5-2018

Usability of Sound-Driven User Interfaces

Zachary T. Roth
University of Arkansas, Fayetteville

Dale R. Thompson
University of Arkansas, Fayetteville

Usability of Sound-Driven User Interfaces

Zachary T. Roth
Department of Computer Science and Computer Engineering
University of Arkansas
ztjohnso@uark.edu

Dale R. Thompson
Department of Computer Science and Computer Engineering
University of Arkansas
drt@uark.edu

ABSTRACT

The model for interacting with computing devices remains primarily focused on visual design. However, sound has a unique set of advantages. In this work, an experiment was devised where participants were tasked with identifying elements in an audio-only computing environment. The interaction relied on mouse movement and button presses for navigation. Experiment trials consisted of variations in sound duration, volume, and distinctness according to both experiment progress and user behavior. Participant interactions with the system were tracked to examine the usability of the interface. Preliminary results indicated the majority of participants mastered every provided test, but the total time spent finding the solution varied highly between participants. Suggestions for expanding the investigation and conducting future work are provided.

DEFINITIONS

Auditory Icons	caricatures of naturally occurring sounds [1]
Earcons	short, abstract sounds used to convey information
Pitch	quality of a sound resulting from its frequency
Timbre	quality of a sound resulting from a combination of its various attributes and distinguishing it from other sound sources

1 Introduction

The significant advances in computational power over the past few decades have enabled increased access to these resources for the general public. These advances also have introduced new form factors such as handheld tablets and smartphones. While these form factors allow for computers to be more accessible in more areas of our daily lives, the model for interacting with these devices remains primarily focused on visual design. There are many reasons for this preference. Previous research has found that human visual perception has a greater data bandwidth than any other sense including hearing [2]. Peripheral vision also permits one to perceive multiple objects simultaneously for as long as desired, making it easy to convey large amounts of information.

However, sound has a unique set of advantages as well. While the human auditory sense is more ephemeral than sight, it is good at picking up relative differences in pitch. Related to this, previous work has shown the inclusion of reference sounds, or beacons, just before playing an earcon improved the ability for participants to accurately discern the pitch and duration of an earcon [3]. Sound can also be effective at quickly distinguishing contrasting situational contexts (i.e. the sound of a busy street is easily distinguished from a prepared speech). Existing consumer products such as desktop computers, mobile phones, and web browsers implement auditory interfaces through features such as text-to-speech to convert visual elements, but verbal explanations may require more time to convey information than a simple nonverbal tone. Existing research into earcons (abstract audio tones) and auditory icons (sounds representative of their real-world counterparts), both forms of nonverbal audio cues, over the past few decades offers an interesting alternative to relying on speech. The difficulty with speech and auditory icon-driven systems is their reliance on users relating the sounds with their real-world equivalents. If those using these systems do not have the necessary experience to relate to, the systems can become unintuitive. In contrast, the use of abstract sounds such as earcons does not have this usability hazard, and having an understanding of what mechanisms define usable systems is essential for developing computer interfaces in the future.

The rest of the paper is organized as follows: Section 2 introduces previous work on auditory interfaces such as earcons and auditory icons. The experimental setup is described in Section 3. Section 4 presents the preliminary results of the experiment. Future work is discussed in Section 5 and the conclusions are in Section 6.

2 PREVIOUS WORK

A strong motivator for exploring alternative computer interface designs in the field of Human-Computer Interaction (HCI) is the pursuit of more usable computer interactions. This is particularly important for instances where an individual cannot interact with a computer in a typical manner. This may arise due to a physical limitation imposed by a sensory deficiency (such as poor eyesight) or the context (such as a crowded public setting). Alternatively, limitations may also be imposed by an individual's previous experience. To interact with a computer effectively, the user must develop an understanding of both the functionality of the computer as well as how to access that functionality through the computer interface. Previous research has shown proficiency with a system is influenced by the amount of training received, dating back as far as 1897, when Bryan and Harter showed repeated training with Morse code significantly improved participant performance [4]. Today, examples of how the functionality and interface of a computer system is conveyed include observing other users of the system or referring to text and verbal explanations. The difficulty with explanations lies with how an individual will interpret identical instructions differently from his or her peers. To avoid this, other approaches have been taken. For example, appropriating design elements users are already familiar with, such as hierarchical menus and folders, may reduce the amount of instruction required. This technique has been used in previous research, such as by Brewster et al. where the paradigm of folders, files, and programs was used [5].

