A Scalable Spatial Sound Rendering System

David Murphy* and Francis Rumsey**

*Computer Science Department, University College Cork, Ireland.
**Institute of Sound and Recording, University of Surrey, Guildford, UK.

ABSTRACT

The spatial rendering of sound in Virtual Reality systems can quickly become a computationally expensive process. The author proposes a Spatial Sound rendering system that allows for the graceful degradation of spatial quality based upon scaling parameters. The parameters are a combination of both physical and perceptual attributes. The Scalable Spatial Sound Rendering system is divided into three User-Profiles; Professional, Prosumer and Consumer, where each profile is composed of a number of varying levels of quality. Typical applications for this scalable framework include Mobile-VR systems and Personal VR systems based upon standard multimedia PCs. One of the main advantages of this scalable architecture is that the audio content is only created once and is appropriately scaled for the end user – write once read many.

INTRODUCTION – THE NEED FOR SCALABLE SOUND

Traditionally research into spatial sound has focused upon high quality renderings of the spatial environment. Spatial rendering has primarily been based upon geometrical properties of environments, physical properties of objects (e.g. reflection and absorption properties), and source characteristics – in other words, the rendering is based upon a Physical Model. This approach, whilst very accurate, requires powerful processing resources and is very difficult to achieve in real-time applications [1].

At the other end of the scale there have been a number of recent projects where the aim is to provide spatial sound rendering on low-end systems [1,2]. Some of these systems are based upon reduced physical models while others focus upon more efficient algorithms for implementation. These projects have been quite successful and are used to render spatial scenes in real-time.

The distance between high- and low-end systems presents developers/content creators with a dilemma – which system should they design the spatial sound scene for? It is envisaged that this research will go some way to solving that problem – an extension of the 'write once run anywhere' philosophy.

The Context – VR

Just as sound enhanced the cinema-going experience, spatial sound increases the sensation of realism in a virtual environment. If a virtual environment merely contained visual objects and scene geometry it would be perceived as bland and fall short in any attempt to immerse the user completely. Ideally the user needs to be enveloped by sound to attain a convincing degree of immersiveness. Hence the importance of spatial sound within VR.

The emphasis in Virtual Reality development has traditionally been on visual processing, dynamic elements (behaviours, interaction, etc.), and scene management. Relatively little consideration has been given to the spatial auditory experience until recently. Several authors have surveyed spatial rendering in the context of Virtual Reality including Lehnert [3], Begault [4], Blauert [5], and Shilling and Shinn-Cunningham [6].

A number of projects, including the DIVA project (Helsinki), the Spatialisateur project (IRCAM), Spatial Sound Framework (Aizu) and the DIVE Auralizer (SICS) have made great advances in the different areas of spatial sound description and presentation. Indeed some of the output from these projects has been incorporated into ISO standards[1].

In Virtual Reality, sound is generally allocated inadequate processing resources especially when compared with resources allocated to visual processing [7]. Generally, in order for an end user to participate in a virtual environment s/he will have to sacrifice some aspect of the sound rendering[2]. This might necessitate basic spatialization, such as simple binaural rendering, or the processing of only a fixed number of sources. The proposed system architecture will use the most appropriate type of spatial rendering at the best quality level available. The Scalable Spatial Sound Rendering System[3] framework enables this process to be carried out based upon criteria that insulate the end-user (listener) from noticeable drops in quality.

Within the context of Multimedia and Virtual Reality Begault has defined four classifications of spatial sound generation: Replication, Creation, Transmutation, and Representation [4]. The first three describe spatial scenes, where Replication is the equivalent of auralization, Creation is the generation of a new auditory experience, and Transmutation is the mixing of two auditory experiences. Representation is the switching of spatial perspectives, for instance a listener hearing the sound from the musician's perspective. Broadly speaking, these categories are used in the creation of auditory scenes in Virtual Reality.

Three models can be used to create these auditory scenes: Physical Model, Perceptual Model, and a hybrid of both [8,9]. Each rendering model has advantages and disadvantages associated with it. For the application of Virtual Reality the author maintains that a hybrid approach is the best solution in terms of processing requirements and authenticity/realism of the environment/source acoustic. Table 1 contains a list of attributes that influence the spatial rendering of a sound source or environment. The three models generally use these attributes to generate a spatial sound scene.

Table 1

Source	Medium	Environment	Listener
Location	Velocity	Reverberation	Shadowing
Directivity	Absorption	Reflection	Filtering
Intensity	Filtering	Occlusion	Cognition
			Visual Association

Quality Aspects

Quality aspects of spatial sound have been researched based upon physical parameters and perceptual factors [4, 5]. Both physical and perceptual models can be used within SSSRS to classify the spatial attributes of a sound source or virtual room. While the emphasis in this research will be upon the perceptual mode some consideration will be given to physical aspects and in particular to the interaction of both modes. An example of a physical determinant might include device limitations such as a lack of support for complex binaural rendering, or insufficient processing power to compute complex HRTF calculations.

Perceptual Model

Our understanding of the perceptual processing of sound has increased in recent years. One of the more active areas of investigation is in the classification of spatial dimensions of sound processing [10, 11]. Relative to the physical models employed in generating sound we are only recently beginning to use perceptual criteria for the rendering of sound and in particular spatial scenes, for instance at IRCAM and MPEG.

