Acoustic Rendering and Auditory-Visual Cross-Modal Perception and Interaction

Vedad Hulusic[1], Carlo Harvey[2], Kurt Debattista[1], Nicolas Tsingos[3], David Howard[4], and Alan Chalmers[1]

[1]**University of Warwick**
[2]**Bournemouth University**
[3]**Dolby Laboratories**
[4]**University of York**

Corresponding author:
Carlo Harvey[2]

Email address: charvey@bournemouth.ac.uk

ABSTRACT

In recent years research in the three-dimensional sound generation field has been primarily focussed upon new applications of spatialised sound. In the computer graphics community the use of such techniques is most commonly found being applied to virtual, immersive environments. However, the field is more varied and diverse than this and other research tackles the problem in a more complete, and computationally expensive manner. Furthermore, the simulation of light and sound wave propagation is still unachievable at a physically accurate spatio-temporal quality in real-time. Although the Human Visual System (HVS) and the Human Auditory System (HAS) are exceptionally sophisticated, they also contain certain perceptional and attentional limitations. Researchers, in fields such as psychology, have been investigating these limitations for several years and have come up with findings which may be exploited in other fields. This paper provides a comprehensive overview of the major techniques for generating spatialised sound and, in addition, discusses perceptual and cross-modal influences to consider. We also describe current limitations and provide an in-depth look at the emerging topics in the field.

1 INTRODUCTION

Hearing is one of the fundamental attributes humans use for a wide variety of reasons: from spatially locating and identifying objects to acting as a reaction mechanism. If virtual environments are to achieve their full potential as a representation of reality, a comprehension of all aspects related to audition is required. This paper focuses on two key areas of acoustics for virtual environments: the correct simulation of spatialised sound in virtual environments, and the perception of sound by the Human Auditory System (HAS) including any cross-modal auditory-visual effects.

The research challenge of spatialised sound is to accurately simulate propagation of sound waves through a 3D environment. This is motivated by possible use in a wide range of applications such as concert hall and architectural design (Naylor, 1993), advanced multimedia applications in Virtual Reality to enhance presence (Calvin et al., 1993; Macedonia et al., 1995) and, more recently, immersive video games (Moeck et al., 2007; Raghuvanshi et al., 2007; Grelaud et al., 2009). The computer graphics community has recently been involved more closely with this research. This is because spatial sound effects can generate an increased sense of immersion when coupled with vision in virtual environments (Durlach and Mavor, 1995) and furthermore can aid a user in object recognition and placement; identification and localisation of disparate sounds; and generating conclusions pertaining to the scale and shape of the environment (Blauert, 1997).

Improved spatialised sound for full immersion is not the sole outcome of computer graphics research into acoustics. An emerging area of computer graphics in the last decade is perceptually based rendering and auditory-visual cross-modal interaction. Limitations of the human sensory system have been used in order to improve the performance of rendering systems. Auditory and visual limitations have been exploited in order to decrease the auditory (Tsingos et al., 2004; Moeck et al., 2007) or visual (Cater et al., 2002; Kozlowski and Kautz, 2007; Ramanarayanan et al., 2007, 2008) rendering complexity with

no or little perceivable quality difference to a user. Moreover, it has been shown that it is possible to increase the perceptual quality of a stimulus in one modality by stimulating another modality at the same time (Mastoropoulou et al., 2005a; Harvey et al., 2010). This can be used for improving the perception of a material quality (Bonneel et al., 2010), Level-of-Detail (LOD) selection (Grelaud et al., 2009) or for increasing the spatial (Mastoropoulou et al., 2005a; Hulusic et al., 2008) and temporal (Mastoropoulou et al., 2005b; Hulusic et al., 2009, 2010a,b) quality of visuals by coupling them with corresponding auditory stimuli.

While there have been surveys on acoustic rendering in the past (Funkhouser et al., 2002; Manocha et al., 2009) in the field of computer graphics and on aspects of cross modality (Shams et al., 2004) within the field of psychology, this is, to the best of our knowledge, the first to bring these fields together and to outline the use of cross-modal perception within computer graphics. The only similar work can be found in the book chapter (Kohlrausch and van de Par, 2005), with the focus on multi-media applications rather than computer graphics.

2 THE HUMAN AUDITORY SYSTEM AND ACOUSTICS

This section serves as a brief introduction on the HAS and sound. It describes the concepts and methods used throughout the rest of the document.

2.1 Human Auditory System

The Human Auditory System (HAS) comprises three parts: the ears; the auditory nerves; and the brain. The ear consists of the outer ear, middle ear and inner ear.

The outer ear is the visible part of the ear. The most noticeable, a shell-like part, is the pinna. The pinna is mostly used for sound localisation. A sound, reflected off of the pinna, is further channelled down the ear (auditory) canal. The ear canal ends with the tympanic membrane, which transmits the incoming vibrations to the middle ear. The middle ear is an air-filled chamber, which connects the outer and the inner ear. On one side, the tympanic membrane closes the “entrance” to the middle ear. Similarly, another tiny membrane, called the oval window, separates the middle ear from the liquid-filled inner ear. The three smallest bones in the human body, called ossicles, bridge these two membranes. The liquid in the inner ear produces more resistance to the wave movement than the air, because of its higher molecular density. Therefore, the ossicles, besides transmitting, also amplify the vibrations from the outer ear into the inner ear. The inner ear consists of few parts and two major functions: maintaining the balance and orientation in space; and frequency and intensity analysis. More details can be found in (Moore, 1982; Blake and Sekuler, 2006).

2.2 Sound

Since sound is an oscillation of pressure transmitted in a wave, modelling sound propagation has been, for the most part, similar to modelling light propagation. However there are several key distinctions that deserve some forethought and expansion upon:

Speed of sound: The speed of sound (c) varies depending on the medium being traversed through. This is approximated by the Newton-Laplace equation, where C is the coefficient of stiffness of the medium and ρ is the density of the medium being traversed given as $c = \sqrt{\frac{C}{\rho}}$. Therefore the speed of sound increases with material stiffness yet decreases with density of the material. However there are more controlling factors that impact the speed of sound depending on the medium, temperature and humidity in gases, temperature and salinity in liquids, shear forces in solids and various states of ions and electrons within plasmas.

Gas is the medium upon which most simulation techniques focus and as such it is important to note the effect of temperature on the speed of sound. Within a normal working range of temperatures (-35°C to 25°C) it is possible to use the following formula to derive the speed of sound in air, where θ is the temperature of the air being propagated within and given as $c_{air} = 331.3\sqrt{1 + \frac{\theta}{273.15}}$. At normal room temperature(20°C) c_{air} works out to be $343.2 m \cdot s^{-1}$. Whilst that is a practical formula for air there is a more general formula for the speed of sound in ideal gases and air where γ is the adiabatic index (the ratio of specific heats of a gas at a constant-pressure to a gas at a constant-volume), p is the pressure and ρ is the density: $c = \sqrt{\gamma \cdot \frac{p}{\rho}}$.

Wavelength: Sound requires a medium to travel through. This is either a solid, liquid, gas or plasma. Sound cannot travel through a vacuum. Through liquids, gases or plasmas, sound travels in longitudinal waves, waves that have the same direction of vibration as direction of travel; oscillations happen in the same plane. This is the case with solids however sound can also travel through solids as a transverse wave; a wave whose oscillations are perpendicular to its direction of travel. Sound waves are often simplified to sinusoidal plane waves, one of whose key properties is wavelength. The wavelength γ of a wave travelling at constant speed v of frequency f is given by: $\gamma = \frac{v}{f}$. Human hearing is limited to frequencies between 20Hz and 20kHz, although the upper limit will decrease with age as the ability to discriminate between sounds, for example speech phones, also worsens. In normal air with a speed of $343.26 m \cdot s^{-1}$ the standard range of wavelength that is audible lies between 17.15 and 0.01715 metres. As a result acoustical propagation tends to reflect specularly. This assertion remains true until a source of distortion to an incoming sound wave is larger than that of the sound signals wavelength impinging upon it. Sound waves also diffract when object size is similar to the wavelength, whilst small objects do not really impact upon the wave-field to a large degree. This means that simulation techniques need to be able to account for and find specular reflections and diffractions and also account for geometry large or small in the environment at a versatile range of wavelengths.

Impulse Gateway: A reverberation from a given sound can be broken down into three distinct parts that a human ear can attribute to a single source: direct sound, early reflections and late reflections. These will be discussed in more detail later in section 3.1. However, it is important that the ear is able to distinguish a sound and attribute it to a source later in the reverberation. The simulation must account for this and typically generates many more time dependant reflection paths than a simulation algorithm for light paths would. This is noticeable in applications such as concert hall design in which Impulse Gateways are typically many seconds in length.

Time and Phase Dependence: Waves which are out of phase can have very distinct impacts on each other should they be superimposed. If two waves with the same amplitude (A), frequency (f), and wavelength(λ) are traveling in the same direction, their amplitude depends on the phase. When the two waves are *in-phase*, they interfere constructively and the result has twice the amplitude of the individual waves (2A). When the two waves have opposite-phase or are *out-of-phase*, they interfere destructively and cancel each other out and the resulting amplitude is 0. As such, acoustical simulations need to consider the phase of the impingent wave upon a receiver when analysing contribution paths. This also means very accurate path lengths need to be computed such that the phase generated is accurate in relation to the wavelength of the impingent wave.

Attenuation: In acoustic attenuation the inverse distance law is always an idealisation in that it assumes a free-field, however when any reflection is involved the points within a previous free-field being traversed by the reflection will have a higher pressure level. However the inverse distance law is the first step in predicting the pressure level attenuation, where R is the position of the receiver in 3D space and S is the position of the sound source in 3D space given by $P(R) = \frac{P(S)}{r}$, where $r = \sqrt{(R_x - S_x)^2 + (R_y - S_y)^2 + (R_z - S_z)^2}$ and $P(\cdot)$ is the sound pressure at a given point in space. In addition to this attenuation, materials that are collided by a sound wave absorb some of the sound wave and this is dealt with via a frequency dependant absorption coefficient in some acoustic simulation techniques. This is shown in Equation 1. C_p is the frequency dependent complex pressure coefficient, Z is the specific acoustic impedance (a ratio of sound pressure to particle velocity at a single frequency) and Z_0 is the characteristic acoustic impedance of the medium (this is $413.3\ N \cdot s \cdot m^{-3}$ for air at room temperature). In contrast to Z, the quantity Z_0 depends only on properties of the medium and the speed of sound and is thus frequency independent.

$$C_p(\theta, f) = \frac{\frac{Z(f)}{Z_0(f)} \cos\theta - 1}{\frac{Z(f)}{Z_0(f)} \cos\theta + 1} \tag{1}$$

More simple, yet acceptable, methods exist using a scalar across frequency octave bands (128, 256, 512, 1024, 2048 and 4096 Hz). The absorption coefficient is given as the energy ratio between the absorbed and the incident energies. When a sound wave in a room strikes a surface, a certain fraction of it is absorbed, and a certain amount is transmitted into the surface. Both of these amounts are lost from the room, and the fractional loss is characterised by a frequency dependant absorption coefficient $\alpha(\omega)$ which can take values between 0 and 1, 1 being a perfect absorber. $A_f(\omega)$ is the pressure of the wave reflected from

the surface at a given frequency: $\alpha(\omega) = 1 - |A_f(\omega)|^2$. Such that an absorption coefficient of 0.9 at a frequency of 4kHz would reflect 10% of the pressure of the incoming wave into the exiting wave at 4kHz. Frequency dependant material profiles can be created for various absorbers, either through industrial or independent measurements or through analytic formula's such as proposed by Sabine: $T = 0.161\frac{V}{A}$ with the duration T of the residual sound to decay below the audible intensity, starting from a 1,000,000 times higher initial intensity, where V is the room volume in cubic metres, and A is the total absorption in square metres.

Part I
Spatialising Sound

3 MODELLING SOUND PROPAGATION

In this section we present a brief overview of the spatialisation pipeline. A set of primitives defining the size, scale and shape of the environment is a necessary input to any sound modelling schema, combined with a source signal and location within that environment for the signal to emanate from, along with a listener position. This information precludes the generation of an *Impulse Response*. This impulse response encodes the delays and attenuations that emulate reverberations to be applied to the source signal. The next step is *Convolution*. Convolving the impulse response with the source signal outputs a spatialised sound signal that can be used via an *Auditory Display* in order for audition.

3.1 Impulse Responses

A Room Impulse Response (RIR) is the output of a time-invariant environment to an input stimulus. This input stimulus attempts to emulate a Dirac Delta or unit impulse function. Auralising a sound for a particular sound source, receiver, and environment can be achieved by convolving an RIR with an anechoic source signal to model the acoustical effects of sound propagation within that environment (Kuttruff, 1991). This auralisation remains accurate only for the particular input position (sound source) and output position (listener) that the RIR simulates. An impulse response can be distinguished by three sub categories: direct sound (R0), early reflection or diffractions (R1|R2) and late reflections or diffractions (R3).

Direct Sound (R0) represents the immediate sound wave reaching the receiver, the first impulse allowing the detection of the presence of a sound. Early Reflections and Diffractions (R1|R2) is the section of an impulse response categorised by the waves that arrive within a time frame such that the number of distinct paths remains discernible by a listener. This is less than 2000 paths. R1 typically contains paths unique to [0:40]ms and R2 (40:100]ms. The early reflection and diffraction phase presents most of the information about wave pressure and directionality (Begault, 1994; Cremer and Müller, 1978; Hartmann, 1997) allowing a listener to discern some information about the shape and scale of the environment that the sound is reverberating within (Begault, 1994; Hartmann, 1983; Nielsen, 1993; Wagenaars, 1990). This section of a response profile must be modelled as accurately as possible due to this.

Late Reflections and Diffractions (R3) form the part of an impulse response that represents an overall decay in the profile of the response whereby the number of paths impinging upon the receiver outweighs the human ability to distinguish unique paths: when the sound waves from the source have reflected and diffracted off and from many surfaces within the environment. Whilst this section is incredibly important to the profile of the impulse response, especially in the case of responses with long gateways such as cathedrals, the modelling techniques used to generate it need not be as accurate as ones used to simulate Early Reflections and Diffractions (Ahnert, 1993; Savioja et al., 1996).

3.2 Convolution

Convolution, in this context, is the process of multiplying each and every sample in one audio file with the samples from another waveform. The effect is to use one waveform to model another. This results in $y_n = i(n) \otimes x(n) = \sum_{k=-\infty}^{\infty} i(k)x(n-k)$, where y is the output waveform, x_n are samples of the audio to be modelled, i_k are samples from the impulse response (the modeller). Whilst typically this process is reserved within the spatialisation pipeline for an anechoic sound source convolved with an impulse response to model the acoustical properties of a particular environment, it should be noted that the

technique is more general than this and can be used in many scenarios; for example statistics, computer vision, image and signal processing, electrical engineering and differential equations.

3.3 Rendering Spatialised Sound

At a fundamental level, modelling sound propagation addresses the problem of finding a solution to an integral equation expressing a wave-field typically at two distinct points, a source to a listener. The computer graphics community will find this transport problem is similar to global illumination, which is described by Kajiya's rendering equation (Kajiya, 1986). Similarly, sound rendering is based on the physical laws of sound propagation and reflection, in this case: the wave equation, described by the Helmoltz-Kirchoff integral theorem (Born and Wolf, 1999).

