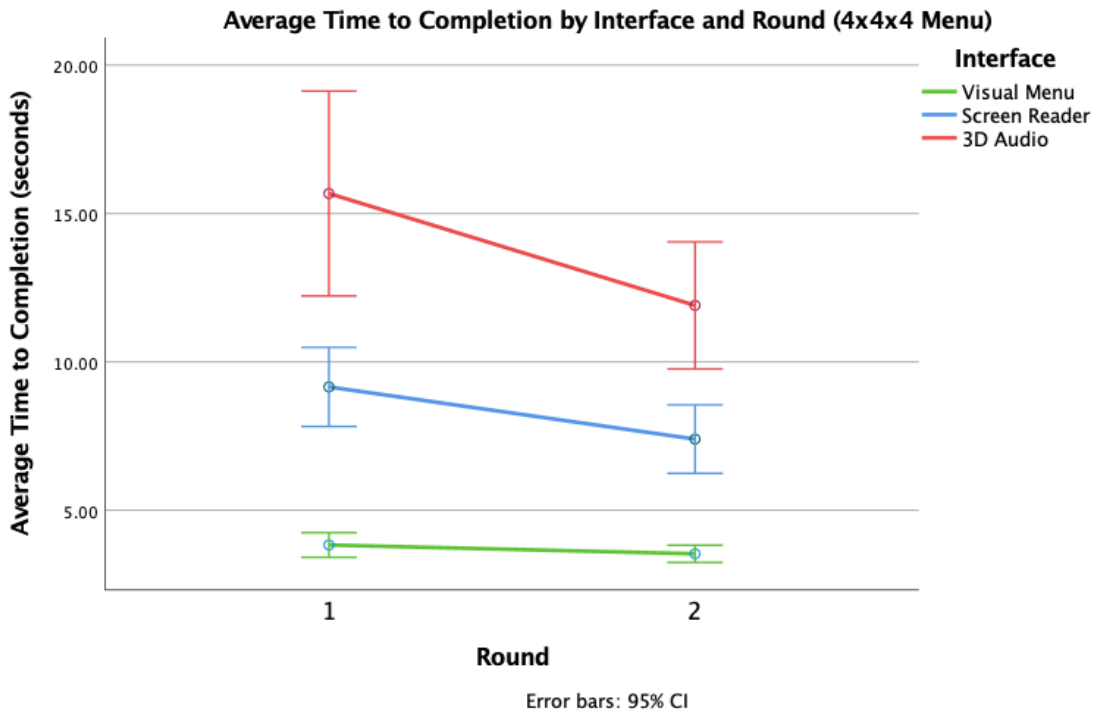

Figure 3.4 Statistically significant improvements in time across all interfaces in the 4^3 menu from round one to round two.

When evaluating the three interface styles irrespective of menu type, the visual interface performed the best in terms of time-to-completion ($M = 3.39$, $SE = 0.12$), with the screen reader performing second-best ($M = 7.64$, $SE = 0.39$), and 3D audio performing the worst ($M = 14.54$, $SE = 1.30$). The ANOVA revealed a significant difference among the three interface styles $F(1.063, 15.949) = 72.889$, $p < .001$, $\eta_p^2 = .829$. A Bonferroni post hoc test was used to determine which interface styles were different from one another. The 3D audio interface was significantly slower than both the screen reader and visual interfaces, with a mean difference of +6.90 seconds, 95% CI [4.17, 9.62] $p < .001$ compared to the screen reader, and a mean difference of +11.14 seconds, 95% CI [7.86, 14.43] $p < .001$ compared to the visual condition. Additionally, the screen reader was significantly slower than the visual interface, with a mean difference of +4.25 seconds, 95% CI [3.41, 5.09] $p < .001$.

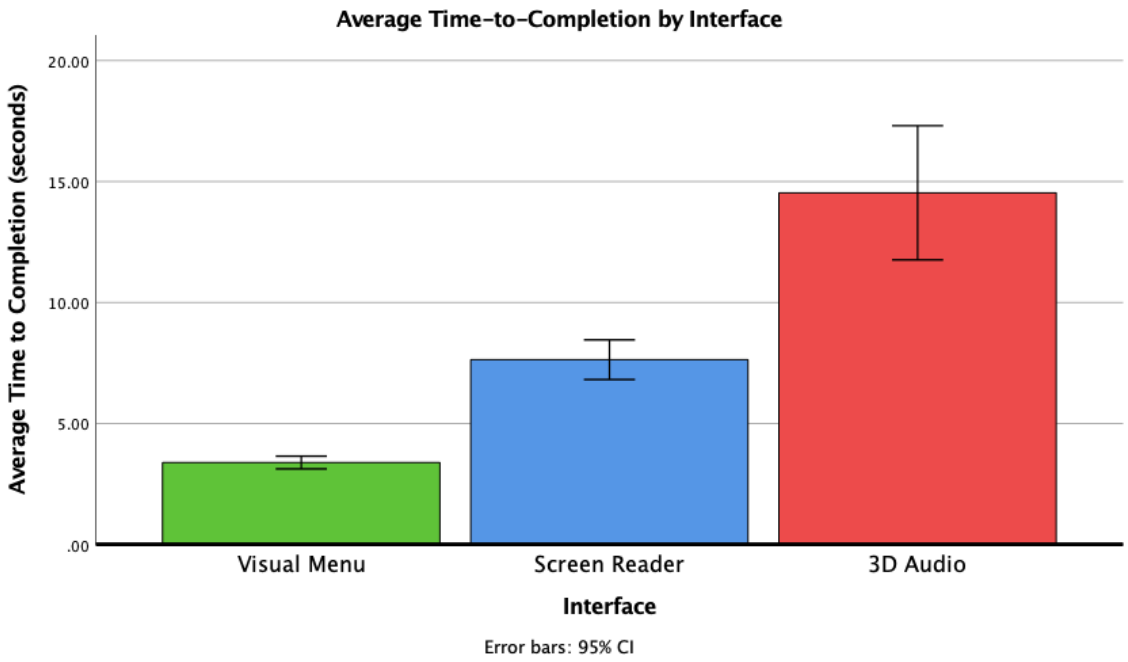

Figure 3.5 Time-to-completion by interface type. Visual menu, on average, was fastest, with the screen reader condition performing second-best and the 3D audio condition performing the worst.