Research undertaken at IRCAM has resulted in a system called Spatialisateur that allows for the rendering of spatial scenes based upon perceptual parameters. Jot et al. [12] developed a high-level abstraction layer that interfaces with underlying algorithms used to generate the physical rendering. The auditory scene is divided into three categories, Source Perception, Room Perception, and Late Room Decay and uses the following perceptual parameters:

Table 2

Source Presence	
	Warmth Brilliance
Room Presence	
	Running reverberance Envelopment
Late Reverberance	
	Heaviness Liveness

MPEG-4

The development of a standardised spatial sound rendering system for scene description languages has been a slow process [8,9]. This culminated with the introduction of version 2 of MPEG-4 which contains a sound spatialization paradigm called 'Environmental Spatialization of Audio' (ESA). At a higher-level, ESA can be divided into a Physical Model and a Perceptual Model.

[1] Specifically, Spatialisateur and DIVA have had an input into the MPEG-4 standard.

[2] This is in the context of a non-distributed VR system.

[3] The development of the Scalable Spatial Sound Rendering System (SSSRS) is currently a work in progress. This paper is a Position Paper that gives an overview of the goals and architecture of the system. Not all aspects of the system have been fully implemented at the time of writing.

Previously, version 1 of the MPEG-4 standard rendered spatial sound using physical criteria only. Whilst this is desirable in a virtual environment it is quite limited. Virtual scenes are not constrained by physical laws and properties; therefore it was necessary to introduce a perceptual equivalence of the physical model. Another motivating factor for the use of a perceptual model is that not all users/listeners desire accurate spatial rendering – some place more emphasise on the ambience of the environment. To this end, MPEG-4 v2.0 introduced two new perceptual Nodes; PerceptualScene and PerceptualSound (see Appendix A for details).

Rault et al, point out the merits of the perceptual approach in a recent document to the MPEG group:

> "A first advantage we see in this concept is that both the design and the control of MPEG4 Scenes is more intuitive compared to the physical approach, and manipulating these parameters does not require any particular skills in Acoustics. A second advantage is that one can easily attribute individual acoustical properties for each sound present in a given virtual scene." [13]

The principles of the perceptual model are drawn from research carried out on the Spatialisateur project (as described above), and additional elements are derived from Creative Lab's Environmental Audio Extensions (EAX) and Microsoft's DirectSound API [14]. Using the perceptual model, each sound source's spatial attributes can be manipulated individually, or an acoustic pre-set can be designed for the environment (only *relative* source positions and orientations are considered in this model).

Fields such as 'Presence', 'Brilliance', and 'Heavyness' are used to configure the room/object's acoustic characteristics. In all, there are nine fields used to describe, in non-technical terms, the spatial characteristics of a room or a sound object. These fields have been derived from psycho-acoustic experiments carried out at IRCAM (Spatialisateur Project). The experiments consisted of listening tests where listeners were asked, "to quantify the perceptual dissimilarity of sound fields reconstructed artificially in an anechoic room with frontal direct sound" [13]. Of the nine subjective fields, six describe perceptual attributes of the environment, and three are perceived characteristics of the source. Table 3 lists the perceptual parameters for both Environment and Source.

Table 3 Perceptual Parameters in MPEG 4 v2.0

Environment Fields	**Source Fields**
LateReverberance	Presence
Heavyness	Warmth
Liveness	Brilliance
RoomPresence	
RunningReverberance	
RoomEnvelopment	

It can also be noted from Table 3 that the last three fields of the Environment Fields and all of the Source Fields are dependent upon the position, orientation and directivity of the source.

The validity of this approach could be questioned in terms of its subjectivity, for example, the choice of words such as 'Warmth' and 'Brilliance'. However, the use of subjective terms as acoustic parameters, in this context, is to enable the non-specialist to create a spatial sound scene with convincing acoustic properties.

More recently, work undertaken by Pellegrini et al. has focused on the creation of low-cost algorithms for Auditory Virtual Environments (AVEs). According to Pellegrini "The aim for a perceptually motivated design of an AVE is to define the most relevant auditory features for an application and then derive the needed physical elements to assure a well-suited representation for that application." [15] Interesting issues are raised particularly between the interplay of 'physical space and perceptual space', including diffusion (both temporal and spatial), distance perception and cognition.

Related Research

Currently there are a small number of research projects focusing on low-cost spatial sound rendering systems, of which the following are note worthy: Mercator, NAVE, and SLAB. However, to the best of the author's knowledge there does not exist any project or research that involves a scalable architecture for the rendering of spatial sound information.

Mercator [1,16]

This research project was established to develop a non-visual interface to the X Window System[4], including its dependant application, for visually impaired programmers [Mynatt & Edwards, 1992]. In an attempt to make spatial sound rendering accessible to the non-research community, the Mercator group designed a system solely on the basis of reducing computational overhead by sacrificing quality. The basic system started off with anechoic recordings captured at 50kHz sample-rate and filter duration of 512 samples (10.24ms). Computational gains were achieved by:

- Resampling audio material to 32kHz, => filter duration is reduced to 328 points
- As the bandwidth between 12kHz and 16kHz in the HRTF was deemed inaccurate allowed for the sampling rate to be further reduced to 24kHz => filter duration of 247 points.
- Long periods of silence (up to 3.4ms) preceding the impulse response and up to .54ms at the end of each filter sample were removed reducing the filter length to 139 points.
- With aggressive filter windowing the duration can be further reduced to 128 points[5].

A combination of all of these options would reduce the computation from circa 200 million convolution points per second to 6.7 million points per second. A considerable saving in processing terms at the expense of audio quality.

