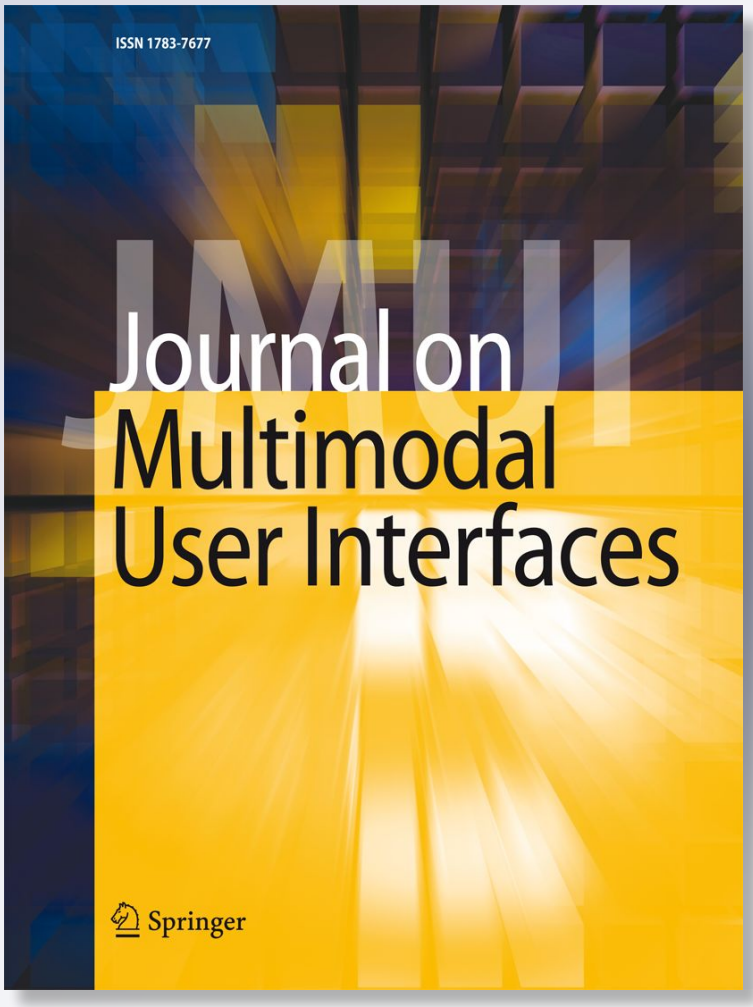
Eyes-free environmental awareness for navigation

Dalia El-Shimy, Florian Grond, Adriana Olmos & Jeremy R. Cooperstock

Journal on Multimodal User Interfaces

ISSN 1783-7677

J Multimodal User Interfaces
DOI 10.1007/s12193-011-0065-5



Springer

J Multimodal User Interfaces
DOI 10.1007/s12193-011-0065-5

ORIGINAL PAPER

Eyes-free environmental awareness for navigation

Dalia El-Shimy · Florian Grond · Adriana Olmos · Jeremy R. Cooperstock

Received: 17 January 2011 / Accepted: 13 August 2011

Abstract We consider the challenge of delivering location-based information through rich audio representations of the environment, and the associated opportunities that such an approach offers to support navigation tasks. This challenge is addressed by In-Situ Audio Services, or ISAS, a system intended primarily for use by the blind and visually impaired communities. It employs spatialized audio rendering to convey the relevant content, which may include information about the immediate surroundings, such as restaurants, cultural sites, public transportation locations, and other points of interest. Information is aggregated mostly from online data resources, converted using text-to-speech technology, and "displayed", either as speech or more abstract audio icons, through a location-aware mobile device or smartphone. This is suitable not only for the specific constraints of the target population, but is equally useful for general mobile users whose visual attention is otherwise occupied with navigation. We designed and conducted an experiment to evaluate two techniques for delivering spatialized audio content to users via interactive auditory maps: the shockwave mode and the radar mode. While neither mode proved to be significantly better than the other, subjects proved competent at navigating the maps using these rendering strategies, and reacted positively to the system, demonstrating that spatial audio can be an effective technique for conveying location-based information. The results of this experiment and its implications to our project are described here.

Keywords Sound spatialization · Auditory maps · Mobile applications · Blind and visually impaired community · Location-based information

D. El-Shimy (✉) · A. Olmos · J.R. Cooperstock
Centre for Intelligent Machines, McGill University, 3480 University Street, Montréal, Québec, H3A 2A7, Canada
e-mail: dalia@cim.mcgill.ca

A. Olmos
e-mail: aolmos@cim.mcgill.ca

J.R. Cooperstock
e-mail: jer@cim.mcgill.ca

F. Grond
Ambient Intelligence Group, Universität Bielefeld, Universitätsstr. 21-23, 33615 Bielefeld, Germany
e-mail: fgrond@techfak.uni-bielefeld.de

1 Introduction

Mobile and wearable computing devices have enjoyed growing popularity over the last decade, allowing users to access an unprecedented volume of content regardless of location. However, the development of applications for mobile devices poses new challenges, the most glaring of which is the limited screen real estate available. To remedy this problem, developers have for some time now focused on exploiting the multimodal capabilities of mobile devices, utilizing the haptic and auditory channels to supplement the visual one. For instance, when visual attention is occupied by other demands while navigating, auditory information can conveniently complement the screen display as a mechanism for information delivery. Moreover, for the visually impaired community, which relies heavily on ambient audio for navigation, a carefully designed auditory display can complement the audio cues provided by the physical environment without interfering or overloading the user's attention.

Our In-Situ Audio Services (ISAS) project has the potential to enhance the awareness of the blind and visually impaired to their environment, and potentially augment that of visually capable users. In comparison with other auditory display applications that are focused primarily on nav-

Published online: 23 November 2011

igation, ISAS is concerned more with providing a mechanism that supports active exploration of the user's environment. The design employs spatialized audio rendering to convey relevant location-based content to users, which may include information about their immediate surroundings, such as restaurants, cultural sites, public transportation and other points of interest. Unlike most screen readers and navigation aids, which provide sequences of spoken instructions or simply read out the names of locations and streets, ISAS renders several simultaneous sounds for users as they navigate through their environments. Not only can users listen to contents accessed from a GIS database through their mobile device, they can also create content themselves by tagging landmarks with corresponding sounds and adding them to the database.

