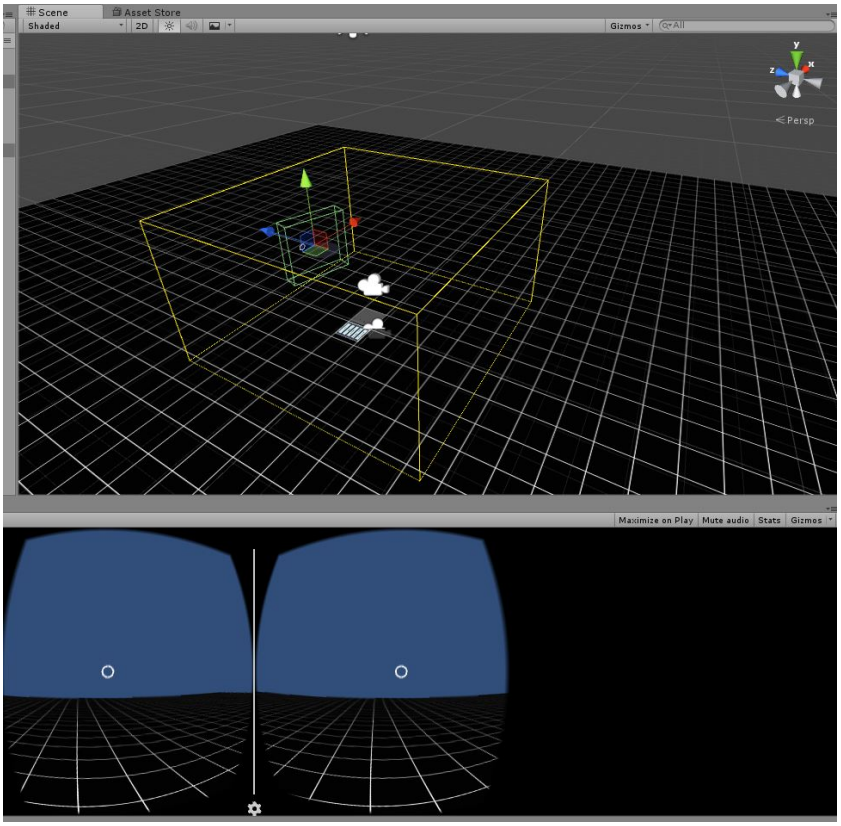


Figure 0.1: A screen-shot of the test interface design within unity, both in third person 'observer' and first person 'subject' views

---

An Attempt to Implement a Listening Test to Quantify the Impact of Diffuse Reverberation Level on Localization in Modern VR Applications

---

Simon Durbridge

January 11, 2017

CONTENTS

LIST OF FIGURES

1 INTRODUCTION

As technology has continued to improve, virtual reality (VR) experiences have become more widely available to the general public. A number of companies [1] have released software development kits(SDK), allowing content creators to begin developing VR applications with little specialist knowledge and minimal tool sets. A significant part of these kits are the spatial audio handling components, that allow content creators to tailor an immersive 'three-dimensional' auralization for their applications with reduced complexity and required skills.

As denoted in various studies in the literature [**?**] [**?**] [**?**] [**?**]; the capacity of humans to analyse a surrounding environment through various auditory analysis methods (including but not limited to Auditory Scene Analysis), is a significant evolutionary trait that may be key to creating truly immersive experiences in VR media. Room reverberation is not only a key component for auditory analysis of local terrain characteristics, but is also a significant contributor to auditory source localisation.

1.1 RESEARCH QUESTION

The aim of this study is to determine if there is a relationship between a subjects capacity to accurately localize a speech sound source in a VR environment, and the level of the diffuse component of room reverberation using a reverberation modelling algorithm for a virtual reality SDK (VRSDK) [**?**]. In this study the Google VRSDK will be used for experimentation.