NAVE [17]

NAVE is a low cost auditory display system by Georgia Tech Virtual Environment Group. The system is based upon a standard multimedia desktop PC and a multiple speaker array. An interesting property of this system is its use of a moving bass, which is steered across four zones embedded in the NAVE floor. This is used to increase the inter-modal interaction by generating audio-tactile effects using the vibrations of the moving bass to reinforce the tactile modality.

SLAB [2]

Sound Lab (SLAB) is a research project undertaken by NASA, Raytheon STX Corporation and the San Jose State University Foundation. Its primary goal is to produce a low-cost software-based tool for developing experiments for the study of spatial hearing. The system achieves processing gains by smarter signal processing algorithms, parameter interpolation (e.g. in the HRTF database) and reduced filter resolution. The emphasis has been on reducing the complexity of physical quantities as opposed to perceptual parameters. According to Wenzel " The goal of the system here, Sound Lab (SLAB), is to provide an experimental platform with low-level control of a variety of signal-processing parameters…"[2].

ARCHITECTURE

The architecture of this system is object oriented by design. This modular approach to design makes the system extensible, easy to maintain, and is compatible with the structure of a scalable

[4] X Window System is a Window Manager that enables developers to put a graphical user interface onto a Unix environment.

[5] Interestingly, filters this short are used with the Convolvotron at 50kHz.

framework. It can also be integrated into existing EAIs (External Authoring Interfaces).

Framework

The framework proposed is dependent upon feedback from both the physical system resources and various perceptual settings as determined by the developer or the user. The feedback, in the form of a set of parameters, is then used to determine the level of spatial rendering sophistication. In trying to establish a framework for spatial rendering, Burgess identified eight cues that influence the localization of sound sources [1]:

- IDT
- Head Shadow
- Pinna Response
- Shoulder Echoes
- Head Motion
- Vision
- Early Echo Response
- Reverberation

Looking closer at these cues, one can identify natural groupings within the set. For instance, Pinna Response and Shoulder Echoes combine to produce the HRTF model, whereas Early Echo Response and Reverberation are characteristics of a room's acoustic signature and source position. IDT (Interaural Delay Time) and Head Shadow correspond to ITD (Interaural Timing Difference) and to IID (Interaural Intensity Difference) respectively.

Scalable System

The use of scalable architectures is not a new idea. Indeed it is a common approach in computer systems design. An obvious example of a scalable architecture is employed in the ISO's multimedia standard MPEG- 4.

"Devices can have differing access speeds depending on the type of connection and traffic. In response, MPEG-4 supports scalable content, that is, it allows content to be encoded once and automatically played out at different rates with acceptable quality for the communication environment at hand." [18]

In computer networking an example of a scalable architecture is the Quality of Service (QoS). In essence, QoS guarantees delivery of a predetermined level of data quality, or as in the case of multimedia, has a mechanism for gracefully reducing the quality of media data. The reduction of media quality is based upon the tolerance levels of the perceptual system. An example of this can be found in media streaming – if the network traffic is high the quality of the visual data is reduced first before any attempt is made to reduce the quality of the audio. The reason for this is that our perceptual system is more tolerant of errors in visual information than it is with audio information [19].

Profiles and Levels

This concept is straightforward and has been implemented in a number of other media delivery systems. Basically, users are classified according to a particular profile. Within each profile the user chooses a level that has a predefined range of quality settings. For instance, if the user is only interested in the accuracy of sound localization then s/he can sacrifice the realism of auralization or room simulation (e.g., the number of reflections, etc.).

Normally, as with MPEG, profiles are implemented with respect to hardware. For instance, the same content can be rendered for different hardware profiles; e.g. 3G[6] enabled phones, or desktop computers. In the context of this research, profiles are applied to the end user type and are therefore subjective in nature. The author has devised a simple three-profile arrangement, shown in Figure 1.

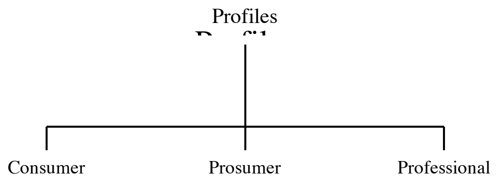

Figure 1

Within each category of user there are three available levels of quality, increasing in rendering complexity/accuracy (Figure 2).

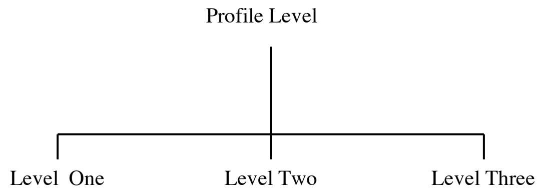

Figure 2

[6] 3G hones are Thrid Generation phones that take advantage of the increased bandwidth available in the UMTS protocol. These devices have been ear-marked for interactive multimedia mobile computing.

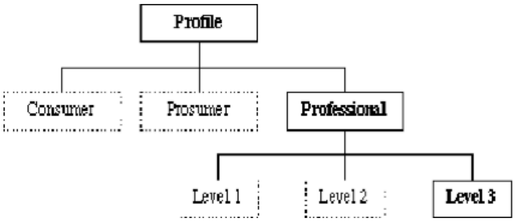

Figure 2 Example of a User Profile & Level

The framework can scale the auditory scene by integrating physical and perceptual quantities. For instance, if the context of the virtual environment is an online meeting space then accurate auralization of the virtual room is not necessarily a priority. Therefore a perceptually motivated representation of the room would suffice whilst the emphasis will be placed on the accuracy of localization (using HRTFs) of the source.