Sound scattering waves from source to a receiver introduce a multitude of different pathways: reflections, refractions, and diffractions upon different surfaces within the environment. For sound simulations these effects are used to generate a filter to apply to a source signal that reconstruct the acoustical properties of the reflection, refraction and diffraction of sound waves upon surfaces within the environment.

Just as radiance transfer is applied in computer graphics to solve the rendering equation for global illumination (Kajiya, 1986), Siltanen et al. (Siltanen et al., 2007) have proposed the *room acoustic rendering equation* as a general model for ray-based methods, shown in Equation 2.

$$L(x',\Omega) = L_0(x',\Omega) + \int_G R(x,x',\Omega) L(x, \frac{x'-x}{|x'-x|})\,\mathrm{d}x \tag{2}$$

where $G \subset \mathbb{R}^3$ is the set of all surface points in the enclosure and $L(x',\Omega)$ is the time dependant outgoing radiance from point x' in direction Ω. In the case of an area source, L_0 represents the radiance emitted by the surface. For point sources, L_0 is considered as the primary reflected radiance instead of defining the source as a part of the surface geometry to make the analysis more convenient.

$$R(x,x',\Omega) = V(x,x')\rho(\frac{x'-x}{|x'-x|}, \Omega : x')g(x,x') \tag{3}$$

where $\rho(\Omega_i, \Omega_o : x')$ is the BRDF term at point x', $g(x,x')$ is a geometry term required in acoustics to describe the effects of the geometry on the acoustic energy flow and is defined as :

$$g(x,x') = \lfloor n(x) \cdot \frac{x'-x}{|x'-x|} \rfloor \lfloor n(x') \cdot \frac{x'-x}{|x'-x|} \rfloor \frac{S_{|x-x'|}}{|x-x'|^2} \tag{4}$$

where $S(\cdot)$ is the operator representing propagation effects on sound radiation over distance r. Finally, $V(x,x')$ is a visibility term to make the model applicable to non-convex geometries:

$$V(x,x') = \begin{cases} 1 \text{ when the ray between } x \text{ and } x' \text{ is unobstructed} \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

3.3.1 Ray-Based Methods

Krokstad et al.'s Ray-Traced Acoustical Room Response (Krokstad et al., 1968): A Ray-Traced method, as first introduced to the computer graphics field in the form of ray casting (Appel, 1968) and recursive ray tracing (Whitted, 1979), finds reverberation paths via tracing rays through an environment from the audio source until a sufficient number of rays have reached the receiver. The receiver is typically modelled as any geometric primitive however a sphere is practically the most widely used and arguably best choice as it serves as an omnidirectional sensitivity pattern and yields the best chance for the listener ray collections to provide a statistically valid result. Indirect reverberation can be accounted for due to ray-surface intersections being able to sample specular reflection, diffuse reflection, diffraction and refraction stochastically. However, the infinitely thin nature of the sampling strategy results in aliasing and miscounted diffraction paths.

To model an ideal impulse response all sound reflection paths should be discovered. A Monte Carlo approach to ray tracing samples these paths to give a statistical approximation to the ideal impulse response and whilst higher order reflections can be discovered by ray tracing, there is no guarantee all the sound paths will be considered. When first published the resources available to the ray tracing algorithm were quite archaic, the algorithm has scaled well with resources and now has some more interactive implementations.

3.3.2 Image Source

Allen et al.'s Image Source Method (Allen and Berkley, 1979) involved mirroring sound sources across all planes in an environment, constructing virtual sources. For each virtual source a specular reflection path is computed by intersecting a line from source to receiver in an iterative manner. Recursively following this method produces specular reflection paths up to an arbitrary order. Thus the contributing images are those within a radius given by the speed of sound times the reverberation time. This guarantees all specular paths will be found; however only specular paths can be found, complexity grows exponentially and the technique is best suited to rectangular rooms. A simple Sabine material absorption coefficient was used originally. In addition, it should be noted that whilst this could have been frequency and reflection angle dependant guided absorption, for computation speed it was ignored.

Borish's extension of the Image Source Method to Arbitrary Polyhedra (Borish, 1984): the basic principle of the image model is that a path of specular reflections can be represented by a straight line connecting the listener to a corresponding virtual source that has been mirrored iteratively over geometry. When this idea was applied to a rectangular room (Allen and Berkley, 1979), a regular lattice of virtual sources ensued. Virtual source position is trivial to calculate in this format of polyhedra. Borish removes the dependency on rectangular shaped rooms with this method by introducing a set of heuristics to guide virtual sound source placement when reflecting across arbitrary polyhedra. Finding the image source within arbitrary geometry required more computation than that of a rectangle. The virtual image source can be found by travelling from the source location a distance 2d in the direction of the planar normal. d, the distance from the point to the plane, is given by $d = p - P \cdot n$ so that R, the position vector of the image point, is: $R = P + 2d \cdot n$. Borish speculated that each virtual source created must adhere to 3 criteria to be valid:

1. **Validity**: an invalid virtual source can be defined to be one created by reflecting across the non reflective side of the boundary.

2. **Proximity**: virtual sources further than a given distance away fail this criteria. This must be specified, else the generation of virtual source would never end.

3. **Visibility**: if the virtual source is visible to the listener it contributes and shouldn't be ignored. This is an involved process of computation especially as the iteration of generation delves levels. For full details on this elimination process please see (Borish, 1984).

Savioja et al. (Savioja et al., 1999): introduced a hybrid time-domain model for simulating room acoustics. Direct sound and early reflections are obtained using the Image Source method. Late reflections of an impulse response are considered generally as nearly diffuse, and are modelled appropriately as exponentially decaying random noise functions.

Late reflection artefacts are modelled using a recursive digital filter and the listener can move freely in the virtual space. This filter consists of n (typically 2,4,6,8 depending on resources) parallel feedback loops. A comb all-pass filter is within each loop which in effect produces an increased reflection density on the input direct sound signal. Whilst the late reverberation artefacts do not need to be modelled using accurate techniques, as in the case of early path reflections with directionality attributes, there are several key aims and heuristics to preserve the integrity of the late reverberation information provided by the feedback reverberater:

1. Produce a dense pattern of reverberations to avoid fluttering in the reproduction acoustic.

2. Simulate the frequency domain characteristics of a high modal density, whilst no mode outweighs another.

3. Reverberations time has to decay as a function of frequency to simulate air absorption effects.

4. Produce partly incoherent signals for the listener's ears to cause interaural time and level differences.

In an extension to Borish's Visibility stipulation this technique improves on this by preprocessing the set of virtual image sources such that $M(i, j)$ where surface i dictates if it is at least partially visible by surface j or not. This eliminates the need for sources reflected over these surfaces to be considered in visibility analysis should it be observed they are not visible. This eliminates a large set of the computation on virtual sources.

3.3.3 Finite Element and Boundary Element Methods (FEM and BEM)

Kludszuweit's Time Iterative Boundary Element Method (TIBEM) (Kludszuweit, 1991): Exact solutions of the wave equation are available only for certain enclosures of simple shape, bounded by rigid walls. These rigid walls have boundary conditions the solution must adhere to in simulation. For more realistic cases of acoustic transmission it is necessary to use one of FEM, BEM or TIBEM which are applicable to various shapes and materials of varying acoustical admittance. TIBEM works within the time domain iteratively calculating sound pressure and velocity on the boundary and at any point within the enclosure.

Kopuz et al.'s Boundary Element Method (Kopuz and Lalor, 1995):The boundary element integral approach to the wave equation can be solved by subdividing solely the boundaries to the environment, whilst also assuming the pressure is a linear combination of a finite number of basis functions on these subdivided bounding elements. By representing boundary surfaces as a set of panels and the boundary functions by a simple parametric form on each panel, the boundary integral equation is reduced to a linear system of equations and a numerical solution becomes possible. The main characteristic of the method is that only a mesh of the boundary of the domain for numerical simulation is required.

Ihlenburg's Finite Element Analysis of Acoustic Scattering (Ihlenburg, 1998): The wave equation is solved using a discrete set of linear equations on elements of subdivided space. At limit, Finite Element Techniques provide an accurate solution to the wave equation. Finite element methods were originally developed for the numerical solution of problems on bounded domains. However, in acoustic scattering applications, often the computational domain may be unbounded. One can either impose that the wave equation is satisfied at a set of discrete points (collocation method) or ensure a global convergence criteria (Galerkin method). This technique presents a problem of how to discretise an infinite domain. The typical approach is to bound the area virtually such that nothing reflects off ∞ and that the work is conducted within a specified region. This introduces bias however as it has to be decided what conditions to adhere to on the virtual boundary space. In addition, as the wavenumber k becomes large the accuracy of standard finite element techniques deteriorates and basis function techniques applicable to higher frequencies are adopted in more generalised FEM approaches.

Gumerov et al. (Gumerov and Duraiswami, 2008) deployed a fast multipole method (FMM) to accelerate the solution of the boundary element method for the Helmholtz equations. Difference in the application of the FMM for low and high kD, k being the wave number and D the domain size, are explored. In addition techniques for quadrature of the boundary shapes using sub-elements are described and a discussion of the tradeoff between accuracy and efficiency with the application of the FMM is given.

3.3.4 Finite Difference Time Domain (FDTD)

Botteldooren et al.'s Finite Difference Time Domain (Botteldooren, 1994, 1995) uses FDTD equations to allow the use of a nonuniform grid to be derived. With this grid, tilted and curved boundaries can be described more easily. This allows a better tradeoff to be defined between accuracy and CPU usage in a number of circumstances. Botteldooren (Botteldooren, 1995) illustrates the use of a numerical time-domain simulation based on the FDTD approximation for studying low and mid frequency room acoustic problems. This is a direct time-domain simulation.

Sakamoto et al. (Sakamoto et al., 2006) extend FDTD by successively solving, step by step, the acoustical quantities at discrete grid points of a closed field according to vector operations. The impulse responses can be obtained directly in the time domain using little computer memory storage. The authors subsequently compare FDTD analytical results with measurements in a hall with material coefficients calculated with the impedance-tube method and show the measured and calculated reverberation were in accordance in the middle-frequency bands. However, in low-frequency bands, large discrepancies occurred because of the difficulties in determining and modelling the boundary conditions using a system comprising masses, springs, and dampers in the analytical form (Sakamoto et al., 2008).

3.3.5 Pseudospectral Time-Domain (PSTD)

FDTD method needs a fine granularity of cells per wavelength in order to give accurate results. PSTD methods suffer from the wraparound effect due to the use of discrete Fourier transform. This effect limited the applicability of PSTD to large-scale problems. Liu (Liu, 1997) applied Berenger's perfectly matched layers (PML) to the pseudospectral method in order to eliminate the wraparound effect. As a result the granularity of cells required per wavelength for computation is much smaller and as a result, solutions of problems 64 times larger than the FDTD method can handle were reported with the same requirement in computer memory and CPU time.

3.3.6 Digital Waveguide Mesh

Campos et al's Mesh (Campos and Howard, 2005): The digital waveguide mesh enables the acoustics of an existing, now ruined or drawing board space to be modelled acoustically. An RIR can be obtained for any combination of source/listener positions to enable the acoustics at different positions to be experienced (Campos et al., 2001).

Mullen et al's Waveguide Mesh Vocal Tract Model (Mullen et al., 2006): This technique enables the synthesis of speech sounds via a two dimensional mesh of the oral tract. Mesh shape variation is made possible by an impedance mapping technique to enable speech articulation to be modelled. Mesh wall reflections can be adjusted to set appropriate formant bandwidths (Mullen et al., 2006). With the addition of a nasal cavity and voiceless excitation, a complete speech synthesis system becomes a possibility.

Murphy et al.'s Digital Waveguide Mesh (Murphy et al., 2007): A digital waveguide mesh is a variant of FDTD methods. The waveguide itself is a bidirectional digital delay line. In one dimensional systems real time applications are easily possible. The mesh is constructed of a regular array of digital waveguides arranged along each dimensional axis and interconnected at each intersection. These are scatterings junctions. Scattering junctions used to construct the mesh enable a RIR to be obtained for a distinct point. Measuring over a number of junctions and post-processing enables an Ambisonic B-format or 5.1-channel RIR to be obtained suitable for surround sound reverberation processing.

The mesh constructed is a rectangular grid in which each node (scattering junction) is connected to its six neighbours by unit delays. The accuracy of the technique is inherent in the granularity of the grid. In addition, it is heavily reliant on the direction dependant dispersion of wave front's such that tetrahedral or triangular mesh extensions (Campos and Howard, 2005) have been implemented to mitigate this. Furthermore, frequency warping (Savioja and Valimaki, 2001) has also been used to deal with this. Due to the dispersion the model is useful for frequencies below the update frequency.

3.3.7 Domain Decomposition

Raghuvanshi et al.'s Domain Decomposition (Raghuvanshi et al., 2008) technique simulates sound propagation with reduced dispersion on a much coarser mesh, enabling accelerated computation. Modal analysis of an entire scene is not usually possible; however using domain decomposition it is possible to shorten the computation time required such that up to an order of magnitude speed up can be gained over standard FDTD models. Extended in 2009 (Raghuvanshi et al., 2009), the authors exploit the known analytical solution of the Wave Equation in rectangular domains, to achieve a hundred-fold performance gain compared to a FDTD implementation with similar accuracy, whilst also bearing an order of magnitude memory efficiency. As a result this method breaks the low frequency barrier to a degree that typically stagnates wave solutions and performs accurate numerical acoustic simulation in the low-mid kilohertz range. This technique is employed in a precomputed offline solution for real time auralisation by the authors (Raghuvanshi et al., 2010). Real-time sound propagation including effects such as diffraction and accounting for moving sources and listeners is accounted for. The offline simulation calculates the scenes acoustic responses and encodes the salient information. The responses are categorised into early and late reflections by the temporal density of arriving wavefronts. Late reflections are calculated once per room and stored in the frequency domain. The early reflections, which provide more spatial information, are recorded by a set of delays in the time domain and a frequency response in various octave bands in every source/receiver pair for the scene. Using an efficient real time frequency-domain convolution filter for the encoded responses binaural reproduction in real time is shown.

3.3.8 Volumetric Methods

Farina's Ramsete - Pyramid Tracer (Farina, 1995): The author employs a completely new pyramid tracer, which avoids the problems encountered with conical beam tracers such as overlapping of cones and multiple detection of the same Image Source.

Rajkumar et al.'s Ray-Beam Tracing (Rajkumar et al., 1996): The method uses a variation of Ray-Tracing dubbed "Ray-Beam Tracing". By introducing the notion of beams while retaining the simplicity of rays for intersection calculations, a beam is adaptively split into child beams to limit the error introduced by infinitely thin rays.