Initially, the Google VRSDK will be introduced, and the reverberation engine will be evaluated. Following this, a listening test implementation using this VRSDK will be discussed. The results of a small preliminary test batch will be discussed, and finally some potential improvements to the method will be noted.

//Increases in available computing power, and great strides in research and development have brought a new surge of ////interest to virtual reality (VR) applications. //With the introduction of improved VR systems, such as Google Jump, Oculus, Vive and facebook360, using both high end //computers, and mobile phone technology to provide immersive visualisation; Further steps are being taken to provide // with immersive sound environments [**?**]. //These environments may be created in an attempt to emulate real places, or to characterise fictional places.

//A significant part of how humans identify with their surroundings, includes the perception of the reverberant characteristics of the surrounding area [**?**]. //Cues such as the timing and strength of early reflections help humans to localize sound sources. //Some VR application development platforms provide a simplified model for how reverberation behaves in an audio system, and though supporting research confirms the link between localization and early reflections [**?**]// [**?**] [**?**] [**?**], there are few examples of research into the required order of reverberation required specifically[2] for headphone based binaural VR.

[1] Notably; Google Jump [**?**], Oculus [**?**] / Facebook 360, HTC

[2] though Corey [**?**] undertook a similar study over loudspeakers

2 HYPOTHOSIS

2.1 SPATIAL PERCEPTION

Research into spatial audio for VR may have flourished in recent past, but the foundation of localization theory remains based on the concepts explored by Rayleigh in 1907 [**?**]. Continued research into the field has matured and evolved understanding of these concepts, and for an in-depth review of auditory localization please review Blauert [**?**]. Localization is the term given to the human capacity to binaurally (with two ears) localize sounds, laterally in the plane of the ears and in the median plane (up and down). A sources perceived location is determined by interaural time differences (ITD), phase differences, interaural level differences(ILD), and the physical listening apparatus itself (pinnae, head shape, torso, head tilting etc).

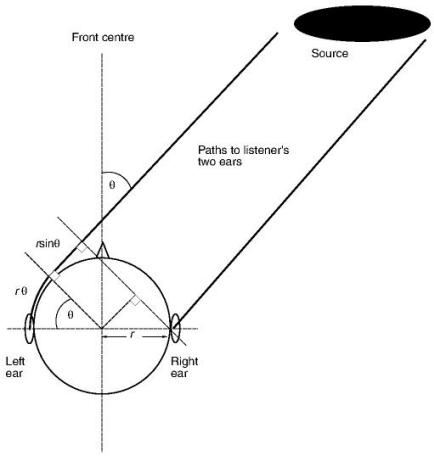


Figure 2.1: A basic model diagram of the ITD from a source to a listener given by $\frac{r(\theta+sin\theta)}{c}$ [**?**]

The ITD is the time difference between a sound arriving at each ear (not to be confused with phase), and is an important cue for localizing at frequencies below the wavelength relative to the size of a listeners head. The phase differences relate to the asynchronous behaviour of sound diffracting around the head also provides temporally motivated localization cues [**?**]. At wavelengths larger than at least half the circumference of the head, sound waves can defract around the head. Above this frequency humans become more dependent on ILDs and phase differences.

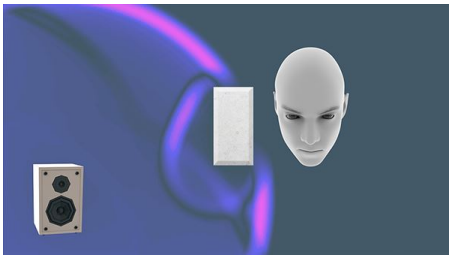


Figure 2.2: A still of visualization of a wave defracing around an obstacle in occlusion [**?**]

The ILD is the level difference between ears of a sound, and is more critical to localization of shorter wavelengths where head and torso shadowing is dominant. The pinnae and torso have a distinct filtering affect that assists humans in localizing sources in the median plane [**?**] [**?**] [**?**]. Another benefit of this filtering is that humans are able to some degree overcome the 'cone of confusion' that would occur when a sound source is in a position that would otherwise produce the same ITD and ILD as another position (illustrated below). Another key cue for localization in the median place, is that humans have a tendency to move and tild their heads, using the changes in what is heard to refine localization [**?**]. Head tracking in headphone based spatial audio has been shown the improve localization accuracy when incorporating early reflections after the floor reflection [**?**].