There are several techniques available for localising the source within an auditory scene (in the context of this research only headphone based binaural rendering is considered). At a most basic level a simple head model using ITDs and IIDs can be used to position the source. This can be further enhanced using reverberation, for instance to help to externalize the source. For a more accurate localization HRTFs can be employed.

The HRTF (Head Related Transfer Function) can be described as a mathematical model of the impulse response of a listener's ear. HRTF filters are based upon Finite Impulse Response Filtering (FIR) and takes the form of:

Equation 1 FIR filtering equation

$$y(n) - \sum_{m=0}^{M} h(m)x(n-m)$$

HRTFs are used in pairs[7] and can be individualized or non-individualized. Individualized HRTFs are captured using probe microphones that record the frequency response of the user's inner ears. This information is then used to build a database of filter coefficients that are used later to filter monophonic signals to give the impression of location. Non-individualized HRTFs are a collection of HRTF sets that were captured using 'expert' ears. These represent the average response of listeners judged to have good hearing. These HRTF measurements are deemed satisfactory for the general population. According to Wenzel, " The main characteristic features of the HRTF are consistent enough such that one such set of filters may be suitable for a large portion of the population" [20].

Individualized HRTFs have been shown to localize sound sources accurately [21]. However, this technique has a number of drawbacks including: difficulties generating individualized HRTFs, density of the database (the more calculations used the greater the density of the database), and the increased processing time required could introduce latencies into the system. To lessen the processing overheads a reduced database set could be considered as an alternative, however this means more interpolations are used which will in turn lead to a less accurate localization.

As with the individualized HRTFs, non-individualized sets have associated problems. It is generally accepted that within particular applications non-individualized HRTFs can increase the confusion in front-back reversals and decrease localization accuracy [4].

Within the context of the Framework[8], localization scaling can be determined by the following example setting:

Table 4

Level	Technique
Professional	Individualized HRTFs
Prosumer	Non-individualized HRTFs
Consumer	Simple Spherical Head Model

Other factors that arise from the use/non-use of these techniques include the phenomenon of 'Inside-the Head' localization (also referred to as lateralization). This is considered to be one of the main drawbacks of using binaural headphone based rendering. This can be overcome by using head tracking and with the addition of a measured amount of reverberation [22].

Inter-Modal Influences

Virtual Reality consists of multiple media types that play on the different perceptual modalities. Environmental and other modal cues can influence our perception of the spatial characteristics of a sound source or the spatial impression of the environment. Visual association is an area where there can be a strong influence/interaction between the aural and visual modalities. Arising from this interaction is the phenomenon known as the 'McGurk Effect' [1].

According to Slaney, "Vision can change the acoustic perception… With our eyes closed, we hear a synthesized voice saying 'ba'. When we open our eyes, and watch the artificial face, we hear 'va'. The acoustic signal is clearly 'ba', yet the lips are making the motions for 'va'. Thus our brains put together these conflicting information sources and, for this sound, trust the information from the eyes."[23]

One can even go so far as to strengthen a weak aural cue with a visual cue. For instance if an inferior localization technique was employed it could be supplemented with strong visual cues. In terms of the framework this would result in reduced processing and relatively little change in the subjective localization of the source.

Ερρορ Τολερανχε

A mechanism for achieving savings in computational costs is to take advantage of the high level of tolerance our perceptual system has to signal errors. Non-professionals, generally, tend to be more tolerant, or less discerning, of systems with reduced quality. This is particularly apparent when those systems rely upon the perceptual resolving powers of the user. VHS, a popular medium for delivering video, compromises the quality of the original video material; this principle is also evident in lossy audio compression schemes such as MP3 (MPEG1, Layer 3).

Within the framework error tolerance is dependent upon the Profile of the user. For instance, if the VR application were simulating a room response and the Profile was that of a Consumer, then the accuracy of localization would be reduced as our perceptual faculties are tolerant of localization errors. In this example, the system, based upon knowledge of localization blur, would sacrifice the accuracy of source location for an approximation of its position.

"The concept of 'localization blur' reflects the fact that auditory space is less differentiated than the space in which sound sources exist. The auditory system possesses less spatial resolution than is achievable using physical measuring techniques." [5] Under optimal conditions, the smallest possible change of position of the

[7] one for each ear

[8] These settings are only considered in relation to headphone reproduction.

sound source that produces a just-noticeable change (JND[9]) of position of the auditory event is 1º (the most precise area of spatial acuity is 0º azimuth and elevation).

Note: There are other considerations, for instance, localization blur is also dependent upon the type of source material (e.g., impulse vs. broadband noise, etc.) and its frequency/spectrum.

However, our localization accuracy is dependent upon the position of the source. For instance, as the source is moved away from the frontal position the JND begins to increase. In the median plane there are very few interaural differences to aid in localization. Without these vital cues JNDs on the scale of 4º to 19º have been recorded. [5]

TECHNIQUES

Prioritisation

In complex environments with more than one sound being rendered the requirements placed upon the system might be too demanding and result in audio drop-outs. Prioritisation is a simple process of prioritising sound sources within the environment. So, for example, speech could be allocated a high priority for spatialization while ear-cons (auditory version of icons, e.g. beeps and clicks) could be realised using simple panning techniques. This approach is very important when dealing with collaborative spaces as the objective is to create an effective communication medium, hence speech must have the highest priority of the various audio signals [24, 6].