Misses

Another two-way repeated measures ANOVA was used to evaluate the misses across interface and menu type. To check for outliers, the studentized residuals were again generated for interface and menu type. As you'll see from Fig. 3.6, there were a significant number of outliers in the misses data. However, because they appeared to be evenly dispersed across all conditions and menu types, the decision was made not to remove these outliers from data analysis.

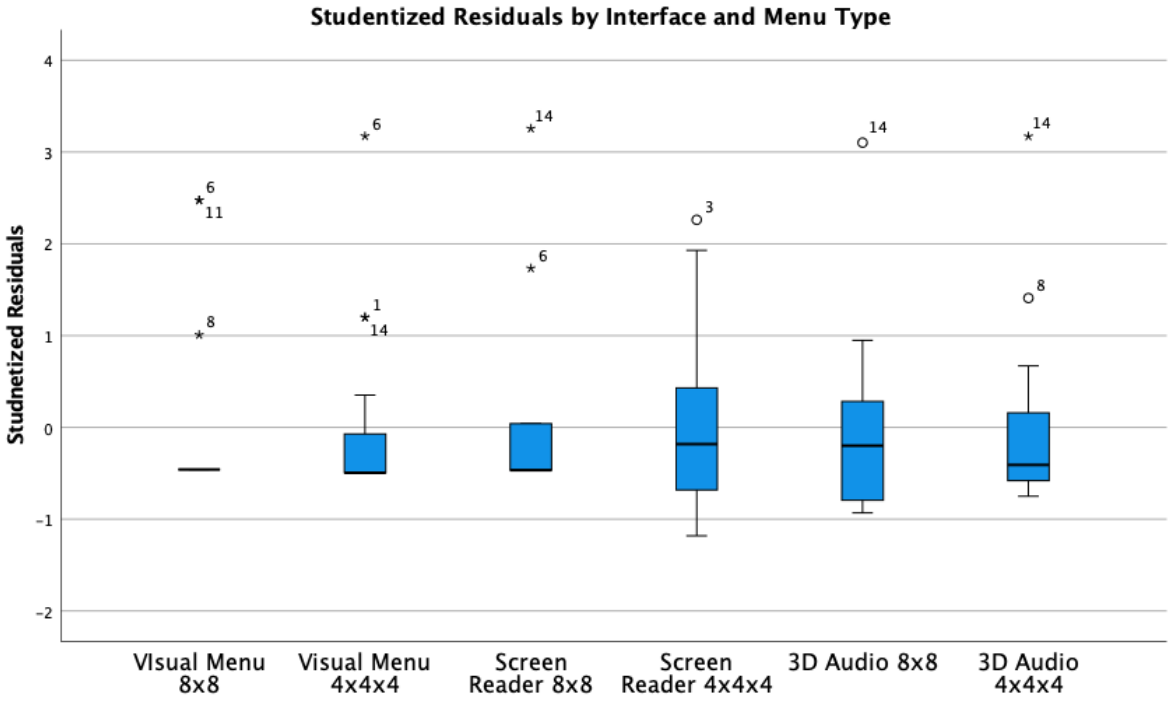

Figure 3.6 Studentized residuals of misses by interface and menu type.

To evaluate the assumption of normality, a Shapiro-Wilk's test of normality was applied to the studentized residuals of the miss data. All cells of the interface and menu types violated the assumption of normality, with four of the six cells being significant at the $p <$ .001 level. The final two cells of residuals were for the screen reader 4^3 menu condition and 3D audio 8^2 menu condition, with p-values of $p = .032$ and $p = .003$, respectively. Because the violations of normality were more severe with the misses data, non-parametric alternatives were conducted to determine if the results were severely affected by this violation.

For misses, there was a statistically significant two-way interaction between interface and menu type, $F(1.400, 20.999) = 7.716$, $p = .006$, $\eta_p^2 = .340$. A one-way repeated measures ANOVA was used to determine the simple main effects. The 3D audio interface had significantly more misses per trial ($M = 0.23$, $SE = 0.06$) than both the visual menu ($M = 0.01$, $SE = 0.01$, $p = .004$) and screen reader ($M = 0.03$, $SE = 0.02$, $p = .002$) conditions within the 8^2 menu $F(1.072, 16.074) = 16.546$, $p < .001$. For the 4^3 menu, there was a significant simple main effect observed between both the 3D audio ($M = 0.16$, $SE = 0.05$, $p = .017$) and screen reader ($M = 0.11$, $SE = 0.02$) conditions and the visual menu ($M = 0.02$, $SE = 0.01$), $F(1.405, 21.075) = 6.842$, $p = .010$.