Funkhouser et al.'s Beam Tracing (Funkhouser et al., 1998, 1999): This approach uses rays, traced in packets through a spatially subdivided data structure stored in a depth-ordered sequence. These packets emulate beam propagation. This application to the acoustic simulation field stems from original beam tracing algorithm for computer graphics by Heckbert et al. (Heckbert and Hanrahan, 1984). This

removes the problems in sampling and aliasing that plague ray traced approaches as first discussed by Lehnert (Lehnert, 1993).
Tsingos et al.'s extension based on the Uniform Theory of Diffraction (UTD) (Tsingos et al., 2001): This builds upon the previous work by Funkhouser et al. (Funkhouser et al., 1998) by incorporating the UTD into the model for propagation within the Beam Tracing architecture.
Lauterbach et al.'s Frustrum Tracing (Lauterbach et al., 2007): Combines the efficiency of interactive ray tracing with the accuracy of tracing a volumetric representation. The method uses a four sided convex frustum and performs clipping and intersection tests using ray packet tracing. A simple and efficient formulation is used to compute secondary frusta and perform hierarchical traversal.
Laine et al.'s Accelerated Beam Tracing Algorithm (Laine et al., 2009): In this method it is shown that beam tracing algorithms can be optimised further by utilising the spatial coherence in path validation with a moving listener. Necessary precalculations are quite fast. The acoustic reflection paths can be calculated in simple cases for a moving source when utilising this approach.

3.3.9 Particle Based Methods

Kapralos et al.'s Sonel Mapping (Kapralos et al., 2004): The authors aim to adapt photon tracing and gear it towards sound simulation by exploiting the synergy of properties between sound and light. The technique dubbed Sonel mapping is a two-pass Monte-Carlo based method that accounts for many of the complex ways in which sound interacts with the environment as opposed to light. It is used to model acoustic environments that account for diffuse and specular reflections as well as diffraction and refraction effects.
Bertram et al.'s Phonon Tracing (Bertram et al., 2005): Inspired by the photorealism obtained by methods such as Photon Mapping (Jensen, 1996); for a given source and listener position, this method computes an RIR based on particle distributions dubbed Phonons, accounting for the different reflections at various surfaces with frequency-dependent absorption coefficients. This does not take into account diffraction effects or low frequency dominated simulations. As such, frequencies on the order $f = \frac{c}{\lambda} \approx \frac{c}{l}$ are limited by this technique, where c is the speed of sound and l is the diameter of the simulation geometry. This technique is similar to that of Kapralos et al. (Kapralos et al., 2004) in that it employs a two pass algorithm for emission of phonons and collection of phonon contributions for generation of the impulse response. Collection of the emitted phonon samples from the map is done via a Gaussian strategy. The Gaussian strategy is used to generate smoother filters since more phonons are collected and contribute weighted by their shortest distance to the collection coordinate.

3.3.10 GPU Accelerated Approaches

Jedrzejewski at al.'s application of ray based methods to programmable video hardware (Jedrzejewski and Marasek, 2006): The method ports ray based methods for sound simulation onto the GPU such that sound source and listener are free to move, producing echograms using simplified acoustic approximation.
Tsingos et al.'s Instant Sound Scattering (Tsingos et al., 2007): This work is a paradigm shift from conventional approaches to sound simulation as it takes advantage of some of the benefits of commodity graphics hardware utilising combined normal and displacement maps for dense sampling of complex surfaces for high quality modelling of first order scattering.
Rober et al.'s Ray Acoustics Using Computer Graphics Technology (Rber et al., 2007): Analyses the propagation of sound in terms of acoustical energy and explores the possibilities of mapping these concepts to radiometry and graphics rendering equations on programmable graphics hardware. Concentrating principally on ray-based techniques this also investigates to a lesser extent wave based sound propagation effects.
Cowan et al.'s GPU-Based Real-Time Acoustical Occlusion Modelling (Cowan and Kapralos, 2010): Provides a computationally efficient GPU-based method to approximate acoustical occlusion and diffraction effects in real time. Ray casting using the GPU from the sound sources perspective into the scene generates a low resolution *listener occlusion map* which provides an occlusion weighting to the listener for use when passed to the audio engine.
Mehra et al. (Mehra et al., 2011) carefully map all the components of the domain decomposition algorithm to match the parallel processing capabilities of GPUs to gain considerable speed up compared to the corresponding CPU-based solver, whilst maintaining the same numerical accuracy in the solver.

For more comprehensive reports and overviews on the topic of using programmable graphics hardware for acoustics and audio rendering, the reader is directed to (Tsingos, 2009b; Savioja et al., 2010).

3.3.11 Classification

Within this section we sum up the common features of methods presented so far. We will also give an indication as to the performance and quality of the various techniques. Included in this will be the principal ideas of the approaches and an analysis of performance and flexibility of various methods.

Table 1 shows which drawbacks are associated with the various spatialisation techniques described in Section 3.3.

Technique	Speed	Accuracy	Comment
Ray Tracing	Very Fast[1]	Inaccurate[2]	Does not natively support diffraction effects. [1]Current implementations, slow on advent due to technology bottleneck. [2]Only accurate without work-arounds for high frequency bands.
Image Source Methods	Fast	Accurate	Only considers specular reflection paths, diffraction and material scattering is ignored. Drawbacks over low frequency bands.
FEM/BEM and FDTD	Very Slow	Very accurate	Computational load grows very fast with frequency, all details must be modelled to achieve full rate of accuracy, Source directivity is difficult to achieve with FEMs. Appropriate only for low frequency simulation and small enclosures.
Digital Waveguide Mesh	Slow	Accurate	The accuracy is inherent in the granularity of the grid used. Only useful for frequencies below the update frequency.
Volumetric Methods	Slow-Fast	Accurate	Scattering effects are not accounted for, geometric clipping techniques have always been a bottleneck. Supports diffraction with work around.
Particle Methods	Slow-Fast	Accurate	Does not natively support diffraction or low frequency sound and typically ignores air absorption.
GPU Approaches	Very Fast	Inaccurate[1]	[1]Tends to trade speed off for simplified acoustic approximation; such as larger granularity of frequency dependant calculation bands.

Table 1. Classification and drawbacks of various Sound Synthesis techniques

The ray based techniques, ray tracing and image source, are the most commonly used algorithms in practise, especially in commercial products. The rays are supposed to be sample points upon a propagating sound wave. This stipulation only remains true when the wavelength of the sound is small when compared to the geometry of the environment but large compared to any defects upon surfaces being impinged upon by the sound wave. The basic distinction between ray tracing and image source techniques is the way paths are found. Generating the IR for a room requires all paths to be found, Image Source techniques find all paths but are limited by the exponential rise in computation as the order of reflection rises. Monte Carlo approaches to Ray tracing on the other hand give a statistical result for the sampled paths, higher order reflections can be considered stochastically but not all paths are guaranteed to be found.

The more computationally demanding wave based models such as FEM and BEM are suitable for the simulation of low frequencies only. Time-domain solutions tend to provide better solutions for auralisation than FEM and BEM which tend to be solved in the frequency domain.

3.4 Generic Models for Environmental Effects (Artificial Reverb)

The study of the perceptual effects of room acoustics and reverberation as well as the physics of sound propagation in rooms lead to the descriptions of the impulse response using simplified models tuned in different time regions. Generally, a first temporal region is devoted to the direct sound, as it is of primary importance for the localisation of the sound source and the perception of its spectral characteristics. The next temporal section comprises a limited set of early reflections, typically contained in a time interval [0:40]ms and that can be individually controlled. Subjectively, they will be integrated in the perception of the direct sound but their temporal and spatial distribution will modify the timbre, spatial position and apparent width of the sound source. As time increases, the density of sound reflection increases and their temporal and spatial distribution can be modelled as a statistical process. While it becomes very difficult to simulate individual late reflections accurately, it is also irrelevant from a perceptual point of view. The late part of the reverberation can be described by the energy decay envelope as well as different parameters related to its finer grain structure such as temporal density of reflections or modal density. A later set of early reflections, generally contained in the time-interval (40:100]ms can also be specifically modelled.

In addition to the temporal description of the reverberation, the frequency and spatial characteristics must also be considered and can be adapted to the desired computational complexity. In particular, the frequential and spatial resolution of the impulse response which must be finely described for direct sound and early reflections can also be simplified for late reverberation effects, using statistical descriptors such as the interaural cross correlation coefficient (Pellegrini, 2001b). In interactive environments, direct sound and early reflections should also be updated at a higher rate than the late reverberation which tends to vary more smoothly.

These formulations lead to the development of efficient artificial reverberators, which are widely used to auralise late reverberation effects in games (Gardner, 1997; Rocchesso, 2002). Artificial reverberators do not model the fine-grain temporal structure of a reverberation filter but assume that reverberated components can be modelled as a temporal noise process modulated by slowly-varying energy envelopes in different frequency sub-bands. These envelopes are often considered as exponentially decaying, which lead to the design of efficient recursive Feedback Delay Network (FDN) filters (Schroeder, 1962; Jot, 1999; Gardner, 1997; Rocchesso, 2002).

In addition from the computational gains, parametric reverberation offers great flexibility and adaptation to the reproduction system, as opposed to directly describing an impulse response that is tied to a particular recording system. Parametric reverberation also offers the flexibility to specify the room effect without geometrical modelling, which is particularly useful for musical applications where the desired effect primarily targets audio perception. For applications where more audio-visual coherence is required, it is possible to model the primary sound reflections using geometry-based models as described in section 3.3.

Parametric reverberation models have been traditionally limited to enclosed space where statistical acoustics models prevail, and are not necessarily a good fit for applications that model outdoor environments such as cities or forests, which may also require significant other acoustical effects. Parametric frequency-domain approaches, that can be driven by geometrical simulations, have recently been proposed supporting more general decay profiles as well as additional parameters for spatial rendering of the reverberation (Vickers et al., 2006; Tsingos, 2009a; Merimaa and Pullki, 2004).

4 SYNTHESISING VIRTUAL SOUND SOURCES

Whilst section 3.3 covers algorithms for generation of sound filters, to yield a particular sound in a given simulation environment, there is a practical need to generate virtual sound effects for other uses.

4.1 Sample-based Synthesis and Sound Textures

A common solution for synthesising signals emitted by virtual sound sources is to process recordings of the desired sound events (i.e., sampling). One or several recordings, generally monophonic, can be combined to re-synthesise complex sound sources as a function of the synthesis parameters. For instance, recent car racing games model the sound of each vehicle by blending tens of recordings corresponding to the engine noise at different speeds, tyre noise and aerodynamic noise. The blending is controlled by higher level parameters, for instance tied to an underlying physical simulation. Several effects, such as pitch shifting, are also generally performed in order to best fit the original set of recordings to the current

parameter state. Sample-based approaches lead to realistic results but generally require a significant effort to record the original material as well as create and fine-tune the synthesis model, which is generally done manually.

It is also desirable to synthesise infinite loops of audio material which lead to the design of audio texture synthesis approaches similar to visual texture synthesis in computer graphics (Lu et al., 2004; Parker and Chan, 2003; J.R.Parker and Behm, 2004; Saint-Arnaud and Popat, 1998; Athineos and Ellis, 2003; Di-Scipio, 2003). Given an example sound, the goal is to synthesise a similar and non-repetitive signal of arbitrary duration. A common approach is concatenative synthesis. The example signal is segmented into a collection of short segments or "grains" and compute transitions probabilities for each pair of grains, thus creating a transition graph (Lu et al., 2004; Jehan, 2005). An infinite signal can be re-synthesised by successively concatenating grains following the transition graph. Other techniques analyse statistics of the example signal, for instance using multi-scale wavelet analysis (Dubnov et al., 2002) or fit parametric models based on the statistics of the input signal (Desainte-Catherine and Hanna, 2000; Bar-Joseph et al., 1999).

A common issue arising with sample-based synthesis is that the source recordings must ideally be free of effects (e.g. Doppler, reverberation) if such effects have to be simulated. This requires using directional microphones or near-field recording of the sources so as to maximise the signal to noise (or direct to reverberation) ratio which is not always possible or requires recording in dedicated anechoic chambers. It is also desirable to remove background noise from the recordings using noise reduction techniques so as to avoid noise build-up when a large number of sources is rendered simultaneously.

4.2 Physically-Based Synthesis

Most of the prior work on sound synthesis in computer graphics has focused on simulating sounds from rigid and deformable bodies (O'Brien et al., 2001; Doel et al., 2001; O'Brien et al., 2002; Raghuvanshi and Lin, 2006; James et al., 2006; Bonneel et al., 2008). Synthesis of natural sounds in virtual environments focuses on noise related to the interactions between objects (shock, rolling, friction), which themselves are a broad category of sound events (M.Rath et al., 2003). Moreover, this category is fundamental for virtual environments since it allows audible user interactions with the environment. These approaches are generally based on an estimate of the vibration modes of objects in the environment and then by a modal synthesis step (Doel and Pai, 1998; van den Doel et al., 2001, 2002, 2004; O'Brien et al., 2002), represented as a sum of dampened sinusoids in time. The frequencies, amplitudes and decay modes are the different parameters of the impulse response of the object. The result varies depending on the geometry of the object, but also the material point impact and contact force. The sound emitted by the object also depends on the outcome of the excitement. In the case of a shock, the impulse response can be directly used. For friction, it is necessary to convolve this response by a representation of the excitation (van den Doel et al., 2001). In the context of rigid bodies, it is possible to first calculate the matrix of vibration modes using a 3D mesh (O'Brien et al., 2002). For deformable objects, the synthesis requires more complex calculations; such as finite element methods, which prevents suitability for real time applications (O'Brien et al., 2001).

An alternative synthesis technique is a combined analysis of recordings and resynthesis. For example, one approach measures the acoustical response of real objects (van den Doel et al., 2001). A robotic arm fitted with a rigid tip is used to excite the surface of an object whose acoustic response is recorded by a microphone. By sampling from the surface of the object, then a 2D texture representing the impulse response of the object at different points on its surface can be constructed. Analysis of recorded results allows extraction of parameters of the main modes of vibration which then allow resynthesis of contact noise and real-time interaction with a virtual model of the object. In particular, these approaches lend themselves well to integration with restitution haptic contacts. Other types of synthesis have also been proposed for natural phenomena such as aerodynamic noise (Dobashi et al., 2003) (wind, swish of a sword) or combustion noise and explosions (Dobashi et al., 2004). In this case, a simulated dynamic fluid, finite element is used to generate synthesis parameters (speed of fluid, etc..). Sound matching is then synthesised by summing sonic textures (usually white noise), modulated by the appropriate parameters for each cell of the space used for simulation. We can therefore consider this approach as a hybrid between purely physical synthesis and synthesis by recordings. Synthesis from fluids was first introduced by Van Den Doel (Doel, 2004; Doel, 2005). This introduced the method for generating liquid sounds using Minneart's formula which makes it possible to synthesise liquid sounds directly from fluid animation. Minneart's

formula approximates the resonant frequency of a bubble in an infinite volume of water as $f = \frac{3}{r}$ which leads to the equation for the formation of the sound of a bubble over time as: $\Lambda(t) = A \cdot e^{-dt} \sin(2\pi ft)$, where $\Lambda(t)$ is the impulse response at time t, e^{-dt} is a decay coefficient, f is Minneart's frequency given in Hertz and r is given in metres. This approach is physically based and relatively simple as it is combined with statistical models to synthesise more complex combinations, which in turn is able to evoke the sound of rain or streams, however the computation time still limits the ability for the technique to derive liquid sounds from real time fluid simulations. More recent work in the field of physical sound synthesis has researched rigid-body fracture (Zheng and James, 2010), modal contact synthesis (Zheng and James, 2011) and animating fire with sound (Chadwick and James, 2011).