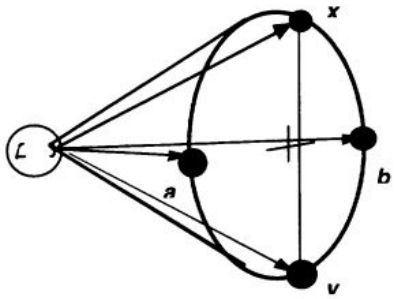


Figure 2.3: A diagram example of the cone of confusion. If a source were in location a, it would have the same ITD and ILD and thus would be ambiguous to position b. Positions x and y share a similar relationship in the vertical plane as oposed to the lateral plane of a and b [**?**]

The total system of localization effects from source to the entrance of the receivers ear canal(s) can be lumped into a head-related transfer function (HRTF), and can be translated into a head-related impulse response (HRIR). Upon synthesizing an HRIR filter, it is possible to reproduce audio with binaural cues over headphones or an appropriate speaker system, that gives listeners the impression of virtual source localization [**?**](and ideally externalization).

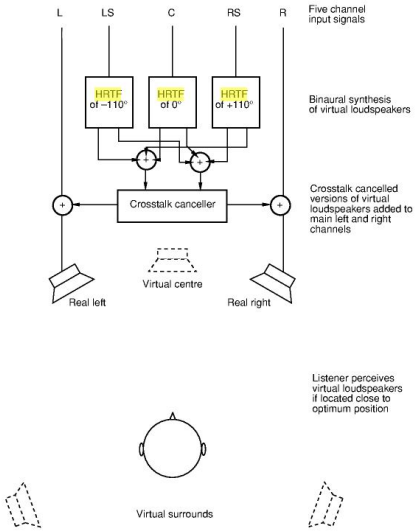


Figure 2.4: A diagram of a loudspeaker system set up for virtualization [**?**]

Auditory Scene Analysis(ASA) is the term given to the brains process of grouping and compartmentalizing sounds into components of the auditory 'scene' [**?**]. Through a complex process of continuous analysis and adaptation, the brain is able to accurately differentiate sounds into specific details such as how large a room is, what the wall materials are or which direction a complex source such as an oboe is facing. Bregman [**?**] suggests that a key part of the ASA process is learned via auditory steaming, which is essentially spectrogram analysis, comparing time and frequency based patterns to those already known, and associating those patterns as another object in the scene. As denoted in anecdotes such as that by Wiggins [**?**], humans use cues from reverberation to categorize the size of an enclosed space.

2.2 AMBISONICS FOR VIRTUAL REALITY

Many of the recent VR systems consist of a viewing headset that may be powered by a mobile phone, a games console or a computer. The headsets viewing system provides a stereoscopic vision of a 3d environment that is linked to a head tracking interface, allowing the user to explore the 'immersive' visual environment in 3d. For many of the new virtual reality platforms, ambisonics has become the signal format of choice [**?**]. This choice may be due to the flexability of a system that can encode and decode arbitrary numbers of input and output channels, while inherently maintaining the spatial nature of the content. Another benefit may be the compatability of ambisonics with concepts such as object based audio [**?**]. For a review of various 'spatial' audio system formats, please refer to [**?**]. Many of the newer VR platforms rely on headphones as the preferred audio system format. This may be in part due to the difficulty in producing HRTFs for multiple listeners over loudspeaker, though research in this area is continually improving [**?**]. Another benefit of the use of headphones is that head tracking can be applied to HRTFs to improve the spatial audio quality [**?**] in a relatively discrete package.