Additionally, the only significant simple main effect observed for menu types was the difference in average misses of the screen reader condition between the broader 8^2 menu (M = 0.03, SE = 0.02) and the deeper 4^3 menu (M = 0.11, SE = 0.02), $F(1, 15) = 13.658$, $p <$ .005. There were no significant differences found for misses between rounds one and two across any of the interfaces or menu types.

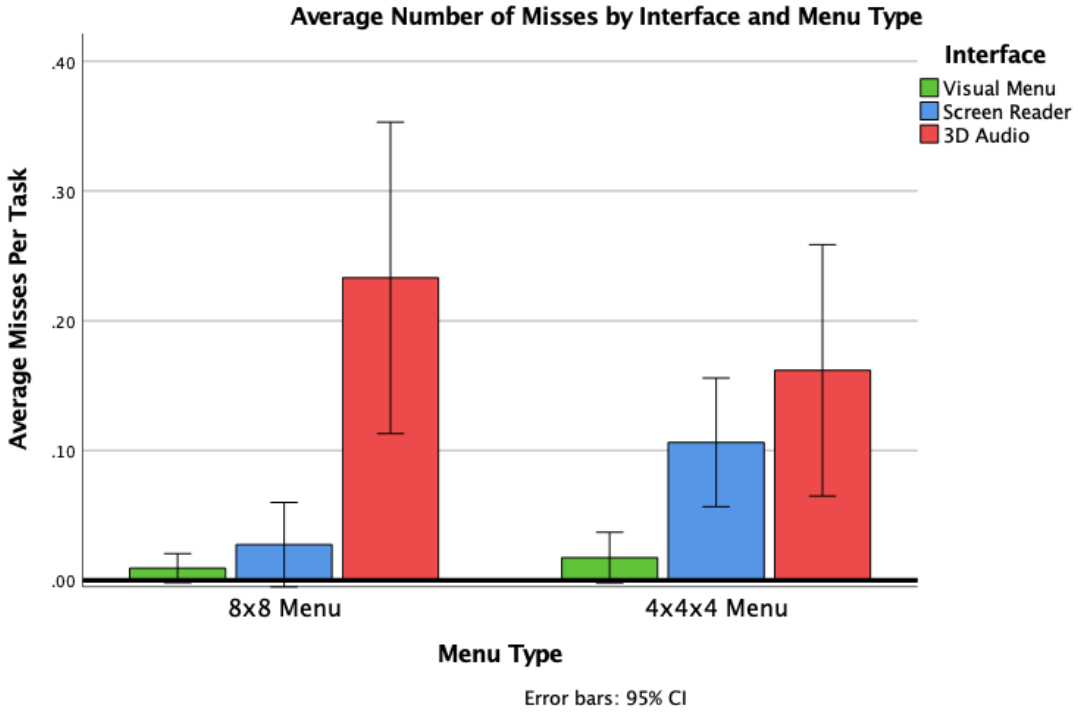

Figure 3.7 In the 8^2 menu, the 3D audio condition had significantly more misses than both the visual menu and screen reader conditions. For the 4^3 menu, the only significant difference in misses were between the 3D audio and visual menu conditions. For menu type, there was a significant simple main effect between the 8^2 and 4^3 menus within the screen reader condition.

When evaluating the three interface styles irrespective of menu type, the visual interface performed the best with the lowest average of misses per task (M = 0.01, SE = 0.01), with the screen reader performing second-best (M = 0.07, SE = 0.02), and 3D audio performing the worst (M = 0.20, SE = 0.05). The ANOVA revealed a significant difference among the three interface styles $F(1.145, 17.175) = 13.703$, p = .001, η_p^2 = .477. A Bonferroni post hoc test was used to determine which interface styles were different from one another. The 3D audio interface contained significantly more misses than both the screen reader and visual interfaces, with a mean difference of +0.13 misses, 95% CI [0.02, 0.23] p = .014 compared to the screen reader, and a mean difference of +0.18 misses, 95% CI [0.06, 0.31] p = .004 compared to the visual condition. Additionally, the screen reader contained

significantly more misses than the visual interface, with a mean difference of +0.05 misses, 95% CI [0.01, 0.10] $p = .009$.

