MULTI-SENSORY RENDERING: COMBINING GRAPHICS AND ACOUSTICS

Jackson Pope and Alan Chalmers
Department of Computer Science, University of Bristol, UK
email : pope@cs.bris.ac.uk

ABSTRACT

Human perception of the geometry and spatial layout of an environment is a multi-sensory process. In addition to sight, the human brain is also particularly adept at subconsciously processing echoes and using these reflected sounds to provide some indication of the dimensions of an environment. This auditory impression of the size of an environment will incorporate surfaces not only to the front, but also to the sides and rear of the person and thus currently hidden from his/her view. So while computer graphics can provide an image of what a person can currently see, the level of perceptual realism may be significantly improved by incorporating auditory effects as well. This paper describes a method for combining the computation of lighting and acoustics to provide enhanced rendering of virtual environments.

Keywords : Computer Graphics, Acoustic Rendering, Virtual Reality, Particle Tracing, Perception.

1 INTRODUCTION

Our perception of the geometry and size of virtual environments is a multi-sensory experience. Both visual and auditory stimuli are collected by our brains and the two types of information are correlated to provide a combined description of our environments [ShiZh93a].

A visual display of any virtual scene limits the perceived realism by only allowing us perception of a small part of the environment - that which lies directly in front of us. When unrelated sound information is simply added to such rendered images it can often be confusing especially when it contains no information about the geometry of the scene or consists of sounds which are at odds with the visual display of the environment.

Despite the significant effect the acoustics has on our perception of the realism of a scene, sound rendering has been studied far less than rendering of the visible spectrum. Previous work on computer modelling of room acoustics has tended to focus on the study of the acoustics of theatres, concert halls and other auditoria, or on noise control in factories. Tools have been developed to aid the design of such buildings, improving the acoustics by removing dead spots and other effects caused by the geometry of the room [Klein93a, Kuttr79a]. Such acoustic tools model the propagation of sound waves through the environment in a very similar manner to the ray tracing technique [Whitt80a] used in computer graphics and calculate impulse responses, such as that shown in fig. 1, which indicate the energy and time-delay associated with the wavefronts when they reach a reference point. Separate impulse responses may be constructed for different incident angles and reference points.

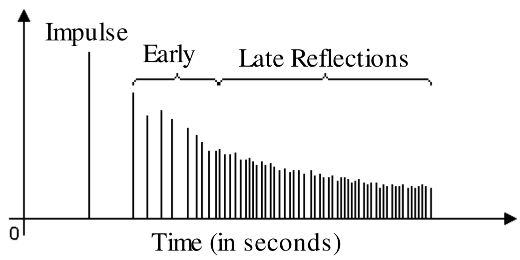


Figure 1. An Example Impulse Response

The rendering of audible information may thus be used for communicating the geometry and more particularly the size of an environment. The speed of travelling sound enables a user to detect subconsciously the delays and directions associated with the echoes which have been reflected off the surfaces delimiting the scene [Klein93a, Kuttr79a]. From this information we can 'hear' the size of the room within which we stand and ascertain where we are with respect to the other surfaces in the environment [Boris84a, Klein93a, Kuttr79a]. Such knowledge is hard to discover from a purely graphical 2D display monitor as no information about the surfaces to the sides or behind the viewing frustrum is shown.

In the past simplified techniques have been used to give some indication of the sounds within an environment, but these, while describing the sound and its position relative to the observer, provided no information concerning the geometry of the room. This can lead to confusion in the perception of the virtual environment. One such technique is the 'ambient' sound method, where sound is treated as global. The only calculation done is to decrease the intensity of the sound relative to the distance from the source. This is the technique used in Virtual Reality Modelling Language (VRML) worlds [Ames 97a]. Another method is `spatialised' sound in which sound is computed according to distance from the sound source and the direction of the source, thus making use of the brain's ability to position a sound according to slight differences in intensity, phase, and time delays between left and right channels [Klein93a, Lehne92a, Mølle92a]. Both the ambient and spatial sound methods only deal with direct sound, that is sound that has reached the receiver without having been reflected off any surfaces. However, it is precisely the reflected sounds that allow us to perceive the geometry and size of the room as we subconsciously process the echoes that are reflected off the surfaces that define the bounds of the environment we are in [Boris84a, Klein93a, Kuttr79a].