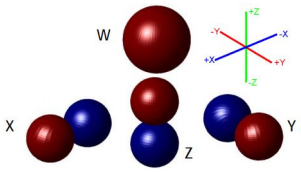


Figure 2.5: A plot of four ambisonic microphone polar patterns (for B format) [**?**]

Ambisonics in a simplistic description, is the spherical harmonic decomposition of a sound field into constituent components [**?**]. That is, by the combination and manipulation of sets of signals from coincident receivers (with figure-of-eight polar responses), with an omnidirectional receiver, it is possible to encode a 3 dimensional spatial sound field into an n channel signal format. For a discussion on different formats such as B, C, D, A and UHJ, please refer to Rumsey [**?**]. Perhaps the greatest benefit of the ambisonic system is that the encoding and decoding channel counts are not mutually exclusive i.e. it is possible to record a sound field in 3rd order using an appropriate sound-field microphone, and decode that signal for two channel headphone playback with the appropriate HRTFs [**?**] as could be used in a VR system [**?**]. This would allow for multiple virtual sound sources to be used to render a 3d sound field for an individual listener, and could be coupled with head tracking to create an auditory scene that changes as a listeners moves their head.

3 Experimental Method and Results

3.1 Reverberation

The reverberant sound field is the steady-state of diffusely scattered sound energy in a space due to the reflection of that energy from boundaries to a high order. Specifically, the amplitude of these reflections are such as to balance in amplitude with the steady state (source - decay) of the acoustic system, at or beyond the critical distance from a source (a classic analogy is given by Everest) [**?**].

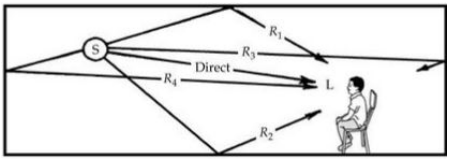
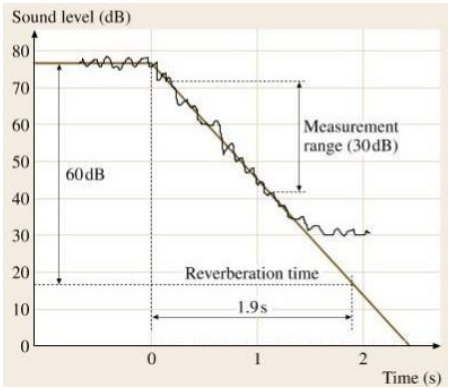
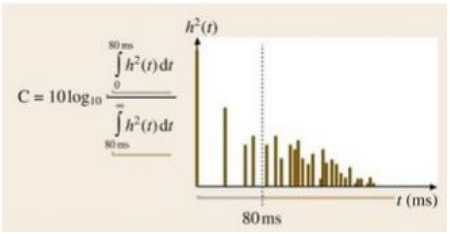


Figure 3.1: A conceptualisation of direct and reflected sound [**?**]

That is in contrast with low order reflections that may occur before the sound contributes to the reverberant field (early reflections), or may occur later and above the steady state amplitude (echos) [**?**]. Early and strong reflections are of significant interest in this study, due to the cues humans receive from perception of them. A reverberant sound field is often quantified by the decay time from steady state, to an amplitude of the steady state level $-60dB$. This is commonly described as the RT_{60} and was proposed by WC Sabine in 1900. For a basic description of reverb time calculation, please see Everest [**?**] page 153. The use of RT_{60} as the preferred metric of decay time is valid, assuming that the acoustics system is linear and time-invariant. A more comprehensive description of reverberation and overview of the associated parameters is given by Rossing [**?**].