For more information on recent work in sound synthesis, we also refer the reader to the work carried out under the European project "SoundObj" (The Sounding Object) (Rocchesso et al., 2003), which offers a very comprehensive overview on the field.

4.3 Properties of Virtual Sound Sources

Describing and acquiring the spatial properties of sound sources is a key factor of audio rendering systems but is still one of the major limitations of current approaches. Most spatial audio rendering systems simulate point sources which simplifies the simulation of propagation phenomena but cannot provide a good representation for more complex or spatially extended sources. A solution is to model spatially extended sources using clusters of elementary point sources. However, as previously discussed, synthesising appropriate signals to feed each elementary source can be challenging. If similar recordings are used, phasing effects can appear due to the difference in propagation delay from the different point sources, which requires decorrelating the signals (Potard and Burnett, 2004). In some cases, it is possible to individually record the different spatial or directional components of the sound source using directional microphones (Allman-Ward et al., 2005; Malham, 2001; Menzies, 2002; Meyer and Elko, 2004a) but these solutions remain hard to implement and are often limited by the transducers and they require processing that can significantly reduce bandwidth and signal-to-noise ratio.

In the case of direct synthesis from physical models, it is generally easier to model complex spatial or directional behaviour of the sound emitters as demonstrated in the recent works covering the sound synthesis of wind, fire or water (Dobashi et al., 2003, 2004; Zheng and James, 2009; Moss et al., 2010).

5 RENDERING FROM SPATIAL RECORDINGS

In this section we discuss methods to capture impulse response filters from real world sources (Kuttruff, 1991; Begault, 1994; Savioja et al., 1999); dirac-delta response capture in environments using ambisonics. We also cover not just the capture of the impulse response but direct capture of soundscapes for re-rendering of spatial scenes. This applies to work in blind source separation, upmixing and source localisation (Gallo et al., 2007; Gallo and Tsingos, 2007). This uses multiple microphones stochastically placed within the sound scape to simultaneously record real world auditory environments. Analysis of the recordings to extract varied sound components through time allows for post-editing and re-rendering the acquired soundscape within generic 3D-audio rendering architectures. In addition, we overview Spatial Impulse Response Rendering (SIRR) (Merimaa and Pulkki, 2005) and the extension, Directional Audio Coding (DirAC) (Pulkki, 2006) which are techniques for the reproduction of room acoustics from analysis of recordings of a soundscape depending on time and frequency. The techniques are applicable to arbitrary audio reproduction methods.

5.1 Coincident Recordings and Directional Decompositions

Processing and compositing live multi-track recordings is a widely used method in motion-picture audio production (Yewdall, 2003). For instance, recording a scene from different angles with different microphones allows the sound editor to render different audio perspectives, as required by the visual action. Thus, producing synchronized sound-effects for films requires carefully planned microphone placement so that the resulting audio track perfectly matches the visual action. This is especially true since the required audio material might be recorded at different times and places, before, during and after the actual shooting of the action on stage. Usually, simultaneous monaural or stereophonic recordings of the scene are composited by hand by the sound designer or editor to yield the desired track, limiting this approach to off-line post-production. Surround recording setups (e.g., Surround Decca Trees) (Streicher, a,b), which historically evolved from stereo recording, can also be used for acquiring a sound-field suitable

for restitution in typical cinema like setups (e.g., 5.1-surround). However, such recordings can only be played back directly and do not support spatial post-editing. Other approaches, more physically and mathematically grounded, decompose the wavefield incident on the recording location on a basis of spatial harmonic functions such as spherical/cylindrical harmonics (e.g., Ambisonics) (Gerzon, 1985; Malham and Myatt, 1995; Daniel et al., 1998; Leese, 1998; Merimaa, 2002) or generalized Fourier-Bessel functions (Laborie et al., 2003). Such representations can be further manipulated and decoded over a variety of listening setups. For instance, they can be easily rotated in 3D space to follow the listener's head orientation and have been successfully used in immersive virtual reality applications. They also allow for beamforming applications, where sounds emanating from any specified direction can be further isolated and manipulated. However, these techniques are practical mostly for low order decompositions (order 2 already requiring 9 audio channels) and, in return, suffer from limited directional accuracy (Jot et al., 1999). Most of them also require specific microphones (Abhayapala and Ward, 2002; Meyer and Elko, 2004b; Laborie et al., 2004), especially when higher-order decompositions must be captured.

5.2 Non-Coincident Recordings

A common limitation of coincident or near-coincident recording approaches is that they sample the environments at only a single location which offers a good solution to record spatial sound ambiences or "panoramas" but makes them impractical for virtual walkthrough applications. Some authors, inspired from work in computer graphics and vision, proposed a dense sampling and interpolation of the plenacoustic function (Ajdler and Vetterli, 2002; Do, 2004) using simpler omnidirectional microphones in the manner of lumigraphs or view interpolation in computer graphics (Chen and Williams, 1993; Buehler et al., 2001; Horry et al., 1997).

Radke and Rickard (Radke and Rickard, 2002) proposed an approach aimed at interpolating in a physically consistent way the audio signal captured along a line joining two microphones. Their work relies on a time-frequency decomposition of the recordings derived from blind source separation (Jourjine et al., 2000). This approach has been extended by Gallo et al. to arbitrary numbers of microphones distributed sparsely throughout the capture environment (Gallo et al., 2007).

Other approaches (Ajdler and Vetterli, 2002; Do, 2004) densely sample the plenacoustic function and interpolate it directly. However, these approaches remain mostly theoretical due to the required spatial density of recordings.

5.3 Extracting Structure From Recordings

A large body of work has been devoted to identifying and manipulating the components of the sound-field at a higher level by performing auditory scene analysis (Bregman, 1990). This usually involves extracting spatial information about the sound sources and segmenting out their respective content. Some approaches extract spatial features such as binaural cues (Interaural Time-Difference (ITD), Interaural Level Difference (ILD), Interaural Correlation) in several frequency subbands of stereo or surround recordings. A major application of these techniques is efficient multi-channel audio compression (Baumgarte and Faller, 2003; Faller and Baumgarte, 2003) by applying the previously extracted binaural cues to a monophonic down-mix of the original content. However, extracting binaural cues from recordings requires an implicit knowledge of the restitution system. Similar principles have also been applied to flexible rendering of directional reverberation effects (Merimaa and Pullki, 2004) and analysis of room responses (Merimaa, 2002) by extracting direction of arrival information from coincident or near-coincident microphone arrays (Pulkki, 2006).

Another large area of related research is Blind Source Separation (BSS) which aims at separating the various sources from one or several mixtures under various mixing models (Vincent et al., 2003; O'Grady et al., 2005). Most recent BSS approaches rely on a sparse signal representation in some space of basis functions which minimizes the probability that a high-energy coefficient at any time-instant belongs to more than one source (Rickard, 2006). Some work has shown that such sparse coding does exist at the cortex level for sensory coding (Lewicki, 2002). Several techniques have been proposed such as independent component analysis (ICA) (Comon, 1994; Sawada et al., 2006) or the DUET technique (Jourjine et al., 2000; Yilmaz and Rickard, 2004) which can extract several sources from a stereophonic signal by building an inter-channel delay/amplitude histogram in Fourier frequency domain. In practise, most auditory BSS techniques are devoted to separation of speech signals for telecommunication applications but other audio applications include upmixing from stereo to 5.1 surround formats (Avendano, 2003).

6 INTERFACES FOR SPATIAL AUDIO REPRODUCTION (AUDITORY DISPLAYS)

The last step in the auralisation pipeline is listening to the sound produced. It is necessary to direct the sound to an auditory device designed to recreate the sound field simulated for the listener. This section overviews a number of techniques and devices used to provide a listener with the auditory cues derived from the spatialisation simulation.

6.1 Binaural Techniques

These techniques use headphones directly at the ears of the listener (binaural) (Jot et al., 1995; Møller, 1989, 1992). In binaural techniques a head related transfer function (HRTF) is applied for each and every path reaching the user. As most HRTFs are ad-hoc and not standardised and almost never measured for a specific person or at the correct distance this only serves as an approximation.

6.2 Perceptual Approaches and Phantom sources

A first family of approaches for spatial sound rendering implements a simple control of basic inter-aural sound localisation cues using a set of two or more loudspeakers located around the listening area. The most widely used model remains stereophony, using a pair of loudspeakers located in front of the listener (Ste, 1989; Streicher and Everest, 1998). By controlling the relative delay and amplitude of the loudspeaker signals, it is possible to re-create a simplified reproduction of the inter-aural sound localisation cues, the ITD and ILD. This reproduction creates a phantom source image, which can be freely positioned (or "panned") along the line joining the two speakers.

6.3 Multi-Channel Techniques

Classic stereophony techniques have been extended in particular in the context of cinema applications to sets of loudspeakers on a plane surrounding the listening area, and recently including elevation. The most widely used configuration is the standardised 5 or 7-channel comprising 3 front channels and 2 to 4 surround channels (Chabanne et al., 2010). This type of configuration is also widely used for 3D interactive applications, such as games. A variety of techniques can be used to drive the different loudspeakers in order to re-create the perception of a sound source positioned at a given direction in 2D or 3D space (Dickins et al., 1999). A commonly used approach for general 3D loudspeaker arrays is Vector-Based Amplitude Panning (VBAP) (Pulkki, 1997) which extends stereo pair-wise panning techniques to triples of speakers in 3D space.

6.4 Holophony and Decomposition on Spatial Harmonics Bases

In a holophonic representation, the acoustical field within the listening area is expressed using the Kirchhoff-Helmoltz theorem, as the sum of secondary sound sources located on an enclosed surface surrounding the area (Berkhout et al., 1993). As opposed to perceptual panning approaches which are generally optimal for a small sweet-spot, the primary interest of holophony is the validity of the obtained reproduction for a large area which is well suited to larger audiences. The holophony principle also implies that a dual recording technique exists by using a set of microphones surrounding the sound scene to capture (Larcher et al., 2004). A simplified practical implementation has been developed by Berkhout and De Vries (Berkhout et al., 1993), who introduced a set of approximations and associated corrective terms to the original theory.

A related set of approaches, such as Ambisonics (Gerzon, 1985; Malham and Myatt, 1995; Leese, 1998) model the sound field to reproduce at a given point using a temporal and directional distribution of the sound pressure, which can be decomposed onto spherical or cylindrical harmonic bases (Hobson, 1955; Larcher et al., 2004). A first order decomposition will require generating 4 signals but the resulting spatial resolution is limited. A primary interest of these representations is that the sound field can be directly manipulated, e.g. rotated, by linearly combining the signals for each basis function and that the sound field can be described independently from the reproduction system.

A major drawback of holophonic or harmonic decomposition approaches is that they require a large number of loudspeakers in order to reconstruct a desired sound field with a sufficient resolution in the reproduction area. While converging towards the same result as the number of speakers and order of the decomposition grows, the two approaches do not suffer from the same artefacts. For harmonic decomposition approaches, a truncation of the decomposition order limits the valid listening area as the frequency increases. However, the sound field is valid across all frequencies in this area. For holophony,

the reproduced sound field is stable in the entire reproduction region independently from the frequency. We refer the reader to (Larcher et al., 2004; Ahrens, 2010) for in depth discussion of these effects and a recent overview of holophony/wave-field synthesis and Ambisonic techniques can be found in (Ahrens, 2010).

7 DISCUSSION

Whilst research has begun to explore much of the synergy between acoustic rendering and computer graphics, the work populating the area between perception and cross-modal rendering is sparse (Tsingos et al., 2004; Kohlrausch and der Par, 2005; Tsingos, 2007; Vamnd Tajadura-Jimz, 2007; Vamnd Soto-Faraco, 2007).

Computer sound synthesis modelling of virtual environment is clearly in a very mature state with Image Source, Ray/Pyramid/Beam Tracing, FEM/BEM, Particle systems, GPU variations and applications to liquid animation synthesis. However there are still some phenomena to take care of within these techniques that are often unaccounted for or worked around such as the seat-dip effect, diffraction, scattering, source directivity, and source or receiver near absorbing surfaces.

A possible next step for the field is to work to develop more universal standards for impulse response encodings, Acoustic BRDFS, material absorption tables and benchmarking for auralisation. Whilst the commercialisation of convolution in the sound effects industry has to some extent helped with this, this area still remains quite ad-hoc within the community possibly serving to stagnate any great leap to the main aim which is physically based spatialised sound in real time.

Part II
Cross Modal Interaction

Our sensory system has complex structure and processing mechanisms. However, it is still not perfect and it has certain limitations. In this part we will give an overview of the Human Sensory System limitations, discuss perceptual limitations and attentional resources, and examine how they have been exploited separately and jointly for the benefit of computer graphics algorithms. For better understanding, related findings in psychology will also be surveyed.

8 HUMAN SENSORY SYSTEM (HSS) LIMITATIONS

Human sensory system consists of multiple senses, including vision, audition, smell, taste, touch, temperature, proprioception and the vestibular system, etc. All those can be examined solely, or the interaction and integration between them can be studied. This section will cover the basics of vision and audition, and the most relevant limitations that might be utilised in computer graphics for enhancing auditory and visual rendering.

8.1 Vision

The Human Visual System (HVS) has extremely complex machinery (Roorda, 2002; Blake and Sekuler, 2006). However, it is still limited and is able to process only certain amount of information at any point in time. The HVS is sensitive to only a portion of the electromagnetic wavelength spectrum. This segment ranges from around 400nm to 700nm and is commonly called the visible spectrum. Additionally, since the highest concentration of the photoreceptors in the eye is in the foveal region, this region has the highest visual acuity, and moving further from the fovea the acuity rapidly decreases. The phenomenon of the foveal vision is also known as the internal spotlight (James, 1890; Humphreys and Bruce, 1989). The area of the foveal vision covers only 2° of the visual field. This low angular sensitivity is compensated by the rapid eye movements called saccades.

There are two aspects of visual perception: spatial and temporal. Spatial perception highly depends on visual attention (discussed in Section 9). However, there are some other factors, such as spatial frequency, which might influence the perception (Loschky and McConkie, 2000). In computer graphics, the spatial frequency is particularly important, as it directly affects the level of details or the image sharpness. Vision has much higher spatial visual acuity (visual angle of one minute (Blake and Sekuler, 2006)) than audition.

However, a threshold of the temporal visual sensitivity is 26Hz (Fujisaki and Nishida, 2005), which is more than three times lower than for the audition. Nevertheless, we perceive visual stimuli as continuous thanks to the phenomenon called flicker fusion. The reason for this is persistence of vision, which is the ability of the retina to retain an image for a period of 1/20 to 1/5 of a second after exposure (Roget, 1825).

Other explanations for the continuous appearance of the stroboscopic display, also called the apparent motion, where two or more distinct flashing stimuli are perceived as one dynamic stimulus can be found in (Staal and Donderi, 1983; Anderson and Anderson, 1993; Steinman et al., 2000; Getzmann, 2007). Alterations in visual appearance over time can affect some other aspects of visual perception. According to Bloch's law, for example, the duration of the stimulus can affect the perception of brightness, even for the stimuli with the same luminance (Macknik and Martinez-Conde, 2009).

