Accessible Spatial Audio Interfaces: A Pilot Study into Screen Readers with Concurrent Speech

Rishi Vanukuru
IDC School of Design
IIT Bombay
Mumbai 400076
India
rishivanukuru@iitb.ac.in

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
http://dx.doi.org/10.1145/3334480.3381440

Abstract
What if interfaces allowed visually challenged users to access more auditory information at a time? This graduate research project explores this question by studying Spatial Audio Interfaces in general, and the use of Concurrent Speech in particular. We present the process of designing an experimental study to measure user performance on web-based search tasks using concurrent speech screen readers, and to understand user preference and perception of more general spatial audio interfaces. The findings from a pilot run of the study, and their implications on future work in this project are discussed.

Author Keywords
Accessibility; Spatial Audio; Concurrent Speech

CCS Concepts
•Human-centered computing → Empirical studies in accessibility; *Auditory feedback;* **Accessibility design and evaluation methods;**

Introduction
The internet has become an indispensable part of modern life. However, internet-based content and services have largely been designed for visual access and consumption. People living with visual impairments access the internet using special forms of assistive technologies

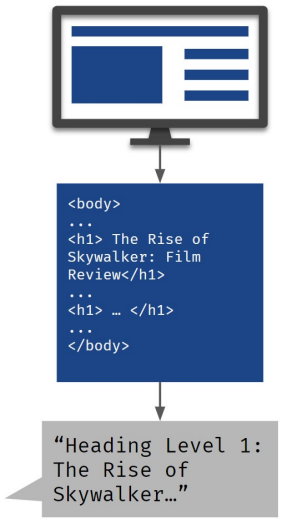

Figure 1: How screen readers use the DOM

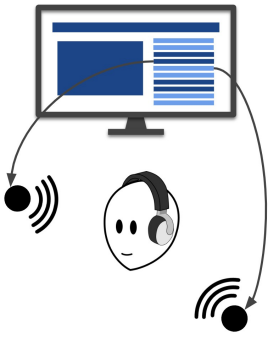

Figure 2: Listening to a list of headlines, two at a time

like screen magnifiers, screen readers, and physical Braille displays. Screen readers are widely used on Desktop and Mobile systems, and help users access content by reading screen-based information out loud. They do so by accessing the syntactic structure, or the Document Object Model (DOM) of websites (Figure 1). This means that they do not have access to the semantic structure of a website, as determined by the layouts and visual metaphors employed. Users attempt to make sense of web content through a single stream of synthesized audio, and this can be slow and cumbersome. Other issues stem from the content and websites that users often attempt to access. Visual content on the web can at times lack corresponding textual descriptions. The textual information that does exist might be interspersed with irrelevant content.

Two major approaches towards tackling these issues have been identified so far. The first deals with online content itself, such as new methods for defining and accessing alternative information (alt-text) for images on the web, and techniques to segment and present information more efficiently [8]. The second approach involves the use of sensory technology to create richer channels of information transfer. It is this approach that we are exploring through this project, by building upon past work done in the field of Spatial Audio Interfaces [3], and human perception of simultaneous or concurrent speech [5]. We aim to tackle the problem space by studying the effects of incremental additions to screen readers (such as Concurrent speech) on user performance, and by developing prototypes of more full-fledged spatial audio interfaces to understand user perception of the same. In the rest of the paper, we describe the process of designing an experimental study to achieve these aims, present the findings from a pilot study, and discuss the implications of the same on future work.

Background and Past Work

Spatial Audio Rendering involves the simulation of audio signals (when presented via headphones or earphones) to appear as if they were located in the physical space around the listener. Spatial sound setups have been used in a number of applications, including representing hierarchical menus [9], and generating sound-fields for audio icons and objects [7]. Most of these ideas take advantage of the spatial separation of sound to present simultaneous auditory information, and many examples involve the concept of Concurrent Speech. In particular, the idea of an Auditory Torch [6] relies on the simultaneous perception of different sound signals around a central focus. The ‘Cocktail Party Effect’ [2] explains how when surrounded by multiple simultaneous conversations and sound sources, people have the ability to focus on a single stream of audio that they deem to be important. Taking advantage of this effect, recent studies have demonstrated the human ability to both distinguish between Concurrent Speech Sources, and simultaneously make sense of them [5].

Despite the accessibility-related issues that exist today, users devise their own strategies to deal with inaccessible content or technological limitations [1]. Search and Browse tasks come up frequently while using the internet, and many users employ the strategy of Heading Navigation while doing so. This involves the use of specific shortcut keys to quickly go through the list of headings or links on a web page. This is useful when content cannot directly be accessed via keyword search, and also helps users gain an understanding of the overall layout of the page.