Although the imperative for an eyes-free interface is obvious for those with visual deficiencies, who cannot, for example, navigate a map visually, the question underlying the design of our ISAS project is whether we can provide a sufficiently powerful representation of the environment using an auditory display. We note that our perceptual and cognitive systems can perceive and interpret many simultaneous audio signals when natural psychoacoustic mechanisms are able to separate each source. This is the case for a "spatialized auditory display", which generates similar perceptual cues of distance and orientation that we exploit in everyday activities to understand our environment. Capitalizing on the innate human ability to localize sound [1] may offer an effective substitute to the visual modality as a representation of the user's surroundings. To provide a concrete example, ISAS makes it possible for a user to hear sounds that sonically identify nearby locations, such as restaurants and theatres, as if they were "emanating" from the places to which they correspond. This informs users of their position and orientation relative to landmarks, and brings to their attention information about their immediate environment, which is otherwise available only through vision. Audio rendering techniques are used to provide appropriate distance cues; e.g., sounds gradually become louder as one approaches their locations, while other sounds become muffled as they drop behind and "out of view". Building on this concept, the experiment described in this paper involves a map exploration task in which the users *steer* their attention in specific directions of interest.

2 Previous work

Achieving the objectives described above, delivering the functionality in a usable manner, requires attention to several technologies and design issues. First, the design of auditory interfaces as a substitute for visual information, with an emphasis on ease of use, learning, and information density, must be considered. In our case, this entails particular attention to aspects of listening and interaction of blind users. Second, the vast abundance of geo-tagged material means that the auditory delivery of such information should be carefully designed in order to avoid overwhelming users with irrelevant content, which would reduce the effectiveness of the auditory display. Third, the use of location and orientation, as these affect the auditory display, is central to the task of exploring one's environment, and possibly, augmenting that environment with virtual content.

Our review of the literature considers each of these topics in turn.

2.1 Auditory interfaces

Auditory displays allow developers to accommodate a subset of users who have long had limited access to modern information technology: the visually impaired. Traditionally, the most widely available tools to assist visually impaired users in their interaction with computing devices have been braille input keyboards and screen readers, such as JAWS. The problem, however, is that such text-centric technologies are poorly suited to the display of spatially organized information, as in a map. In particular, using text alone, it is difficult to provide an overview of spatial relations that allows for effective exploration.

As Mountford and Gaver suggest [7], "Sound exists in time and over space, vision exists in space and over time". In other words, information delivered via sound tends to indicate changes over time, but can be picked up over a wide range of spatial locations. Visual displays, on the other hand, can only be perceived at specific locations in space, but are less transient. In addition, sound is omnidirectional and alerting, suggesting that it can be used to provide information over and above visual display and the limitations of visual attention [12].

Taking advantages of such principles, Frauenberger et al. [3] conducted an experiment in which a sample grocery shopping application was rendered in high-definition audio and presented to a mixed group of users: some visually impaired, and others sighted. The results showed that typical applications with the most common interaction tasks such as menus, text input and dialogs can be presented effectively using spatial audio, proving the claim of Walker et al. [13] that "Audio display space is not wed to the disappearing resource of screen space". Surprisingly, the authors found no significant differences in effectiveness between normal sighted users and visually impaired one.

Brewster [2] conducted a number of experiments to investigate the possibility of using structured nonspeech audio messages called "earcons" to provide navigational cues in a menu hierarchy. A hierarchy of 27 nodes was created, each

node associated with a particular earcon. Initially, the timbre, register and spatial location of each earcon was varied according to the position of its corresponding node within the hierarchy. After a training period, users were presented with various sounds and asked to identify their position within the hierarchy, achieving a recall rate of 81.5%. The author then used compound earcons: each node in the hierarchy was treated as a chapter, section or subsection, denoted by a set of numbers separated by dots as is common in book structures (for instance, Chapter 1, Section 1.1, Subsection 1.1.1). When a set of simple motifs was used to denote each of the numbers and the dot, recall rate increased to a remarkable 97%, proving that, with careful design, performance with auditory interfaces can match that of visual interfaces.

Yalla et al. [15] developed auditory menus to allow sighted users to multitask and, more importantly, make traditional visual menus accessible to a wider range of users that included the visually impaired. In particular, the authors investigated the design of an auditory scrollbar to navigate such menus. Pitch polarity was used to indicate the scrollbar's position, with tones increasing in pitch as the user scrolled up, and vice-versa. Test subjects, both sighted and visually impaired, had a favorable reaction to the interface, finding it both "helpful and informative". In particular, some test subjects were excited to recover the same information they used to receive while navigating menus before they had lost their vision.

In an attempt to improve usability of non-GUI interfaces, such as menus, for visually impaired users, Walker introduced speech-based earcons, or *spearcons* [14]. These are created by speeding up a spoken phrase until it is no longer recognized as speech. Their efficiency is mostly due to play speed, but they exhibit the further advantage of forming acoustically similar groups. For example, the spearcons for *Save* and *Save As*, are of different length, but they start with the same sounds. Similarly *metro station* and *bus station* would belong to one group due to acoustic similarity. For ISAS, we based the sound design on spatialized spoken words recorded by a human voice, although the latter may be replaced by modified text-to-speech. However, we are not yet employing compression or otherwise altered speech.

2.2 Location-based audio

Location-based audio dates back at least to the late 1960s when Schafer founded the *World Soundscape Project* [10], and started to collect the sounds of the environment. Today, as wireless data transmission and real-time position tracking become ubiquitous, we are seeing a growth in this domain of projects and initiatives that address multi-party locative audio collaborative applications. Examples include the COST Action on Sonic Interaction Design (www.cost-sid.org), which aims to explore new interactive technologies for auditory display, sonification, modelling, and sound/music computing, and the SAME project (www.sameproject.eu), which intends to "create new end-to-end systems for active, experience-centric, and context-aware active music listening".

