Design and user evaluation of a spatial audio system for blind users

S H Kurniawan[1], A Sporka [2], V Nemec[2] and P Slavik[2]

[1]Department of Computation, UMIST
PO Box 88, Manchester M60 1QD, UK

[2]Department of Computer Science, Czech Technical University in Prague
Faculty of Electrical Engineering, Karlovo náměstí 13, Praha 2, 12135, CZECH REPUBLIC

s.kurniawan@co.umist.ac.uk, *sporkaa@fel.cvut.cz*, *nemec@fel.cvut.cz*, *slavik@fel.cvut.cz*

ABSTRACT

The paper reports on the design and evaluation of a spatial audio system that models the acoustic response of a closed environment with varying sizes and textures. To test the fit of the algorithms used, the system was evaluated by nine blind computer users in a controlled experiment using seven distinct sounds in three environments. The statistical analysis reveals that there was insignificant difference in user perception of room sizes between sounds in real and simulated scenes. This system can contribute to the area of VR systems used for training blind people to navigate in real environments.

1. INTRODUCTION

Virtual Reality (VR) has been an important and exciting field for many years. Recently, its potential for people with disabilities has picked up. VR systems have been applied in the areas of education, training, rehabilitation, communication and information technology for people with disabilities (Colwell, et al., 1998).

One important application of VR is to train blind users to navigate and move around in real environment, also known as the orientation and mobility (O&M) training (Inman and Loge, 1999). O&M training is important because it helps blind people develop the skills and techniques to overcome travel difficulties created by blindness and to maximise their ability to move around in different environments independently, safely and confidently (The Royal Blind School, 2003).

Conventional O&M training involves exposing trainees to various environments to train them to detect the sound variation caused by environmental factors, e.g., the floor textures, the room size, the location of the closest obstacle, etc., or instructing a blind trainee to approach a wall or an obstacle to show the sound variation caused by the presence of object, known as the obstacle perception training (Seki and Ito, 2003). This method is very time consuming and in may pose some danger to the trainees (e.g., when training them to cross a busy road). This is an area where VR may be beneficial. However, this also means that the acoustic system used for the training must be able to simulate the sound variation caused by the environments.

This paper reports on the design and evaluation of one component of the O&M training system for blind and visually impaired people: a spatial audio system that is capable of modelling the acoustic response of a closed environment with varying sizes and textures (e.g., a small carpeted room vs. a large marble hallway).

2. SPATIAL SOUNDS IN THE REAL AND VIRTUAL ENVIRONMENTS

2.1 Spatial Sound

Sound is a vibration of particles around their equilibrium positions. The vibration of the particles causes small local adiabatic variations of pressure in the medium, referred to as the *acoustic pressure*. Through the environment, these variations are propagated by means of waves of acoustic pressure.

In the real world, containing obstacles among the sound sources and receivers, only some of the sound wave may travel directly between the source and the receiver (*the direct sound*). Signal 1 in Figure 1 is an example of a direct sound. Except for the change of its intensity due to energy dissipation, and temporal

displacement due to the finite phase velocity of the sound waves, the shape of the signal of a direct sound is unchanged. Other parts of the original sound energy will be reflected or diffracted by the obstacles of the environment before reaching the receiver.

Combined from all contributions, the receiver obtains the acoustic response of the environment to the sound emitted from its source. As described later, the acoustic response may be understood as a compound of the early echoes and late reverberation. Figure 1 illustrates sound propagation in a closed environment. The acoustic response of Figure 1 can be represented as a diagram with the time of arrival (at the receiver) of all echoes on the X-axis and the intensity of sound on the Y-axis, known as the Impulse Response (IR) diagram.

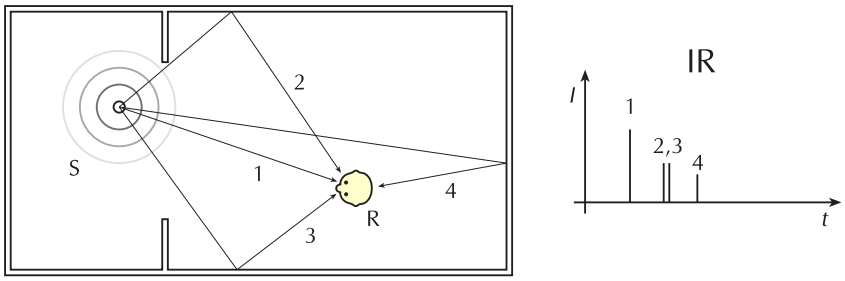

Figure 1. *Propagation of sound in a closed environment. S = sound source, R = sound receiver. The diagram on the right hand side is the corresponding IR diagram.*

