Using Virtual Environments to Prototype Auditory Navigation Displays

Bruce N. Walker and Jeffrey Lindsay

Georgia Institute of Technology

Author Contact Information:

For information regarding this article, please contact:
Bruce N. Walker, Ph. D.
School of Psychology
Georgia Institute of Technology
654 Cherry Street, Atlanta, GA, USA, 30332-0170.
Phone: (404) 894-8265 Fax: (404) 894-8905
Email: bruce.walker@psych.gatech.edu
Abstract

There is a critical need for navigation and orientation aids for the visually impaired. Developing such displays is difficult and time consuming due to the lack of design tools and guidelines, the inefficiency of trial-and-error design, and experimental participant safety concerns. We discuss using a virtual environment (VE) to help in the design, evaluation, and iterative refinement of an auditory navigation system. We address questions about the (real) interface that the VE version allows us to study. Examples include sound design, system behavior, and user interface design. Improved designs should result from a more systematic and scientific method of assistive technology development. We also point out some of the ongoing caveats that researchers in this field need to consider, especially relating to external validity and over-reliance on VE for design solutions.

Key words: Auditory display, blind navigation, sonification
Using Virtual Reality to Prototype Auditory Navigation Displays

There are approximately 11.4 million people with vision loss in the United States, 10% of whom have no usable vision; and by 2010 these numbers will nearly double (De l’Aune, 2002; Goodrich, 1997; National Center for Veteran Analysis and Statistics, 1994). As the population of the United States ages, there will continue to be more persons with visual impairments resulting from, for example, glaucoma, macular degeneration, and diabetic retinopathy. For a person with vision loss, the two fundamental tasks of navigating through a space and knowing what is around her can be a great challenge. Spatial orientation is the major mobility problem encountered by all individuals with profound vision loss (LaGrow & Weessies, 1994; Welsh & Blasch, 1997), but is especially difficult for people whose onset of vision loss occurs later in life (Levy & Gordon, 1988; Welsh & Blasch, 1997). This includes a growing sector of the aging workforce. Wayfinding (the ability to find one's way to a destination) is dependent on the ability to remain oriented in the environment in terms of the current location and heading, and the direction of a destination. Even highly experienced blind pedestrians exhibit random movement error large enough to occasionally veer into a wall or into a parallel street when crossing an intersection (Guth & LaDuke, 1995). These problems can be compounded when the person is indoors by the lack of external orienting cues such as the sound of traffic, noise from the flow of other pedestrians, or the chirping of birds in a particular tree. While there has been a great deal of research in the area of electronic travel aids for obstacle avoidance there has not been comparable research in the development of orientation devices that keep one apprised of both location and heading (Blasch, Wiener, & Welsh, 1997). Thus, there is a critical need for navigation and orientation aids for the visually impaired.
In addition to persons with vision loss, there are whole classes of persons who have normal vision but for whom navigation is still a challenge. This can be due to poor balance or other mobility difficulties, or to simply forgetting the route or destination, due to cognitive impairments. For these individuals, their visual attention is often dominated by other simple tasks, such as maintaining balance or attempting to identify landmarks.

It is therefore highly important to develop a system that communicates a range of information about the environment in a non-visual manner, to allow a person greater knowledge and enjoyment of, connection to, and more effective navigation through the space. An appropriately developed auditory display (or sonification) can enhance our ability to (1) keep track of our current location and heading as we move about, (2) find our way around and through a variety of environments, (3) successfully find and follow an optimally safe walking path to our destination, and (4) be aware of salient features of the environment.

Developing such displays is difficult and time consuming for a number of reasons. To start with, there are few tools available to create and evaluate auditory displays of any type, and, as Walker (2002) points out, there are virtually no guidelines for what they should be like, in any case. Much of the development time to date has been spent on inefficient trial and error design, which also requires access to many listeners who can test a display design and provide suitable feedback. Further, the hardware needed to make such an auditory navigation display feasible is currently bulky and expensive, which makes rapid testing difficult. Finally, any organization that might have the infrastructure and participant populations to be able to develop and test such an audio navigation system will have an Institutional Review board (IRB) with real and legitimate concerns about the safety of any system meant to guide people along the sidewalks of a city, without being thoroughly tested first in a more controlled environment. An increasingly effective
and viable alternative is to prototype the auditory display in a virtual environment (VE), and take advantage of the many benefits that approach brings. This is the type of VE research that Durlach et al. (2000) categorized as “Type 4C”, in which VEs are used to assist in the development of real-world performance aids. It is similar, but clearly distinct from using VEs to train a person to perform in the real world (“Type 4A”, according to Durlach et al., 2000).