8.2 Audition

Hearing is an important modality which helps us to learn about an environment and to identify surrounding objects and their features (Moore, 1982; Blake and Sekuler, 2006). For sensing, perceiving and processing auditory cues the Human Auditory System (HAS) is used. Although being a highly complex sensory organ, it has certain limitations.

The main factors that affect sound localisation are: binaural and monaural cues, reverberation and inter-sensory interaction. Binaural cues comprise Interaural Intensity Difference (IID) and Interaural Time Difference (ITD). Although being a powerful tool for sound localisation, binaural cues do not provide sufficient information about the sound source elevation. Monaural cues, however, can provide us with that information using HRTFs. As the sound travels it reflects off the head, body and pinna. During these reflections some of the energy is lost which leaves the sound spectrum suitable for sound localisation. In certain ambiguous positions, such as from ahead or from the behind of the head, where the IID and ITD are the same, head movement breaks the symmetry and resolves the confusion. Another important element of sound localisation is distance perception. This ability evolved as we had to know if a prey or a predator is nearby or far away. When listening to a sound indoors, we rely on the reverberation. However, this cue is missing in outdoor environments, and it is substituted by sound intensity and movement of the sound source. Although this can be useful in sound localisation, it behaves rather poorly for unfamiliar sounds.

Despite these localisation techniques, the spatial auditory resolution is very limited. According to Perrott and Saberi, minimum vertical audible angle without change in elevation is 0.97° and minimum horizontal audible angle without change in azimuth is 3.65° (Perrott and Saberi, 1990). This makes hearing substantially weaker than vision in spatially related tasks. However, the temporal resolution of the HAS is rather high compared to the visual, and according to Fujisaki et al. it is 89.3Hz (Fujisaki and Nishida, 2005).

9 ATTENTION AND PERCEPTION

Human sensory information processing can be divided into three stages: sensation, perception and cognition. Sensation is the physical stimulation of the sensory organs. Perception is a set of processes by which we deal with the information sensed in the first stage. Cognition may be considered the most complicated stage in which the information has been fully processed and possibly used for learning, decision making, storing into memory, etc. (Maragos et al., 2008). Closely linked is the attention, which enables focusing on a particular event or location, which will be sensed, perceived and possibly processed.

William James in his book Psychology defines perception as "the consciousness of particular material things present to sense" (James, 1892). Research in psychology has considered the perception of individual senses separately (Broadbent, 1958; Blake and Sekuler, 2006; Pylyshyn, 2006; Scholl, 2001), and across different modalities (Driver and Spence, 1998; Dufour, 1999; Bertelson and de Gelder, 2004). Although the understanding of the perception of individual senses is crucial, in reality, we are rarely exposed to stimuli affecting solely one modality. Instead, few or all of the senses are stimulated simultaneously, where even if one modality "fails", the information is received and processed unmistakably, due to the effect of cross-modal integration (see Section 11.3). Additionally, stimulation in one sensory modality can affect the perception in other. This will be discussed in Section 11.

Perception can also be affected by other factors, e.g. by user's beliefs and experience, or by their value and need. This was described in 1947 by Jerome Bruner and initiated a movement later named "new look

in perception" (Pylyshyn, 2006). This paper inspired hundreds of experiments, which proved that e.g. poor children perceive coins as bigger than rich and that a hungry person is more likely to see food.

During the sensation stage, our senses are exposed to a number of different stimulations. However, even though they affect our sensory organs, due to attentional limitations they may never get processed so that we experience them. This mostly depends on our consciousness and the focus of the senses and our mind, which is called attention. It can be described as a filter to perception, which helps us to process only relevant information and ignore the rest. The attention can be: completely concentrated, where even the body injuries can remain unnoticed due to the extreme focus of interest; dispersed attention, where the mind is emptied and a person is thinking of nothing - we look and listen but none of what we "see" and "hear" is being absorbed and processed; and the attention that is between these two extremes (James, 1890, 1892). Depending on the intent, the attention can be intentional, endogenous, top-down attention, where the observers voluntarily orients attention towards a spatial location relevant to the task or action they are undertaking; and unintentional, exogenous, bottom-up attention, in which it is involuntarily captured by a certain event (Theeuwes, 1991).

Endogenous attention is selective, which means that it is possible to focus the attention in order to process some stimuli more than other. The exogenous attention is mostly attracted by salient objects or their salient features, or by a sudden motion (Yarbus, 1967; Itti and Koch, 2001; Scholl, 2001). This means that if there is a red ball on a white background, the gaze will be shifted towards it, or if in the static display an object starts moving, our attention will be unintentionally shifted towards the moving object. According to Koch and Ullman, exogenous visual attention depends on colour, intensity, orientation and direction of movement, which form topographical, cortical maps called featured maps (Koch, C. and Ullman, S., 1985).

9.1 Resources and Limitations

Attention and perception in humans have limited resources and certain limitations. One such limitation, caused by the selectiveness of the endogenous attention, is innatentional blindness, firstly introduced by Rock et al (Rock et al., 1992). This phenomenon demonstrates the inability to detect salient objects in the centre of our gaze, when performing a task irrelevant to the distracting object (Rock et al., 1992; Mack and Rock, 1998). In the experiment, participants were asked to judge the size of the arms of a cross briefly presented on a computer screen. The majority of the participants failed to notice unexpected objects appearing on the screen along with the cross. The research was extended with more natural displays by Simons and Chabris in 1999, confirming the same hypothesis (Simons and Chabris, 1999).

Similar limitations of not being able to process all the incoming stimuli at one time exists in the HAS. Moore reported a phenomenon called auditory masking, also known as the cocktail party effect (Moore, 1982). This is the ability to pick out and listen to a single sound in a noisy environment. Another HAS limitation is the continuity illusion (Warren et al., 1988; Kelly and Tew, 2002). The authors showed that, when under suitable conditions a sound A is switched off for a short time, while being replaced by sound B, a listener perceives the A as being continuous.

Pashner characterised attention as capacitively limited and effortful (Pashler, 1999). The latter means that continuous processing of an even stimulus, even if it is enjoyable, may lead to fatigue. Although it is well known that our attentional capacity is limited, it has not been confirmed to what level. There are two parallel, though opposing views on the matter. The first one claims that these resources are inter-modal, shared between modalities (Driver and Spence, 1994; Spence et al., 2000; Strayer and Johnston, 2001), and the second that resources are individual, intra-modal, where each modality has its own attentional pool (Allport et al., 1972; Duncan et al., 1997; Bonnel and Hafter, 1998; Alais et al., 2006; Burr and Alais, 2006). However, there are a number of parameters affecting the evaluation of this kind, such as the detection versus discrimination paradigm and forgetting in short-term memory (Massaro and Warner, 1977). Furthermore, there is an example of how cross-modal attentional links depend on type of attention, such as covert versus overt and endogenous versus exogenous attention (Driver and Spence, 1998). The paper shows that shifts of covert attention in one modality induce the attentional shift in other modalities. Similar results can be found in (Spence and McDonald, 2004).

9.1.1 Inter-modal

Some models of attention propose that our attention operates on a global level and is not divided across multiple senses. This means that the performance of a task requiring attention for one modality will be affected by a concurrent task in some other modality. For example, speaking on a mobile phone

can disrupt the performance of driving a car, due to attention diversion (Strayer and Johnston, 2001). Additionally, there is a difficulty in attending to different locations in the two modalities (Driver and Spence, 1994). In this study, recorded audio was used, played from either left or right side, with active (synchronous) and passive (meaningless) lip-movement on either same or opposite side of the audio. In another study, Spence et al. showed that the further the positions of auditory and visual stimuli are, the easier it is to selectively attend to a particular modality (Spence et al., 2000).

9.1.2 Intra-modal

On the other hand, Alais et al. (Alais et al., 2006), in a study dealing with attentional resources for vision and audition, claim that there are no attentional dependencies between modalities, at least for low-level tasks, such as discrimination of pitch and contrast. In their experiment, they showed that there was no significant difference in performance between single stimulus and multi-modal dual task. Nevertheless, when two tasks within the same modality were assigned, the performance was significantly reduced, which indicated that there might be some attentional limitations within the modality when performing a dual task. Similar results can be found in (Allport et al., 1972; Bonnel and Hafter, 1998; Duncan et al., 1997; Burr and Alais, 2006).

Nevertheless, when observing visual and spoken letters presented simultaneously, there is no significant difference in performance when both letters along with the modalities must be reported or when either visual or auditory letter has to be reported regardless of the modality (Larsen et al., 2003). As reported in the same study, the modality confusion is often experienced, where the spoken letter is reported to be seen or visual letter to be heard.

10 PERCEPTUALLY-BASED AUDITORY RENDERING

In the previous two sections the limitations of the HSS, attention and perception have been introduced. This section will demonstrate how these limitations have been utilised in computer graphics with relation to auditory rendering.

The introduction of auditory cues associated to the different components of a virtual scene together with auditory feedback associated to the user interaction enhances the sense of immersion and presence (Hendrix and Barfield, 1996; Larsson et al., 2002). However, the rendering of a 3D sound source requires a large number of signal processing operations. Even in the case of simplified models, performing all of these processes for a number of sound sources remains taxing on computation time (Bregman, 1990; Best et al., 2005; Brungart et al., 2005). Moreover, the solutions using rendering hardware support only a limited number of simultaneous sound sources, also called “channels”.

The main computational bottlenecks are a per sound source cost, which relates to the different effects desired (various filtering processes, Doppler and source directivity simulation, etc.), and the cost of spatialisation, which is related to the audio restitution format used (directional filtering, final mix of the different sources, reverberation, etc.). Although a realistic result can be achieved through physical modelling of these steps (Pellegrini, 2001a; Lokki et al., 2001), the processing of complex sound scenes, composed of numerous direct or indirect (reflected) sound sources, can take advantage of perceptually based optimisations in order to reduce both the necessary computer resources and the amount of audio data to be stored and processed. One such example exploiting the perceptual audio coding (PAC), where prior work on auditory masking (Moore, 1997) has been successfully utilised, is the well known MPEG I Layer 3 (mp3) standard (Painter and Spanias, 2000). In the context of interactive applications, this approach is thus also linked to the continuity illusion phenomenon (Kelly and Tew., 2002), although current work does not generally include explicit models for this effect. This phenomenon is implicitly used together with masking to discard entire frames of original audio content without perceived artefacts or “holes” in the resulting mixtures.

Since our nervous system is not capable of processing all input stimuli at once, the attention is biased towards more salient stimuli. A salient stimulus is that which is more likely to be noticed and therefore attract attention, such as a red fruit in a green bush or an emergency siren. The proposed auditory saliency map, based on the visual saliency model discussed in the following section, consists of three features: intensity, frequency contrast and temporal contrast, combined into a single map (Kayser et al., 2005). Saliency maps can be used to predict the events that will attract our attention, so that more resources in rendering process could be assigned for their computation. This method has been adapted by Moeck et al. (Moeck et al., 2007) in acoustic rendering, by integrating saliency values over

frequency subbands. Although the approach showed certain limitations, Moeck et al. suggest using audio saliency for the clustering stage. Recent work on the synthesis phase of sound, showed that combining the instantaneous energy of the emitted signal and attenuation is a good criteria (Gallo et al., 2005; Tsingos, 2005). Properties of the signal can also be pre-calculated. MPEG7 and other similar standards and work in audio indexing databases (Herrere et al., 1999; Logan, 2000; Peeters, 2004) are descriptors that can be stored in a wide range of sound signals with a very limited impact on the memory required (Tsingos et al., 2004). Ultimately, this method remains very inefficient while adapting to the characteristics of signals to be processed.

Additionally, spatial rendering of auditory environments with hundreds of sound sources can be significantly simplified using interactive sound masking and spatial LOD, without any perceivable difference (Tsingos et al., 2004). This greedy algorithm starts by sorting sources by importance (In (Tsingos et al., 2004) an indicator of loudness is used). Then the sources are considered in order of decreasing importance until their sum masks the sum of the remaining sources. Another indicator determines whether the signal is close to a noise or close to a harmonic signal and can also be used to more finely adjust the sound masking thresholds (Rangachar, 2001; Kurniawati et al., 2002). The algorithm then dynamically determines the number of audible sources. Lagrange and van den Doel, for example, propose using a model of an acoustic masking algorithm to speed modal synthesis methods by removing inaudible artefacts (van den Doel et al., 2002; Lagrange and Marchand, 2001; van den Doel et al., 2004). However, the measure of the importance of a sound source is not limited necessarily to energy properties within the sounds profile. Other criteria (Edworthy et al., 1991; Haas and Casali, 1995) can also be used to quantify the relative importance of different sound sources from the environment to adapt the signal processing techniques.

For the sake of compatibility with standard rendering approaches, impostor sounds can be constructed as a subset of point sources representing the scene's original sound. Each group of sources is then replaced by a representative whose sole source position, generally the centroid of the group, can be adapted over time depending on the importance of various sources in the group (Tsingos et al., 2004). It is also necessary to determine a signal equivalent to the impostor noise, e.g. the sum of the signals from each source group. This combination of sources can be put into practice in a number of different ways in particular using a fixed directional or spatial subdivision (Herder, 1999; Strar and Wand, 2004) or by adaptive clustering, k-means clustering algorithms (Tsingos et al., 2004).

Another similar example of such a technique is "Binaural Cue Coding (BCC)" (Baumgarte and Faller, 2003; Faller and Baumgarte, 2003; Faller and Merimaa, 2005), which extracts indices of spatial location from a multi-channel recording and encodes the result as a mixture positions in space that evolves over time. Upon arrival each frame is decoded and re-spatialised according to the position determined by the encoding. Such a strategy can be evolved over time, in a manner similar to (Tsingos et al., 2004). Obviously, in the case of BCC that solves an inverse problem, starting from the final mix is not feasible directly from the source sound position as is the case in a traditional system of spatialisation. Attaching a 3D position registration is a problem that can also intervene for rendering 3D audio directly from a set of recordings. The sound scene analysis (Bregman, 1990) proposes other criteria for grouping of sound (simultaneity, close to the principle of Gestalt theory). Other approaches exploit mathematical representations that encode the directional properties of the sound field, for example by decomposition on a basis of spherical harmonics. Implemented within the encoding technique and restitution Ambisonics (Malham and Myatt, 1995), these approaches allow a level of detail by truncation of the harmonic decomposition, which results in a decreased precision of spatial rendering (ie, a low pass spatial filter). They also allow global operations such as turning on a group of sources encoded in this representation. This type of representation can be used to represent non-point sound with variable spatial resolution or recreate the sonic background of a scene (Foley et al., 1990).

Large scale sound signals can be rendered utilising level-of-detail, progressive sources. A large range of signal operations is required for all sources. Due to the possibility of a large number of signals, it is possible to define a computational cut-off, such that each source only contributes to the final result in proportion to its importance. One possibility is to encode the signal and the wavelet (Darlington et al., 2002), or to use a frequency representation in Fourier space (Tsingos, 2005). Another family of approaches performs processing on signals directly compressing with the help of a perceptual codec (MPEG I Layer 3 (mp3) standard (Painter and Spanias, 2000)), which may be more effective than a decoding, processing and re-encoding cycle. Nevertheless, a partial decoding should generally be done

and treatments in area codes are generally more delicate and require adapted filters (Touimi, 2000; Touimi et al., 2004). The separation between compression and audio signal processing tends to blur approaches in which the representation of signals is adapted both to the transmission and processing. This problem is particularly important for applications in audio rendering, a distributed massively multi-user application framework, for example.