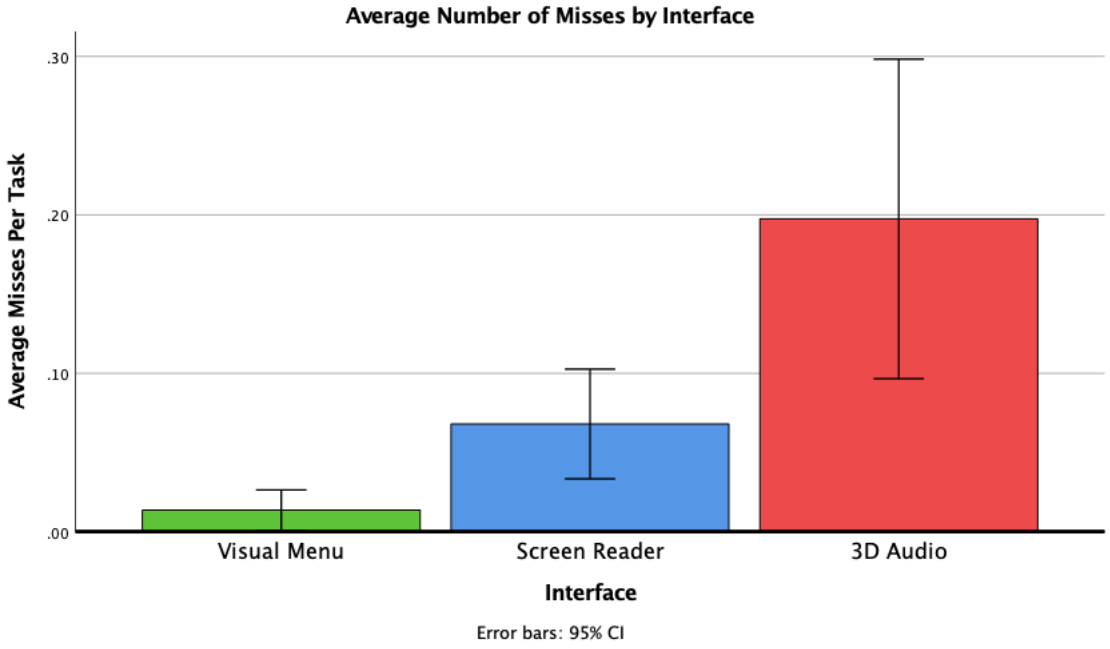

Figure 3.8 Average misses per task by interface type. Visual menu, on average, was the least error-prone, with the screen reader condition performing second-best and the 3D audio condition performing the worst.

Friedman tests were conducted to assess how severely these results were affected by their violations of normality. The Friedman tests came back with nearly identical results. In the 8^2 menu, the 3D audio interface was significantly different than both the screen reader and visual menu $\chi^2(2) = 28.255$, $p < .001$. Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons, which indicated that differences were all significant at the $p < .001$ level.

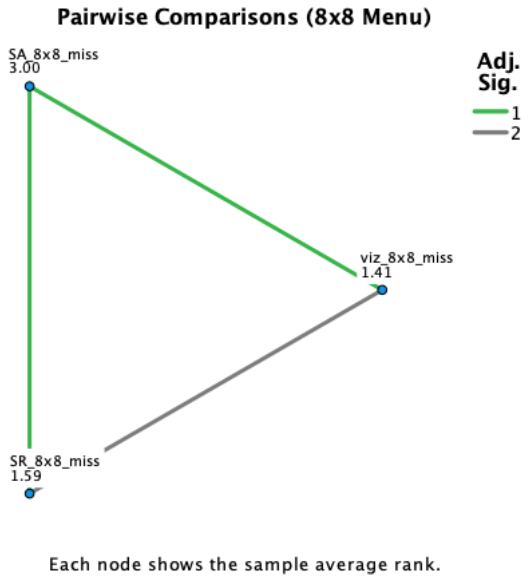

Figure 3.9 Pairwise comparisons of the different interface types in the 8^2 menu. The 3D audio condition was significantly different than both the visual and screen reader interfaces, while the differences between the screen reader and visual interfaces did not reach significance. Green lines indicate significant differences between nodes.

In the 4^3 menu, the 3D audio and screen reader interfaces were both significantly different than the visual menu $\chi^2(2) = 18.633$, $p < .001$. Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons. The 3D audio interface was significantly different from the visual menu at the $p < .001$ level, while the screen reader was significantly different from the visual menu at the $p = .003$ level.

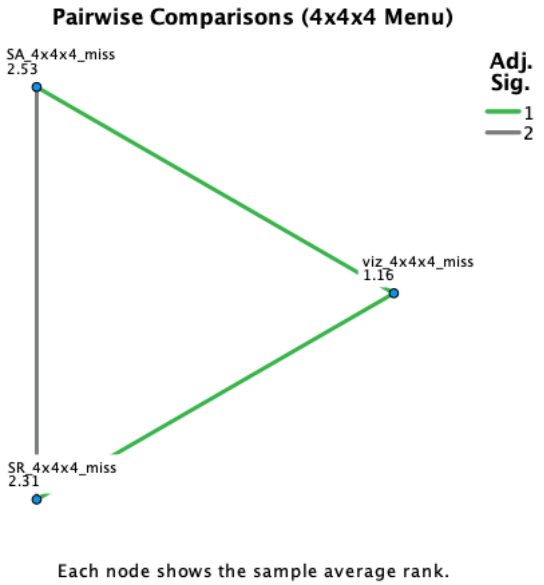

Figure 3.10 Pairwise comparisons of the different interface types in the 4^3 menu. The visual menu was significantly different than both the 3D audio and screen reader interfaces, while the differences between the screen reader and 3D audio did not reach significance. Green lines indicate significant differences between nodes.

When comparing the interface types irrespective of menu type, the 3D audio interface was significantly different than both the screen reader and visual menu $\chi^2(2) = 28.222$, $p < .001$. Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons. The 3D audio was significantly different from the visual menu at the $p < .001$ level and the screen reader at the $p = .003$ level. The only Friedman results that deviated from the ANOVA was that the differences between the screen reader and visual menu in total average misses did not reach significance.

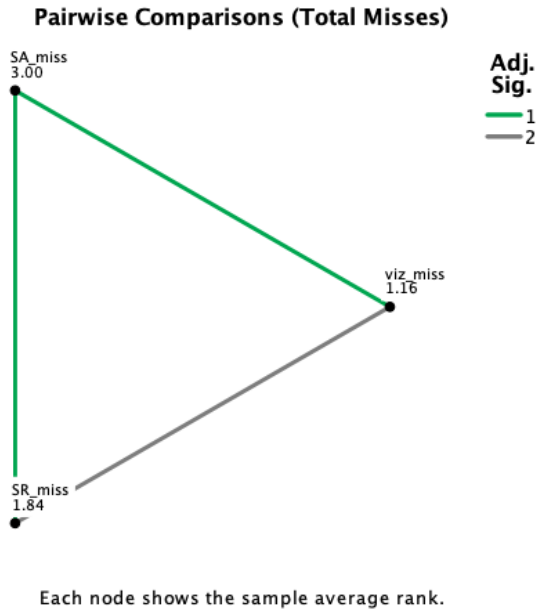

Figure 3.11 Pairwise comparisons of the different interface types. The 3D audio condition was significantly different than both the visual and screen reader interfaces, while the differences between the screen reader and visual interfaces did not reach significance. Green lines indicate significant differences between nodes.