As part of the development of the Georgia Tech System for Wearable Navigation (SWAN), we have encountered exactly the hurdles described above, and have, not surprisingly, begun to make use of VE to keep our primary project on track. We discuss here our very successful use of a VE to help in the design, evaluation, and iterative refinement of our auditory navigation system. In particular, we address several key questions about the (real) interface that the VR version allows us to study. We also point out some of the ongoing challenges that researchers in this field need to continue to consider.

The SWAN Audio Interface

The SWAN interface utilizes a repertoire of non-speech auditory icons and earcons within a specific framework to allow users to navigate successfully. The sounds in SWAN include navigation beacon sounds, object sounds, and surface transitions. These sounds are presented in a 3D audio environment, with each sound source being spatialized to seem as if it were located at the corresponding real-world location. For example, if the environmental feature the sound represents (e.g., a water fountain) is ahead and to the right of the user, the sound will appear to emanate from a location in front and to the right of the user. SWAN is able to spatialize these sounds by first determining where the user is located and then placing the sounds in relation to the desired destination and the surrounding environmental features it has detected. Beacon sounds are used to accomplish the primary wayfinding task, while the others are used to
convey knowledge about the features in the world and allow exploration of the immediate environment.

A complete path that a user might wish to travel is broken down into shorter, straight, unobstructed path segments, joined by waypoints. The beacon sounds are spatialized to emanate from the location of waypoints along the path the user is traveling. In order to move along a preset path the user listens for the beacon of the next waypoint, and simply walks toward its apparent location. Once the user reaches the waypoint indicated by the beacon, the sound shifts to represent the location of the next waypoint, the user reorients, and then sets off on the next path segment. Thus, a crucial element of the system is the ability of the user to localize the beacon sounds in the 3D audio space. Since SWAN uses generalized head-related transfer functions (HRTFs) to spatialize the sound, the more we can do to help the listener in this auditory localization task, the better.

The SWAN VR Sandbox

In order to study a variety of aspects of the SWAN interface we have developed a virtual reality-based prototyping environment. This allows us to implement and rapidly evaluate our sounds, menus, and interaction devices, in a safe and controlled lab environment before testing with the full SWAN system. Our VR environment was constructed using the Simple Virtual Environments (SVE) software package developed by the College of Computing at the Georgia Institute of Technology (GVU Virtual Environments Group, 1997). SVE is run on a Dell Optiplex PC running at 1.7Mhz, with 528 MB of RAM. The beacon sounds are played through headphones. To change direction participants rotate on the spot where they are standing, and use two buttons on a joystick to control forward and backward movement in the VE (they do not actually walk forward). Their orientation within the environment is tracked by an Intersense
InertiaCube 2 head-mounted tracking cube attached to the headphones. We have ensured that the auditory interaction, including physically orienting to 3D audio sounds, is identical in the VE environment and full SWAN systems. Other than the movement method, our testing with participants to date has revealed that users of both systems do not report any major differences in the experience.

**Key Questions**

The use of the virtual SWAN allows us to address several important questions in the development of the full SWAN system. The flexibility of the VE testbed allows us to be very thorough and exhaustive, testing a much greater number of sounds and configurations than would be possible without the VE.