11 CROSS-MODAL INTERACTION

Since the temporal sensitivity of vision and audition are not the same, the synchrony detection between auditory and visual stimuli was investigated using psychophysical experiments. Results revealed that it is not just a temporal lag between stimuli that influences the discrimination task, but also the temporal frequency. For temporal frequencies higher than 4Hz the synchrony-asynchrony discrimination becomes impossible even when the lag between stimuli is large enough to discriminate it with single pulses. Above this frequency the auditory driving effect occurs (Gebhard and Mowbray, 1959; Shipley, 1964). This effect is described in Section 11.1.

These differences in spatial and temporal sensitivities of vision and audition are the basis of the modality appropriateness hypothesis (Howard and Templeton, 1966; Welch and Warren, 1980). This hypothesis advocates that the modality that is more appropriate for a certain task will dominate the perception of that particular task. In other words, human vision is more accurate in spatial judgements, while audition dominates in temporal domain.

Research in psychology has shown that strong cross-modal interactions exist (Guttman et al., 2005; Recanzone, 2003; Burr and Alais, 2006) and that these cross-modal effects must be taken into consideration when the perception of distinct sensory modalities is investigated (Shimojo and Shams, 2001; Shams et al., 2004).

The auditory-visual cross-modal interaction can be divided in two ways: according to target modality into auditory influence on vision and visual influence on audition; and according to the domain into spatial and temporal domains.

11.1 Auditory Influence on Vision

In order to better understand and appreciate the cross-modal research in computer graphics, the examples from psychology are first presented. The most relevant work in the field is described below. These findings could be applied in multi-modal rendering, where graphics rendering is demanding, requiring significant amount of time and processing power.

Several researches have shown that if frequency of the auditory flutter, initially presented simultaneously with the flickering light, changes, then the perception of the visual flicker changes accordingly, i.e. the flicker "follows" the flutter. This phenomenon is known as the auditory driving effect (Gebhard and Mowbray, 1959; Shipley, 1964; Wada et al., 2003; Recanzone, 2003). Initially, the experimental results did not show the reverse effect (Gebhard and Mowbray, 1959; Shipley, 1964). However, Wada et al. proved that, if auditory stimuli are ambiguous, the change in the visual flicker can change the perception of the auditory flutter (Wada et al., 2003), which is in collision with the modality appropriateness.

Audition can not only change the temporal perception of the visual stimuli, it can even create the perception of additional visual stimuli. When a single visual flash is presented simultaneously with two or more auditory beeps, an observer perceives two flashes. This illustrates how illusory flash can be induced by a sound beep (Shams et al., 2000, 2002). Nevertheless, when a single beep is accompanied by multiple flashes, only one beep is perceived (Shams et al., 2002).

An analogue phenomenon to the visual ventriloquism effect (see Section 11.2) is the temporal ventriloquism (Morein-Zamir et al., 2003; Bertelson and Aschersleben, 2003; Aschersleben and Bertelson, 2003; Burr et al., 2009). If two visual stimuli are observed, the temporal order judgement can be affected if auditory stimuli are presented in a certain order. Namely, when the first flash is preceded by an auditory beep and the second followed by another beep, the visual perception is affected as if the sounds pulled the lights further in time. Analogously, if the sounds are presented between the visual stimuli, the perceived temporal distance between the visuals seems to be decreased (Morein-Zamir et al., 2003). Aschersleben and Bertelson (Aschersleben and Bertelson, 2003) showed that the temporal ventriloquism works in the opposite direction, but to a much lesser extent.

11.2 Visual Influence on Audition

Similarly, as audio can influence visual perception, audition is the subject of visual influence. The findings from this area may be utilised for enhancing the performance and quality of audio rendering in a multi-modal virtual environment.

An example of such influence is the ventriloquism effect (Howard and Templeton, 1966; Choe et al., 1975; Vroomen et al., 1998; Vroomen and de Gelder, 2004). The effect was named by Howard Templeton after the illusion created by ventriloquists when producing the words without moving their lips (Howard and Templeton, 1966). The effect is apparent while watching TV or a puppet show. Although the audio is originating from the audio speakers or ventriloquist's mouth, remote from the observed visual location, the spectator perceives it as if it was emanating from the mouth of the actor or puppet respectively. Vroomen and de Gelder demonstrated the robustness of the effect, proving that attention towards the visual cue is not needed to obtain the effect (Vroomen and de Gelder, 2004).

Although speech is generally considered as a purely auditory process, the visual influence on auditory perception cannot be neglected. McGurk and MacDonald reported that pronunciation of *ba* is perceived as *da* when accompanied by the lip movement of *ga* (Mcgurk and Macdonald, 1976). This phenomenon is know as the McGurk effect.

11.3 Multisensory integration

Cues in different modalities do not always "compete" against, but they can be complementary as well. This generally happens when a stimulus of a dominant sense is ambiguous or corrupted. The cross-modal integration in this case enhances the overall experience of the observer's stimulation. A study by Stein et al. (Stein et al., 1996) demonstrated that a simultaneous auditory stimulus can increase the perceived visual intensity. The authors showed that the effect is present regardless of the auditory cue location. However, it persisted only at the location of visual fixation. Furthermore, Van der Burg et al. (Van der Burg et al., 2008) showed that in a visual search task, a single synchronised auditory *pip*, regardless of its position, significantly decreases the search time. Another study demonstrated that a single auditory click can change the meaning of the visual information (Sekuler et al., 1997). When two identical disks, moving towards each other, coinciding and moving apart, are presented on a display with no sound, they are perceived as they streamed through each other. However, when a brief click was introduced at the time of the collision, the disks appeared as if they bounced off of each other.

Burr and Allais proposed a framework in which a cross-modal information can be optimally combined as a sum of all individual stimuli estimates weighted appropriately (Burr and Alais, 2006). The optimal estimate can be calculated as $\hat{S} = w_A\hat{S}_A + w_V\hat{S}_V$, where w_A and w_S are weights by which the individual stimuli are scaled, and $\hat{\mathrm{S}}_A$ and $\hat{\mathrm{S}}_V$ are independent estimates for audition and vision respectively. The weights are inversely proportional to the auditory and visual variances (σ^2) of the underlying noise distribution $w_A = 1/\sigma_A^2, w_V = 1/\sigma_V^2$. This has been tested using different visual stimuli with different level of blurriness (Alais and Burr, 2004). An example where audition captures the sight occurs when visual stimuli are corrupted by blurring the visual target over a large region. The blurring, however, has to be significant i.e. over about 60°, which makes most scenes unrecognisable. Nevertheless, auditory localisation was performed only by interaural timing difference without time varying, which is around one-sixth of the total cues used in regular hearing. Chalmers et al. proposed to extend this to multiple senses (Chalmers and Debattista, 2009).

12 CROSS-MODAL INTERACTION IN COMPUTER GRAPHICS

In previous sections, findings on related work in psychology have been summarised. In this section, work in computer graphics, that uses these findings is presented.

12.1 Auditory Rendering

Usually, in virtual environments, it is not enough to deliver only high-fidelity graphics. For a more complete experience and higher degree of immersion, the other senses should be stimulated. Most often, sound is presented along with the video. However, as discussed in Part I, some auditory stimuli need to be rendered in real-time, which requires significant processing power, especially if multiple sound sources are present in a complex virtual environment. Different techniques have been explored in order to enhance this process, while maintaining equal perceptual quality.

Phenomenon	Used in
angular sensitivity (James, 1890; Humphreys and Bruce, 1989)	(Yee et al., 2001; Cater et al., 2003; Mastoropoulou et al., 2005a; Mastoropoulou, 2006; Chalmers et al., 2006; Longhurst et al., 2006)
inattentional blindness (Rock et al., 1992; Mack and Rock, 1998; Simons and Chabris, 1999)	(Cater et al., 2003; Mastoropoulou et al., 2005a; Mastoropoulou, 2006)
modality appropriateness hypothesis (Howard and Templeton, 1966; Welch and Warren, 1980)	(Mastoropoulou and Chalmers, 2004; Mastoropoulou, 2006; Hulusic et al., 2009, 2010a,b)
auditory driving effect (Gebhard and Mowbray, 1959; Shipley, 1964; Wada et al., 2003; Recanzone, 2003)	(Mastoropoulou and Chalmers, 2004; Mastoropoulou, 2006; Hulusic et al., 2009, 2010a,b)
temporal ventriloquism (Morein-Zamir et al., 2003; Bertelson and Aschersleben, 2003; Aschersleben and Bertelson, 2003; Burr et al., 2009)	(Hulusic et al., 2009, 2010a,b)
illusory flash induced by sound (Shams et al., 2000, 2002)	(Hulusic et al., 2009, 2010a,b)
stimuli weighting (Burr and Alais, 2006)	(Chalmers and Debattista, 2009)
ventriloquism effect (Howard and Templeton, 1966; Choe et al., 1975; Vroomen et al., 1998; Vroomen and de Gelder, 2004)	(Moeck et al., 2007)

Table 2. The cross-modal phenomena found in psychology (left column) and the studies that were inspired by within the computer graphics (right column)

To date there has not been much work done on cross-modal interaction in auditory rendering. In this section we will give an overview of the work using this phenomenon. The majority of the work on this topic has been done within the CROSSMOD project (CRO, 2010). One of the first studies, conducted by Moeck et al. investigated sound source clustering (Moeck et al., 2007). In their approach the authors used hierarchical clustering algorithm and a metric for cross-modal audio clustering, which encourages creating more clusters within a view frustum.

Grelaud et al. developed an audio-visual level-of-detail (LOD) selection algorithm (Grelaud et al., 2009) based on the work of Bonneel et al. (Bonneel et al., 2010). Bonneel et al. demonstrated that both audio and video stimuli influence the material perception during impact, when many objects produce sound at the same time. Nevertheless, Grelaud et al. in their study used both pre-recorded and impact sounds. The energy for the recorded audio was pre-computed, while for the impact sound a quick energy estimate was calculated. This way the rendering process was significantly speeded up. The experimental results indicate that it is possible to increase audio LOD while decreasing visual LOD without significant perceived visual difference.

12.2 Visual Rendering

Cross-modal interaction has also been used to enhance visual rendering. An early study on auditory-visual cross-modal interaction demonstrated that the quality of the realism in virtual environments depends on both auditory and visual components (Storms, 1998). The author showed that high-quality audio further increases the perceptual quality of the high-quality video. Furthermore, high-quality video further decreases the perceived quality of a low quality audio.

Auditory-visual cross-modal interaction in video rendering is mostly oriented towards the auditory influence on visual perception. This influence can be divided into two domains: spatial and temporal. The former investigates how audition can be utilised in order to enhance video rendering by decreasing the spatial quality of the generated imagery, without any perceivable degradation in overall user experience. Below are the examples of work done on auditory-visual cross-modal interaction in computer graphics,

both in temporal and spatial domain.

In the context of the spatial domain, Mastoropoulou et al. showed that selective rendering technique for sound emitting objects (SEO) in animation rendering can be efficiently used for decreasing the rendering time (Mastoropoulou et al., 2005a; Mastoropoulou, 2006). The authors, via user studies showed that participants attended to the sound emitting object (SEO). Having in mind the angular sensitivity and inattentional blindness, it is necessary to render in high-quality only the SEO, while computing lower quality for the rest of the scene. This approach could be used in conjunction with the Aleph map, described above (Yee et al., 2001). Harvey et al. (Harvey et al., 2010) showed this phenomena further extended to spatial sound impacting gaze directions to attend a directional sound source when present.

The human visual system can perceive quality improvements up to a certain level, which is called the perceived quality threshold. When rendering visual imagery this threshold is important, since any quality improvement above this threshold is considered as a waste of time and resources. Hulusic et al. investigated how the rendering quality threshold is influenced by audio (Hulusic et al., 2008). The authors examined how related and unrelated audio influences visual perception for the presented scenes and showed that unrelated sound can be used for increasing the perceptual quality of graphics, while related audio has no significant effect on perceived rendering threshold.

Auditory-visual cross-modal interactions have been explored in the temporal domain also. According to the modality appropriateness hypothesis, audition is the dominant modality in temporal judgements. Hence, researchers tried to find a perceptual model which will allow for lower frame rates, while playing adequate sound, maintaining the same perceptual visual quality. Mastoropoulou et al. investigated how music can affect temporal visual perception (Mastoropoulou and Chalmers, 2004; Mastoropoulou, 2006), based on modality appropriateness hypothesis and the auditory driving effect. For auditory stimuli two music types were used: slow tempo / relaxing and fast tempo / exciting music, both compared with the no sound condition. The results showed no significant effect for either slow or fast tempo music on the perceived frame rate of the observed animations. According to the authors, this may be due to a couple of factors: the frame rate difference between compared animations (4fps) might have been too small; animation clips lasted for 40 seconds, which is beyond the human working memory.

In another study, walk-through animations with related (sound source visible in the scene) or unrelated sound effects were compared with silent animations played at higher frame rates (Mastoropoulou et al., 2005b). The experimental results showed that sound effects, e.g. a phone ringing or a thunder clap, can attract a part of a viewer's attention away from the visuals and thus allow the frame rate of the presented animated content to be decreased without the user being aware of this reduction. Furthermore, users familiar with computer graphics were found to have more accurate responses to the frame rate variations. There was no effect of camera movement type found to be significant in the experiments.

Hulusic et al. investigated the relationship between the audio beat rate and video frame rate on static (objects static - camera moves) and dynamic (object move - camera static) animations (Hulusic et al., 2009). More specifically, the effect of beat rate, scene and familiarity on the perception of frame rate was investigated. The results showed that the correlation between the beat rate and frame rate exists. For example, in the case of static scenes lower beat rates had a significant effect on the perception of low frame rates. Additionally, the results reveal that there is no effect of familiarity, and that scene complexity and animation dynamics affect the visual perception. However, since this is the first study examining this correlation, further investigation is needed for more conclusive results.

Hulusic et al. investigated the influence of the movement related sound effects on temporal visual perception (Hulusic et al., 2010a,b). The results indicate that introducing the sound effect of footsteps to walking animations in the presented scenes increased the animation smoothness perception. For example, animations played at 10 frames per second (fps) with sound effects have been found as significantly smoother than animations played at 30 or 60 fps without sound. Additionally, the same test showed that animations presented at 20 fps with audio were rated as significantly smoother than silent animations played at 30 or 60 fps. However, no significant influence of sound effects was found for the fast - running animations.

13 SUMMARY AND DISCUSSION

The demand for the improvement of quality in auditory and visual rendering is constantly increasing. Despite the advances in both graphics and general purpose hardware, and algorithmic advancements, it is still not possible to render high-fidelity audio and graphics in real-time. Therefore, perceptually-based

rendering and cross-modal interactions have great, yet to be fulfilled, potential for improving the quality of virtual environments. While researchers in computer graphics and interactive methods have begun to explore the interaction of different modalities and how to exploit them, many of the phenomena discussed in Sections 9.1 and 11 remain unexplored. Some of the psychological phenomena are directly mapped and some of them extrapolated into computer graphics applications, see Table 2. However, there are still some to be investigated, and potentially utilised in computer graphics, such as: modality confusion (Larsen et al., 2003), the McGurk effect (Mcgurk and Macdonald, 1976), the audio effect on visual intensity (Stein et al., 1996) / colour perception, the effect of audio on visual search (Van der Burg et al., 2008) and bouncing targets / circles (Sekuler et al., 1997).