Other techniques have been used to render both direct and reflected information. The ray tracing technique, where sound rays are traced from a sound source through the environment until the energy they carry is sufficiently small to be negligible is one such method [Krok68a]. The ray tracing technique works in a very similar way to graphical ray tracing, but with the inclusion of time dependence. It has been used to model both specular and diffuse reflection and remains one of the most popular methods in use. Another method, the mirror image source technique, generates accurately the reflections rather than relying on a stochastic process to determine the reflections [Boris84a]. In the mirror image source method the sound source is reflected in the surfaces of the environment to provide an attenuated reflected source. High order reflections can be generated by reflecting sources that have already been reflected in another surface. This method does not scale well for very high order reflections or scenes with complex furnishings although image source culling methods exist. The image source method does not take into account diffuse reflection, and cannot model non-planar surfaces.. Shi and Zhang and others have modified the radiosity method to model the diffuse reflection of acoustic energy [ShiZh93a]. However, due to the far greater wavelengths concerned when modelling sound waves, the reflections are more specular than when modelling light, hence surfaces cannot be treated as perfectly diffuse. Beam tracing has also been used to model sound fields; Funkhouser et Al. used beam tracing to generate a tree of valid mirror image sources in an off-line stage to enable real-time rendering of the sound field for small numbers of reflections [Funkh98a].

2 COMBINED RENDERING OF GRAPHICS AND SOUND

The method described in this paper uses an augmented particle tracer [Patta92a] to render sound in addition to light propagation. Particle tracing models the emission of light particles from a source and tracks them as they are reflected around the environment, until they are finally absorbed. As particles are reflected form a surface the brightness of that surface is incremented, which leads to calculation of the global illumination of the scene. The use of a single model and single program for both modelling procedures ensures that the two renderings are coherent, and enables time and computer processing to be saved by only performing some tasks once for both models.

Two methods can be used to trace the sound rays: tracing the wavefronts themselves from the source, or tracing from the receiver back to the source. For a scene containing many receivers it is more efficient to trace from the source as these rays can be traced through each receiver in turn.

The light modelling using particle tracing computes global information throughout the scene, calculating the illumination of every surface in the scene, this enables reuse of several of the algorithms for sound rendering which is also a global rendering of the scene. The particle tracer supplies us with the ability to model ideal diffuse, ideal specular and partly specular reflections and participating media (which would enable areas of different temperature or humidity to be modelled). The main adjustment that must be made for rendering sound is the treatment of the time it takes sound the traverse the environment. When rendering light information we assume that the light distribution within the environment has reached equilibrium, so time can be ignored, but the time delay associated with audible echoes enables us to gain some indication of the room's dimensions.

The scene is partitioned into a grid voxels for the purpose of optimising the object intersection tests, these voxels can be used in sound rendering for housing sound receivers. The sound responses are then calculated at the centre of each voxel in a manner similar to the generation of irradiance volumes done by Greger et Al. [Grege98a]. By positioning the cells adjacent to each other, in a regular grid, the perceived sound field may be calculated throughout the environment. The advantage of this method as opposed to the conventional ray tracing method is the independence gained from listener position that is achieved by generating a global model of the sound field.

2.1 Modelling

In our environment, the sound is modelled using an approximation of an impulse - an infinitely short sound composed equally of all frequencies. This is treated as a single sample of unit strength in all the frequency bands being modelled. The spherical wavefront of the impulse emanating from the sound source is modelled using a large number of rays normal to the sound wavefront, as shown in fig. 2 which are then followed as they traverse the environment interacting with the surfaces and objects within it.