Another approach to training on a computer system is allowing the user to freely explore the functionality and interface. As the user spends time with the system and receives feedback from interactions, he or she develops a personal understanding more personal to his or her actual experiences. This understanding allows users to develop strategies for accomplishing tasks efficiently and/or effectively, including ways not foreseen by the system designers. Teo provides an overview of the general studies on exploration for computer systems in his dissertation

[6], including Rieman in 1996 [7], who found exploration to already be a common strategy for learning about unfamiliar computing environments when there is a specific goal. For visual interfaces, Teo suggests the use of models to predict how users will interact with the interface and inform the design of these interfaces. In future work, these concepts may be extended to improve the designs of audio-only interfaces as well. An in-depth study of using exploration as a training mechanism for audio-only user interfaces is not as well-developed as visual interfaces.

Research into whether auditory icons or earcons are more appropriate with audio-only interfaces is still ongoing, but the choice on which is more appropriate depends on the application. Auditory icons are interesting due to their imitation of sounds experienced in daily life. Because individuals naturally focus on the event causing a sound rather than the pitch and timbre of the sound itself [1], auditory icons can convey a complex amount of information quickly. Gaver et al. [1] found the sounds did not need to be a perfect representation of the original sound, but they did need to contain the original sound's important aspects. However, a drawback of this is the requirement for the listener to already be familiar with what would naturally cause the sound. If the listener is not, the sound will have no inherent meaning and may increase confusion. Earcons, on the other hand, do not rely on prior experience for understanding what the sound relates to outside of the computing environment. Investigations by Blattner et al. [8] led to their suggestions to design earcons using Western musical conventions such as key and rhythm due to their familiarity to listeners.

3 EXPERIMENTAL DETAILS

3.1 Experiment Design Motivations

The goal of the experiment was to test the usability of an abstract, audio-only user interface. The factors impacting the usability are complex; this experiment was designed to provide insight into these factors rather than a thorough analysis.

Earcons were chosen as the mechanism for providing feedback to the participant because they do not rely on previous life experiences as with auditory icons or a set speed of interaction as with speech. Additionally, because common operating systems such as Windows, Mac OS, and many variants of Linux primarily rely on visual interfaces, it is less likely for a participant to have used a non-speech auditory interface. This provides a good opportunity to observe the strategies participants develop as they explore an unfamiliar computing environment.

3.2 Experiment Setup

The experiment consisted of a participant sitting at a desk with only a mouse and a pair of headphones. The mouse was the only input device; right and left mouse movements moved left and right in the experiment, left mouse clicks selected the current element, and right mouse clicks repeated instructions. The headphones provided non-speech feedback in response to user actions and used text-to-speech functionality to provide instructions throughout the experiment.

The experiment was divided into four segments, and each segment asked the participant to locate a specific element from a randomly generated list of earcons. This list was divided into three or four categories such that all elements in a category were located consecutively in the overall list as seen in Figure 1. A random number of items was placed in each category and the relative size differences between the categories was held constant. Each category used a unique frequency randomly chosen from five frequencies evenly distributed between 280 Hz and 440 Hz. All elements in the same category used the same frequency. When the participant moved the mouse to move through the list, a 500-millisecond tone at that frequency would play to represent the element at the current position. Thus, the only information encoded in the elements was their category, and elements within a category were indistinguishable. Each experiment segment asked the participant to locate the left-most element of a category because doing so required the precision to identify both the correct category and correct element. Each segment consisted of three tests corresponding to list sizes of 30, 80, and 120 elements. If the element could not be found within four attempts, the software would begin skipping tests on each consecutive miss. Figure 1 depicts the arrangement of elements in each test, and Table 1 describes the parameters used to generate each test. The first and second experiment segments consisted of identifying the left-most element in the category with the most or least elements, respectively. The third and fourth experiment segments mirrored the first and second for both the tasks to accomplish and the element list was randomly generated, but a continuous background tone was introduced. This tone used the same frequency as the currently selected element. The volume of the tone was also dynamic, becoming quieter with slower mouse movements and louder with faster movements. The purpose of the tone was to see if it impacted the participants' abilities to correctly identify tones.