The main focus of interest in computer graphics so far was on the perceptual and attentional limitations such as angular sensitivity, inattentional blindness or modality appropriateness hypothesis, and on auditory influence on visual perception in the temporal domain, e.g. auditory driving effect, temporal ventriloquism and illusory flash induced by sound. Additionally, auditory influence on visual perception in the spatial domain and visual influence on audition have been briefly explored. The cross-modal interaction in computer graphics has been investigated for less than a decade, and therefore, there is a substantial amount of work still to be done. Although this is a long and effortful process, the findings presented in this report promise a bright future for the field.

14 CONCLUSIONS

Sound remains a fundamental component if virtual environments are to deliver high-fidelity experiences. This paper has focussed on two key aspects of audio for virtual environments: The correct simulation of spatialised sound in virtual environments, and, the perception of sound by the HAS including any cross-modal auditory-visual effects. As this report shows, there has been a significant amount of previous work in both these areas. Despite this, current spatialised sound systems are still some way from achieving full physical accuracy with key real phenomena, for example diffraction or scattering often not considered. Similarly, perceptual solutions have come a long way in the last few years. However there is still more research required, for example to investigate interesting issues, such as synaesthesia and the “colour of sound” (Ward et al., 2006). As more co-ordinated multi-disciplinary efforts are made to provide physically accurate audio and visuals in virtual environments in real-time, this paper should provide a valuable resource from which this future research can build.

REFERENCES

(1989). *Stereophonic Techniques - An anthology of reprinted articles on stereophonic techniques*. Audio Engineering Society.

(2010). CROSSMOD project, cross-modal preceptual interaction and rendering. www-sop.inria.fr/reves/CrossmodPublic/index.php.

Abhayapala, T. and Ward, D. (2002). Theory and design of high order sound field microphones using spherical microphone array. *ICASSP*.

Ahnert, W. (1993). Ears auralization software. In *J. Audio Eng. Soc.*, volume 11, pages 894–904.

Ahrens, J. (2010). *The Single-layer Potential Approach Applied to Sound Field Synthesis Including Cases of Non-enclosing Distributions of Secondary Sources*. PhD thesis, Technische Universitat Berlin.

Ajdler, T. and Vetterli, M. (2002). The plenacoustic function and its sampling. *Proc. of the 1st Benelux Workshop on Model-based processing and coding of audio (MPCA2002), Leuven, Belgium*.

Alais, D. and Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, 14(3):257–262.

Alais, D., Morrone, C., and Burr, D. (2006). Separate attentional resources for vision and audition. *Proc Biol Sci*, 273(1592):1339–1345.

Allen, J. and Berkley, D. (1979). Image Method For Efficiently Simulating Small Room Acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.

Allman-Ward, M., Balaam, M., and Williams, R. (2005). Source decomposition for vehicle sound simulation. *www.mts.com/nvd/pdf/source_decomp4veh_soundsim.pdf*.

Allport, D. A., Antonis, B., and Reynolds, P. (1972). On the division of attention: a disproof of the single channel hypothesis. *Q J Exp Psychol*, 24(2):225–235.

Anderson, J. and Anderson, B. (1993). The Myth of Persistence of Vision Revisited. *Journal of Film and Video*, 45(1):3–12.

Appel, A. (1968). Some techniques for shading machine renderings of solids. In *AFIPS '68 (Spring): Proceedings of the April 30–May 2, 1968, spring joint computer conference*, pages 37–45. ACM.

Aschersleben, G. and Bertelson, P. (2003). Temporal ventriloquism: crossmodal interaction on the time dimension: 2. evidence from sensorimotor synchronization. *International Journal of Psychophysiology*, 50(1-2):157 – 163. Current findings in multisensory research.

Athineos, M. and Ellis, D. P. (2003). Sound texture modelling with linear prediction in both time and frequency domains.

Avendano, C. (2003). Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppresssion and re-panning applications. *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2003), New Paltz, NY, USA*.

Bar-Joseph, Z., Lischinski, D., Werman, M., El-Yanniv, R., and Dubnov, S. (1999). Granular synthesis of sound textures using statistical learning.

Baumgarte, F. and Faller, C. (2003). Binaural cue coding - part i: Psychoacoustic fundamentals and design principles. *IEEE Trans. on Speech and Audio Proc*, 11(6).

Begault, D. (1994). 3-D Sound for Virtual Reality and Multimedia. *Academic Press Professional*.

Berkhout, A., de Vries, D., and Vogel, P. (1993). Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764–2778.

Bertelson, P. and Aschersleben, G. (2003). Temporal ventriloquism: crossmodal interaction on the time dimension: 1. evidence from auditory-visual temporal order judgment. *International Journal of Psychophysiology*, 50(1-2):147 – 155. Current findings in multisensory research.

Bertelson, P. and de Gelder, B. (2004). *Crossmodal Space and Crossmodal Attention*, chapter The Psychology of Multimodal Perception. Oxford University Press, USA.

Bertram, M., Deines, E., Mohring, J., Jegorovs, J., and Hagen, H. (2005). Phonon tracing for auralization and visualization of sound. *In Proceedings of IEEE Visualization*, pages 151–158.

Best, V., van Schaik, A., Jin, C., and Carlisle, S. (May/June 2005). Auditory spatial perception with sources overlapping in frequency and time. *Acta Acustica united with Acustica*, 91:421–428(8).

Blake, R. and Sekuler, R. (2006). *Perception*. McGraw-Hill Higher Education, 5th edition.

Blauert, J. (1997). *Spatial Hearing : The Psychophysics of Human Sound Localization*. M.I.T. Press, Cambridge, MA.

Bonneel, N., Drettakis, G., Tsingos, N., Viaud-Delmon, I., and James, D. (2008). Fast modal sounds with scalable frequency-domain synthesis. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*, 27(3).

Bonneel, N., Suied, C., Viaud-Delmon, I., and Drettakis, G. (2010). Bimodal perception of audio-visual material properties for virtual environments. *ACM Transactions on Applied Perception*, 7(1):1–16.

Bonnel, A. M. and Hafter, E. R. (1998). Divided attention between simultaneous auditory and visual signals. *Percept Psychophys*, 60(2):179–190.

Borish, J. (1984). Extension of the image model to arbitrary polyhedra. *The Journal of the Acoustical Society of America*, 75(6).

Born, M. and Wolf, E. (1999). *Principles of Optics*. 7th Edition, Pergamon Press.

Botteldooren, D. (1994). Acoustical finite-difference time-domain simulation in a quasi-cartesian grid. *The Journal of the Acoustical Society of America*, 95(5):2313–2319.

Botteldooren, D. (1995). Finite-difference time-domain simulation of low frequency room acoustic problems. *Acoustical Society of America Journal*, 98:3302–3308.

Bregman, A. (1990). *Auditory Scene Analysis, The perceptual organisation of sound*. The MIT Press.

Broadbent, D. E. (1958). *Perception and communication*. Oxford : Oxford University Press.

Brungart, D. S., Simpson, B. D., and Kordik, A. J. (May/June 2005). Localization in the presence of multiple simultaneous sounds. *Acta Acustica united with Acustica*, 91:471–479(9).

Buehler, C., Bosse, M., McMillan, L., Gortler, S., and Cohen, M. (2001). Unstructured lumigraph rendering. *Proc. of ACM SIGGRAPH*.

Burr, D. and Alais, D. (2006). Combining visual and auditory information. *Prog Brain Res*, 155:243–258.

Burr, D., Banks, M., and Morrone, M. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research*, 198:49–57. 10.1007/s00221-009-1933-z.

Calvin, J., Dickens, A., Gaines, B., Metzger, P., Miller, D., and Owen, D. (1993). The Simnet Virtual World Architecture. *In Proceedings of the IEEE Virtual Reality Annual International Symposium*, pages 450–455.

Campos, G. and Howard, D. (2005). On the computational efficiency of different waveguide mesh topologies for room acoustic simulation. In *IEEE Transactions on Speech and Audio Processing*, volume 5, pages 1063–1072.

Campos, G., Howard, D., and Dobson, S. (2001). Acoustic reconstruction of ancients structures using three-dimensional digital waveguide mesh models. In G. Burenhult, B. I. S. . ., editor, *Computer Applications in Archaeology (CAA2001)*, pages 173–176.

Cater, K., Chalmers, A., and Ledda, P. (2002). Selective quality rendering by exploiting human inattentional blindness: Looking but not seeing. In *Symposium on Virtual Reality Software and Technology 2002*, pages 17–24. ACM.

Cater, K., Chalmers, A., and Ward, G. (2003). Detail to attention: exploiting visual tasks for selective rendering. In *EGRW '03: Proceedings of the 14th Eurographics Workshop on Rendering Techniques*, pages 270–280, Leuven, Belgium. Eurographics Association.

Chabanne, C., McCallus, M., Robinson, C., and Tsingos, N. (2010). Surround sound with height in games using dolby pro logic iiz. In *Audio Engineering Society Convention 129*.

Chadwick, J. N. and James, D. L. (2011). Animating fire with sound. *ACM Transactions on Graphics (SIGGRAPH 2011)*, 30(4).

Chalmers, A. and Debattista, K. (2009). Level of realism for serious games. *Games and Virtual Worlds for Serious Applications, Conference in*, 0:225–232.

Chalmers, A., Debattista, K., and Santos, L. P. (2006). Selective rendering: computing only what you see. In *GRAPHITE '06: Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 9–18, New York, NY, USA. ACM Press.

Chen, S. and Williams, L. (1993). View interpolation for image synthesis. volume 27, pages 279–288.

Choe, C. S., Welch, R. B., Gilford, R. M., and Juola, J. F. (1975). The "ventriloquist effect": Visual dominance or response bias? *Perception and Psychophysics*, 18(1):55–60.

Comon, P. (1994). Independent component analysis: A new concept. *Signal Processing*, 36:287–314.

Cowan, B. and Kapralos, B. (2010). Gpu-based real-time acoustical occlusion modelling. *Virtual Reality*, 14:183–196.

Cremer, L. and Müller, H. (1978). Principles and Applications of Room Acoustics. *Applied Science*, 1.

Daniel, J., Rault, J.-B., and Polack, J.-D. (1998). Ambisonic encoding of other audio formats for multiple listening conditions. *105th AES convention, preprint 4795*.

Darlington, D., Daudet, L., and Sandler, M. (2002). Digital audio effects in the wavelet domain. In *Proceedings of COST-G6 Conference on Digital Audio Effects, DAFX2002, Hamburg, Germany*.

Desainte-Catherine, M. and Hanna, P. (2000). Statistical approach for sound modelling. pages 91–96.

Di-Scipio, A. (2003). Synthesis of environmental sound textures by iterated non linear functions.

Dickins, G., Flax, M., McKeag, A., and McGrath, D. (1999). Optimal 3D-speaker panning. *Proceedings of the AES 16th international conference, Spatial sound reproduction, Rovaniemi, Finland*, pages 421–426.

Do, M. (2004). Toward sound-based synthesis: the far-field case. *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada*.

Dobashi, Y., Yamamoto, T., and Nishita, T. (2003). Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics. *ACM Transactions on Graphics*, 22(3):732–740. (Proceedings of ACM SIGGRAPH 2003).

Dobashi, Y., Yamamoto, T., and Nishita, T. (2004). Synthesizing sound from turbulent field using sound textures for interactive fluid simulation. *Computer Graphics Forum (Proc. EUROGRAPHICS 2004)*, 23(3):539–546.

Doel, K. V. D. (2004). Physically-based models for liquid sounds. *ICAD*.

Doel, K. V. D. (2005). Physically based models for liquid sounds. In *ACM Trans. Appl. Percept.*, volume 2, pages 534–546.

Doel, K. V. D., Kry, P., and Pai, D. (2001). Foleyautomatic: physically based sound effects for interactive simulation and animation. In *In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, ACM,*, pages 537–544.

Doel, K. V. D. and Pai, D. (1998). The sound of physical shapes. *Presence*, 7(4):382–395.

Driver, J. and Spence, C. (1994). *Attention and Performance XV*, chapter Spatial Sy, pages 311–331. MIT Press.

Driver, J. and Spence, C. (1998). Crossmodal attention. *Curr Opin Neurobiol*, 8(2):245–253.

Dubnov, S., Bar-Joseph, Z., El-Yanniv, R., Lischinski, D., and Werman, M. (2002). Synthesis of sound textures by learning and resampling of wavelet trees.

Dufour, A. (1999). Importance of attentional mechanisms in audiovisual links. *Exp Brain Res*, 126(2):215–222.

Duncan, J., Martens, S., and Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature*, 387(6635):808–810.

Durlach, N. and Mavor, A. (1995). Virtual reality scientific and technological challenges. Technical report, National Research Council Report, National Academy Press.

Edworthy, J., Loxley, S., and Dennis, I. (1991). Improving auditory warning design: Relationship between warning sound parameters and perceived urgency. In *Human Factors*, volume 33, pages 205–231.

Faller, C. and Baumgarte, F. (2003). Binaural cue coding - part ii: Schemes and applications. *IEEE Trans. on Speech and Audio Proc*, 11(6).

Faller, C. and Merimaa, J. (2005). Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *JASA*, 116(5):3075–3089.

Farina, A. (1995). Ramsete - a new pyramid tracer for medium and large scale acoustic problems. *In Proceedings of EURO-NOISE*.

Foley, J. D., van Dam, A., Feiner, S. K., and Hughes, J. (1990). *Computer graphics, principles and practice*. Addison Wesley.

Fujisaki, W. and Nishida, S. (2005). Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Exp Brain Res*, 166(3-4):455–464.

Funkhouser, T., Carlbom, I., Elko, G., Pingali, G., Sondhi, M., and West, J. (1998). A beam tracing approach to acoustic modeling for interactive virtual environments. *ACM Computer Graphics, SIGGRAPH'98 Proceedings*, pages 21–32.

Funkhouser, T., Jot, J., and Tsingos, N. (2002). Sounds good to me! computational sound for graphics, vr, and interactive systems. *SIGGRAPH 2002 Course Notes Number 45*.

Funkhouser, T., Min, P., and Carlbom, I. (1999). Real-time acoustic modeling for distributed virtual environments. *ACM Computer Graphics, SIGGRAPH'99 Proceedings*, pages 365–374.

Gallo, E., Lemaitre, G., and Tsingos, N. (2005). Prioritising signals for selective real-time audio processing. In *Proceedings of Intl. Conf. on Auditory Display (ICAD) 2005, Limerick, Ireland.*

Gallo, E. and Tsingos, N. (2007). Extracting and re-rendering structured auditory scenes from field recordings. *30th International Conference: Intelligent Audio Environments*.

Gallo, E., Tsingos, N., and Lemaitre, G. (2007). 3d-audio matting, postediting, and rerendering from field recordings. *EURASIP J. Appl. Signal Process.*, pages 183–183.