2.2 The Human Auditory System

Because the human auditory system is capable of the detection of the reverberation in the sound received and the analysis of the spatial description of the surrounding environment contained in it, it is necessary to include the acoustic response of the environment to make the rendered sounds in the VR systems sound natural to their users. As the sounds interact with different obstacles while propagating, the sounds that arrive to the listener contain multiple echoes of the original signal emanated from the source (a combination of sounds with varying delays and magnitudes of attenuation), and also the information about the directions of arrival.

The echoes may be divided into the following groups:

- The first echo received by the listener is interpreted by their auditory system as *the direct sound*. Its direction of arrival gives the most significant information about the position of the source. Its intensity gives a hint on the distance of the source.
- *The early echoes* (within the first 100 ms) allow the listeners to tell the position of the nearest obstacles from the analysis of their incoming direction and intensity (Funkhouser, et al., 2002).
- *The late reverberation* carries the information about the overall layout of the environment (the size of the room, the textures of the floor/wall, etc.)
- The human receives the information about the direction of the incoming sound by analyzing the difference of the sound signals received by each ear. The following effects are of special interest:
- Interaural Time Difference. As sound travels at relatively low speed and human ears are spatially separated, there is a slight time difference between the sound arriving to the left ear and to the right ear.
- Interaural Intensity Difference. Because of its mass, human head becomes an obstacle for the ear at the opposite side of the sound source.
- Pinnae Response. Due to their high geometric complexity, ear auricles (pinnae) act as very sophisticated filters whose characteristics are highly dependent on the direction of the incoming sound. The analysis of attenuations of signal in certain frequencies allows human auditory system to determine the vertical component of the direction of incoming sound signal.

All these phenomena are modelled as the Head-Related Transfer Function (HRTF), a response function describing the acoustic filter of the environment near a listener's head. The HRTF depends on the position of a sound source relative to a listener and the size and shape of the torso, head and pinnae (Algazi, et al., 2002).

2.3 Spatial Sound in Virtual Environment

A sound propagation in any environment may be described using an acoustic wave equation (Kuttruff, 2000). However, its solution is very complex even for simple configurations and virtually impossible for more

complicated scenes where many obstacles are involved. Therefore, some alternative ways to describe sound waves are necessary. We may distinguish three different types of approaches to the approximation of the wave equation solution: numerical, geometrical, and statistical.

2.3.1. The numerical approaches. These approaches provide an approximation of the wave equation by reducing the problem to estimation of energy transfers among finite elements specified within the modelled scene. There are the following two principal approaches:

1. *The finite and boundary element methods* provide a solution to the wave equation by dividing the space of the modelled environment into distinct elements. The continuous wave equation is converted into a discrete set of linear equations. The underlying computation of these methods is generally very complex. As the necessary resolution of the spatial subdivision increases with the highest frequency of the sound to be modelled, these methods are suitable only for simulation of low-frequency energy transfers within simple scenes.
2. *The waveguide mesh* is a regular array of elements with its neighbours connected by unit delays. Each element describes the sound energy of a finite part of the modelled environment. Each sound source and receiver is modelled as an element from the mesh either with input or output of the signal. The simulation itself is iterative. During each iteration, every element updates its energy status based on the previous energy status of all its neighbours following the energy conservation law. The IR is then described by the development of sound energy in the receiver (Lokki, et al., 2002).