A "geotag" is an association (tag) of a specific location to its related content stored in a GIS database. Geotagged audio databases are limited largely by the tedium of uploading new content, typically accomplished via browser-based submission, the limited appeal of audio-only playback, usually one sound file at a time, and the lack of general public interest in the available content. However, the use of mobile devices for recording and listening to audio, while actually *at* a specific location, breaks through many of these problems, and presents an opportunity to achieve more spontaneous, human-centered audio interactivity, as seen in systems such as the iPhone-based *GeoGraffiti* (www.geograffiti.com) and *Woices* (www.woices.com), the latter that enables user-contributed audio guides or walks, and *Ocarina* (ocarina.smule.com), one of Apple's "All-time top 20 apps".

An early example of a map exploration scenario for the blind was the Auditory Information Seeking Principle (AISP) [16], modeled on insights from visualization strategies. This supports functions of gist, navigate, filter, and details-on-demand. The authors propose that data sonification designs for exploration should conform to this principle. To improve access for the blind to geo-referenced statistical data, the authors developed two sonifications, an enhanced table and a spatial choropleth map. Their pilot study offered evidence that AISP conforms to people's information-seeking strategies, and demonstrated that in both designs, people can recognize geographic data distribution patterns on a real map with 51 geographic regions.

Another recent prototype of location-based audio that refers to the AISP is the real-time underground disruption map [8]. The authors of that system describe a conceptual strategy for providing an overview of disruptions in the London Underground based on what information is perceived as most crucial to the user. Positive feedback was reported from informal user-testing. As we integrate user- and location-specific information processing and filtering in later stages of ISAS development, we anticipate following the AISP framework as well.

Among the navigational aid tools designed to provide members of the blind and visually impaired community with location-based auditory information is Humanware's Trekker Breeze GPS device.[1] This speaks the names of streets and intersections based on the user's GPS location.

[1] www.humanware.com.

Typically, routes are entered beforehand, but users can additionally customize their device by tagging landmarks and places of interest along their paths.

Another GPS-based tool, the Intersection Explorer, was developed for mobile devices with touch screens running the Android operating system. This provides a virtual map to help blind users explore their neighborhood. As users move their fingers along the surface of their mobile device, the system speaks each street and its associated compass direction. Additionally, the presence of streets is cued by a slight vibration as one traces the circle.

2.3 Spatial sound for orientation

Humans have a well developed ability to pick up spatial and distance cues in sound. Within the blind and visually impaired community, where sound is the primary means of orientation, individuals make active use of these cues to navigate their environment. Examples include locating and orienting oneself through the specific reverberant atmosphere of a room, or identifying directions by the directed soundstream of urban traffic [9]. Various other sonic orientation aids are often exploited by the blind, including sounds from technological artifacts in our surroundings, such as the hard disk noise from a booting computer [11]. Some blind individuals master the art of echolocation [5], which involves sending out short impulses by clicking the tongue and interpreting the first reflections of nearby objects. Similarly, simply tapping the cane on the floor or listening to ones own footsteps gives valuable acoustic feedback about the environment and potential obstacles. Even if sound is not actively emitted, just turning the head for differential listening situates the blind and visually impaired person in a closed auditory action-perception loop.

There has been some work exploiting this natural disposition for echolocation and making it more accessible for untrained users. Morland et al. [6] designed a human "sonar system", which uses echolocation for this purpose. Ultrasonic clicks with a broad bandwidth are emitted from transmitters, reflected by surrounding objects, then recorded through microphones and mapped to the human audible range through heterodyning. A non-individualized HRTF-like transform is applied to give the user the sensation that the objects themselves are emitting the sound. Use of open ear headphones ensured that the device integrated natural sounds together with the synthetically generated ones.

While such systems may support untrained echo-locators, they are limited to conveying information related to the material configuration and properties of the environment. In contrast, our project seeks to exploit spatialized audio to convey richer information content that is typically, for non-blind users, extracted from the interpretation of visual information.

3 Design

ISAS is anticipated to provide three major modes of operation, "walking", "listening", and "tagging". In walking mode, users are presented passively with an occasional auditory display of their surroundings, taking advantage of GPS and compass data, as well as knowledge of category preferences, to select salient information.

Tagging mode supports the addition of new audio data by users, both for their own benefit and as a means of social networking with other users. The focus of the experiment described in this paper is on the "listening" mode. In this mode, users query their surroundings in a focused direction, pointing the mobile device in the direction of interest to hear what objects lie ahead. For this purpose, audio need not be spatialized, since the direction is chosen explicitly. However, spectral and reverberation effects can be employed to convey a sense of distance, e.g., the sounds of more distant objects resonate and are attenuated in their high frequencies.

The early phases of design were primarily concerned with the audio rendering strategy, as we considered this the central factor that was critical to the usability and success of our system. To this end, we included a target user in the design process, exposing him to early prototypes and soliciting frequent feedback on various aspects of the spatialized audio rendering techniques we were testing. Through this process, we learned first that blind users are typically highly accustomed to specific parameters of their preferred TTS system, such as the voice and speed of speech. To avoid a potential bias by selecting only one set of TTS parameters, which may be more familiar to some users than others, we decided to use recorded human speech for the evaluation of our prototype. Second, since an auditory display evolves in time as users explore their environment, blind users tend to become impatient if they have to wait too long to receive relevant information. Locations of interest must be repeated sufficiently often so that their information is updated according to the actual listening position, which can result in a dense display. Third, the relatively recent introduction of touch screen technology in consumer hardware has not yet led to widespread applications and interaction paradigms that are familiar to blind users. Thus, we introduce our audio touch interface to subjects through an instructive learning session. Finally, although bone conducting headphones were appreciated by our target user as an "ears free" listening device, these nevertheless present an initially unfamiliar listening experience, leading us to continue exploring alternatives.