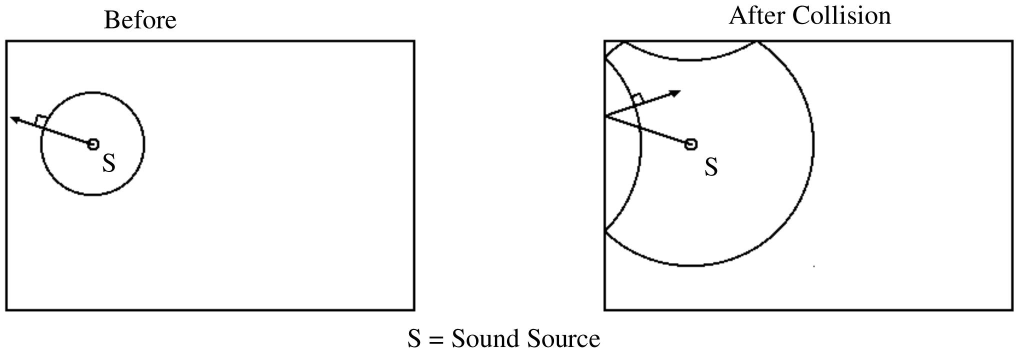


Figure 2. Spherical Wavefront and Normal Ray

If, (for now) we ignore the effects of diffraction in the air, then the sound rays may be assumed to travel in straight lines.
The directions of the modelled sound rays are chosen stochastically over either a sphere for isotropic sources or a cone for directional sources. At present, each ray is then traced until either there is a 60dB drop in its energy (a factor of 1 x 10^6) or until absorbed. The sound wavefronts arriving at a receiver fall into two categories: early and late reflections. Early reflections are distinct enough for the brain to distinguish them individually, while late reflections, on the other hand, appear as one long audible blur giving some indication of the size of the room. Most current systems model these two separately, using beam tracing, ray tracing or the image source method to calculate the early reflections, and use a function to calculate an approximation of the late reflection. Using particle tracing to model all the reflections results in an accurate model for both.

A ray is traced until it collides with either a surface (such as a wall) or a receiving cell. The two cases are significantly different so are described separately. On collision with a surface, the distance from the source of the ray to the collision is calculated and the effect of air attenuation on the energy of the sound is subtracted from the incident energy. It is unnecessary to decrease the energy according to the distance travelled (according to the inverse square law) as this decrease in energy is treated by the diversification of the rays with distance[Kulow85a]. The resultant energy is then further decreased by the surface absorption for the specific material. Finally, the ray is specularly or diffusely reflected depending upon the surface properties. The ray has its source reset to the current collision point and its direction changed to the new reflected direction. The ray is then traced further. The collisions with receiving cells are considerably simpler to compute since there is no need to calculate a reflection. The distance from the source is calculated and the air attenuation effect is subtracted. The resultant energy is stored with the receiver together with the time delay and incident angle and the ray continues on its path. This procedure is repeated for a suitable number of rays until the solution for the environment is achieved. The choice of the number of rays is determined by the absorption coefficients of the surfaces and the dimensions of the scene. Assuming an average absorption coefficient of 0.2 (stone = 0.04, carpet = 0.3 at 2kHz [Kuttr79a]), then it will take approximately 65 reflections for the sound energy of the reflections to become negligible (a drop in level of 60dB). From this an average distance between objects (taken as the largest dimension of the scene) a number or rays can be calculated. For the small scene in the results in the region of 3 million rays need to be traced for accuracy up to 65 reflections, for the large scene 48 million rays.

2.2 Collection And Correlation At The Receiving Cells

As the results arrive at the receiving cells (which are treated as spheres for the purposes of receiving sound) impulse responses are constructed according to the incident direction of the sound ray and the frequency of the sound. It is necessary to model the sound receivers as spheres to ensure that collisions between the rays and the receivers occur. Were the receivers modelled as points then the chance of a ray intersecting with that receiver would be negligible.

In our implementation, only the azimuth angle is used to describe the direction from which the sound ray came. (Our current playback technology precludes rendering the elevation of an incident sound wavefront.) Each receiver stores its own impulse responses, which correspond to the sound field within the cell in which the receiver resides. Separate impulse responses are created for each of eight 45 degree azimuth bands, fig. 3, and additional impulse responses within those directional bands are also created if more than one frequency band is being considered.

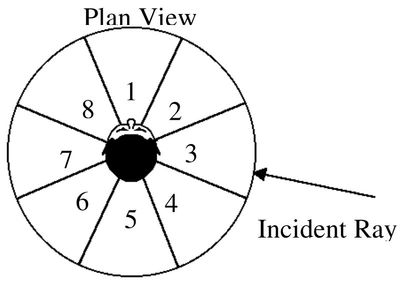


Figure 3. Azimuth Angle Filter Bands

Each impulse response describes a time-delay filter, such as shown in fig 4, through which the played sound is directed. The sound samples enter from the left in the diagram. In this example, the current sample and the previous three samples are multiplied by the four filter values which are then added together and played through the speakers. In our implementation we use a filter bank of length 3840 (80ms at 48kHz). Each sample is multiplied by a factor in the range (0,1) which indicates the percentage of energy of that reflected ray. For example, 0 indicates that no ray was received at that time, while 0.78 describes the arrival of a ray that has lost 22% of its energy through attenuation and absorption before reaching the receiver.