2.3.2 Geometrical approach. In these approaches it is assumed that the sound wavelengths are smaller than the obstacles in the scene and therefore they are usable only for the simulation of sounds of high frequencies. However, their smaller computational costs compensate for their inaccuracy and make them usable in various VR systems. The paradigm of these approaches is the sound wave simulation through the investigation of the sound rays. The audibility of sound sources in the position of the listener is determined and quantified through finding the rays that represent audible echoes of the emitted sound. The following two methods should be mentioned:

1. *Ray tracing*, adopted from the well-known method in the 3D computer graphics, is based on the concept of sound rays tracing as shown in Figure 2. Each ray of the initial set of rays emanating from the sound source *S* is traced and compared with the position of sound receivers *R1* and *R2*. The tracing is stopped when a certain condition is met (e.g., the maximum order of reflection or the minimum level of energy has been exceeded, or the ray hits the receiver).
2. *Beam tracing* is based on the concept of tracing the sound beams. A beam is a cone defined by its apex (sound source) and its base (a closed environment). as illustrated in Figure 3. A beam represents all rays that would originate in the beam's apex and intersect the beam's base. Using this method, larger areas of the space are searched at once as illustrated in Figure 4.
3. The beam tracing is a generalization of the ray tracing method. The result of the tracing process of a single beam is a beam tree in which each beam is represented by a node. The children of these nodes represent the beams that originate from the collisions of their parent beam with the obstacles. Calculating the reflections of a beam is more computationally expensive than calculating the reflections. Since the number of beams increases exponentially in reflections of higher order, this method is only usable for early echoes. Consequently, it is impossible to use this method to simulate the long reverberations, as the reflections of high order (e.g., ≥20) are not computed in reasonable time.

Formally, the beam tracing may be described as (Heckbert and Hanrahan, 1984):

```
algorithm beamtracing
input: scene in boundary representation
output: beam tree
method
        create an initial set of beams
        for each beam from the initial set
            call function trace_beam(scene, beam)
end_algorithm
```

In our designed system, the **trace_beam** function was implemented as follows:

1. All obstacles within the volume of the beam *B1* to trace are stored into a list ordered by their increasing distance from the beam's apex.
2. Each obstacle *F* from the list (from the nearest to the farthest) is tested to see whether any of its part occludes a part of the *B1* beam volume but is not yet occluded by any other obstacle. If it does, a

reflecting beam, *B2*, is created and inserted into the beam tree as a child of *B1* and is of the following properties:

- The apex of *B2* is obtained as the planar reflection of the *B1*'s apex about the plane of the obstacle *F*.
- The base shape of *B2* is obtained as the intersection of the projection of the *B1*'s base into the plane of the obstacle *F* and the obstacle *F*.

3. For each newly created beam *B2* from the children of *B1*, the **trace_beam** function is performed if a limit has not been reached (e.g., the beam *B2* is of smaller depth in the tree than a certain limit, the energy represented by *B2* is greater than a certain limit, etc.).

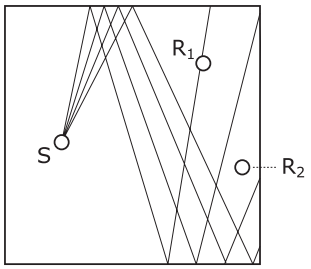

Figure 2. *Ray tracing.*

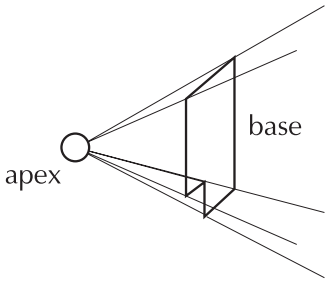

Figure 3. A beam.

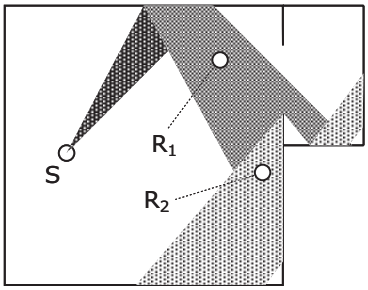

Figure 4. *A single beam; S –source, R_1, R_2 –receivers.*

2.3.3. The statistical approaches. The human auditory system is able only to distinguish the early echoes. As the late reverberation phase only gives the information about the size of the environment, it is possible to model the late reverberation phase using a statistical model where the echoes contributing to the simulated IR are randomly generated.(Funkhouser, et al., 2002).

2.3.4. The Convolution. The process of applying the IR to the sound signal for spatialization is usually modelled as the convolution of the sound signal and the IR. The convolution of two discrete signals in the Digital Signal Processing (DSP) is usually defined as:

$$f_1[t] * f_2[t] = \sum_{u=0}^{t} f_1[u] - f_2[t-u]; \qquad f_1 f_2 \text{ are the input signals} \tag{1}$$

3. THE SPATIAL SOUND SYSTEM

The main function of the spatial audio system we designed is to perform off-line (non real-time) simulations of the sound propagation between sources and receivers, taking into account the acoustic response of the environment. Our system employs a hybrid sound propagation model built as a combination of a beam tracing algorithm (for the early echoes) and a statistical model (for the late reverberation).