(a) An example graph of reverberation decay time RT_{60}



(b) An example of the calculation of Clatiry score

Figure 3.2: Graphics pertaining to quantification of reverberant sound [**?**]

Further to this, the ratio of direct sound amplitude from a source to the amplitude of the reverberant field at any place in the sound field is also of interest [**?**] and is described as the **D**irect to **R**everberant (D/R) enegry ratio. This can be quantified in terms of amplitude at a point in the sound field using the modified Hopkins-Stryker equation [**?**]:

$$L_T = L_W + 10log\left(\frac{QMe}{4\pi D_x^2} + \frac{4N}{S\bar{a}Ma}\right) + K$$

where:

1. L_T = total sound pressure
2. L_W = sound power level of source
3. $\frac{QMe}{4\pi D_x^2}$ is a description of the sound source direct radiation properties
4. $\frac{4N}{S\bar{a}Ma}$ is a description of the reverberant field properties
5. $K = \frac{\rho c}{400}$ relating to the transfer of sound through air

This equation is particularly useful, as it can be rearranged to calculate important parameters such as the amplitude of the direct and reverberant sound fields, but also presents the intrinsic relationship between source parameters, receiver location and sound field geometry(including sound absorption properties). The D/R ratio is intrinsically linked in this equation to the **critical distance**(D_c), the distance from sound source at which the reverberant sound field is equal in amplitude to direct sound energy from the source. At distances beyond D_c the reverberant sound field dominates the direct sound in amplitude. A third parameter of interest is the clarity score, that is determined by the ratio between direct sound and an integration of the sound received directly from the source for some arbitrary [**?**] time (commonly 80ms). Another parameter of note in calculation of the reverberant nature of a geometry is the mean free path (**MFP**), that is the average distance between reflections in space [**?**]:

$$\textbf{MFP} = 4\frac{Volume}{Surfacearea}$$

As a wave must propagate further between reflective surfaces, the amplitude of higher order reflections is increasingly diminished. This is compounded by the absorption characteristics of the boundaries of the geometry. Thus, a higher mean free path may result in a reduction in a a more periodic impulse response(IR) with a reverberant tail of lower amplitude, compared to a much smaller space with the similar absorption characteristics.

3.2 PERCEPTION OF REFLECTIONS

There is a plethora of research and understanding of the objective quantification of reverberation [**?**], less so about the subjective effects [**?**] beyond the seminal studies of early reflections by early pioneers (even before Haas) [**?**] [**?**]. It may be argued that many facets of human auditory perception relate in some way to the perception of reverberation. The concepts of interest in this paper are those linked with localization and perception of the auditory scene. Within these concepts are Haas/Precedence effect, and auditory masking.

The Haas effect describes that the first wave-front perceived by a listener, determines the source localisation in an auditory scene. The following reflections received within a 0.7 to 35ms window are integrated with the initial sound, such that the sound is given the impression of increased loudness, an increase in source width and tonal shifts of the percieved sound [**?**]. Beyond 40mS, following strong reflections are heard as echoes [**?**]. Begault [**?**] suggests that the precedence effect is the intrinsic mechanism of the human auditory system that allows for the localization of sound sources in a reverberant sound-field. Begault also suggests that there is a link between inter-aural time difference(ITD) perception for lateralization, and changes in apparent sound source 'width', suggesting a 5us to 1.5ms ITD 'window' within which a sound sources lateral location is determined. Another ascpet of note, is that if a sound source is occluded and the reflected sound is louder than the direct sound, the precedence effect is diminished [**?**].

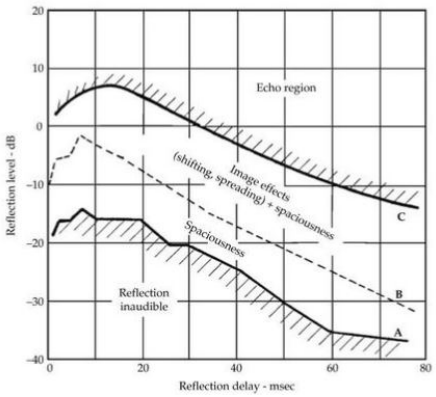


Figure 3.3: A plot of the window regions of precedence between reflection 'level' and delay [**?**]