Positioning the Project

Recent studies by Guerreiro et al. [5] have shown that there is scope for incorporation of 2 to 3 concurrent channels of speech in screen reader applications. There is a need to

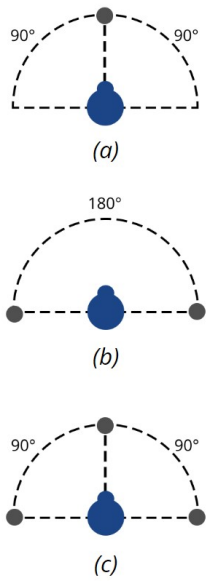

Figure 3: Speaker locations for the various configurations (from top to bottom, one speaker, two speaker and three speakers)

explore the integration of concurrent speech with screen readers [4]. In particular, concurrent speech could be used to support tasks that involve heading navigation. Building upon this, we aimed to design an experimental study to evaluate the effect of concurrent speech on list-based search tasks using screen readers. More full-fledged spatial audio interfaces such as the Auditory Torch also rely on the perception of concurrent speech. Spatial Sound also helps to understand the layout of information in such a scenario. In order to convey the possibilities that spatial audio might allow in the near future, and understand user perception of the same, we also decided to create a few simple interface prototypes that demonstrated these concepts.

Experiment Design

Research Question

What is the effect of (a) number of speakers and (b) voice differentiation on the time, errors committed, and preference, when performing a list searching or scanning task using a screen reader capable of providing concurrent speech information. In addition, we wanted to understand user acceptance and perception of spatial audio interfaces in general.

Variables

The primary independent variable considered in this study was the number of concurrent speakers. The study by Guerreiro [5] compared 2, 3 and 4 simultaneous talkers. The 4 talker condition proved to be too distracting for most participants. We therefore decided to consider the 2 speaker and 3 speaker conditions, and also compare these to the base condition of 1 speaker at a time. The spatial positioning of the various concurrent speech sources was done in a manner similar to Guerreiro [5] (Figure 3). In the 2 and 3 speaker conditions, we decided to study two levels of the Text-to-speech (TTS) voice characteristics of the speakers - same voice versus different voices. This resulted in 5 main configurations for the experiment: 1 Speaker, 2 Speaker Same Voice, 2 Speaker Different Voices, 3 Speaker Same Voices, 3 Speaker Different Voices.

Task

The use-case of searching for news on a website with a list of headlines is a direct extension of the task in the study by Guerreiro [5]. The task would involve presenting a list of headlines that may be scanned either individually, in pairs (two voices at a time), or in triplets (three voices at a time), in search of a goal item. We decided to have 12 headlines per list, in order to ensure that scanning would take place solely in those three conditions for all headlines.

Content

We began by collecting 400 headlines from popular Indian news websites at random over the span of a month. We curated a series of lists of 12 headlines, each of which contained exactly one pre-defined goal item. We ensured that this goal item was not of a topic similar to any of the other headlines in the same list. Each list was designed to have a balanced distribution of news headlines from various topics. The tasks were staged as follows. We would read out a descriptive prompt corresponding to a given list of headlines, and then ask the participants to navigate and scan through the list in search of the headline that was most related to the prompt. Once participants feel like they have found the relevant headline, they will be asked to speak it out loud, and indicate the characteristics of the speaker.

Apparatus

Generation of Headline Audio Clips:

Initially, we attempted to follow the speech synthesis procedure given in [10] and followed by [5] with an Indian-English female TTS voice from the Amazon Polly speech synthe-

Figure 4: Screenshot from the experiment application

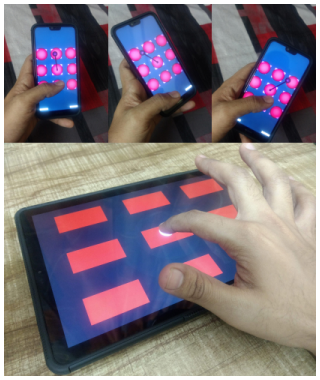

Figure 5: Spatial Audio Interface prototypes (gyroscope and touchscreen controlled)

Figure 6: The experimental setup. Participants sat to the right, and researchers to the left of this figure

sizer, in order to eliminate the effects of varying prosody and pronunciation across many speakers. However, the resultant voices seemed quite unnatural, perhaps because the calculations used in the original paper were for a male speaker in a different language and accent. We then used two other readily available TTS voices on Amazon Polly - one male voice and one female voice, both speaking American English. We assumed that many screen reader users in India are familiar with western accent TTS voices (which are usually the default option on screen readers). In the 1 speaker condition, we used the Indian English Female voice. In the 2 speaker condition, we used this voice along with the American English Male voice. The 3 speaker condition involved all 3 TTS voices. The resultant audio clips of the headlines were between 6 and 12 seconds long.