The process of modelling the acoustic response of the environment consists of the two fundamental steps:

1. IR is computed as a result of simulation of the propagation of the sound from the source to the receiver in the environment.
2. The sound signal representing the acoustic activity of the sound source is taken to the convolution with the IR obtained in the previous step. The result is the spatialized audio signal.

3.1 The Architecture of the Spatial Audio System

The architecture of the designed system is described in Figure 5. The main module of the system is the Task Controller. The Task Controller initiates the calculation by reading the configuration of the rendering task from an external task description file. The format of the task files was chosen to be a proprietary derivative from the XML language for easy readability by both humans and computers. After all parameters are fetched (i.e., the scene description, the position and sound stimuli of the sound sources, the position of the sound receivers, and the acoustic properties of materials), they are stored into the Scene Representation module.

The Task Controller then activates the Sound Propagation Model module whose main purpose is to perform the simulation of the sound propagation in the environment of the scene. The module implements both the beam tracing and the statistical reverberation techniques. It is done by using the scene description

(room size and texture, the obstacles' locations, etc) and the location of the sources and receivers to compute the IR profile (i.e., the individual echoes). The IR profile is stored in the Scene Representation module.

Finally, the Convolution module performs the spatialization of the raw sound signals using the echoes generated by the Sound Propagation Model module. The results are written to the output sound files through the Output Channels. An Output Channel is assigned to each elementary receiver. The resulting sound file is stored into an external output file of a standard format, which is presented through selected target device.

In the designed system, a multi-channel model of the sound receivers is employed. This model assumes that every receiver consists of several elementary receivers (e.g., a headphone consists of 2 elementary receivers: the right and the left receivers). Each elementary receiver contains a unique directivity filter. This model enables the system to render spatial audio signals for multiple target devices, e.g., headphones (where the HRTF of the listener is used as the directivity filters) or multi-speaker systems (quadraphonic, 5.1, etc.).

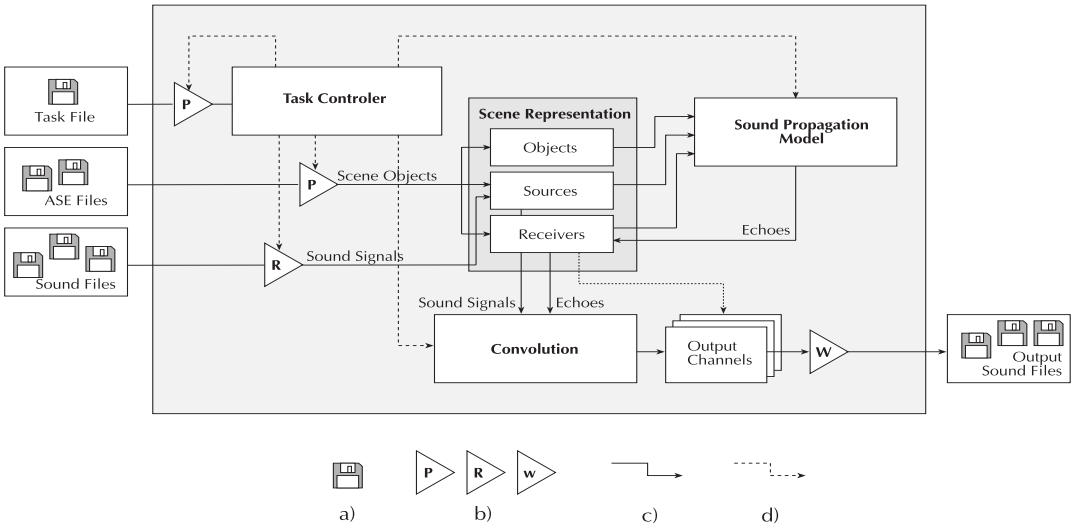

Figure 5. *The architecture of the designed system. a) Permanent data structure (file), b) Parser, reader of binary files, writer of binary files, c) Data flow in the direction of arrow, d) Spawn of a new instance of a block pointed to by the arrow.*

Because of the performance issues and the limited availability of the required 3rd party libraries, the system was implemented in the C++ language using Microsoft Visual C++ 6.0. The modules that involve DSP were implemented using the Intel Signal Processing Library version 4.5 (Intel, 2000) as it offers useful signal processing primitives such as the superposition of signals, FIR filtering, convolution, etc. The binary operations with polygons were implemented using The General Polygon Clipping Library (Murta, 2003).