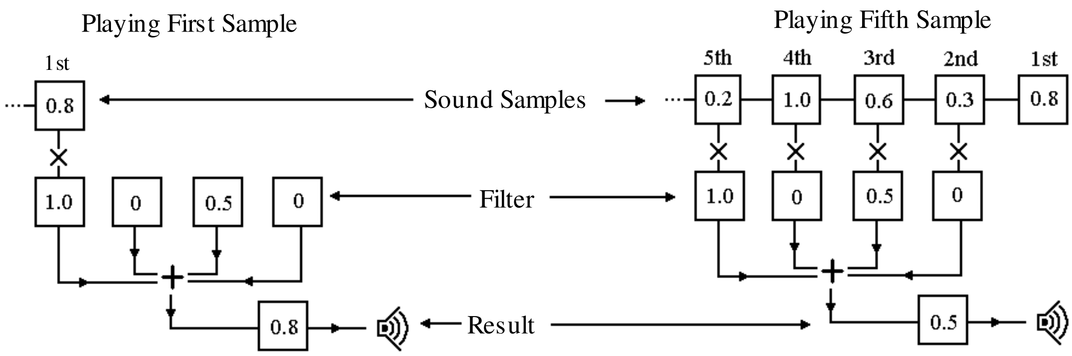


Figure 4. A Simple Time Delay Filter

2.3 Playback

In order to perform a walk-through of the scene, the auditory response at each point in the scene is described by the time-filters in each cell. If only one frequency band was considered then the sound file is simply piped through the appropriate time-delay filters and to the speakers. However, if multiple frequency bands have been used then the sound must first be translated into the frequency domain using a suitable transform (such as Fast Fourier Transforms). The multiple frequency responses can now each be piped through the appropriate time-delay filter before being reconstructed into the time domain using an inverse transform before being outputted to the speakers.

The resulting time-delay filters constructed from the impulse responses are played back through an eight loudspeaker array. The eight loudspeakers are arranged in a cube, approximately 2 metres in all dimensions. The room in which the loudspeakers have been positioned has been damped to limit the effect of the room's impulse response on the presented results.

As the user traverses the scene, the cell within which the user is positioned is determined from their position. The impulse responses for the cell in which the user resides are then convolved with the input signal. It is important that the input signal, which will be presented to the listener, is clean and has no echo information of its own. Otherwise the two sets of echoes will provide conflicting information. For the results presented here all the input signals used were either computer generated, or recorded in an acoustically dead room. Once the correct impulse responses have been chosen and convolved with the input signal the resulting signals (one for each of the azimuth bands in fig. 3 if only one frequency is being modelled) are presented to the user. To map the eight azimuth bands to the loudspeakers the signal from each band is assigned to a virtual sound source. The virtual source is positioned on the unit sphere with zero elevation and azimuth angle equal to the centre of the azimuth band in with the impulse response was recorded, fig. 5.

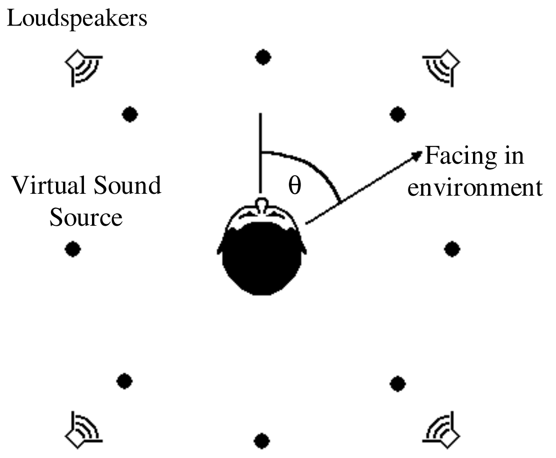


Figure 5. Positions of Virtual Sound Sources

These virtual sound sources are then rotated by angle θ to account for the direction in which the user is facing in the environment, and their output is played through the relevant speakers.