Many of the texts referenced in this section of the report do not discuss auditory masking in great detail[3], though it may be of considerable importance when considering perception of noise-like sound (the late portion of reverberation). Auditory masking can be conceptualised as the time-envelope dependent blocking of new sounds being perceived, due to some other sound having already excited the inner ear. Specifically, sounds having already excited a portion of the bascilar membrane are thus stopping other sounds from being perceived as separate, that would otherwise excite the same portion of the bascilar membrane [**?**]. The basilar membrane is often regarded as to be discretized into critical bands or equivalent rectangular bandwidths (ERB), and these bandwidths are different depending on sound level as well as centre frequency. General works by Moore [**?**] should be reviewed for a more in-depth discussion of masking. The relevance of masking in this context, is that continuous noise is

[3]Masking is noted several times by Begault, but not discussed

often used as a test signal in masking experiments, and steady-state reverberation may be similar to noise and thus cause masking of reflections and general reverberation. As such, creation of realistic reverberation may be important for masking in a natural way as opposed to having a spectrally dense or high amplitude artificial reverb that may cause excessive masking.

Sigfried Linkwitz [**?**] also discussed the idea that human perception of sound in enclosed spaces is powerful enough that given an appropriate pair of sound sources, a person can subconsciously differentiate between reflected sounds and the stereo sound-stage created by the two sound sources. This is not disimilar to abstractions made in much of the literature such as Begault [**?**] or Wiggins [**?**], who describe typical scenarios pertaining to walking into large spaces and using cues from reflections to determine that the space they have entered is actually large. The relevant link between reverberation and associated perceptual faculties is with auditory scene analysis. Reverberation perception (particularly precedence) is intrinsically linked to humans capacity to analyse the auditory characteristics of their surroundings. This is compounded by Begault [**?**] who notes the significant breadth of reverberation research for concert hall acoustics, suggesting that the quality of reverberation may be an important part of audio for VR and multimedia.

Is it possible that more realistic reverberation modelling may provide a more realistic scene for our brains to analyse? It has been shown by Corey [**?**] that for a 5 channel system, adding simulated early reflections are enough to allow listeners to accurately localize sound sources with greater certainty. This experiment utilised a 5 channel loudspeaker system. A traditional panning law was used to pan direct sound between 2 speakers, and a polarity restricted cosine law was used for distributing early reflections through all speakers in the system. Although this study showed a positive relationship between added early reflections and certainty, it was suggested that under the test circumstances localization blur was not significantly effected. Karjalainen *et al* [**?**] also undertook a study into the perception of late reverberation, using a perceptual model to reinforce listening test data. In a similar way some of the VR packages, Corey's early reflections were generated using an image source method. As noted in acoustic modelling research such has those by Mourik and Murphy [**?**], geometric or ray based simulations do not produce accurate results at low frequencies. The image based method described by Begault [**?**], and used by Corey is further limited as will be discussed below. Some reverb calculation methods are discussed in the next section.

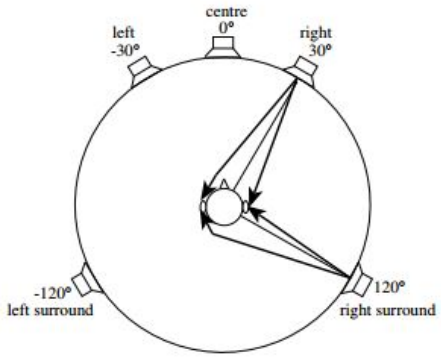


Figure 3.4: A diagram of the layout for Corey's listening test [**?**]

3.3 REVERBERATION ALGORITHMS

3.3.1 FILTER BASED REVERB

A relatively simple electronic method of simulating reverberation is that derived by a feedback loop. An analogy of this system is that of a tape deck with a looped reel of tape [**?**]. A delay presented between the read and write heads of the tape deck, with the write head writing not only new signals to the tape but an attenuated version of the signal read by the read head producing a decaying echo. In the frequency domain, such a system presents a comb filtered response as a version of a signal is interfering with itself but staggered in time.
When a series of these reverberation filters are cascaded together with varying delay times, it is possible to create a slightly metallic sounding reverb [**?**]. An improvement to this method is to add a single pole low pass filter to the delay block for simulating attenuation, and to add a number of all-pass filters in series after the delays have been summed in order to reduce periodicity of the delay signal [**?**].