4. USER EVALUATION

A group of prospective users has evaluated our spatial audio system and the fit of the algorithms used in it.

4.1 The Stimuli

We have acquired two different sets of stimuli: The sounds in the real scenes (the recorded scene stimuli) and the sounds spatialized by our system (the simulated scene stimuli.)

The real scene sounds were recorded using a stereophonic microphone PHILIPS SBC 3050 and a SoundBlaster 16 compatible sound card. Seven distinct sounds: guitar, flute, mobile phone ringing, human voice, cane tapping, glass tinkling and handclapping, were recorded in three different room conditions, coded small (S), medium (M) and large (L). The characteristics of these rooms are summarized in Table 1.1.

To create the spatialized sounds, the dry sounds (the signals without the effects of the surrounding environment) were recorded in a music studio with a very short reverberation time (less than .05s). We have used the AKG C1000S microphone and the Midiman Delta 1010 sound card. The sound signals were stored as a set of 44.1 kHz PCM files. Later on, the effects of the environments were simulated using the designed spatial audio system. It was a two-step process:

1. A model of the real rooms was created using the 3D Studio Max (and stored in its native format ASE.)
2. Each room model has been used with each recorded dry sound signal as an input to our system.

Table 1. *The approximate characteristics of the real scenes*

Environments	Dimensions	Surfaces	Reverberation length
Bedroom (S)	4 × 4 × 2.5 m	Plaster, carpet, wood	.2 s
Hallway (M)	8 × 3 × 5 m	Plaster, marble	1 s
Stairway (L)	12 × 12 × 10 m	Plaster, marble, tiles	3.5 s

4.2 The Evaluation Method

Nine registered blind participants (8M, 1F; mean age 29.3 with a S.D. of 6.76 years) listened to 42 sound files (7 sound types x 3 environments x 2 scenes) through a headphone. The sequence of the sounds played was controlled so that no adjacent sounds shared any similarity (e.g., the following stimulus to the flute sound in a simulated small room had to be a recorded scene, being not a flute sound, nor a small room). Each participant went through two sessions of evaluations with no other participant around. In the first session, each participant listened to one of the two sets of sounds. The order of the second set of sounds is the reversed order of the first set. After listening to each sound, the participants answered in writing three questions:

1. What sound was it?
2. Was that sound more likely to be from a small (S), medium (M) or large (L) room?
3. Was that room more likely to be a real room (R) or simulated using computer (C)?

After taking a short break, each participant listened to pairs of sounds in the second session (these pairs represent all possible combinations of room sizes and scenes, e.g., flute sounds in recorded S and M rooms, handclaps in recorded and simulated L room etc). Each participant listened to either one set of sound pairs or its reversed order. After listening to each pair of sounds, they answered in writing two questions:

1. Did you hear any difference between sound 1 and 2? Y/N
2. If Y, describe the difference.

4.3 Results and Analysis

4.3.1. The first session. The first question in the first session was intended to encourage the participants to listen carefully. Therefore, in this paper the answers were not analysed.

The answers to the second question were scored 0, 0.5 or 1. When the participants answered correctly, they were scored 1. A score of 0.5 was given when the difference between the correct and the wrong answers was one room size (e.g., a participant answered S for a sound in an M room). When the difference was two sizes (a participant answered S for L or vice versa) then a score of 0 was given.

The one-way Analysis of Variance (ANOVA) reveals that across all participants, the sum of scores for the room size question were not significantly different between the recorded and simulated scene groups, with $F(1,376) = 0.03$, $p = 0.862$. This might mean that the designed system successfully simulates various room sizes. Further analysis shows that the difference was not significant in any room size (see Figure 6).

The answers to the third question were scored 0 (wrong) or 1 (correct). In essence, this question asked the participants to perform a signal detection task. Signal detection theory (SDT) is a method of assessing the decision making process of detecting two classes of item (in this study, R and C). In an SDT task, the answers are described in terms of hit (H) and false alarm (FA) rate. In this study, the hit rate is defined as the proportion of the correct C answers in the C room conditions as explained in Table 2.