An exact sign test was conducted to assess how severely the results between the two menu types were affected by the violations of normality. The results mimicked that of the ANOVA, with the only differences reaching significance being those between the 8^2 and 4^3 menus in the screen reader condition, $p < .001$. Largely, the results from all of the non-parametric tests confirmed the results of the ANOVAs.

System Usability Scale

For the screen reader interface, the average SUS score was 88.75. For the 3D audio interface, the average SUS score was 64.22. According to the recommendations put forth by Bangor et al. (2008), this puts the screen reader in the highest category of "acceptable", with 3D audio going in the next lowest category of "marginally acceptable".

Section 4: Discussion

Main Findings

The original hypotheses were that, 1) the 3D audio interface would outperform the screen reader in both average time-to-completion and error frequency, 2) the broader 8^2 menus would outperform the deeper 4^3 menus across all interface styles, and 3) the 3D audio interface would score more favorably on the SUS measure of subjective usability than the screen reader. The hypothesis that the 3D immersive audio interface to outperform the screen reader interface on both time-to-completion and number of misses was not supported. Both the average time-to-completion data and the occurrence of errors showed a clear disadvantage of the 3D audio. The hypothesis that the 3D audio condition would score higher on the SUS than the screen reader was not supported, with the screen reader landing in the "acceptable" range of usability and the 3D audio falling into the "marginally acceptable" range. The hypothesis that the broader 8^2 menus would outperform the deeper 4^3 menus was partially supported, as the general trend was that the visual menu and screen reader tasks were performed faster and with less errors in the 8^2 menus. Interestingly, the opposite was true for the 3D audio condition, with the general trend indicating performance on both time-to-completion and number of errors was better in the 4^3 menus.

The differences in overall completion times between the three interfaces were apparent early on in data collection. I believe part of the performance decrements in the 3D audio condition can be explained by Fitts' Law. Most often referenced when talking about the time it takes to move a mouse pointer to a particular point on a computer screen, Fitts' Law tells us that the time required for a person to move a pointer to a desired target is a function of the size and distance of said target (Fitts, 1954). This comes into play when you realize that the selection of targets in the 3D audio condition is dependent on the precision of the user's head movements. With this condition, the participant had a roughly 90 degree window (45 degrees to the left and right) to explore the items in that menu structure. The precision requirement is made more obvious in the 8^2 menu, where the menu items are forced to be smaller (about 11 degrees vs. 22 degrees in the 4^3 menu) in order for them all to fit in the 90 degree half-circle envelope around the user's head. This may explain why differences in performance between each interface were more pronounced in the 8^2 menus. The screen reader interface gets around this problem by only requiring arrow key presses to move about

the menu structure. Once the user's finger is on the arrow key, there is very little precision needed to move to the next item. One limitation of this study was the lack of actual blind users as test subjects and/or pilot testers. Had they been involved early on in the pilot testing, a different direction of the subsequent iterations of the 3D audio interface could have potentially found a solution to this Fitts' Law problem

The SUS data was particularly surprising, but not because the screen reader was rated as more usable than the 3D audio. Rather, it's because the nature of the SUS questions would lead you to expect *both* of the auditory interfaces to score very poorly on this measure. When someone has been using visual interfaces their entire life, one would think an auditory only interface would be such a stark contrast that it would be hard to score well on a subjective usability scale such as the SUS. The fact that neither auditory interface scored in the unacceptable range for usability seems to suggest that the underlying functionality of each can be built upon to try bridge the performance gap between auditory interfaces and visual ones. Or, it may suggest that scoring marginally acceptable in the SUS is a rather low bar to clear.

In regards to the menu types with varying degrees of depth and breadth, the results of the visual condition replicated past research with similar menu structures (Dray et al., 1981; Kiger, 1984; Miller, 1981; Shneiderman, 1986). This same trend held up for the screen reader condition, with the broader menu outperforming the deeper menu. While the opposite was true for the 3D audio condition, I don't think this is because spatial audio provides any unique benefits to interacting with deeper menus. Rather, I think this effect occurred because the time saved in the deeper menu by requiring less precision in head movements was enough to outweigh the extra time needed to gaze at the correct item in the broader menu. This says more about the lack of usability of the 3D audio interface with the broader menu than it does about any improved performance in deeper menus. If the Fitts' Law issue can be solved for broader menus utilizing head movements as the input selection, I would expect broader menus to then outperform deeper ones.

Future Research

A promising finding with the 3D audio condition was that the improvements in completion times from round one to round two were more pronounced than both the screen reader and visual menu. This may indicate that more practice with this interface would close

the performance gap even more between 3D audio and the screen reader and visual menu. One limitation of this study was that the experiment needed to be built programmatically in Unity's game development program. As I had no programming experience going into this project, that was a time-consuming process that limited the amount of investigation I could put towards evaluating the practice effects. With that said, a combination of exposure to VR 3D audio and improvements in the implementation of the spatial audio by changing the target size, target location, and how many targets are audible could provide enough improvements in performance to become similar to screen reader performance. Perhaps a longitudinal study spanning weeks to months would show this to be true.