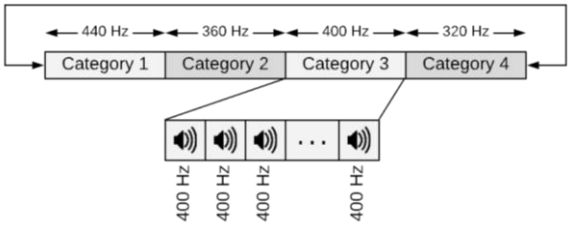

Figure 1: Example Arrangement of Generated Elements

Table 1: Parameters for Generating Elements

Segment	Test	Elements Generated	Category Sizes	Size Difference
1 Find Largest	1	30	4, 10, 16	6
	2	80	5, 15, 25, 35	10
	3	120	15, 25, 35, 45	10
2 Find Smallest	4	30	4, 10, 16	6
	5	80	5, 15, 25, 35	10
	6	120	15, 25, 35, 45	10
3 Find Largest (Tone)	7	30	4, 10, 16	6
	8	80	5, 15, 25, 35	10
	9	120	15, 25, 35, 45	10
4 Find Smallest (Tone)	10	30	4, 10, 16	6
	11	80	5, 15, 25, 35	10
	12	120	15, 25, 35, 45	10

As Table 1 describes, the generated lists consisted of between 30 and 120 elements. Participants were only told that pitches corresponded to

the category the element belonged to and that the list looped if the end was reached. Participants were not told how many elements were in a list or what the continuous background tone in experiment segments three and four signified. Throughout the experiment, the timing and results of participant actions were recorded. The software used synthesized speech to specify the element to find as well as to inform the participant about whether element selections were correct or incorrect.

Each participant completed two surveys. Before the experiment, he or she rated his or her perceived competency with common computing tasks as well as perceived ability to use a computer. Responses were formatted as a Likert scale. Questions regarding the degree of the participant's musical background and previous experience with non-visual user interfaces were also asked. These questions are listed in Appendix A.

After the experiment, the participants completed a survey consisting of a modified version of the NASA Task Load Index [9] to rate various workload demands experienced during the experiment, how these demands changed over the course of the experiment, and the usability of the experiment's design. Each category was rated from 1 (very low) to 5 (very high), and from the suggestion of research by Brewster [10], the annoyance category was also included. These questions are listed in Appendix B.

Participants in the experiment consisted of undergraduate students in the computer science and computer engineering programs on the University of Arkansas's Fayetteville campus.

3.2 Technical Implementation

To implement the software program that conducts the experiment, existing audio research programs were first considered. Programs such as NASA's SLAB Spatial Audio Renderer [11] had advanced features, were open to modification, and were intended for audio research, but they were found to have a steep learning curve or poor documentation. Most of these programs also required substantial modification to meet the needs and scope of the experiment. As a result, a custom software implementation was deemed to be more practical.

For the custom implementation, HTML 5 and the Angular JavaScript framework was found to be the best solution. The web platform allowed the development, testing, and experiment to take place on any computer on the network, and the ability to use a single browser across each computer greatly improved the software's portability and compatibility. Other advantages of this platform included the widespread availability of documentation, developer resources, and access to advanced feature implementations in modern browsers. The latter was especially important, as the ability to capture mouse input, implement tone oscillators, and include custom text-to-speech functionality were all native features of browsers and did not need to be designed by hand. The unforeseen difficulty in this approach was the lack of standardization on text-to-speech functionality; the behavior of these systems is dependent on features implemented by the browser and operating system. However, standardizing on an operating system and browser resolved these issues.