With regard to the design of our rendering techniques, this feedback eventually narrowed our choices to two prototypes. To evaluate these techniques further and gain both qualitative and quantitative insight into potential use of our system, we then conducted a formal experiment with visually impaired users.

3.1 Interaction

Although in practice, the system will likely be used with bone conducting headphones, as described above, we conducted the experiment with closed headphones to reduce the impact of external sound sources on the results and avoid the need for user accustomization. Users held an iPod Touch in their non-dominant hand. Its surface was meant to represent a two-dimensional map containing ten randomly distributed places of interest: a bank, park, toilet, restaurant, pharmacy, metro, supermarket, sports centre, bus stop and school. Users could navigate through the map by moving the index finger of their dominant hand along the surface of the device. At any given point, spearcons describing nearby places of interest were played according to the audio rendering technique used.

3.2 Auditory display, design and implementation aspects

Since the goal of ISAS is to create a cross-platform mobile application for rendering spatialized sound, we considered OpenAL, FMOD and Pd-ZenGarden as potential candidate libraries. We settled on Pd-ZenGarden, a portable and embeddable stand-alone library that can interpret Pure Data patches. Implementing various audio effects in Pure Data offers the greatest level of control and flexibility, allowing us to adapt the sound rendering easily to our needs.

State of the art spatialization and externalization of sound sources through headphones is usually achieved by rendering sounds through head-related transfer functions (HRTF), ideally tailored to each individual user. The alternative of generic HRTFs, intended for the average human head, provides suboptimal spatial cues, often leading to front-back confusion, which was critical for us to avoid. Another important constraint was the computational limit imposed by the smart phones and possible battery drain when carrying out simultaneous polyphonic rendering of several sounds with different HRTF spatializations. To avoid these issues, we choose to use simple binaural rendering instead, thereby allowing the ISAS application to provide a reasonable quality auditory display for generic users. Although the resulting auditory cues that we generate may appear slightly unnatural, this is an acceptable trade-off to ensure that users are readily able to discriminate and interpret them.

Our binaural rendering strategy employs level differences and inter-aural time differences with an effective inter-aural distance of 0.63 ms as directional cues. Front-back discrimination is supported through the Pure Data bandpass filter object *bp~*, with the filter width $q = 5$, and the center frequency based on angular position of the sound source, the lowest one with a midinote of 68 according to the MIDI Tuning Standard (MTS) ($\sim$ 415 Hz) behind the listener and the highest in front with a midi note of 100 ($\sim$ 2637 Hz). The mapping of the source listener angle to the center frequency was according to the function $f_{MTS}(\alpha) = 68 + (100 - 68)\sqrt{\cos(\alpha/2)}$.

The iPod's multi touch display has an aspect ratio of 2:3 with a diagonal of 89 mm. The sound rendering provided the listener with two distance cues, a linear decay that reached -6 dB at a distance of 13.5 mm, and a simple reverberation that increased with distance. The reverberation consisted of one delay line of 120 ms and a feedback, whose gain started at 0.8 and decreased linearly with time, reaching 0 gain after 1.5 s in order to limit the reverberation time. The gain of the dry signal decreased linearly from 1 to 0 while the wet part increased, both reaching a gain of 0.5 at a distance of 9.0 mm. Nearby sound sources thus appeared louder and gave the impression of a small, almost anechoic room, whereas distant sounds were quieter, with a small but noticeable reverberation.

We developed two models for the sequential order in which sounds were rendered, as described below. In both cases, the rendering of the scene repeated continuously, both taking approximately 1 s for each cycle to trigger all surrounding sound sources. Once triggered, the recorded speech samples would play at a moderate pace from start to end.[2]

3.2.1 Shockwave mode

The shockwave mode was inspired by the datasonogram of Hermann et al. [4] with the rendering order determined by the radial distance of each sound source from the listener's position, starting from the nearby sounds, followed by the more distant ones. This design can be understood as a shockwave that triggers sonic responses from the surrounding objects. It offers the benefit that the order conveys an additional distance cue and has some conceptual similarities with principles of echolocation, in which distance is gauged by response time of early reflections. However, for a typical scene, there is a drawback of increased sound sources with distance as the circumference of the shockwave expands, as seen in Fig. 1. As the number of simultaneously rendered sources increases, the display in turn becomes less intelligible.

3.2.2 Radar mode

In this method, the rendering order is determined by the angular position of each sound source. To prioritize the rendering of nearby sounds, the space is divided into three zones according to radial distance, with the closest zone 0–10.9 mm "swept" at a frequency of 1.2 Hz, the middle

[2]Recorded audio samples from scene explorations using these rendering methods can be retrieved from http://isas.cim.mcgill.ca/.