**Sound Design**

Since sound localization is so crucial for the SWAN, the first question we addressed was what types of sounds would lead to the most effective navigation performance. Based on a number of human factors and perceptual issues, as well as the findings of Tran, Letowski, & Abouchacra (2000), we have decided to use non-speech beacon sounds (for more discussion on this issue, see Walker & Lindsay, under review) (see also Stokes, Wickens, & Kite, 1988). The question then becomes one of how to design the non-speech sounds. We investigated this by having listeners navigate through three separate maps in the virtual space, using one of three different beacon sounds. The VE in which these maps were located was essentially a large empty (virtual) room with four walls. No visual fidelity is required, since the participant has no visual display. This makes the use of a virtual testing environment even more compelling, since relatively little work is required to obtain a suitable simulation. The three maps differed simply
in the layout of the waypoints and in overall length (see Walker & Lindsay, 2003). The three beacon sounds were each one second long, with a center frequency of 1000 Hz and equal loudness. The sounds differed greatly in timbre, however. The first sound beacon was a burst of broadband noise centered on 1000 Hz. The second beacon was a pure sine wave with a frequency of 1000 Hz. The third beacon sound was a sonar pulse, similar to the sound that Tran et al. (2000) found to be one of the best sounds for use as a navigation beacon. Each participant navigated using the same sound throughout his or her three maps. At the start of a map the beacon sound played in an on-off pattern, where the sound was on for one second and off for one second. As the listener moved closer to the next waypoint the silence was shortened to effectively make the beacon tempo faster. Hence, increasing proximity to the waypoint was mapped to increasing tempo, which is consistent with our findings for population stereotypes or preferred mappings between proximity and tempo (Walker, in preparation).

The results of that study (Walker & Lindsay, 2003) showed a significant difference in performance for the different sounds, such that the broadband noise beacon led to the best performance, overall. This is not surprising, since a key to success in the navigation task is being able to localize the sound in space. The broadband noise should be most localizable, based on its multiple spectral cues. However, we also found that the best sound to use depended on the other details of the audio interface (discussed next). In some cases, the noise beacon was not the best performer; rather, a sonar-like ping was more effective. This finding highlighted the complexity of developing a novel interface, and pointed out the need for further experiments. Using the VE testing system we were able to be systematic and scientific, and include a full range of capture radius conditions to probe the details of the beacon sound results. We discuss these investigations, next.
**Capture Radius**

In addition to the sounds used to represent a waypoint, the particular interaction a listener has with the waypoint is important to consider. Each waypoint is specified by exact x, y, and z coordinates. However, the precise location of the user might never exactly “reach” the waypoint’s location. Consider the following analogy: A person starts at point A, walks down the sidewalk to point B on the corner, turns left, and continues along the sidewalk to point C. There is a penny on the sidewalk at the corner, indicating the exact place to turn (the waypoint, or point B). The typical user walks the first path segment, reaches the waypoint, turns the corner, and completes the path successfully. However, she might never actually step right on top of the penny at the waypoint, despite passing pretty much right over it. A computer system might say that she failed to traverse the path correctly, since she never technically arrived at the penny-sized point. A human observer would, on the other hand, say she was definitely “close enough” to each of the points. This points to the need for a *capture radius*. That is, there must be a radius around the waypoint that is considered close enough, so that the next beacon sound can appear, and the user can carry on down the next path segment. If the capture radius is too small, a person might *overshoot* the waypoint, walk past the corner, off the sidewalk, and into the street. If the capture radius is too large, the user may be told she has reached the turning point too soon, and as a result either cut across the grass or run into the corner of a building at the intersection. Thus, to keep the person on the intended path—neither missing the marks nor turning too soon—an optimal capture radius needs to be determined.

Rather than deciding on an optimal capture radius simply by trial and error with the real system, we were able to use the VE to go a step beyond and try out a number of capture radii in controlled conditions. We explored small (30 cm), medium (1.5 m), and large (15 m) values, and
allowed dozens of participants to move through paths, as described above. An example of the results is presented in Figure 1. Note how with the smallest radius (top panel), participants spend a considerable amount of time hunting for the waypoint. This would clearly be unacceptable in the real world. In the large radius condition (bottom panel), listeners often never reached the waypoint, and generally deviated from the prescribed path too much for practical navigation success. Participants using the medium waypoint (middle panel) showed successful navigation, neither straying too far from the path, nor spending much time trying to find the waypoint. The results discussed here are part of a much larger study (see Walker & Lindsay, under review). Without the VE, data collection in such a large study would have meant transcribing hours of videotape, estimating scores of movement paths, assessing inter-rater reliabilities, and so on. The VE testing environment allowed us to gather a considerable amount of precise quantitative data, thereby allowing for a careful statistical analysis of both speed and accuracy. Since navigational accuracy is our primary concern in the development of the SWAN system, we can use these results to determine that the medium capture radius was most appropriate, and based on that decision, we can consider that the sonar beacon may, in some cases, be better than the noise beacon. Thus, the scientific method, in conjunction with effective VE tools, allows us to perform iterative investigations to converge on a more sophisticated understanding of user interaction with the system we are developing.