3 RESULTS

To validate our approach, we have initially considered a single sound source in a (closed) room. During the pre-processing required for particle tracing, the scene is divided up into voxels for object intersection optimisation, those at approximately "head-height" which we have taken as 1.5m, as shown in fig. 6, are given sound receivers. The location of the sound (and light) sources are known from the modelling phase.

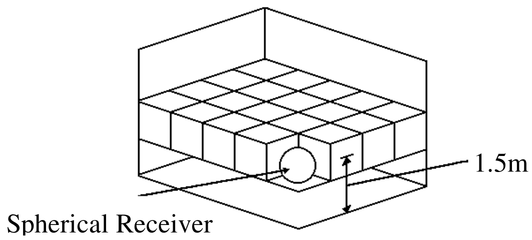


Figure 6. Cut-away Diagram of a room showing cell distribution

Fig. 7 shows the results for three similar rooms with identical proportions, the only difference being a scaling in all three dimensions, the first room is 6m long, 2.5m wide and 2m high, the second twice those dimensions and the third twice again. In each case the walls and ceiling reflect 98% of the light that hits them and 95% of the sound, and the floor 30% of light and 70% of sound. The light source is a narrow strip light, which scales with the room. The sound source is in the centre of the room 1.5m high in all rooms and represents a person standing in the centre of the room talking. To this end the impulse responses considered are from the central receiver (centred in the same place as the sound source).

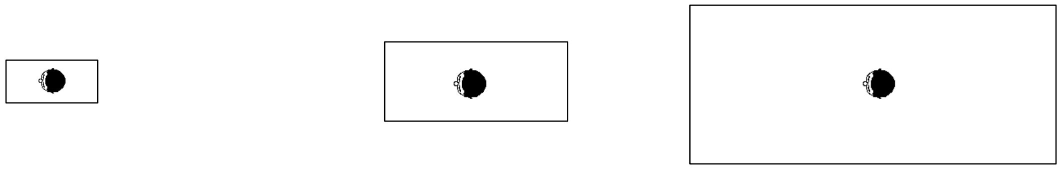
Due to the lack of items with inherent scale (such as chairs or tables) in the room, the size of the room is difficult to judge, the only visual clue being the height of the viewing position from the floor (which remains at head height - 1.5m - throughout all three rooms). The impulse responses on the other hand show clear indication of scale, the smallest room receives reflections far quicker than the others, and the reflections die out much quicker as well, and the front and right impulse responses show discernible differences as expected from the different geometry in those directions. These differences are also clearly audible from the speaker output. The slight differences in the impulse responses between rooms (other than time scaling) are caused by the fact that both source and the receiver we are testing remain at a constant height from the ground, and hence don't scale in accordance with the geometry.

3.1 Hardware

The results presented in fig. 7 were rendered on a Sun Ultra Sparc 1, with 160 megabytes of memory. Rendering took approximately 6 minutes per million rays traced from the sound source. The memory requirements for rendering were 3 megabytes per receiver, but these impulse responses were compressed to contain only non-zero samples for saving and during playback.

The playback section was performed on a Silicon Graphics Octane with 128 megabytes of memory, which drives an ADAT interface to the loudspeakers.