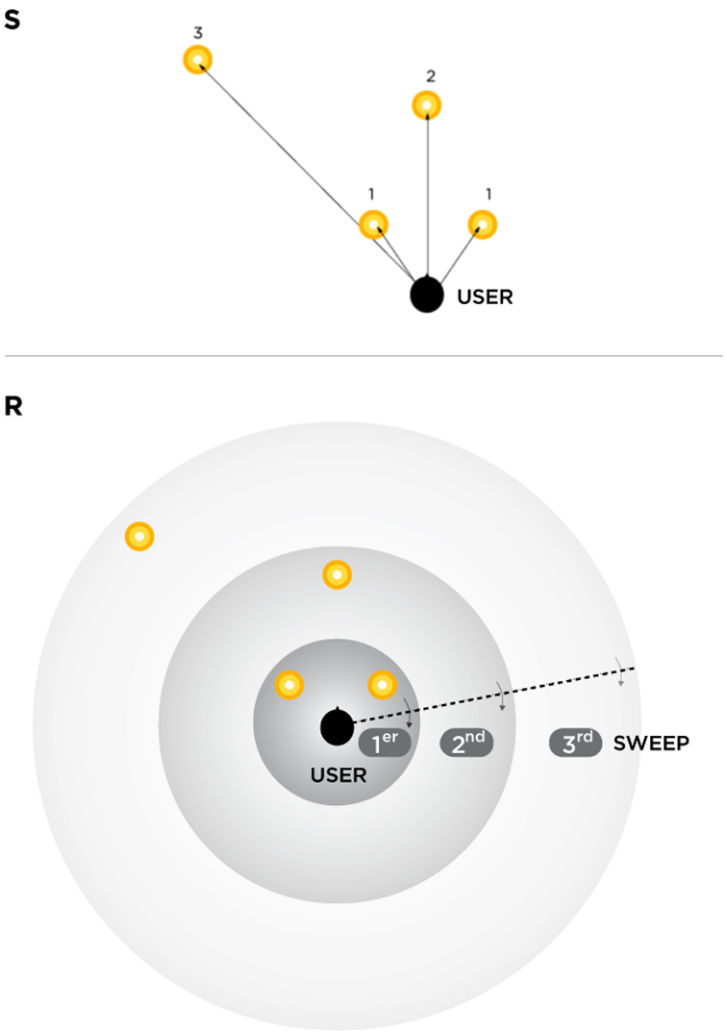


Fig. 1 Rendering modes. In the shockwave (S) mode the sounds are played according to the radial distance of each sound source from the listener's position, followed by the more distant ones (this is represented by the *numbers next to the circles*). In the radar (R) mode the space around the subject is divided into three concentric zones according to radial distance. Three circular sweeps are made one after the other. In the first sweep, only the sounds in the innermost ring are played. In the second sweep, all of the sounds in the inner and middle rings are played, and in the third sweep, all of the sounds are played. Thus, the closest sounds are played three times as often as the furthest sounds

one 10.9–21.7 mm at 0.8 Hz and the furthest >21.7 mm at 0.4 Hz, as illustrated in Fig. 1. Thus, further sound objects are heard less often, which helps avoid crowding of the auditory display.

4 Experiment

The goals of our experiment were, first, to evaluate the suitability of two audio rendering techniques, described above, for delivering spatial 2-D content through an eyes-free interface, and second, to expose our users to the fundamental concepts of our system, thereby, gaining some critical early feedback. The experiment thus involved testing in blocks to compare the two conditions, which were presented in balanced order across ten visually impaired participants. To avoid learning effects biasing the results, half of the participants were presented first with the shockwave mode first, and the other half first experienced the radar mode.

4.1 Experiment design

Each block consisted of two parts: a training session and the actual experimental session.

The purpose of the training session was to expose our users to the ISAS prototype, familiarize them with the audio rendering technique, and elicit some qualitative feedback. Participants were encouraged to take their time and express any questions or concerns they may have. In particular, we wanted to ensure each subject felt completely comfortable with the interaction technique and audio rendering mode before moving on to the experiment session. To achieve this objective, each subject was exposed to three very simple maps: one where all places of interest were laid out horizontally, one where all places of interest were laid out vertically, and one where all places of interest were laid out in two vertical but offset lines, creating a "zig-zag" pattern. Such configurations were useful in testing our subjects' ability to distinguish the left vs. right audio cues, as well as the front vs. back audio cues. Users were told in advance that each scene could be only one of those three configurations.

The formal experiment session consisted of ten trials, in which the participants were instructed to situate themselves between two sound objects in the map, e.g., pharmacy and park, using the interaction technique described earlier. This forced the participants to develop a simple mental model of the scene, rather than simply scanning the map for a single point of interest. Participants were asked to answer as accurately and as quickly as possible, but without any time restriction. At the end of each trial, participants were asked to rate (from 0 to 5) how confident they were with their answer. This rating, the path followed to solve the task, the completion time per trial, and distance between target and participants' final location were recorded. The experiment was concluded with a post-test questionnaire, exploring the overall preferred rendering mode to complete the task and ease of task completion. On average, the entire experiment and questionnaire took each participant one hour to complete.

4.2 Subject pool

Ten subjects, nine male and one female, were recruited to participate in the experiment. All were volunteers involved with the Institut Nazereth et Louis-Braille of Montreal. Subjects ranged from 19 to 69 years in age. Five of the subjects were legally blind and able to see some light, and the remaining five were totally blind. Five out of the ten subjects were congenitally blind. One subject was a converted left-hander, and the rest were all right-handed. The subjects were reimbursed for their transportation expenses but otherwise not offered any compensation for their participation in this experiment.

4.3 Quantitative results

The performance and responses of the participants in the two rendering modes were compared using paired t-tests. The distance between target and participant response, the confidence rating, as well as the total distance traveled by the finger were log-transformed to achieve normality. Completion time was compared using a Wilcoxon signed-rank test since this data could not be normalized by log transformation.

The two audio rendering modalities did not yield significant differences in the distances between target and participant response ($t = 0.505$, $df = 9$, $p = 0.6254$), confidence ratings ($t = 0.418$, $df = 9$, $p = 0.6686$) and task completion times ($T(N = 10) = 26$, $p = 0.9219$). This was also reflected in the overall answers at the end of the experiment (Table 1). However, the distance traveled by the participants' fingers was significantly shorter in the radar mode than in the shockwave mode ($t = 2.581$, $df = 9$, $p = 0.02964$), as seen in Fig. 2.

Table 1 Summary of the post-test questionnaire responses

Question	R	S	No answer
Overall preferred mode	4	1	5
Ease of task completion	2	3	5

4.4 Analysis of heat maps

Figure 3 presents heat maps for three scenes, selected as representative of different densities and distributions of the sounds. The heat maps are shown for each rendering mode, averaged over all subjects. These allow for a qualitative interpretation of how subjects made use of the auditory display to accomplish the given task. We can thus investigate whether the utility of a given auditory rendering depends on such variables as the number and density of the sound sources. During the experiment we observed that subjects occasionally tried to identify the two sounds of interest, and then positioned their finger at their estimate of the midpoint between them, without extensively verifying their final finger position by listening. We speculate that subjects sometimes used their listening skills for adjusting their position, and at other times, relied on their muscle memory of finger locations to solve the given task. The heat maps offer some evidence for such speculation.