Practice and Training Effects

The issue of how use of the system changes with practice is crucial to understanding the effectiveness and utility of the device. Though practice has not been specifically isolated and investigated yet, our findings to date show large improvements in performance by participants over a relatively small number of trials. Over the course of only three trials, during which the
virtual path navigated became progressively more complex and difficult, participants showed a remarkable degree of improvement in both rate and efficiency. In fact, the movement traces shown in Figure 1 represent only the second map for participants. That is, after only a few minutes of experience with the system users are comfortable and quite proficient, given the constraints of the interface. It is important also to consider that after the initial instructions on how to navigate using the auditory cues in the SWAN system, no more guidance was provided. Hence a small amount of initial instruction coupled with only three trials worth of practice (typically less than a total of a half hour of experience with the system) resulted in very large increases in proficiency by the participants. At the conclusion of those trials, performance was still improving, indicating continued practice could lead to even greater gains. Studies currently underway with the VE system are examining the asymptotes of proficiency.

Given the improvements noted in the VE system, it makes sense to use the VE to help train users of the outdoor system. The VE allows the task to be broken down into component parts, such as the analytic listening, sound localization, sound recognition, and movement components. This task-decomposition training has been used effectively in other auditory display training approaches (e.g., Smith, 2003; Smith & Walker, under review). However, this is a different kind of use for the VE, namely using the VE testbed to train a person to perform in the real world (“Type 4A”, according to Durlach et al., 2000). This is not specifically what the SWAN VE test system was developed for. All indications with participants and researchers who have used the two versions (virtual and real) lead us to believe that practice in the VE transfers to the outside. However, this remains to be systematically evaluated. In any VE, it is important to at least admit the possible limitations of this “Type 4A” activity. We are confident of the transfer, but must still exercise caution. Even if the whole task does not completely transfer (though we
view this as very unlikely), it is still clear that the specific listening skills gained in the VE (e.g., analytic versus holistic listening; localization of spatialized sounds; etc.) will transfer, since the tasks are identical in the two environments. The veridicality of the auditory display in both situations is crucial to this effective transition between environments. It is considerably more difficult to develop a visual interaction in a VE to the same level of realism as an auditory VE can be developed.

*Front-Back Confusions*

Despite the use of generally well-localizable sounds, and then within that class of sounds experimentally determining the best sounds, front-back confusions remain a potentially serious problem for an audio navigation interface. This typically arises from the slight discrepancies between the listener’s exact individual HRTF and the generalized HRTF employed by the system for practical reasons. For many purposes the generalized HRTF is sufficient, which is good because the effort involved in measuring an individualized function is considerable, and can in some cases outweigh the benefit of using the VE for testing and development. In any case, it is obvious that a person not knowing whether to move forwards or backwards is a significant issue for many reasons. In the original version of the SWAN, the only methods to resolve front-back confusions involved turning the head and listening for any spectral cues, or moving forwards or backwards and listening to the tempo change in the beacon. However, our early results suggested that a more formal method of resolving front-back confusions might improve performance. For example, the hunting behavior seen in the top panel of Figure 1 show that when walking past a waypoint, a certain amount of time passes before the listener determines the overshoot has occurred, and turns to correct the error. If such an overshoot could lead to stepping off the sidewalk and into the street, this could be a very serious problem (even though it is somewhat
rare in the medium capture radius condition). One such method currently being examined in the VE testbed to make an overshoot more salient, and to reduce front-back confusions in general, is the use of beacon sounds that change categorically based on whether they are in front of or behind the user. This is just one of many possible ways to reduce front-back confusions, and with the use of the VE it is much easier to quickly test multiple possibilities and determine what is the most useful disambiguation scheme. In recent studies, first with the VE and then in the real SWAN system, we have actually determined, somewhat to our surprise, that the disambiguation sounds tend to harm performance. It turns out that the listeners begin to attend too much to the shift in the sound from “in front” sounds to “behind” sounds, and end up not attending to the location of the sounds themselves, which is a more reliable cue. Without the VE studies, we may have assumed that the disambiguation sounds would help, and implemented them directly. Instead, we are now completing follow-on studies to determine if other forms of front-back disambiguation methods can be effective. It is important to understand whether it was the concept of disambiguation that was ineffective, or just the specific implementation. We suspect the latter, and the VE will allow us to explore a range of alternative options before implementing any successful approaches in the full SWAN.