The resulting software served a webpage consisting of a blank canvas item. When the page loads, text-to-speech functionality guides the user to click on the canvas item, allowing for interacting with the system without a monitor. Once clicked, the canvas item captures the mouse and the experiment begins. At the end of the experiment, the recorded mouse movements and corresponding timestamps are encoded in a JSON format and displayed at the bottom of the page for later analysis.

4 PRELIMINARY RESULTS

Overall, eleven male undergraduate students between the ages of 18 and 24 participated in the experiment, and the speed and accuracy with which they accomplished the tasks were recorded. None of the participants had previous experience with audio interfaces outside of voice-dictation interfaces such as Apple's Siri or Amazon's Alexa, but five considered themselves to be musicians and ten had played an instrument regularly for over a year at some point in their life. The results of the participants including correctly identified elements, highest percentage of moves in a single direction, and time to complete all tests are shown in Table 2.

Table 2: Participant Results by Number Correct

Participant	Elements Identified	Total Moves in Single Direction	Total Time
1	100%	86%	5m 51s
2	100%	84%	7m 08s
3	100%	83%	7m 54s
4	100%	91%	8m 19s
5	100%	61%	9m 53s
6	100%	64%	11m 14s
7	100%	74%	18m 56s
8	58.3%	51%	8m 02s
9	25%	91%	6m 33s
10	8.3%	89%	5m 03s
11	0%	100%	3m 58s

Some interesting trends are immediately apparent. The first is the variance in total completion time for those who were able to locate every element successfully, calculated to be approximately three minutes and fifty-seven seconds. This is not completely unexpected as the experiment relied on participants developing their own strategy for locating elements. The slower times were likely due to counting the number of elements in each list while those with faster times could judge the relative time to scroll through each category. The participants who did well also developed a preference for moving through the list in one direction, and this was more prominent for those who completed the tests the fastest. This was possible because of how the list loops from one end to another. Moving in a consistent direction helps if the participant uses a strategy of timing the relative lengths of each category. Because the software decides to skip tests after too many failed attempts, the shorter completion times for the remaining participants is less significant.

The final interesting result from the data was participants tended to be divided into groups who did very well or very poorly. While seven out of the eleven participants managed to identify the correct element in every test, the remaining participants identified around half or less of the elements. It is suspected this is due to how the instructions were interpreted, the experiment design of skipping tests, the individual's chosen problem-solving strategy, or a combination of these causes. If the participant did not use an effective strategy in the beginning, he or she may not have had enough time to develop a better strategy before the experiment skipped to the next stage. A more detailed analysis of the participants' interactions with the system requires further study of the experiment data.

The responses to the survey after the tests was also interesting. Using the categories of the modified NASA Task Load Index to rate the demands of the experiment, all but one of the participants described the mental demand between somewhat high and very high. The majority of participants rated the time pressure experienced, effort required, and performance level achieved to be between very high to neither high nor low. The ratings for the frustration and annoyance experienced were mixed, landing between somewhat low and somewhat high with both

categories leaning toward somewhat low. The results of the post-experiment survey are in Appendix B.

A follow-up question was posed about how each of these demands changed over the course of the experiment. The majority of participants indicated the mental demand and effort required rose somewhat while the physical demand and time pressure experienced remained the same. The responses were split between whether the frustration and annoyance increased or decreased somewhat. However, the participants rated their performance level and ease of completion as increasing over the course of the experiment. This is most likely an effect of increased training, where participants develop a better understanding of how the software interacts and become more confident of their strategy as the experiment progresses.

Finally, participants were asked to rate the usability of the interface with regards to the intuitiveness and ease of learning of the system. The majority of the participants rated both to be between somewhat high and very high, and all but one participant rated both to be between very high and neither high nor low. Because more participants rated the system highly in these categories than those that mastered finding all of the items, this may imply the discrepancy is due more to the lack of training than the design of the system itself.

5 FUTURE WORK

5.1 Data Analysis

While the preliminary results of the experiment are interesting, a more detailed analysis of the participants' interactions with the system is required to identify other underlying trends in the data. Using logs the system collected from participant interactions, examining the difference in performance between including and omitting a constant reference tone or beacon, whether accuracy improved or diminished over the course of the experiment, the impact of frequency differences between generated categories, and the types of mouse movements is possible. These investigations will both add to the discussion of the system design and provide suggestions for further research in the area.