While the current iteration of the 3D audio interface was clearly outperformed by the screen reader, there are enough changes that could be made to the spatial audio's implementation where future research could again be considered. Potentially, a combination of spatial audio and arrow key navigation could be used to evaluate any performance benefits over a traditional screen reader. Also, with past research showing spatial audio (absent of VR) could be better than screen readers at providing contextual information (Sodnik et al., 2012) and performing complex computer interactions (Frauenberger et al., 2004), another avenue for future research could be to attempt to replicate these findings while utilizing VR to administer the spatial audio aspects of the interface.

References

Abdolrahmani, A., Kuber, R., & Branham, S. M. (2018, October). " Siri Talks at You" An Empirical Investigation of Voice-Activated Personal Assistant (VAPA) Usage by Individuals Who Are Blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 249-258).

Azenkot, S., & Lee, N. B. (2013, October). Exploring the use of speech input by blind people on mobile devices. In *Proceedings of the 15th international ACM SIGACCESS conference on computers and accessibility* (pp. 1-8).

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, *24*(6), 574-594.

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1-24.

Bogers, T., Al-Basri, A. A. A., Rytlig, C. O., Møller, M. E. B., Rasmussen, M. J., Michelsen, N. K. B., & Jørgensen, S. G. (2019, March). A study of usage and usability of intelligent personal assistants in Denmark. In *International Conference on Information* (pp. 79-90). Springer, Cham.

Bouck, E. C., Flanagan, S., Joshi, G. S., Sheikh, W., & Schleppenbach, D. (2011). Speaking math—A voice input, speech output calculator for students with visual impairments. *Journal of Special Education Technology*, *26*(4), 1-14.

Bourne, R. R., Adelson, J., Flaxman, S., Briant, P., Bottone, M., Vos, T., ... & Taylor, H. R. (2020). Global Prevalence of Blindness and Distance and Near Vision Impairment in 2020: progress towards the Vision 2020 targets and what the future holds. Investigative Ophthalmology & Visual Science, 61(7), 2317-2317.

Branham, S. M., & Mukkath Roy, A. R. (2019, October). Reading between the guidelines: How commercial voice assistant guidelines hinder accessibility for blind users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 446-458).

Brewster, S., Raty, V., & Kortekangas, A. (1996). Earcons as a method of providing navigational cues in a menu hierarchy. *People and computers XI: Proceedings of HCI '96,* 169-183. doi:10.1007/978-1-4471-3588-3_12

Brooke, J. (1996). SUS: A “quick and dirty” usability. *Usability evaluation in industry*, *189*.

Brookes, J., Warburton, M., Alghadier, M., Mon-Williams, M., & Mushtaq, F. (2020). Studying human behavior with virtual reality: The Unity Experiment Framework. *Behavior research methods*, *52*(2), 455-463.

Brophy, P., & Craven, J. (2007). Web accessibility. *Library trends*, *55*(4), 950-972.

Byrne, M. D., Anderson, J. R., Douglass, S., & Matessa, M. (1999). Eye tracking the visual search of click-down menus. *Proceedings of the SIGCHI conference on human factors in computing systems the CHI is the limit - CHI '99*. doi:10.1145/302979.303118

Cantoni, V., Lombardi, L., Setti, A., Gyoshev, S., Karastoyanov, D., & Stoimenov, N. (2018, July). Art masterpieces accessibility for blind and visually impaired people. In *International Conference on Computers Helping People with Special Needs* (pp. 267-274). Springer, Cham.

Card, S. K., Moran, T.P., & Newell, A. (1983). *The psychology of human-computer interaction*. L. Erlbaum Associates.

Castle, H., & Dobbins, T. (2004). Tactile display technology. Technology And Innovation.

Chae, M., & Kim, J. (2004). Do size and structure matter to mobile users? An empirical study of the effects of screen size, information structure, and task complexity on user activities with standard web phones. *Behaviour & Information Technology, 23*(3), 165-181. doi:10.1080/01449290410001669923

Chen, M. L., & Wang, H. C. (2018, March). How personal experience and technical knowledge affect using conversational agents. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion* (pp. 1-2).

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975-979. doi:10.1121/1.1907229

Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., ... & Bandeira, N. (2017, September). " What can I help you with?" Infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 1-12).

Crispien, K., Würz, W., & Weber, G. (1994). Using spatial audio for the enhanced presentation of synthesised speech within screen-readers for blind computer users. *Computers for handicapped persons lecture notes in computer science,* 144-153. doi:10.1007/3-540-58476-5_117

Davidson, T., Ryu, Y.J., Brecknell, B., Loeb, R., & Sanderson, P. (2019). The impact of concurrent linguistic tasks on participants' identification of spearcons. *Applied Ergonomics.*, *81*. https://doi.org/10.1016/j.apergo.2019.102895

Di Blas, N., Paolini, P., & Speroni, M. (2004, June). Usable accessibility" to the Web for blind users. In *Proceedings of 8th ERCIM Workshop: User Interfaces for All, Vienna*.

Dingler, T., Lindsay, J., Walker, B.N., (2008). Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. *Proceedings of the 14th international conference on auditory display*.

Dray, S. M., Ogden, W. G., & Vestewig, R. E. (1981). Measuring performance with a menu-selection human-computer interface. *Proceedings of the Human Factors Society annual meeting, 25*(1), 746-748. doi:10.1177/107118138102500119 4

Efthymiou, C., & Halvey, M. (2016, November). Evaluating the social acceptability of voice based smartwatch search. In *Asia Information Retrieval Symposium* (pp. 267-278). Springer, Cham.