LOD

Level of Detail is a process borrowed from 3D graphics systems [25]. VRML implements LOD (in visual rendering) as part of its scalable model. This is based on the principle that from a distance the level of detail an object possesses is rather limited; as one moves closer to the object the LOD is increased and so on until eventually the user is presented with an object that is high in visual detail. VR languages manipulate many of the deficiencies in our visual perceptual system [25] and it is only recently that this approach is being considered in an audio context [26].

This is a rather simple concept but one that integrates effectively into a scalable system. In this model one can assign different levels of localization accuracy or, as in the case of auralization, acoustic realism. For instance, if the distance of the user is far from the source then a low level of detail is realised/rendered. As the user moves towards the source a higher LOD is rendered. This is done in discrete steps, with different 'snap-shots' or settings being rendered. As there is no interpolation of distance values in LOD a coarse transition from one level to the next is produced (this would only be used for the most basic of Profiles and Levels).

Scheduling & Space Subdivision Techniques

This technique applies to a dynamic VR environment. The principle is that each area of a user's space (360ºs) is partitioned for optimal rendering. For instance, sound emanating from behind a user would not be rendered with the same fidelity as a source positioned in front of the user. This can be realised, for instance, by reducing the density of the HRTF set for a particular region and by using coarser interpolation points.

As our rate of movement is slower than our rate of audition, areas of our spatial environment can be scheduled, or cached, in anticipation of the user traversing a neighbouring subdivision. This technique is very common in the visual domain particularly in the field of computer games.

IMPLEMENTATION

The development system for the framework is be based upon the standardised Virtual Reality Modelling Language (VRML). VRML does support sound rendering and includes a basic spatial sound model, however it is generally accepted that it is too basic for most interactive VR applications [8,9]. VRML is highly extensible and is well suited to proprietary extensions. The author is aware of two research initiatives that are addressing the audio limitations of VRML[10]. However both projects are based upon predefined non-scaling systems.

HRTF

When choosing the type of HRTF set for this research a number of considerations were taken into account. These are concisely summed up by Shinn-Cunningham in the paper 'Learning Reverberation: Considerations for Spatial Auditory Displays' [24]. "When designing a spatial auditory display, there are many tradeoffs to consider: whether it is necessary to employ individualized HRTFs, the sampling density of the HRTFs to be stored and used, the sampling rate and length of the HRTFs to be used, whether to include realistic, distance-dependent HRTFs in the simulation, etc."

For non-individualised HRTF sets the binaural reproduction cannot be perfectly exact. The main perceptual consequence on the auditory impression over headphones is that sounds positioned in the frontal sector will often be perceived more elevated and possibly somewhat closer than intended. This problem can be overcome by using a Head Tracker (discussed later) and techniques based upon visual association.

As this framework is designed to work within the context of a collaborative virtual environment, where the number of participants varies it was decided to use non-individualized sets of HRTFs (MIT[11] set).

OTHER CONSIDERATIONS

Head Tracking

Head Tracking is an important tool in a dynamic virtual environment. Apart from the obvious advantages it brings to the visual presentation it is also important in the spatial rendering of sound. According to Burgess "The lack of these [head-related] cues can make spatial sound difficult to use. We tend to move our heads to get a better sense of a sound's direction. This 'closed-loop' cue can be added to a spatial sound system through the use of a head-tracking device." [1]

Recent research has shown that the use of Head Tracking reduces reversals by a ratio of 2:1 [22] and there is evidence that it assists in the externalisation of sources that would otherwise be located 'inside-the-head'. Another area where Head Tracking is helpful is in the simulation and control of the Doppler Effect and to resolve source-listener movement ambiguities. Blauert terms this 'persistence' - "In connection with spatial hearing, the term 'persistence' refers to the fact that the position of the auditory event can only change with limited rapidity. Under appropriate conditions the position of the auditory event exhibits a time lag with respect to a change in position of the sound source. Persistence must always be taken into consideration when using sound sources that change position rapidly." [5]

EVALUATION

The research is currently at the stage where various testing methodologies are being examined. The author is currently devising a test suite that requires the user to (a) determine the intelligibility of the speech content [28], (b) localise source position, (c) assess the effectiveness of cross-modal synergy [29] and (d) assess the affect of reduced auralization upon the

[9] Just Noticeable Difference

[10] DIVE (Distributed Interactive Virtual Environments) and SSF (Sound Spatialization Framework)

[11] Created by Bill Gardner and Keith Martin.

communication experience. The context for this test suite is a multi-user collaborative environment where the emphasis is upon communication.

One of the primary aspects of a scalable system is that it is deterministic. The quality of the output must be predicable at all times for the system to function properly. As there are two modes for affecting quality, the evaluation of the system will be divided into formal (system or physical) and perceptual evaluation. As stated previously, maintaining the intelligibility of speech is the main consideration of a collaborative VR system.

Formal Testing

Formal testing, or System testing, involves verification of physical quantities, such as the order of reflections rendered and the spectral content of the audio. Other calculations undertaken include the computation of system latencies and the complexity of the rendering system, for instance calculating the number of taps used by a filter to spatialize a sound event. This type of testing is easily verifiable and does not involve any form of user response in its results.

Perceptual Testing

The Perceptual evaluation will include subjective listening tests - and will examine both spatial characteristics of the sound and inter-modal influences – the author is currently researching this area.

Applications

With the increasing growth of the Internet and CyberSpace[12], Virtual Reality in its many forms is set to become as pervasive as television. There are as many hardware configurations possible as there are users, hence it is imperative that a robust system can handle the many permutations of configurations without requiring a rewrite of the scene description for each arrangement. SSSRS facilitates this by dynamically scaling the content based upon the derived parameters.