Comparison of Sighted and Blind Users

Visually impaired individuals could obviously benefit from audio navigation systems such as the SWAN. However, there is prior evidence (Walker & Lane, 2001) that auditory interface characteristics may not be interpreted the same by sighted and visually impaired users. In that study, participants responded to sounds designed to represent data values such as temperature, velocity, or distance. There were many similarities, but some important differences in the way sighted and visually impaired listeners interpreted the sounds. Hence it is evidently
necessary to test auditory interfaces with members of both groups, and look for any differences in usage. A study is currently underway at the Atlanta VA hospital examining these questions, using the VE environment. This allows us to safely determine if and where there are differences in the way the interface needs to be designed, depending on the listener. The visually impaired participants are just as able to utilize the spatialized sounds to navigate, and in fact the most proficient participant we have had to date (across all studies, all groups) is visually impaired. Preliminary results seem to indicate that there are few if any differences in use between sighted and non-sighted listeners overall, but full conclusions will not be able to be made until more data has been collected and analyzed. Regardless, this step is necessary, and would not be possible without the VE. Further, we can take feedback from the visually impaired participants and make changes to the system on the fly, if necessary. It is interesting to note that the visually impaired listeners seem to have very little difficulty adapting to the paradigm of moving in a virtual environment that has no visual component. Anecdotal evidence thus far indicates that they might have an easier time adapting to such an interface since members of this group are already used to navigating with no or little visual information and using auditory cues as important sources of information. This has to be balanced, however, with the generally lower level of experience with virtual environments, computer games, 3D drawings, and so on, as compared to the sighted participants in our studies. Generalizations of the results always need to be very careful in such circumstances.

*Human-SWAN Interface: Input*

In addition to the auditory display components of the system, it is crucial that the user be able to input her intent and instructions to the system. It is not a trivial matter to develop a successful interface methodology for a system that is both mobile and intended for use without
vision. We can approach design decisions from the analytical perspective (e.g., using task analysis or needs analysis), or from the empirical perspective (i.e., employing the results of systematic studies). With the VE, we have the luxury of utilizing both methods.

Consider, for example, the decision of whether or not to implement speech recognition as a means of commanding and controlling the system. It is a complex decision process, involving a variety of opinions, perhaps split on the issue. Proponents would consider speech to be a natural, hands-free, eyes-free interaction method, highly appropriate to the task. On the other hand, due to the wear-anywhere design goal, task analysis would suggest that a speech interface (i.e., voice recognition) is not the most appropriate, because of privacy concerns, among others. To clarify, if the (visually impaired) user wished to withdraw some cash from the bank, speaking aloud a phrase like, “System, take me to the nearest ATM” while standing on the street corner could have serious personal security issues. Further, understanding the technical limitations of speech recognition technology “in the wild”, we would predict that the issues of background noise, varying sound levels, wind, and so on, make the speech recognition considerably more error-prone, which is unacceptable in an assistive technology of this sort. Thus, a priori we might have theoretically motivated reasons for avoiding speech recognition in this particular design.

We might wish to resolve this issue (speech or no speech) by testing it in the VE. We could implement a speech recognition interaction system, and see how well it performs. We did just this, and it can perform (technically) quite effectively with the virtual system, in the lab. However, in the end we decided against that approach for the real system. The reason for bringing this up is that the use of the VE to develop a system could lead a team to implement something like speech recognition without realizing that the actual usage context may be quite different from the test environment. Clearly, this discussion shares much with the discussion of
transfer, above. In that case, the recommendation was to make the actual listening task as identical as possible in both practice and live situations. In the present discussion, the key is to do two things: first, conduct a careful task analysis to determine the environmental context and situational factors that will impact on the system usage, and therefore, system design. In general, these considerations need to be given substantial weight in the decision process. Second, make sure that additional “environmental” sounds are present in the VE when appropriate. For example, we are able to add sounds in the virtual world that move and change as a real sound source would. The VE allows us to place arbitrarily many complex sounds into the world so that they become audible when the listener comes close enough to the source, or when some other condition is satisfied (including randomly timed sound occurrences, like virtual birds chirping). The VE handles all the associated calculations, attenuations, interaural difference cues, and so on. This approach preserves scientific control while introducing a considerable measure of ecological validity. This is in keeping with the growing trend toward more ecological psychoacoustic research in the development of auditory displays (see Walker & Kramer, 2004).