Espinosa, M. A., Ungar, S., Ochaíta, E., Blades, M., & Spencer, C. (1998). Comparing methods for introducing blind and visually impaired people to unfamiliar urban environments. *Journal of environmental psychology*, *18*(3), 277-287.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*(6), 381–391. doi: 10.1037/h0055392

Frauenberger, C., Putz, V., Holdrich, R. (2004). Spatial auditory displays – a study on the use of virtual audio environments as interfaces for users with visual disabilities. *In DAFx04 proceedings*, 5-8.

Gaver, W. (1989). The SonicFinder: An interface that uses auditory icons. *Human-Computer Interaction, 4*(1), 67-94. doi:10.1207/s15327051hci0401_3

Geven, A., Sefelin, R., & Tscheligi, M. (2006). Depth and breadth away from the desktop – the optimal information hierarchy for mobile use. *Proceedings of the 8th conference*

on human-computer interaction with mobile devices and services - MobileHCI '06. doi:10.1145/1152215.1152248

Goodhue, D. (1986). IS attitudes: Toward theoretical and definition clarity. *ICIS 1986 Proceedings*. 26. https://aisel.aisnet.org/icis1986/26

Goodwin, N. C. (1987). Functionality and usability. *Communications of the ACM*, *30*(3), 229-233.

Goose, S., & Möller, C. (1999). A 3D audio only interactive web browser. *Proceedings of the seventh ACM international conference on multimedia (part 1) - Multimedia '99*. doi:10.1145/319463.319649

Harper, S., Bechhofer, S., & Lunn, D. (2006, October). Taming the inaccessible web. In *Proceedings of the 24th annual ACM international conference on Design of communication* (pp. 64-69).

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26. https://doi.org/10.1080/17470215208416600

Hochheiser, H., & Lazar, J. (2010). Revisiting breadth vs. depth in menu structures for blind users of screen readers. *Interacting with Computers*, *22*(5), 389-398.

Hornof, A. J., & Kieras, D. E. (1997). Cognitive modeling reveals menu search is both random and systematic. *Proceedings of the SIGCHI conference on human factors in computing systems - CHI '97*. doi:10.1145/258549.258621

Hutchins, E.L., Hollan, J.D., & Norman, D.A. (1986). Direct manipulation interfaces. In D.A. Norman & S.W. Draper (Eds.), *User centered system design: New perspectives on human-computer interaction* (pp. 87-124). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology, 45*(3), 188-196. doi:10.1037/h0056940

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human–Computer Interaction, 12*(4), 391-438. doi:10.1207/s15327051hci1204_4

Kiger, J. I. (1984). The depth/breadth trade-off in the design of menu-driven user interfaces. *International Journal of Man-Machine Studies, 20*(2), 201-213. doi:10.1016/s0020-7373(84)80018-8

Klein, D., Myhill, W., Hansen, L., Asby, G., Michaelson, S., & Blanck, P. (2003). Electronic doors to education: Study of high school website accessibility in Iowa. *Behavioral sciences & the law*, *21*(1), 27-49.

Landauer, T. K., & Nachbar, D. W. (1985). Selection from alphabetic and numeric menu trees using a touch screen: breadth, depth, and width. *ACM SIGCHI Bulletin*, *16*(4), 73-78.

Larson, K., & Czerwinski, M. (1998, January). Web page design: Implications of memory, structure and scent for information retrieval. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 25-32).

Lazar, J., Allen, A., Kleinman, J., & Malarkey, C. (2007). What frustrates screen reader users on the web: A study of 100 blind users. *International Journal of human-computer interaction*, *22*(3), 247-269.

Leuthold, S., Bargas-Avila, J. A., & Opwis, K. (2008). Beyond web content accessibility guidelines: Design of enhanced text user interfaces for blind internet users. *International Journal of Human-Computer Studies*, *66*(4), 257-270.

Lewis, J. R. (2018). The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, *34*(7), 577-590.

Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., ... & Martinez, A. (2019). Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science*, *51*(4), 984-997.

Lorho, G., Marila, J., Hiipakka, J., (2001). Feasibility of multiple non-speech sounds presentation using headphones. *Proceedings of the international conference on auditory display*, 32-37.

Lorho, G., Hiipakka, J., & Marila, J. (2002). Structured menu presentation using spatial sound separation. *Human computer interaction with mobile devices lecture notes in computer science,* 419-424. doi:10.1007/3-540-45756-9_51

Luger, E., & Sellen, A. (2016, May). "Like having a really bad PA" The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5286-5297).

Miller, D. P. (1981). The depth/breadth tradeoff in hierarchical computer menus. *Proceedings of the Human Factors Society annual meeting, 25*(1), 296-300. doi:10.1177/107118138102500179

National Federation of the Blind (2009). *The braille literacy crisis in America.* https://www.nfb.org/images/nfb/documents/pdf/braille_literacy_report_web.pdf

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic books.