5.2 Experiment Refinements

The most apparent refinement to the experiment would be to increase the number and diversity of the participants. This would help mitigate the bias the selection of participants had on the results. This would also allow for the identification of statistically significant trends in how participants approach exploring an unfamiliar audio-only user interface.

Informal comments the participants shared both during and after the experiment suggested the provided instructions were not helpful for completing the experiment. While minimizing the instructions given was a key aspect of the experiment, improving the structure of the experiment could improve how the participants developed an understanding of how to use the system. To avoid long verbal instructions, including a dedicated training phase with very simple scenarios before the experiment could improve the results. This would ensure participants could understand how to interact with the system and know what the expectations were as well as help distinguish whether the difficulty or experiment methodology led to poor performance. A final improvement may be to include dedicated sounds to indicate when the edge of the list has been reached as it may help participants distinguish between new and repeated elements. Varying the sizes and frequency distinctions between categories in future work could determine a practical maximum possible speed for interacting with an audio-only interface.

5.3 Beyond the Experiment

There are several avenues to expand this area of study beyond this experiment. Because the preliminary results show the majority of the participants being able to navigate the audio-only interface effectively, it suggests simple, single-tone earcons can be used to navigate a long series of elements quickly. While the items within each category were not distinguishable, minimizing the complexity of the sounds may be one way to increase the speed one can navigate audio-only interfaces. This may be effective as an alternative navigation mode paired with more complex audio interfaces. To quickly assess a large set of elements such as files, data entries, or a webpage, the simple earcons could be used when complex ones are impractical or unwieldy. When a more detailed view is required, the audio mode could be switched to convey more complex information about each item.

While the majority of participants were able to complete every test successfully, it is not known if changing the volume of the interface in response to the speed of interaction improved or degraded participant performance. A study with more participants and a control group is suggested to determine this.

During the experiment, the sensitivity of mouse was constant without a way to adjust it. It is likely the sensitivity was perceived to be either too high or low and may have impacted the usability of the interface. Including a mechanism for the responsiveness of an interface to adapt to the participant's preferences could be a major improvement. Mouse acceleration, where faster mouse movements result in more distance traveled, was enabled for the experiment. This may be a factor to consider in future work as it can impact the participants' kinesthetic sense.

Finally, while this experiment relied on finding the left-most element of a category—an element located on the edge of a change in frequency, locating an element in a different location within the category was not investigated.

6 CONCLUSIONS

While this experiment was a small-scale study with several factors to consider, the preliminary findings suggest using simple earcons to represent elements in long lists is an effective form of audio-only navigation. The majority of the experiment participants were able to locate every element successfully in lists of between 30 and 120 items. Further investigations are suggested to focus on the impact of frequency and volume on participant performance.

7 ACKNOWLEDGMENTS

This research has been supported by a University of Arkansas Honors College research grant.

8 APPENDIX

Appendix A: Participant Responses to Pre-Experiment Survey

Computer Competency Responses

	No Level	Low Level	Average Level	Moderately High	High Level
Managing folders, files, and programs	0	0	1	2	8
Customizing a computer to my needs	0	0	1	5	5
Using word processors such as Microsoft Word	0	0	2	3	6
Using spreadsheet software such as Microsoft Excel	0	0	2	6	3
Using presentation software such as Microsoft PowerPoint	0	0	2	4	5
Using database software such as Microsoft Access	3	3	3	2	0
Managing email	0	0	3	5	3
Using a web browser	0	0	0	1	10
Web design	0	1	4	4	2
Software Development/ Programming in a language such as Java, C, C++, C#, Python, Ruby, etc.	0	1	1	4	5
Microsoft Windows Operating System	0	0	0	4	7
Apple MacOS Operating System	4	2	2	3	0
Linux Operating System	1	4	2	4	0