Room A | Room B | Room C

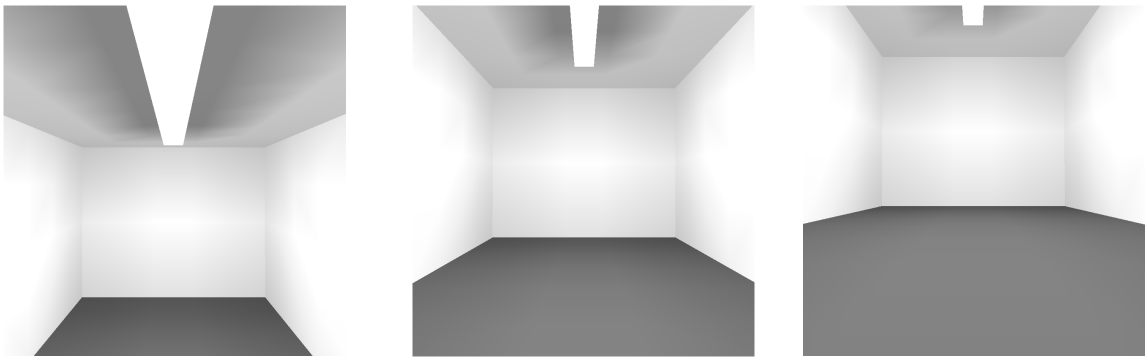


Length 6m
Width 2.5m
Height 2m

Length 12m
Width 5m
Height 4m

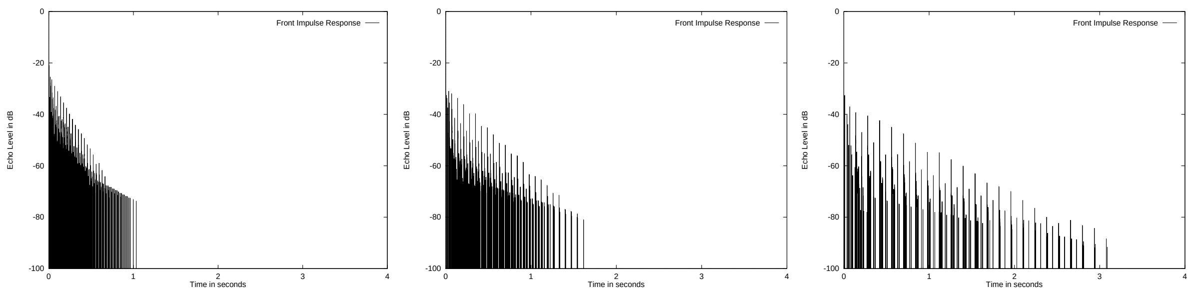
Length 24m
Width 10m
Height 8m

View from the receiver position



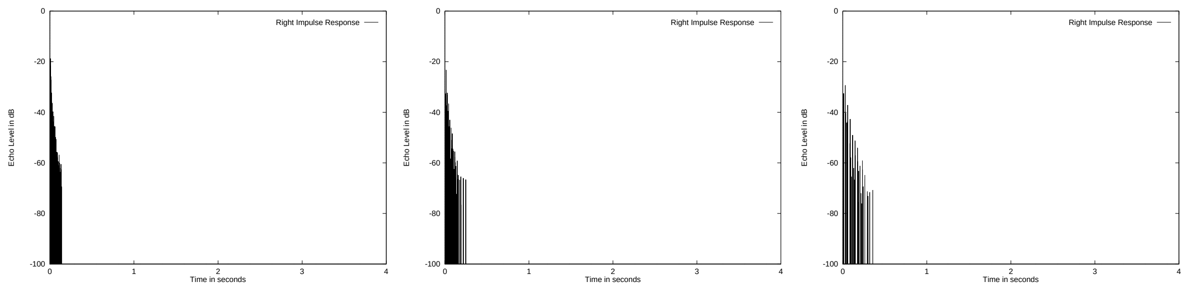
Room A | Room B | Room C

Impulse Responses received from the front



Room A | Room B | Room C

Impulse Responses received from the right



Room A | Room B | Room C

Figure 7 Results for three rooms of differing sizes when the source and receiver are in the centre of the room at a height of 1.5m

4 CONCLUSIONS

Realistic rendering is an important part of most virtual/augmented reality environments. While lighting simulations, such as the particle tracing method, can provide an acceptable approximation of the illumination within environments, the image they present to the user is restricted by the viewing frustrum. We have presented a method for combining both graphics and sound in the display to provide multi-sensory feedback to the user. Sound information, in the form of impulse responses and speaker output, can improve the user's perception of the size and geometry of the whole environment.

The results presented here show graphically how the size of a room affects the room impulse response. These effects are also clearly audible, the reverberation being noticeably different between the test environments presented in fig. 7. Due to the high computational overheads of the accurate impulse response calculations we are unable to recalculate them in real-time. This limits this approach to the accurate modelling of sound fields for static scenes where the geometry of the scene and the position of the sound source are unchanging, such as is required for simulations of concert halls, architectural planning and noise control.

The ability to calculate and in real-time convolve and spatialise the accurate impulse responses, which contain all reflections with more than a millionth of their initial energy is the main advantage of the pre-processing stage presented.

There are many possible improvements to this preliminary work in the field, one is a level of detail algorithm to allow the sound particles to ignore the small highly-detailed surfaces necessary for graphically realistic rendering depending on the size of the surface compared to the wavelength of the sound.