Developing the Experiment Application

The spatial sound rendering was implemented within the Unity3D Live Development Engine, using Google's Resonance Audio API to place sound sources virtually around the listener (Figure 4). The scripting for the same was done in C#. The same Unity program was used to accept user input through a standard mechanical keyboard. The program also provided experimenters with visual feedback about the state of each task in the experiment.

Interface Prototype Design

The Unity Live Development Engine was also used to create two mobile-based prototypes of an Auditory Torch, where users can control a cursor to explore a series of auditory icons (in this case news headlines), while also being able to hear information from the icons immediately around the one in focus. In the first prototype users controlled the cursor using a touchscreen, and in the second they could point and move their phone around the space directly in front of them. These prototypes are depicted in Figure 5.

Experiment Method

We aimed to conduct two within-subjects studies (with visually challenged and sighted participants) to compare task time, errors, number of repetitions per set of headlines, and preference of users. Each participant would go through all five screen reader configurations (1 speaker, 2 speaker same voice, 2 speaker different voice, 3 speaker same voice, 3 speaker different voice), with appropriate counterbalancing. We limited the number of tasks for each screen reader configuration to 2, with one additional training task, in order to keep each experiment session under 60 minutes. This results in 15 tasks per participant per session, 10 of which were timed and measured. We created 15 curated lists of headlines using the method described earlier in this paper. Task time, number of repetitions per group of headlines in a list, and task errors were to be measured. After the tasks are completed, we would ask users to rank the five configurations on perceived ease of use, and also rate each configuration on Effort using an ordinal scale. Subjective feedback on the various configurations and the interface prototypes would also be collected.

Pilot Study

Having defined the study protocol, we applied for ethical approval from our Institute's Ethics Review Board. We earlier conducted multiple pilots with sighted participants to refine the study protocol. Once ethical approval was obtained, we conducted a short pilot study to test the final protocol with 4 Visually Challenged and 4 Sighted Participants. The participants were recruited through open calls on social media, and their travel to and from the experiment venue was arranged for. Figure 6 shows the experimental setup used.

Findings and Discussion

Participants response to the idea of Concurrent speech was mixed. One Visually Challenged user was particularly in-

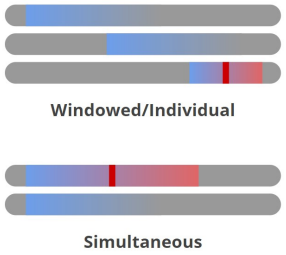

Figure 7: Representation of Listening Strategies - sequential focus shifting versus collective listening

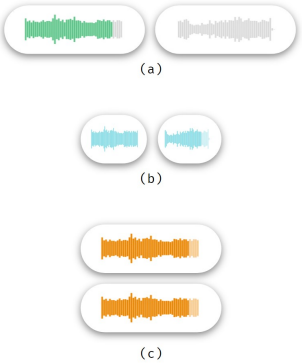

Figure 8: Representation of future work - comparing faster speech with concurrent speech

trigued by the idea, and performed progressively better as speakers were increased, while some others were left unconvinced. While quantitative data was collected for all 8 participants, the purpose of the pilot study was not to make claims on the basis of the data or responses, but rather to validate the experiment method and identify any issues with the same. The main findings from the pilot with Visually Challenged participants are classified into the following broad categories:

Listening Strategies
In multiple speaker conditions, some participants began by focusing on one particular speaker at a time. When they felt the content was irrelevant, they tuned out and focused on the next speaker. For some participants, this resulted in them listening to each headline individually, essentially reducing the multiple voice condition to a single voice one. Few participants stated that they did not focus on any given headline until they heard a particular word relevant to the prompt, at which point they focused in on that specific speaker to listen more carefully (Figure 7). These nuances in participant cognition were affected by the length of headlines, the position of the relevant words in a headline, as well as their ability to focus on and shift focus away from speakers.

Effect of Space
Most participants were able to clearly distinguish between headlines in the two speaker condition. However, participants found it particularly difficult to understand the forward source in the three speaker condition. This made it difficult to focus on the left and right speakers while attempting to listen to specific headlines.