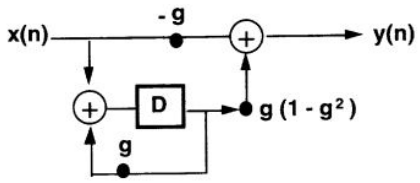


Figure 3.5: A block diagram of a Schroeder and Logan all pass filter [**?**]

A third method of creating a reverb effect is to convolve a signal with the IR of the system to be emulated [**?**], or with a decaying 'shaped' noise function [**?**]. This process in discrete digital systems involves multiplying a current sample and $n-1$ previous samples with n coefficients, thus adding varying amounts of the previous portion of the signal to the 'current' sample. This method is also required for digitally applying the measured IR of a real space to some audio for simulation purposes, and the same again for a modelled or calculated IR using one of the modelling techniques discussed below.

3.3.2 MODELLING

Geometric based modelling methods are those based on assumptions of plane wave propagation and ray based geometry to calculate the time of flight and of some sound between a source and a receiver. These methods include image source modelling, ray tracing, cone tracing and variations [**?**]. Geometric modelling can be used for creating IRs suitable for auralization [**?**]. This report will focus on the previously mentioned image source technique.

The image source(IMS) technique uses a rectilinear geometry with some basic frequency-independent absorption characteristic, and conceptually multiplies that geometry in symmetric fashion around the origin geometry [**?**]. The effective time of flight and number of virtual boundaries passed indicates some absorption, reflection order and time delay between the source and receiver. Thus, the IR of the geometry can be calculated with accuracy up to the order of the simulation i.e. how many times the

geometry is mirrored. This method cannot handle an irregularly (not piecewise linear [**?**]) shaped geometry, any obstacles, and assumes that the ray represents part of a plane wave whose wavelength is much smaller than any part of the geometry [**?**]. This method is however relatively simple, fast to compute and is potentially a good method for calculating arrival times and amplitudes of early reflections for a binaural pair of receivers down to the rooms Schroeder frequency.

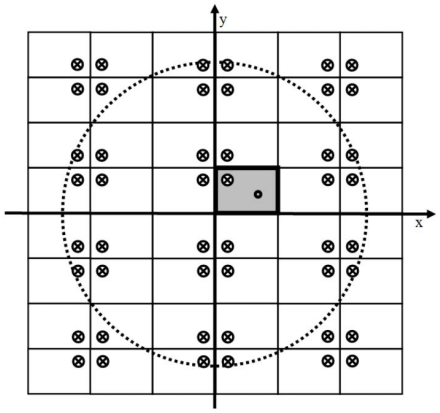


Figure 3.6: A diagram of a cascaded geometry in the IMS method [**?**]

Using the calculated reverb decay time RT_{60} to set up an IIR reverb and the early reflection results from an IMS simulation [**?**], it is potentially possible to combine the two responses to create and artificial reverberation for VR environments that provides early reflections for localization, but is not to such a high order as to become computationally expensive [**?**].

Wave based modelling methods directly model spaces by calculating solutions to a discretized partial differential equation across the geometry [**?**] [**?**]. For linear and room acoustics, the equation used is often the linearised Navier-Stokes or Euler diffusion style equations that satisfy the conservation laws. Depending on the solving scheme used, the geometry will need to be discretized to at least 2 points per wavelength (in line with Nyquist theorem) and potentially beyond 10 points[4] per wavelength to satisfy the Courant condition [**?**]. The result is that problems become exponentially large and thus expensive to solve up to high frequencies. However, some methods such as the PTSD have been used to calculate up to low-mid frequencies in real time [**?**] [**?**] and may in future become suitable to calculating low frequency portions of hybrid models in VR [**?**].