The heat maps for Scenes 2 and 4, representative of many other cases we observed, demonstrate a better concentration of activity at the target in the shockwave mode than the radar mode. This is most evident in Scene 4, for which the target appears in a fairly dense area of activity for the former but not the latter mode.

Our impression was that Scene 2 (shown on the left, distance between locations 16.3 mm, both shown in zone 2) was particularly challenging for the subjects, regardless of rendering mode. This is evident from the strongly colored spot in the target area in the corresponding heat maps. We later determined that a source of difficulty in this scene was a distracting sound located close to the *sport centre* location. Exploration density was more concentrated between the locations with the shockwave mode, whereas in the radar rendering, subjects had a more pronounced tendency to position themselves around the *supermarket* sound. We hypothesize that the distracting sound prevented the subjects from recognizing the auditory display of the *sport centre*, resulting in their moving on to the *supermarket* location. From there, the radar mode would render the *sport centre* less often, hence exacerbating the difficulty of locating it.

Scene 4 was also challenging since the *bus stop* and *bank* locations were far from each other (distance between locations of 22.4 mm, both shown in zone 3, although the target point between the sound locations is in zone 2). Due to the

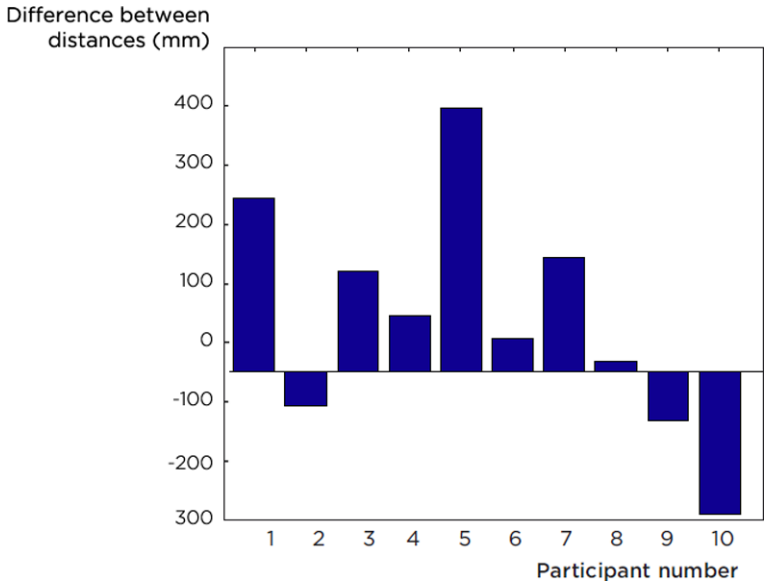


Fig. 2 Difference between the distance covered while exploring the scene under the two audio rendering modes

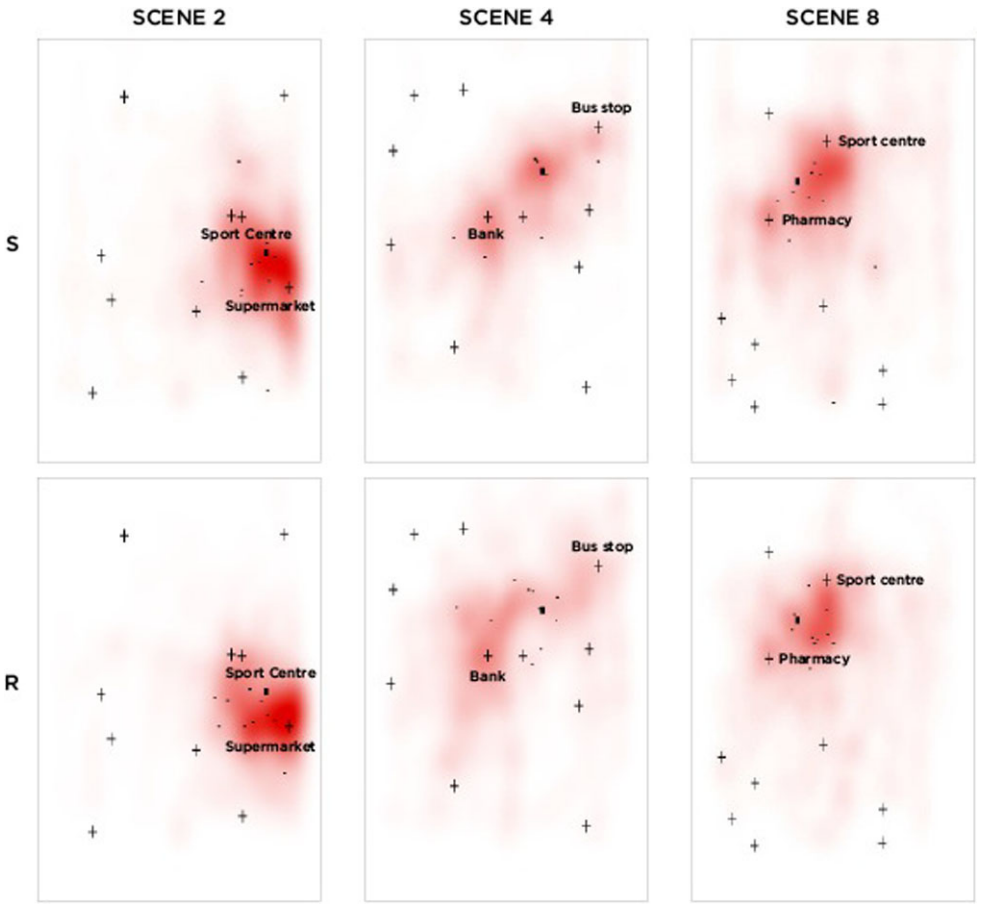


Fig. 3 Three different scenes explored with the two different rendering modes: shockwave (*top row*) and radar (*bottom row*). Each *rectangle* presents the entire scene, and its representative heat map indicates the regions (in *red*) most visited by the participants while solving the task. The *cross symbols* represent the locations or "sound objects". The *black square* indicates the target between the two locations specified by the task, for example the target "between the bank and bus stop" in Scene 4

distance of the *bus stop* and *bank* from the user, the radar rendered their sounds less often than the shockwave mode. In the latter, users performed more finger traveling back and forth from one of the sound locations as they explored for the second one or sought to confirm that they had found the approximate mid-point between the two, i.e., the target position. However, in the radar mode, once users found both sound locations, they seemed to require less travel to identify the target position. This behavior lends support to our speculation, above.