Effect of Voices
The various TTS voices used were unfamiliar for most participants to varying extents. In the multiple speaker condition, preference affected their ability to change focus from one voice to the next. We realised that the TTS voice used by blind participants plays an integral part of their browsing experience, and while we attempted to control for voices across participants, this resulted in the experiment being perceived differently by each participant.

General comments and perception of Prototypes
Participants raised questions pertaining to the relevance and applicability of this research to other real-world problems. Some of these were: *"To get to something quickly, yes the voices help, but I'd be more confident with one screen reader and increasing the speed", "This may help with news headlines, but where else?", "... would require lot of effort in longer tasks", "This doesn't seem practical, what if I don't have earphones, or if they are broken"*.

Presenting the Auditory Torch prototypes helped convey the intention behind the experiment, and also provided a more believable use case for concurrent speech. All participants expressed a desire to test out the concept with more well-developed prototypes.

Next Steps

One way to address the apparent disconnect between concurrent speech and the task being studied, would be to compare it with the standard method of increasing the rate of speech of screen reader voices to perform tasks faster. This increase in speech rate is also often associated with loss in information transfer. We plan to conduct a study that compares concurrent speech with faster speech (Figure 8). Positioning Concurrent Speech as a feature of screen readers for specific tasks, much like how users can increase the speech rate, would potentially ground the technological component in an easily understandable real world context. We also plan to allow participants to choose a TTS voice

that they are comfortable with (from a given set of options), to reduce the effect of unfamiliarity with the voice.

We also plan to design more such full-fledged spatial audio prototypes, and approach the problem space from a Design-based-research perspective as well. Deploying real designs would allow for more relevant feedback to be captured, and perhaps make the significance and the contributions of this kind of experimental research more evident to Visually Challenged participants.

To the best of our knowledge, this project is among the first few studies to ground concurrent speech research for accessibility in a real-world task. We designed an experimental study, developed the apparatus to conduct the same, created simple interface prototypes, and conducted a pilot study. This study demonstrates the value of coupling the approach of incremental quantitative studies in HCI with more design-based research, in order to effectively communicate the value of research to stakeholders. This is an ongoing project, and work towards both the approaches mentioned above is currently underway.

Acknowledgements

I would like to thank all the participants in the study, Prof. Anirudha Joshi for guiding me through this project, Mr. Vikas Dabholkar, Dr. Charudatta Jadhav, and Prof. Sameer Patil for their advice along the way.

REFERENCES

[1] Yevgen Borodin, Jeffrey P Bigham, Glenn Dausch, and IV Ramakrishnan. 2010. More than meets the eye: a survey of screen-reader browsing strategies. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*. ACM, 13.

[2] E Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America* 25, 5 (1953), 975–979.

[3] Hilko Donker, Palle Klante, and Peter Gorny. 2002. The design of auditory user interfaces for blind users. In *Proceedings of the second Nordic conference on Human-computer interaction*. ACM, 149–156.

[4] João Guerreiro. 2016. Towards screen readers with concurrent speech: where to go next? *ACM SIGACCESS Accessibility and Computing* 115 (2016), 12–19.

[5] João Guerreiro and Daniel Gonçalves. 2016. Scanning for digital content: How blind and sighted people perceive concurrent speech. *ACM Transactions on Accessible Computing (TACCESS)* 8, 1 (2016), 2.

[6] Wilko Heuten, Niels Henze, and Susanne Boll. 2007. Interactive exploration of city maps with auditory torches. In *CHI'07 extended abstracts on Human factors in computing systems*. ACM, 1959–1964.

[7] Jörg Müller, Matthias Geier, Christina Dicke, and Sascha Spors. 2014. The boomRoom: mid-air direct interaction with virtual sound sources. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 247–256.

[8] IV Ramakrishnan, Vikas Ashok, and Syed Masum Billah. 2017. Non-visual Web Browsing: Beyond Web Accessibility. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 322–334.

[9] Jaka Sodnik, Grega Jakus, and Sašo Tomažič. 2011. Multiple spatial sounds in hierarchical menu navigation for visually impaired computer users. *International journal of human-computer studies* 69, 1-2 (2011), 100–112.

[10] Martin D Vestergaard, Nicholas RC Fyson, and Roy D Patterson. 2009. The interaction of vocal characteristics and audibility in the recognition of concurrent syllables. *The Journal of the Acoustical Society of America* 125, 2 (2009), 1114–1124.