In addition to adding sounds in the VE itself, additional sounds may need to be added to the testing room, external to the VE. For example, recorded traffic noise can be played over loud speakers surrounding the participant. The reason for these external sounds is that the internal sounds in the VE, especially if played over headphones, will not impact on the speech detection capabilities of the system. The addition of external sounds within a testing room is an important example of where the VE is limited, and needs to be supplemented in order to maintain careful evaluation of the input capabilities of the system. In addition to the task-analytical methods already discussed, these “noise supplement” studies with the VE are the sorts of assessments that have led us to use manual (button-press) input methods, and to use speech messages (presented
privately to the user) only for output. Audio menus built in this manner have yielded effective and, perhaps most importantly, user-acceptable input with the SWAN system. The user selects the desired functionality from a menu structure, using buttons on a small wireless handheld device kept in a pocket or attached to a belt loop. The menus are presented aurally. The details of this system are discussed elsewhere, but the key here is that the need for such a system came about partially through testing alternative (speech) input methods, in an enhanced VE that incorporates “environmental” noises, and partially through task-analytic approaches that led to conclusions that the VE alone would never have produced. The virtual system alone may not have pointed to the need for privacy or the difficulties of speech recognition in the open usage environment. We should also point out that, once developed, the user interface can be tested entirely separately from the navigation interface, utilizing user centered design and usability testing methodologies commonly employed in the study of human-computer interaction (see, e.g., Wickens, Lee, Liu, & Gordon Becker, 2004). Such investigations do not require the use of the VE, though in practice it is often just as efficient to use the same equipment setup, especially in the case of an audio VE, where no complex and resource-limited CAVE or multi-projector display rigs are necessary.

Discussion

There is a clear need for assistive technology to help persons with vision loss continue to navigate and learn about the environment. The use of a virtual environment can be extremely useful in developing, evaluating, and iteratively improving technological solutions, such as the System for Wearable Audio Navigation (SWAN) being developed at Georgia Tech. The key is that the VE enables rapid, safe, and systematic studies in ways that are simply impossible in a trial-and-error design approach. Examples of the benefits include sound design for maximum
localizability, details of how the system interoperates with the user (e.g., capture radius), and an examination of practice effects. When novel or unexpected findings emerge, such as the interaction of capture radius and beacon sound in our system, the flexibility of the VE, and its utility for gathering data enable researchers to probe the details of how the system performs.

It is crucial to recognize, however, that becoming entirely reliant on the VE can lead to design flaws, and potential safety issues. We discussed here how the exclusive use of the simulator might lead a designer to implement, for example, a speech recognition system, simply because it is relatively simple to do so in the virtual system. It should be stressed that the full consideration of task needs have to be included in the design process. That is, the VE must never rise above the status of being just one of the many tools employed by the assistive technology developer.

On a related note, it cannot be over-emphasized that while there are many similarities between indoor VE test systems and outdoor full use devices, ultimate validation studies definitely have to be part of the development plan. Further, wherever possible an ecologically valid testing environment needs to be utilized. Having said that, however, it is also important to recognize that displays using certain modalities (notably auditory displays) are generally easier to duplicate exactly, or nearly so, in a VE test system. For this reason, with careful considerations always in mind, developers of auditory assistive technologies may feel quite comfortable in making the most of what is proving to be a very effective way to enhance the science in assistive technology design. Virtual environments, used in conjunction with wise traditional design methods such as task analysis and user centered design, play an important role in ensuring not only that assistive technology is available, but that it is also effective, usable, and acceptable.
References


Figure Captions

Figure 1. Navigation performance with the three different capture radius conditions, when using the sonar beacon. Note the hunting behavior with the smallest radius, and the erratic behavior with the largest radius.
Figure 1