This method has, for the moment, ignored the effects of diffraction within the sound field. The diffraction of light within a scene can generally be ignored due to the very short wavelengths of visible light. However, the longer wavelengths of sound waves lead to more diffraction of sound waves. The method presented in this paper could be improved to model diffraction, as this would only effect the offline rendering stage, and not the real-time convolution. The effect of diffraction could be modelled by either warping or splitting rays when they encounter diffraction.

The computational effort required to compute both the particle tracing method and the acoustics is high, although they share pre-processing routines. Future work will investigate a parallel implementation of this multi-sensory rendering to improve the result times with a view to including this approach in an interactive augmented reality system.

ACKNOWLEDGEMENTS

We would like to thank David Creasey for his valuable input to this project, and Erik Reinhard for his useful comments.

REFERENCES

[Ames 97a] Ames, A, Nadeau, D, Moreland, J: *VRML 2.0 Sourcebook, Second Edition, John Wiley & Sons*, 1997.

[Boris84a] Borish, J: Extension of the Image Model to Arbitrary Polyhedra*, Journal of the Acoustical Society of America* Vol. 75, No. 6, pp. 1827-1836, 1984.

[Dance97a] Dance, S, Shield, B: The Complete Image-Source Method for the Prediction of Sound Distribution In Non-Diffuse Enclosed Spaces, *Journal of Sound and Vibration,* Vol. 201, No. 4, pp. 473-489, 1997.

[Funkh98a] Funkhouser, T, Carlbom, I, Elko, G, Pingali, G, Sondhi, M, West, J: A Beam Tracing Approach to Acoustic Modeling for Interactive Virtual Environments, Computer Graphics (ACM SIGGRAPH '98 Proceedings), pp. 21-32, 1998.

[Grege98a] Greger, G, Shirley, P, Hubbard, M, Greenberg, D: The Irradiance Volume, *IEEE Computer Graphics and Applications*, Vol. 18, No. 2, pp. 32-43, 1998.

[Heinz93a] Heinz, R: Binaural Room Simulation Based on an Image Source Model with Addition of Statistical Methods to Include the Diffuse Sound Scattering of Walls and to Predict the Reverberant Tail, *Applied Acoustics* Vol. 38, pp. 148-159, 1993.

[Klein93a] Kleiner, M, Dälenback, B, I, Svensson, P: Auralization - An Overview, *Journal of the Audio Engineering Society* Vol. 41, No. 11, pp. 861-875, 1993.

[Krok68a] Krokstad, A, Strøm, S, Sørsdal, S: Calculating the Acoustical Room Response by the Use of a Ray Tracing Technique, *Journal of Sound and Vibration*, Vol. 8, No. 1, pp. 118-125, 1968.

[Kulow85a] Kulowski, A: Algorithmic Representation of the Ray Tracing Technique, *Applied Acoustics*, Vol. 18, pp. 449-469, 1985.

[Kuttr79a] Kuttruff, H: *Room Acoustics, Second Edition, Applied Science Publishers*, 1979.

[Lehne92a] Lehnert, H, Blauert, J: Principles of Binaural Room Simulation, *Applied Acoustics* Vol. 36, pp. 259-291, 1992.

[Mølle92a] Møller, H: Fundamentals of Binaural Technology, *Applied Acoustics* Vol. 36, pp. 171-218, 1992.

[Naylo92a] Naylor, G: Treatment of Early and Late Reflections in a Hybrid Computer Model for Room Acoustics, *Proceedings of the 124th Acoustical Society of America meeting*, 1992.

[Patta92a] Pattanaik, S, Mulder, S: Computation of Global Illumination by Monte-Carlo Simulation of the Particle Light, *Proceedings of 3rd Eurographics Workshop on Rendering*, pp. 71-83, 1992.

[Patta93a] Pattanaik, S: Computational Methods for Global Illuminations and Visualisation of Complex 3D Environments, *PhD Thesis*, Birla Institute of Technology and Science, Pilami, India, 1993.

[ShiZh93a] Shi, J, Zhang, A, Encarnação, J, Göbel, M: A Modified Radiosity Algorithm for Integrated Visual and Auditory Rendering, *Computers & Graphics,* Vol. 17, No. 6, pp. 633-642, 1993.

[Whitt80a] Whitted, T: An improved illumination model for shaded display. *Communications of the ACM*, Vol. 23, No. 6, pp. 343-349, 1980.