Typical applications for this scalable framework include Mobile-VR systems and Personal VR systems based upon standard multimedia PCs. This will enable a user with a basic system, for instance a 3G Mobile PDA to engage in a virtual environment alongside a user with a fully immersive VR system. The ability to participate in a virtual environment can extend to more specific applications such as Virtual Teleconferencing, Collaborative Spaces, Telemedicine, etc.

CONCLUSIONS

The author has established a need for a scalable spatial sound rendering system. This system is based upon both physical and perceptual parameters. Having established the system architecture and the Profile-based framework the project is now progressing onto its development stage, this will be followed by a series of evaluations that should produce some meaningful results. The applications of this system are many and should prove very beneficial to VR scene developers.

ACKNOWLEDGEMENTS

The author would like to thank the following for their advice and direction: Durand Begault, NASA Ames Research Centre and Jyri Huopaniemi, Nokia Research Centre.

REFERENCES

[1] Burgess, D., "Techniques for Low Cost Spatial Audio", 1994, Located at IRCAM ftp site

[2] Wenzel, E. et al., Sound Lab: A Real-Time, Software-Based System for the Study of Spatial Hearing, AES, 108th Convention Paris, France, February, 2000

[3] Lehnert, H., 'Fundamentals of Auditory Virtual Environments', in Artificial Life and Virtual Reality, Thalmann N. and Thalmann. (eds.) 1994, D., Wiley & Sons, Chichester, UK.

[4] Begault, D.R., 3-D Sound for Virtual Reality and Multimedia, AP Professional, 1994.

[5] Blauert, J., Spatial Hearing, The Psychophysics of Human Sound Localization MIT Press (Rev. Ed.), 1997

[6] Shilling, R. and Shinn-Cunningham, B.G., ' Virtual Auditory Displays', in Handbook of Virtual Environment Technology", 2000, K. Stanney (ed.), Lawrence Erlbaum, Associates, Inc., NY.

[7] Brice, R., Multimedia & Virtual Reality Engineering, Newnes, Oxford, 1997

[8] Murphy, D., Spatial Sound Description in Virtual Environments, Proceedings of the Cambridge Music Processing Colloquium, September, 1999.

[9] Väänänen, R. and Huopaniemi, J., Spatial Presentation of Sound in Scene Description Languages, 17th AES International Conference, Munich, April 1999

[10] Bregman, A.S. Auditory Scene Analysis:Perceptual Organization of Sound, Bradford Books, 1994

[11] Berg, J. & Rumsey, F., In Search of the Spatial Dimensions of Reproduced Sound: Verbal Protocol Analysis and Cluster Analysis Of scaled verbal descriptors, AES, 108th Convention Paris, France, February, 2000

[12] Jot, J-M., Efficient models for reverberation and distance rendering in computer music and virtual audio reality, Proc. 1997 Int. Computer Music Conference, pp. 236-243, September 1997.

[13] J-B. Rault et al., "Audio Rendering of Virtual Room Acoustics and Perceptual Description of the Auditory Scene", ISO/IEC JTC1/SC29/WG11, M4222, 1998

[14] Jot, J-M., Ray, L. and Dahl, L., "Extensions of Audio BIFS: Interfaces and Models Integrating Geometrical and Perceptual Paradigms for the Environmental Spatialization of Audio", ISO/IEC JTC1/SC29/WG11, M4223, 1998

[15] Pellegrini, R.S., Comparison of data- and model-based simulation algorithms for auditory virtual environments, Proceedings of the 107th AES Convention, Munich, 1999.

[16] Mynatt, E. & Edwards, W.K., The Mercator Environment: A non-visual interface to X Windows and Unix workstations, ACM Symposium on User Interface Software and Technology, USIT '92, 1992

[17] Wilson, J. et al., The NAVE: Design and Implementation of a 3D Audio System for a Low Cost Spatially Immersive Display, http://www.cc.gatech.edu

[18] Konen, R., MPEG-4 Multimedia For Our Time, IEEE Spectrum, Vol. 36, No. 2 February, 1999

[19] Dworetzky, J.P., Psychology, 4th Ed., West Publishing Company, New York, 1991

[20] Wenzel, E., Localization in Virtual Acoustic Displays, Presence: Telepresence & Virtual Environments Vol.1 Num.1 pp 80-107, 1992

[12] A term used to denote an online virtual environment where participants can interact. William Gibson first used the term in his book Neuromancer. Gibson is also credited with the invention of the term Virtual Reality.

[21] Begault, D.R. and Wenzel, E.M., Headphone Localization of Speech. Human Factors, Vol. 35, pp361-376, 1993

[22] Begault, R. et al. Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source, AES, 108th Convention Paris, France, February, 2000

[23] Slaney, M., A Critique of Pure Audition Computational Auditory Scene Analysis, http://developer.apple.com, 1997.

[24] Shinn-Cunningham, Barbara, 'Spatial Auditory Displays', in International Encyclopedia of Ergonomics and Human Factors, 2000, W. Karwowski, Ed. London: Taylor and Francis, Ltd.

[25] Reddy, M., Perceptually Modulated Level of Detail for Virtual Environments, Ph.D. Thesis, 1997, University of Edinburgh.

[26] Martens, W., Understanding 3D Audio Rendering Through Analogy to 3D Graphic Rendering, http://www.u-aizu.ac.jp

[27] Shinn-Cunningham, Barbara, Learning Reverberation: Considerations for Spatial Auditory Displays, ICAD, 2000, Atlanta, GA.