In Scene 8 (distance between locations of 18.7 mm, both shown in zone 2) we found that some subjects tried first to identify the sound locations precisely and then place themselves in between. However, most subjects explored the map with a bias towards the right of these locations. This can be explained by the bandpass filtering technique used to distinguish front and back. Due to the filter adjustment mentioned above, some subjects positioned themselves such that they heard the *pharmacy* sound slightly to their left and behind rather than straight behind. This made the sound appear less muffled, and in turn, the users seemed better able to locate the mid-point between the two locations. Again, in the radar mode, while starting the approach toward the *pharmacy* from the *sportcenter*, the former sound was not rendered frequently, whereas the rendering frequency was constant in the shockwave mode. We believe that these differences influenced the targeting strategy, motivating subjects to optimize their position to hear the sounds more clearly. In the radar mode exploration of Scene 8, deviation from the connecting line between the two locations of interest is apparent. Interestingly, this unexpected strategy, similar to that taken by users in Scene 2, appears to be more efficient in terms of finger travel than that used with the shockwave mode.

4.5 User feedback

We conducted three rounds of post-test questionnaires with all subjects: one after each of the two rendering techniques in order to gather some immediate impressions, and one at the end of the experiment to compare the two techniques and elicit some suggestions from subjects regarding how they would use the system in a real context.

Interestingly, four of the five users who expressed a preference between the two rendering techniques in terms of ease of task completion (Table 1) chose the last one presented, reporting that they had become "more used to the system", "had developed tricks to complete the task", and "felt more comfortable by that point". The remaining users who indicated no preference stated that they could not really perceive the difference between both modes. Only one user had pronounced difficulty using the system, while the rest had an overall positive reaction, noting that both rendering techniques were useful in helping them understand the layout of the maps and, thus, adequately complete the task at hand.

The participants in our experiment offered very interesting suggestions regarding how they would make use of ISAS themselves. We had deliberately chosen to not tell them exactly how ISAS was intended to be used, so as not to bias their impressions of the system. Given that seven out of our ten subjects owned a Humanware Trekker navigational device, it came as no surprise that many were tempted to compare ISAS to traditional GPS systems. One noted that, while he did not encounter the same level of "confusion" with the Trekker as he did with ISAS, the Trekker did not provide concurrent information about his surroundings the way ISAS did. Another subject observed that while GPS systems can provide location information, they do not provide any insight regarding whether the final destination is to the left, right, front or back of the user the way that ISAS does.

Several of the comments we received also validated our planned additions to the system, as described in Sect. 3. One noted that once she found a target on her two-dimensional map, it would be nice to be able to "click" on it to hear more information about this place of interest, as well as directions for how to reach it. Another subject found ISAS to be more interactive and interesting than tactile maps, saying that "it helps you visualize where things are, and could help you map out a route beforehand." Yet another raised the idea of filtering by category as a means of reducing the crowding of the maps. That same participant also suggested the addition of cardinal points and street names as places of interest, and noted that information about street direction, public transport and hazards would also be quite useful.

4.6 Discussion

The total distance traveled was significantly shorter for the radar mode, making it more "economical" overall, than the shockwave mode. However, qualitative analysis of the heat maps also suggests that auditory distance cues may be better conveyed by the shockwave mode, since distance is continuously mapped to time delay in the playback. Further experiments and longitudinal studies will have to be conducted to better understand how the different auditory rendering modes affect interaction quality. On the initial experiments, our quantitative results indicated no statistically significant difference in performance between the two rendering modes, both for completion time and error, i.e., distance between the participant's response and the actual target. Pairing this information with the fact that four out of the five users who indicated a preference for any mode chose the one presented second, neither mode seems convincingly superior to the other. Rather, user feedback suggests that either mode could successfully deliver clear, spatial audio content.

Thus, although it did not determine a "winning" rendering strategy, the experiment served to confirm that our sound spatialization approach provides effective auditory cues for blind users to explore a map through a small (mobile device) touch screen. For obvious reasons, this result is itself critical to the success of our project.

From our observations during the experiment, we also found that the traditional angular control of spatialization parameters such as equal power panning and interaural time differences can be problematic if the source listener relation is changing quickly because of a small lateral or longitudinal

movement in the immediate proximity of the source which causes a big angular change. This can lead to sound sources that appear to jump from left to right, or back to front, in response to a small movement of the finger. Alternative mappings of the source listener in relation to the spatialization parameters in the future might prove to be beneficial in helping users efficiently construct a mental model of the scene they are exploring.

Finally, several suggestions from our test subjects reinforce our confidence in the development plans we laid out for upcoming releases, confirming our overall vision for the project.

5 Future work

Further work remains before ISAS reaches the maturity of a ready-to-use application by the blind community. Our current efforts are divided into two streams. First, we are continuing development efforts on the iPhone, such that the full application can migrate to operation on a mobile device. This entails expansion of our database server from which the application draws for 2-D maps contents and associated tags, and to which users will eventually be able to contribute content of their own, such as new places of interest, and attach additional information (such as reviews) to existing locations.