Perceived Computer Use Responses

	Never	Rarely	Sometimes	Frequently	Always
Frequency of computer use	0	0	0	1	10
Ability to use a computer effectively	0	0	0	1	10
Understanding of computer interface design	0	1	0	3	7
Understanding of what the computer is showing on the monitor	0	0	1	2	8
Understanding of how to accomplish my tasks	0	0	0	7	4
Understanding of how to interact with a computer	0	0	0	4	7
Frustration in typical computer use	2	4	5	0	0
I find computer interfaces difficult to use	1	8	2	0	0

Previous Experience Responses

Have you had experience with an audio-only interface (other than voice-dictation)?	
Yes	No
0	11
Do you consider yourself to be a musician?	
Yes	No
5	6

Do you or have you previously played a musical instrument regularly?				
No	Yes, <1 year in duration	Yes, 1-3 years in duration	Yes, 3-5 years in duration	Yes, >5 years in duration
1	0	5	0	5

If you played a musical instrument regularly, did you play in a group setting?	
Yes	No
8	3

How often do you use sound (notifications, feedback) when using a computer?				
Never	Rarely	Sometimes	Frequently	As Often As Possible
2	4	4	1	0

Appendix B: Participant Responses to Post-Experiment Survey

Responses on Overall Demands of the Experiment – Modified from NASA Task Load Index Categories [9]

	Very Low	Somewhat Low	Neither High nor Low	Somewhat High	Very High
Mental Demand	0	0	1	9	1
Physical Demand	5	3	1	2	0
Time Pressure Experienced	0	1	4	4	1
Effort Required	0	0	5	5	1
Performance level Achieved	0	3	5	1	1
Frustration Experienced	1	4	3	3	0
Annoyance Experienced	1	3	3	4	0

Responses on How the Demands of the Experiment Changed Compared to the Beginning – Modified from NASA Task Load Index Categories [9]

	Very Low	Somewhat Low	Neither High nor Low	Somewhat High	Very High
Mental Demand	2	2	0	7	0
Physical Demand	1	2	7	1	0
Time Pressure Experienced	0	3	6	2	0
Effort Required	1	2	2	6	0
Performance level Achieved	0	3	4	3	1
Frustration Experienced	1	3	3	4	0
Annoyance Experienced	1	4	2	4	0
Ease of Completion	0	3	2	6	0

Responses on Usability of the Experiment Design

	Very Low	Somewhat Low	Neither High nor Low	Somewhat High	Very High
Intuitiveness	0	0	3	7	1
Ease of Learning	0	1	2	5	3

REFERENCES

[1] Gaver, W. 1986. Auditory Icons: Using Sound in Computer Interfaces. Human-Computer Interaction. 2, 2 (1986), 167-177.

[2] Kokjer, K. 1987. The Information Capacity of the Human Fingertip. IEEE Transactions on Systems, Man, and Cybernetics. 17, 1 (1987), 100-102.

[3] Watson, Marcus & J. Gill, T. 2004. Earcon for intermittent information in monitoring environments.

[4] Bryan, W. and Harter, N. 1897. Studies in the physiology and psychology of the telegraphic language. Psychological Review. 4, 1 (1897), 27-53.

[5] Brewster, S., Wright, P. and Edwards, A. 1993. An evaluation of earcons for use in auditory human-computer interfaces. Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '93. (1993).

[6] Teo, L. 2011. Modeling Goal-Directed User Exploration in Human-Computer Interaction. Carnegie Mellon University.

[7] Rieman, J. 1996. A field study of exploratory learning strategies. ACM Transactions on Computer-Human Interaction. 3, 3 (1996), 189-218.

[8] Blattner, M., Sumikawa, D. and Greenberg, R. 1989. Earcons and Icons: Their Structure and Common Design Principles. Human-Computer Interaction. 4, 1 (1989), 11-44.

[9] NASA 1986. NASA Task Load Index (TLX) v. 1.0. NASA Ames Research Center.

[10] Brewster, S., Wright, P. and Edwards, A. 1995. The Application Of A Method For Integrating Non-Speech Audio Into Human-Computer Interfaces. (1995).

[11] Wenzel, E., Miller, J. and Abel, J. 2000. A software-based system for interactive spatial sound synthesis.