[28] Evans, M.J., The Perceived Performance Of Spatial Audio For Teleconferencing, Ph.D. Thesis, 1997, University of York.

[29] Storms, R.L., Auditory-Visual Cross-Modal Perception, Army Research Laboratory, Georgia Institute of Technology, 1999.

BIBLIOGRAPHY

Baldis, J.J., Effects of Spatial Audio on Communication During Desktop Conferencing, M.Sc. Thesis, 1998, University of Washington.

Begault, D.R., Challenges to the Successful Implementation of 3-D Sound, J. Audio Eng. Soc. Vol. 39, No. 11, November, 1991

Billinghurst, M., & Kato, H., Collaborative Mixed Reality. In Proc. 1st Int. Symp. on Mixed Reality, ISMR'99

Billinghurst, M., et al., WearCom: A Wearable Communication Space, http://www.hitl.washington.edu

Cohen, E., Technologies for Three Dimensional Sound Presentation: Issues In Subjective Evaluation of The Spatial Image, http://www.cudenver.edu/aes/tech/TECH3D.HTML, 1999

Cohen, M., and Wenzel, E.M., The Design of Multidimensional Sound Interfaces, Aziu University, Japan

Dalenbäck, B., Kleiner, M. and Svensson. P., Auralization, virtually everywhere. In the 100th Convention of the AES, Copenhagen, May 1996. Preprint 4228 (M-3).

Ellis, D.P.W., A Perceptual Representation Of Audio, M.Sc. Thesis, 1992, Massachusetts Institute of Technology.

Dale, W., A Machine-Independent 3D Positional Sound Application Programmer Interface To Spatial Audio, The Proceedings of the AES 16th International Conference, April 1999, p160.

Evans, M.J., Tew, A.I., and Angus, J.A.S., Spatial Audio Teleconferencing – Which Way Is Better?, 1997, University of York.

Gardner, W.G., 3D Audio and Acoustic Environment Modeling, http://www.wavearts.com, March, 1999

Gardner, W.G., The Virtual Acoustic Room MSc Thesis, MIT, 1992

Gaver, W., The Affordances of Media Spaces for Collaboration, in Proc. CSCW'92, Toronto, Canada, Oct.31-Nov.4, 1992, New York: ACM Press, pp.17-24

Gibson, William. Neuromancer, Victor Gollancz Ltd., London, 1984

Herder, J., Sound Spatialization Framework: An Audio Toolkit for Virtual Environments, http://www/u-aizu.ac.jp/~herder, 1998

Hindus, D., Ackerman, M., Mainwaring, S., Starr, B., Thunderwire: A Field study of an Audio-Only Media Space. In Proc. CSCW'96, Nov. 16th-20th, 1996, New York: ACM Press.

ISO/IEC FDIS 14496-3 sub5, Structured Audio, 1999.

ISO/IEC FDIS 14496-3, 1999.

ISO/IEC JTC/SC24 IS 14772-1 "The Virtual Reality Modeling Language (VRML97) Information technology – Computer graphics and image processing – The Virtual Reality Modeling Language (VRML)", http://www.vrml.org/Specifications/VRML97/, April1997

Jot, J-M., J-B. Rault, ISO/IEC JTC1/SC29/WG11, N2578, "Extensions of Advanced AudioBIFS: a Perceptual Paradigm for the Environmental Spatialization of Audio", 1998

J. Signès, "Binary Format for Scene (BIFS): Combining MPEG-4 media to build rich multimedia services". http://www.cselt.stet.it/mpeg/documents/koenen/signes.zip

Java 3D API Specification, version 1.2_alpha, Sun Microsystems, Inc., August 1999

Jot, J-M. and Rault, J-B., "Extensions of Advanced AudioBIFS: a Perceptual Paradigm for the Environmental Spatialization of Audio", ISO/IEC JTC1/SC29/WG11, M4449, 1999

Jot, J-M., Synthesizing Three-Dimensional Sound Scenes in Audio or Multimedia Production and Interactive Human-Computer Interfaces, 5th International Conference: Interface to Real & Virtual Worlds, May 1996.

Jot, Jean-Marc, Efficient Models for Reverberation and Distance Rendering in Computer Music and Virtual Audio Reality, ICMC (International Computer Music Conference), Greece, 1997

Jot, Jean-Marc, Synthesizing Three-Dimensional Sound Scenes in Audio or Multimedia Production and Interactive Human-Computer Interfaces, 5th ICMC, France, 1996

Lee, M.D. and Burgess, D.A., The Perception of Location Using Synthetic Auditory Localization Cues: Accuracy and the Effects of Stimulus Bandwidth http://www.cc.gatech.edu/~burgess, 1994

Mandeville, J., et al., GreenSpace – Creating a Distributed Virtual Environment for Global Applications, 1995, HILT, University of Washington.

Moore, J., An Introduction To The Psychology Of Hearing, Academic Press, London, 1997.

Murphy, D., A Review of Spatial Sound in the Java 3D API Specification (v.1.2), Technical Paper, Institute of Sound

Recording (IOSR), University of Surrey, UK, 1999.

Murphy, D., Audio Quality and Capacity Issues in Network Design. The Proceedings of the AES UK Conference Moving Audio, May 2000, p9-17.

Pellegrini, R.S., Perception-Based Room Rendering for Auditory Scenes, Proceedings of the 109th AES Convention, Los Angeles, USA, 2000.