The most commonly used SDT measure is d' (discriminability), which is the standardized difference between the means of the distribution of the two classes. Larger absolute value of d' means higher distinctiveness between the two classes and d' near zero indicates chance performance. The maximum value of d' is around 5. The formula of d' is $d' = z(H) - z(FA)$ where z is the performance in terms of the number of standard deviations above or below the mean. Table 3 lists the H and FA or each room condition.

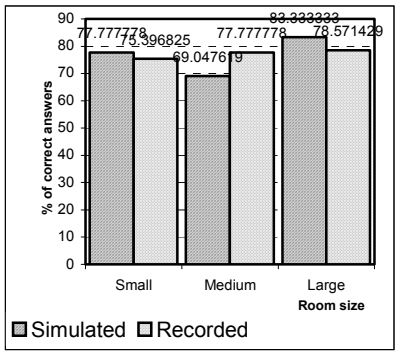

Figure 6. *The percentage of correct answers to the room size questions.*

Table 2. *The decision matrix of the user test.*

	C room	R room
Answer = C	Hit (H)	False Alarm (FA)
Answer = R	Miss (M)	Correct Rejection (CR)

Table 3. *Hit, false alarm and d' for each room condition.*

	S	M	L
H (%)	49	41	71
FA (%)	17	41	43
d'	0.935	0	0.38

Table 3 shows that in the medium room condition, the participants were not able to distinguish between the recorded and simulated scenes. The participants were most able to distinguish between those scenes in small room condition, followed by the large room condition. However, even in the latter two room conditions, the d' values were still quite small. These results might mean the designed system was successful in modelling the medium environment (hence, the simulated and recorded scenes could not be distinguished), and was less successful in simulating the small or large room conditions, although taking the absolute values of d', there was some degree of success in simulating these two conditions.

Looking at the types of sounds, there was no significant difference between various sounds in terms of the scores of the room size ($F(6,371) = 0.498$, $p = 0.106$) or the nature of scenes ($F(6,371) = 1.763$, $p = 0.810$). However, some interesting observations emerged from the sounds used to question the participants.

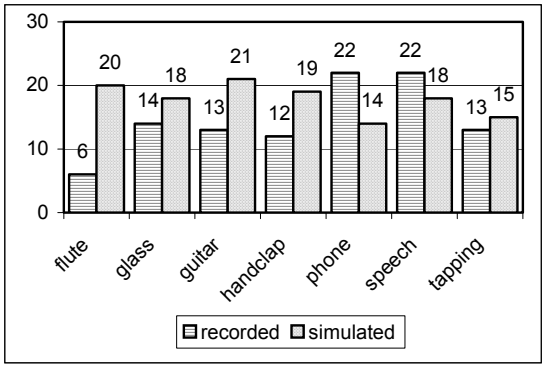

Figure 7. *The scores for the recorded vs. simulated scene question by type of sound.*

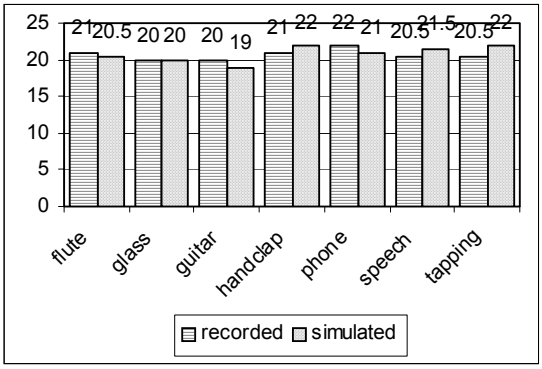

Figure 8. *The scores for the room size question by type of sound.*

Figure 7 and 8 depict the scores for each sound. Figure 10 shows that when the participants were asked whether the sounds were from recorded or simulated scenes, the highest scores were from the recorded phone ringing and human's speech sounds (both were 22 out of 27), while the highest for the simulated scenes were from the guitar and flute sounds. This finding may indicate that when musical instruments were used, the participants could detect that the sounds came from simulated scenes quite accurately. Similarly, when the sounds are ones that the blind people may hear in their daily life (in this case, the human voice and the phone ringing), they were able to detect quite accurately that these sounds came from real environments.