In parallel, we are continuing to conduct user experiments to help understand the use context in which ISAS may be employed most advantageously and to refine various aspects of the system. Of immediate interest, we are exploring the use of more expressive gestural paradigms than the current "move your finger along the screen" technique, comparing the effectiveness of these options through carefully designed user tests. In addition, we are preparing an experiment to gain qualitative feedback regarding user preference for the amount of information conveyed through audio playback that defines each place of interest. For example, possibilities for conveying to the user that a Starbucks location is in their vicinity include playing the words "Coffee shop", "Starbucks coffee", and "Starbucks", among others. While detailed information is sometimes desirable, we have learned from discussion with members of the blind community that lengthy utterances are often annoying. This encourages the optimization of delivery of auditory information in the most compact form that retains its informational value, which may dictate the use of auditory icons in place of speech sounds.

Finally, it bears mention that the development of an eyes-free interface such as ISAS holds promise not only for the blind but also suggests applications for the wider community. In particular, such a system may be valuable for the large number of mobile situations—such as cycling, driving or sightseeing—in which users' visual attention should be focused on navigation, and for which it can be dangerous or distracting to interact with a computer display to obtain information. We hope to explore such uses cases as the project advances.

6 Conclusion

We introduced "In Situ Audio Services" (ISAS), a system that provides members of the blind and visually impaired community with a rich auditory representation of their surroundings. ISAS can operate in one of three modes "walking", "listening" and "tagging". As part of our development efforts for the walking mode, we designed a technique to represent the content of a 2-D map through spatial audio. To test our model, an experiment was conducted comparing two modes of sequential delivery of the spatial audio content: the shockwave mode and the radar mode. Ten blind or visually impaired users were asked to navigate through a map by moving their finger along the surface of an iPod Touch, and position themselves as accurately as possible between two target locations. While neither mode prevailed significantly over the other, users had an overall positive reaction to the system. They found the spatial audio content useful in helping them locate various places of interest on a map, and expressed a strong interest is using ISAS in the future to help them gain awareness of their surroundings. As we deploy ISAS as a stand-alone mobile device application in the near future, a series of planned experiments will help us refine other aspects of the system, such as the gestural interaction and the type of the auditory display describing the various places of interest. We hope that ISAS can become a tool that is beneficial not only to the blind and visually impaired community, but also sighted users whose visual channel may be constrained during navigation.

Acknowledgements The authors would like to thank our co-developers, Mike Wozniewski, Zack Settel, and Jeff Blum, whose efforts have been critical to the project and Thomas Hermann, whose input at various stages has proved most useful, and Tamar Tembeck for recoding the human speech sounds used in the prototype. The authors are deeply indebted to Stephane Doyon of Google Montreal, who has been an invaluable resource in commenting on our various rendering strategies and overall vision for the project, to Lucio D'Intino of the Montreal Association for the Blind, who has shared so much of his knowledge and experience in exploring his surroundings as a blind user, and to Marie-Chantal Wanet, Claire Trempe, and Carole Zabihaylo of the Institut Nazareth et Louis-Braille, our liaisons with the large pool of gracious participants in our experimental studies. This project is being supported by research funding from the Ministère des services gouvermentaux and Google Research.

References

1. Bregman AS (1994) Auditory scene analysis: The perceptual organization of sound. The MIT Press, Cambridge

2. Brewster SA (1998) Using nonspeech sounds to provide navigation cues. ACM Trans Comput-Hum Interact 5:224–259
3. Frauenberger C, Putz V, Höldrich R (2004) Spatial auditory displays—a study on the use of virtual audio environments as interfaces for users with visual disabilities. In: DAFx04 proceedings, Naples, Italy, October 5–8 2004. 7th Int. Conference on Digital Audio Effects (DAFx'04), 7th Int. Conference on Digital Audio Effects (DAFx'04)
4. Hermann T, Ritter H (1999) Listen to your data: Model-based sonification for data analysis. In: Lasker GE (ed) Advances in intelligent computing and multimedia systems, pp 189–194, Baden-Baden, Germany, 08 1999. Int Inst for Advanced Studies in System Research and Cybernetics
5. Kish D (2003) Sonic echolocation: A modern review and synthesis of the literature. http://www.worldaccessfortheblind.org/sites/default/files/echolocationreview.htm
6. Morland C, Mountain D (2008) Design of a sonar system for visually impaired humans. In: Proceedings of the 14th international conference on auditory display, Paris, France.
7. Mountford SJ, Gaver WW (1990) Talking and listening to computers. Addison-Wesley, Massachusetts
8. Nickerson LV, Stockman T, Thiebaut J (2007) Sonifying the ndon underground real-time-disruption map. In: Scavone GP (ed) Proceedings of the 13th international conference on auditory display (ICAD2007), pp 252–257, Montreal, Canada, 2007. Schulich School of Music, McGill University
9. Saerberg S (2010) Just go straight ahead" how blind and sighted pedestrians negotiate space. Senses Soc 5(3):364–381
10. Schafer RM (1977) The Tuning of the World, 1 edn. Random House, New York
11. Stockman T (2010) Listening to people, objects and interactions. In: Proceedings of ISon 2010, 3rd interactive sonification workshop. KTH Stockholm Sweden, April 2010
12. Tannen RS (1998) Breaking the sound barrier: designing auditory displays for global usability. In: 4th conference on human factors and the web, Basking Ridge, USA
13. Walker A, Brewster S (2000) Spatial audio in small screen device displays. Pers Ubiquitous Comput 4:144–154. doi:10.1007/BF01324121
14. Walker BN, Nance A, Lindsay J (2006) Spearcons: speech-based earcons improve navigation performance in auditory menus. In: Proceedings of the international conference on auditory display, pp 95–98
15. Yalla P, Walker BN (2008) Advanced auditory menus: design and evaluation of auditory scroll bars. In: Proceedings of the 10th international ACM SIGACCESS conference on computers and accessibility, Assets '08. ACM, New York, pp 105–112
16. Zhao H, Plaisant C, Schneiderman B, Duraiswami R (2004) Sonification of geo-referenced data for auditory information seeking: Design principle and pilot study. In: Barrass S, Vickers P (eds) Proceedings of the 10th international conference on auditory display (ICAD2004), Sydney, Australia International Community for Auditory Display (ICAD)