Pope, T. et al., The Use of 3-D Audio in a Synthetic Environment, http://www.sics.se/~stp, 1993

Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R., Creating Interactive Virtual Acoustic Environments, J. Audio Eng. Soc., vol. 47, No. 9, 1999

Sawhney, N. and Schmandt, C., Design of Spatialized Audio in Nomadic Environments., http://media.mit.edu, 1997

Scheirer, E., et al., "AudioBIFS: The MPEG-4 Standard for Effects Processing", COST-G6 Workshop on Digital Audio Effects Processing (DAFX'98), Barcelona, Nov. 1998

Schmandt, C., Mullins, A. AudioStreamer: Exploiting Simultaneity for Listening. In Proc. CHI'95 Conference Companion, May7-11, Denver Colorado, 1995, ACM: New York pp. 218-219

Shen. L., et al., "Virtual Playground: Architectures for a Shared Virtual World", In Proc. ACM Symp. on Virtual Reality Software and Technology, 1998 (pp. 43-50), New York: ACM.

Steuer, I., Telepresence is defined as the experience of presence in an environment by means of a communication medium (Defining VR: Dimensions Determining Telepresence), Journal of Communication, 1993

Thalmann, N. and Thalmann, D., Artificial Life and Virtual Reality, John Wiley & Son, W. Sussex, London, 1994.

Thiede, T. et al., PEAQ – The ITU Standard for Objective Measurement of Perceived Audio Quality, J. Audio Eng. Soc., Vol. 48, No.1/2, January, 2000

Väänänen, R. and Huopaniemi, J., Update of advanced Audio BIFS: The Physical Approach, ISO/IEC JTC1/SC29/WG11, M4590, 1999

Väänänen, R., "Verification Model of Advanced BIFS (Systems VM 4.0 subpart 2)", ISO/IEC JTC1/SC29/WG11, N2525, 1998

Wenzel, E. et al., A software-based system for interactive spatial sound synthesis, ICAD, 1997

Wenzel, E.M., Localization in Virtual Acoustic Displays, Presence, vol.1, 1992, pp. 80-107.

APPENDIX A

MPEG-4 version 2 Advanced Audio Nodes

Physical Nodes

```
AcousticScene        {
exposedField         SFFloat    paramfs               0
field                SFVec3f    3DVolumeCenter        0, 0, 0
field                SFVec3f    3DVolumeSize          -1, -1, -1
exposedField         MFFloat    reverbtime            0
}

AcousticMaterial     {
exposedField         SFFloat    reffunc               0
exposedField         SFFloat    transfunc             1
exposedField         SFFloat    ambientIntensity      0.2
exposedField         SFColor    diffuseColor          0.8, 0.8, 0.8
exposedField         SFColor    emissiveColor         0, 0, 0
exposedField         SFFloat    shininess             0.2
exposedField         SFColor    specularColor         0, 0, 0
exposedField         SFFloat    transparency          0
}

DirectiveSound       {
exposedField         SFVec3f    direction             0, 0, 1
exposedField         SFFloat    intensity             1
field                MFFloat    directivity           1
exposedField         SFFloat    speedOfSound          340
exposedField         SFFloat    distance              100
exposedField         SFVec3f    location              0, 0, 0
exposedField         SFNode     source                NULL
exposedField         MFBool     useAirabs             FALSE
exposedField         SFBool     spatialize            TRUE
exposedField         SFBool     roomEffect            TRUE
}
```

Perceptual Nodes

```
PerceptualScene {
eventIn              MFNode     AddChildren           NULL
eventIn              MFNode     RemoveChildren        NULL
exposedField         MFNode     Children              NULL
Field                SFVec3f    BboxCenter            0, 0, 0
Field                SFVec3f    BboxSize              -1, –1, –1
exposedField         MFBool     UseAirabs             FALSE
exposedField         MFBool     UseAttenuation        TRUE
exposedField         SFFloat    RefDistance           1
exposedField         SFFloat    Latereverberance      TBD
exposedField         SFFloat    Heavyness             TBD
exposedField         SFFloat    Liveness              TBD
exposedField         MFFloat    RoomPresence          TBD
exposedField         MFFloat    RunningReverberanceTBD
exposedField         MFFloat    RoomEnvelopment       TBD
exposedField         SFFloat    Presence              TBD
exposedField         SFFloat    Warmth                TBD
exposedField         SFFloat    Brillance             TBD
exposedField         SFFloat    Fmin                  250
exposedField         SFFloat    Fmax                  4000
}

PerceptualSound {
exposedField         SFVec3f    direction             0.0, 0.0, 1.0
exposedField         SFFloat    intensity             1.0
exposedField         MFFloat    directivity           1.0
exposedField         MFFloat    omniDirectivity       1.0
exposedField         SFFloat    speedOfSound          340.0
exposedField         SFFloat    distance              1000.0
exposedField         SFVec3f    location              0, 0, 0
exposedField         MFFloat    relPParams            1.0, 1.0, 1.0, 1.0,
1.0, 1.0, 1.0, 1.0, 1.0
exposedField         MFFloat    directFilter          1.0, 1.0, 1.0
exposedField         MFFloat    inputFilter           1.0, 1.0, 1.0
exposedField         MFBool     useAirabs             FALSE
exposedField         MFBool     useAttenuation        TRUE
exposedField         SFInt      spatialize            FALSE
exposedField         SFInt      roomEffect            FALSE
exposedField         SFNode     source                NULL
}
```