Figure 7 shows that the score from each sound was very similar when the participants were asked about the room size in both the recorded and simulated scenes (the scores ranges from 19-22 out of 27). This may simply mean that the types of sounds do not have any effect in helping or hindering blind people in guessing

the size of room in both recorded and simulated scenes. In other words, the type of sounds does not affect users' perception of room sizes in both room scenes.

4.3.2. The second session. The second session aims to gather ideas about users' terms when describing the differences between various experimental conditions. In general, most answers to the questions about the differences between the sounds in different room sizes contain the words: echoes, reverberation, resonance, closer/further and depth. Most answers to the question about the difference between the sounds in simulated vs. recorded scenes contain the words: more/less natural/realistic, crispier/sharper and metallic/ digital sounds. As these words represent the terms that blind users may use to describe the sound variations in different environments, these terms may be useful in the context of the O&M training for blind people (e.g., "Please listen carefully to the sound variation. The first sound is from a larger room than that of the second one as the first sound has more resonance.")

5. CONCLUSIONS

The results of the user studies indicated that the algorithms behind the designed spatial audio system were able to simulate the environments to certain extent. The system was able to simulate the sound variation in different room sizes successfully, indicated by the lack of significant differences between the sum of scores in the simulated and recorded scene groups. When simulating a medium room, the difference between the reverberation of the simulated and recorded scenes were not noticeable to our blind participants. Based on these results, we can speculate that the designed audio system is potentially useful as a part of the O&M training suite for blind and visually impaired people, preferably to simulate sounds in a medium room. Our immediate future work is to implement a scene audibility graph, that would describe the acoustic relations among different places in the scene, regardless on the actually employed method of the spatial audio simulation. We also plan to integrate this system into a training suite and testing the suite with its prospective users. Further studies are also needed to investigate users' mental model of various room conditions.

6. REFERENCES

V R Algazi, R O Dudaa, R Duraiswami, N A Gumerov and Z Tang (2002), Approximating the head-related transfer function using simple geometric models of the head and torso, *Journal of Acoustics Society of America* **112**, pp. 2053-2064.

C Colwell, H Petrie, D Kornbrot, A Hardwick and S Furner (1998), Haptic Virtual Reality for Blind Computer Users, *Proc. ASSETS 1998*, pp. 92-93.

R O Duda, C Avendado and V R Algazi (1999), An Adaptable Ellipsoidal Head Model for the Interaural Time Difference, *Proc. ICASSP 1999*, pp. 965-968.

K M Franklin and J C Roberts (2003), Pie Chart Sonification, *Proc. of 7th International Conference on Information Visualization*, pp. 4–9.

T Funkhouser, J M Jot and N Tsingos (2002), Sounds Good to Me!, Computational Sound for Graphics, Virtual Reality, and Interactive Systems, *SIGGRAPH 2002 Course Notes*. http://www.cs.princeton.edu/gfx/papers/funk02course.pdf.

T Funkhouser, P Min, I Carlbom (1999), Real-Time Acoustic Modeling for Distributed Virtual Environments, *Proc. SIGGRAPH 1999*, pp. 365–374.

P S Heckbert and P Hanrahan (1984), Beam Tracing Polygonal Objects, *Computer Graphics* **18**, *3*, pp. 119–127.

D P Inman and K Loge (1999), Teaching orientation and mobility skills to blind children using simulated acoustical environments, *HCI* **2**, pp. 1090-1094.

Intel® Signal Processing Library (2000), *Software library documentation 630508-012*.

H Kuttruff (2000), In Room Acoustics, 4th ed., Spon Press, London, U.K.

T Lokki, L Savioja, R Vaananen, J Huopaniemi and T Takala (2002), Creating Interactive Virtual Auditory Environments, *IEEE Computer Graphics & Applications* **22**, pp. 49-57

A Murta (2003), A General Polygon Clipping Library. http://www.cs.man.ac.uk/aig/staff/alan/software/gpc.html

Y Seki and K Ito (2003), Study on Acoustical Training System of Obstacle Perception for the Blind. In *Assistive Technology - Shaping the Future* (Craddock, McCormack, Rielly & Knops, Eds.), pp. 461-465.

The Royal Blind School (2003), Orientation and Mobility. http://www.royalblindschool.org.uk/Departments/Mobility.htm