Norman, K. (1992). The psychology of menu selection: Designing cognitive control of the human/computer interface. *Displays, 13*(4), 206. doi:10.1016/0141-9382(92)90066-z

Paterson, E., Sanderson, P.M., Salisbury, I.S., Burgmann, F.P., Mohamed, I., Loeb, R.G., & Paterson, N.A.B. (2020). Evaluation of an enhanced pulse oximeter auditory display: a simulator study. *British Journal of Anaesthesia : BJA.* https://doi.org/10.1016/j.bja.2020.05.038

Pradhan, A., Mehta, K., & Findlater, L. (2018, April). " Accessibility Came by Accident" Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Sauro, J., PhD. (2018, September 19). 5 ways to interpret a SUS score. Retrieved March 04, 2021, from https://measuringu.com/interpret-sus-score/

Screen reader user survey #8 results. (2019, September 27). Retrieved March 03, 2021, from https://webaim.org/projects/screenreadersurvey8/#usage

Shneiderman, B. (1986). Designing menu selection systems. *Journal of the American Society for Information Science, 37*(2), 57-70. doi:10.1002/(sici)1097-4571(198603)37:23.0.co;2-s

Snowberry, K., Parkinson, S. R., & Sisson, N. (1982). Computer display menus. *Ergonomics, 26*(7), 699-712. doi:10.1080/00140138308963390

Sodnik, J., Jakus, G., & Tomažič, S. (2012). The use of spatialized speech in auditory interfaces for computer users who are visually impaired. *Journal of Visual Impairment & Blindness, 106*(10), 634-645. doi:10.1177/0145482x1210601007

Sumikawa, D.A. (1985). Guidelines for the integration of audio cues into computer user interfaces. United States.

The WebAIM million annual accessibility analysis of the top 1,000,000 home pages. (2020, March 30). Retrieved February 20, 2021, from https://webaim.org/projects/million/

The WebAIM millionth 2022 report on the accessibility of the top 1,000,000 home pages. (2022, March 31). Retrieved July 27, 2022, from https://webaim.org/projects/million/

Theofanos, M. F., & Redish, J. (2003). Bridging the gap: between accessibility and usability. *interactions*, *10*(6), 36-51.

Thinus-Blanc, C., & Gaunet, F. (1997). Representation of space in blind persons: vision as a spatial sense?. *Psychological bulletin*, *121*(1), 20.

Vtyurina, A., Fourney, A., Morris, M. R., Findlater, L., & White, R. W. (2019, May). Bridging screen readers and voice assistants for enhanced eyes-free web search. In *The world wide web conference* (pp. 3590-3594).

Walker, B.N., Nance, A., & Lindsay, J. (2006). Spearcons: Speech-based earcons improve navigation performance in auditory menus. *Proceedings of the 12th international conference on auditory display*.

Watson, M., & Sanderson, P. M. (2001). Intelligibility of sonifications for respiratory monitoring in anesthesia. *Proceedings of the Human Factors and Ergonomics Society annual meeting, 2*, 1293. Retrieved from https://uidaho.idm.oclc.org/login?url=https://search-proquest-com.uidaho.idm.oclc.org/docview/235462286?accountid=14551

Wickens, C. D. (1991). Processing resources and attention. Multiple-task performance, 1991, 3-34.

Wickens, C. D., Hollands, J.G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (Fourth edition.) (pp. 329-330). Pearson.

Yoon, K., Dols, R., Hulscher, L., & Newberry, T. (2016). An exploratory study of library website accessibility for visually impaired users. Library & Information Science Research, 38(3), 250-258.

Zhang, Z., Basili, V., & Shneiderman, B. (1999). Perspective-based usability inspection: An empirical validation of efficacy. *Empirical Software Engineering*, *4*(1), 43-69.

Appendix A: System Usability Scale

All questions scored on a scale from 1 to 5, with 1 indicating “strongly disagree” and 5 indicating “strongly agree”.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Appendix B: 8^2 Menu Items

Utensils	States	Music	Geography	Presidents	Automobiles	Trees	Technology
Spatula	North Dakota	Rap	Mountain	Trump	Volvo	Cottonwood	Computer
Fork	Vermont	Blues	Ocean	Clinton	Honda	Spruce	Monitor
Knife	Texas	Funk	Jungle	Obama	Cadillac	Sycamore	Printer
Spoon	Louisiana	Disco	Crater	Biden	Chevy	Walnut	Television
Grater	Wisconsin	Jazz	Forest	Washington	Dodge	Mahogany	Headphones
Colander	Nebraska	Punk	Volcano	Jackson	Porsche	Chestnut	Router
Tongs	Florida	Pop	Glacier	Reagan	Toyota	Maple	Mouse
Whisk	Ohio	Soul	Island	Lincoln	Ferrari	Redwood	Keyboard

Appendix C: 4^3 Menu Items

Levels 1 and 2

Sports	Animals	Clothes	Countries
Football	Fish	Shoes	USA
Gymnastics	Reptiles	Shirts	Canada
Hockey	Birds	Pants	Japan
Baseball	Bugs	Hats	Italy

Sports

Football	Gymnastics	Hockey	Baseball
Touchdown	Beam	Powerplay	Bases
Sack	Floor	Puck	Pitcher
Quarterback	Vault	Icing	Outfield
Receiver	Bars	Goalie	Innings

Animals

Fish	Reptiles	Birds	Bugs
Pike	Snake	Woodpecker	Ant
Walleye	Turtle	Robin	Beatle
Trout	Crocodile	Sparrow	Caterpillar
Goldfish	Iguana	Penguin	Mosquito

Clothes

Shoes	Shirts	Pants	Hats
Sneakers	Tank Top	Jeans	Stetson
Heels	T-shirt	Leggings	Beanie
Sandals	Polo	Shorts	Fedora
Boots	Blouse	Slacks	Bowler

Countries

USA	Canada	Japan	Italy
Detroit	Toronto	Tokyo	Rome
Seattle	Vancouver	Yokohama	Milan
Chicago	Calgary	Osaka	Naples
Houston	Ontario	Nagoya	Pisa