[4] particularly for explicit finite different and finite element methods

4 RESULTS ANALYSIS AND EVALUATION

4.1 EXPERIMENT AIMS

The aim of subsequent research to this report, is to define how much of an impact the accuracy of reverberation synthesis has on the perception of source localization in a VR application. Is there a direct correlation between the order of accuracy of a simulation, and test subjects capacity to localize sounds in the space? The core method behind this experiment is to determine if there is an average improvement in localization i.e. a reduction in localization blur, by comparing the change in error for a group of test subjects across varying stimulus and environments. Comparing group error for localizing sources per instance of 'large hall with speech with 3rd order early reflections' with the same of 5th order reflections, may give an idea of how well early reflections order can improve a subjects capacity to localize in that environment. The difference between this method and other examples, is that the VR system will allow subjects to turn and localize sources in the way best for them. Ideally, this should give for negligible localization error in all instances, and also discounts any effects from sources being in the cone of confusion of a subject.

Similar experiments have been carried out by numerous groups, two prime examples being Corey [**?**], Angel [**?**], both of which showed added early reflections improved localization in a loudspeaker based localization test. In similar experiments; Boerum [**?**] undertook a series of listening tests to show that early reflections improved perception of lateral movement, and Saji [**?**] showed that a process of incorporating reflections into HRTFs improved localization under anechoic conditions.

4.2 EXPERIMENT METHOD

Users will be asked to evaluate 3 different listening environments, with two stimuli, of two people talking in the space and a stereo pop music track respectively. The test results will be analysed for localization error, front-back reversals. There will be five potential reverb scenarios to evaluate in each environment:

- convolving with an IR of a real space
- a derived reverb with no early reflections
- a low order image source IR combined with a derived reverb
- a high order image source IR
- a control instance of the signals without spatial reverb

The locations of sources in the scene will be randomized so that source location patterns cannot be learned, thus skewing perception. The source amplitudes will remain consistent for all tests, to allow for more realistic distance biasing. The distance between the sources will be fixed, to ensure that the source width remains consistent. This may allow for more consistent analysis of localization error. Due to this source size restriction, the three environments will be relatively large e.g. a barn, a theatre and a small warehouse. The sources will not have directional sound propagation characteristics, as this is not handled in the IMS model.

5 Experiment Review

Subjects will be asked to perform an evaluation test similar to those mentioned above, in which they use a simple VR headset and headphones to evaluate an auditory scene with a visual component such as a 360 degree image. Subjects will be able to look around the scene, but not move from that position within the scene. The visual scene will be used to give viewers a bias towards room size and thus reverb time. The subjects will be asked to identify in which directions the sources are situated, without a visual cue as to the location of the sources. Subjects will be shown an initial 'training' scene without reverb, that will prepare them for the upcoming stimulus, but the test will be blind to what is being tested for until the post-test debrief. The tests will be taken in real-world conditions i.e. in variable environments and with non-strict hardware conditions, this is to ground the experiments to scenarios where these basic VR systems are intended to be used.

6 Conclusion

The intention of this paper was to explore some of the key psychoacoustic and physical principals behind the localization of sounds in relation VR, so that a listening test may be designed for evaluating the effect of different reverb synthesis methods on localization in VR applications. Key texts by Begault [**?**] , Wiggins [**?**], Blauert [**?**] and Rumsey [**?**] denote the importance of reverb in the context of localization and perception with relation to spatial audio.

A number of sources in the literature such as Corey [**?**], Johnson [**?**], Usher [**?**] examine the influences of reverberation on localization, and aspects related to localization and reverberation perception are discussed above. A listening test is described that would aim to test the influence of reverberation on localization in a VR context. Further work will go on to implement the listening test described above in more detail, along with implementing the reverb synthesis methods above and further research into the leading of localization perception by VR imagery.

7 References