Gardner, W. (1997). Reverberation algorithms. In Kahrs, M. and Brandenburg, K., editors, *Applications of Digital Signal Processing to Audio and Acoustics*, pages 85–131. Kluwer Academic Publishers.

Gebhard, J. W. and Mowbray, G. H. (1959). On discriminating the rate of visual flicker and auditory flutter. *Am J Psychol*, 72:521–529.

Gerzon, M. (1985). Ambisonics in multichannel broadcasting and video. *J. Audio Eng. Soc.*, 33(11):859–871.

Getzmann, S. (2007). The effect of brief auditory stimuli on visual apparent motion. *Perception*, 36(7):1089–1103.

Grelaud, D., Bonneel, N., Wimmer, M., Asselot, M., and Drettakis, G. (2009). Efficient and practical audio-visual rendering for games using crossmodal perception. In *I3D '09: Proceedings of the 2009 symposium on Interactive 3D graphics and games*, pages 177–182, New York, NY, USA. ACM.

Gumerov, N. A. and Duraiswami, R. (2008). A broadband fast multipole accelerated boundary element method for the three dimensional helmholtz equation. *The Journal of the Acoustical Society of America*, 125(1):191–205.

Guttman, S. E., Gilroy, L. A., and Blake, R. (2005). Hearing what the eyes see: Auditory encoding of visual temporal sequences. *Psychological Science*, 16(3):228–235.

Haas, E. C. and Casali, J. C. (1995). Peceived urgency of and response time to multi-tone and frequency-modulated warning signals in broadband noise. *Ergonomics*, 38(11):2313–2326.

Hartmann, W. (1983). Localization of sound in rooms. *Journal of the Acoustical Society of America*, 74(5):1380–1391.

Hartmann, W. M. (1997). *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates.

Harvey, C., Walker, S., Bashford-Rogers, T., Debattista, K., and Chalmers, A. (2010). The Effect of Discretised and Fully Converged Spatialised Sound on Directional Attention and Distraction. pages 191–198, Sheffield, United Kingdom. Eurographics Association.

Heckbert, P. S. and Hanrahan, P. (1984). Beam tracing polygonal objects. *SIGGRAPH '84: Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pages 119–127.

Hendrix, C. M. and Barfield, W. (1996). Presence within virtual environments as a function of visual display parameters. *Presence*, 5(3):274–289.

Herder, J. (1999). Optimization of sound spatialization resource management through clustering. *The Journal of Three Dimensional Images, 3D-Forum Society*, 13(3):59–65.

Herrere, P., Serra, X., and Peeters, G. (1999). Audio descriptors and descriptors schemes in the context of mpeg-7. *Procedings of International Computer Music Conference (ICMC99)*.

Hobson, E. (1955). *The Theory of Spherical and Ellipsoidal Harmonics*. Chelsea Pub Co.

Horry, Y., Anjyo, K.-I., and Arai, K. (1997). Tour into the picture: using a spidery mesh interface to make animation from a single image. In *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 225–232, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Howard, I. P. and Templeton, W. B. (1966). *Human spatial orientation [by] I.P. Howard and W.B. Templeton*. Wiley, London, New York,.

Hulusic, V., Aranha, M., and Chalmers, A. (2008). The influence of cross-modal interaction on perceived rendering quality thresholds. In Skala, V., editor, *WSCG 2008 Full Papers Proceedings*, pages 41–48.

Hulusic, V., Czanner, G., Debattista, K., Sikudova, E., Dubla, P., and Chalmers, A. (2009). Investigation of the beat rate effect on frame rate for animated content. In Hauser, H., editor, *Spring Conference on Computer Graphics 2009*, pages 167–174. Comenius University, Bratislava.

Hulusic, V., Debattista, K., Aggarwal, V., and Chalmers, A. (2010a). Exploiting audio-visual cross-modal interaction to reduce computational requirements in interactive environments. In *Proceedings of the IEEE conference on Games and Virtual Worlds for Serious Applications*. IEEE Computer Society.

Hulusic, V., Debattista, K., Aggarwal, V., and Chalmers, A. (2010b). Maintaining frame rate perception in interactive environments byexploiting audio-visual cross-modal interaction. *The Visual Computer*, pages 1–10. 10.1007/s00371-010-0514-2.

Humphreys, G. W. and Bruce, V. (1989). *Visual Cognition: Computational, Experimental and Neuropsychological Perspectives*. Lawrence Erlbaum Associates Ltd, East Sussex, BN3 2FA, UK.

Ihlenburg, F. (1998). *Finite Element Analysis of Acoustic Scattering*. SpringerVerlag, New York.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nat Rev Neurosci*, 2(3):194–203.

James, D. L., Barbič, J., and Pai, D. K. (2006). Precomputed acoustic transfer: Output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM Transactions on Graphics (SIGGRAPH 2006)*, 25(3).

James, W. (1890). *The principles of psychology*. Holt, New York.

James, W. (1892). *Psychology*. McMilan and Co.

Jedrzejewski, M. and Marasek, K. (2006). Computation of room acoustics using programmable video hardware. In *Computer Vision and Graphics*, volume 32 of *Computational Imaging and Vision*, pages 587–592. Springer Netherlands.

Jehan, T. (2005). *Creating Music by Listening*. PhD thesis, M.I.T.

Jensen, H. W. (1996). Global illumination using photon maps. *Proceedings of the eurographics workshop on Rendering techniques '96*, pages 21–30.

Jot, J., Larcher, V., and Warusfel, O. (1995). Digital signal processing issues in the context of binaural and transaural stereophony. *Proc. 98th Audio Engineering Society Convention*.

Jot, J.-M. (1999). Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. *Multimedia Systems*, 7(1):55–69.

Jot, J.-M., Larcher, V., and Pernaux, J.-M. (1999). A comparative study of 3D audio encoding and rendering techniques. *Proceedings of the AES 16th international conference, Spatial sound reproduction, Rovaniemi, Finland*.

Jourjine, A., Rickard, S., and Yilmaz, O. (2000). Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP00)*.

J.R.Parker and Behm, B. (2004). Generating audio textures by example. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP04)*.

Kajiya, J. T. (1986). The Rendering Equation. *SIGGRAPH Comput. Graph.*, 20(4):143–150.

Kapralos, B., Jenkin, M., and Milios, E. (2004). Acoustic modeling utilizing an acoustic version of photon mapping. *In Proc. of IEEE Workshop on HAVE*.

Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15(21):1943–1947.

Kelly, M. and Tew., A. (2002). The continuity illusion in virtual auditory space. *Proc. of the 112th AES Conv., Munich, Germany*.

Kelly, M. C. and Tew, A. I. (2002). The continuity illusion in virtual auditory space. In *in proc. of AES 112th Convention*, Munich, Germany.

Kludszuweit, A. (1991). Time iterative boundary element method (TIBEM) - a new numerical method of four-dimensional system analysis for the calculation of the spatial Impulse Response. *Acustica*, 75:17–27.

Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227.

Kohlrausch, A. and der Par, S. V. (2005). Audio-visual interaction in the context of multimedia applications. In *Communication Acoustics*, pages 109–138. Springer Verlag.

Kohlrausch, A. and van de Par, S. (2005). *Communication Acoustics*, chapter AudioVisual Interaction in the Context of Multi-Media Applications, pages 109–138. Springer.

Kopuz, S. and Lalor, N. (1995). Analysis of interior acoustic fields using the finite element method and the boundary element method. *Applied Acoustics*, 45:193–210.
Kozlowski, O. and Kautz, J. (2007). Is accurate occlusion of glossy reflections necessary? In *APGV '07: Proceedings of the 4th symposium on Applied perception in graphics and visualization*, pages 91–98, New York, NY, USA. ACM.
Krokstad, A., Strøm, S., and Sørsdal, S. (1968). Calculating the acoustical room response by the use of a ray tracing technique. *J. Sound Vib.*, 8:118–125.
Kurniawati, E., Absar, J., George, S., Lau, C. T., and Premkumar, B. (2002). The significance of tonality index and nonlinear psychoacoustics models for masking threshold estimation. In *Proceedings of the International Conference on Virtual, Synthetic and Entertainment Audio AES22*.
Kuttruff, H. (1991). *Room Acoustics (3rd edition)*. Elseiver Applied Science.
Laborie, A., Bruno, R., and Montoya, S. (2003). A new comprehensive approach of surround sound recording. *114th convention of the Audio Engineering Society, preprint 5717*.
Laborie, A., Bruno, R., and Montoya, S. (2004). High spatial resolution multi-channel recording. *Proc. 116th convention of the Audio Engineering Society, preprint 6116*.
Lagrange, M. and Marchand, S. (2001). Real-time additive synthesis of sound by taking advantage of psychoacoustics. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland, December 6-8*.
Laine, S., Siltanen, S., Lokki, T., , and Savioja, L. (2009). Accelerated beam tracing algorithm. *Applied Acoustics*, 70:172–181.
Larcher, V., Laborie, A., Bruno, R., and Montoya, S. (2004). Techniques de spatialisation des sons. In Pachet, F. and Briot, J.-P., editors, *Informatique Musicale - du signal au signe musical*. Hermès science.
Larsen, A., McIlhagga, W., Baert, J., and Bundesen, C. (2003). Seeing or hearing? perceptual independence, modality confusions, and crossmodal congruity effects with focused and divided attention. *Percept Psychophys*, 65(4):568–574.
Larsson, P., Västfjäll, D., and Kleiner, M. (2002). Better presence and performance in virtual environments by improved binaural sound rendering. *proceedings of the AES 22nd Intl. Conf. on virtual, synthetic and entertainment audio, Espoo, Finland*, pages 31–38.
Lauterbach, C., Chandak, A., and Manocha, D. (2007). Interactive sound propagation in dynamic scenes using frustum tracing. *IEEE Trans. on Visualization and Computer Graphics*, 13:1672–1679.
Leese, M. J. (1998). Ambisonic surround sound FAQ (version 2.8). http://members.tripod.com/martin_leese/Ambisonic/.
Lehnert, H. (1993). Systematic errors of the ray-tracing algorithm. *J. Applied Acoustics*, 38:207–221.
Lewicki, M. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363.
Liu, Q. (1997). The pstd algorithm: A time-domain method combining the pseudospectral technique and perfectly matched layers. *The Journal of the Acoustical Society of America*, 101(5):3182–3182.
Logan, B. (2000). Mel frequency cepstral coefficients for music modelling. *Proceedings of the International Symposium on Music Information Retrieval (Music IR 2000)*.
Lokki, T., Hiipalla, T., and Savioja, L. (2001). A framework for evaluating virtual acoustic environments. *AES 110th convention, Berlin, preprint 5317*.
Longhurst, P., Debattista, K., and Chalmers, A. (2006). A gpu based saliency map for high-fidelity selective rendering. In *Proceedings of the 4th international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*, AFRIGRAPH '06, pages 21–29, New York, NY, USA. ACM.
Loschky, L. C. and McConkie, G. W. (2000). User performance with gaze contingent multiresolutional displays. In *ETRA '00: Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 97–103, New York, NY, USA. ACM.
Lu, L., Wenyin, L., and Zhang, H.-J. (2004). Audio textures: Theory and applications. *IEEE Transactions on Speech and Audio Processing*, 12(2):156–167.
Macedonia, M. R., Zyda, M. J., Pratt, D. R., Brutzman, D. P., and Barham, P. T. (1995). Exploiting Reality with Multicast Groups. *IEEE Computer Graphics and Applications*, 15:38–45.
Mack, A. and Rock, I. (1998). *Inattentional Blindness*. The MIT Press.
Macknik, S. and Martinez-Conde, S. (2009). *Encyclopedia of Perception*, chapter Vision: te, pages 1060–1062. SAGE Press.
Malham, D. (2001). Spherical harmonic coding of sound objects - the ambisonic 'O' format. *Proc. of the 19th AES Conference, Surround Sound, Techniques, Technology and Perception, Schloss Elmau, Germany*.
Malham, D. and Myatt, A. (1995). 3d sound spatialization using ambisonic techniques. *Computer Music Journal*, 19(4):58–70.
Manocha, D., Lin, M., Calamia, P., Savioja, L., and Tsingos, N. (2009). Interactive sound rendering. *SIGGRAPH2009 Course Notes*.
Maragos, P., Gros, P., Katsamanis, A., and Papandreou, G. (2008). *Cross-Modal Integration for Performance Improving in Multimedia: A Review*. Springer-Verlag.
Massaro, D. W. and Warner, D. S. (1977). Dividing attention between auditory and visual perception. *Perception & Psychophysics*, 21(6):569–574.
Mastoropoulou, G. (2006). *The Effect of Audio on the Visual Perception of High-Fidelity Animated 3D Computer Graphics*. PhD in Computer science, University of Bristol.
Mastoropoulou, G. and Chalmers, A. (2004). The effect of music on the perception of display rate and duration of animated sequences: An experimental study. In *TPCG '04: Proceedings of the Theory and Practice of Computer Graphics 2004 (TPCG'04)*, pages 128–134, Washington, DC, USA. IEEE Computer Society.
Mastoropoulou, G., Debattista, K., Chalmers, A., and Troscianko, T. (2005a). Auditory bias of visual attention for perceptually-guided selective rendering of animations. In *GRAPHITE '05: Proceedings of the 3rd international conference on Computer graphics and interactive techniques in Australasia and South East Asia*, pages 363–369, New York, NY, USA. ACM Press.
Mastoropoulou, G., Debattista, K., Chalmers, A., and Troscianko, T. (2005b). The influence of sound effects on the perceived smoothness of rendered animations. In *APGV '05: Proceedings of the 2nd symposium on Applied perception in graphics and visualization*, pages 9–15, New York, NY, USA. ACM Press.
Mcgurk, H. and Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.
Mehra, R., Raghuvanshi, N., Savioja, L., Lin, M. C., and Manocha, D. (2011). An efficient gpu-based time domain solver for the acoustic wave equation. *Applied Acoustics*, 29(3).
Menzies, D. (2002). W-Panning and O-format, tools for object spatialization. *ICAD*.
Merimaa, J. (2002). Applications of a 3D microphone array. *112th AES convention, preprint 5501*.
Merimaa, J. and Pulkki, V. (2005). Spatial impulse response rendering I: Analysis and synthesis. *J. Audio Eng. Soc.*, 53:1115–1127.
Merimaa, J. and Pullki, V. (2004). Spatial impulse response rendering. *Proc. of the 7th Intl. Conf. on Digital Audio Effects (DAFX'04), Naples, Italy*.
Meyer, J. and Elko, G. (2004a). Spherical microphone arrays for 3d sound recording. *Chap. 2 in Audio Signal Processing for next-generation multimedia communication systems, Eds. Yiteng (Arden) Huang and Jacob Benesty, Bosten, Kluwer Academic Publisher*.
Meyer, J. and Elko, G. (2004b). Spherical microphone arrays for 3d sound recording. *chap. 2 in Audio Signal Processing for next-generation multimedia communication systems, Eds. Yiteng (Arden) Huang and Jacob Benesty, Bosten, Kluwer Academic Publisher*.
Moeck, T., Bonneel, N., Tsingos, N., Drettakis, G., Viaud-Delmon, I., and Alloza, D. (2007). Progressive perceptual audio rendering of complex scenes. In *I3D '07: Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 189–196, New York, NY, USA. ACM.
Møller, H. (1989). Reproduction of artificial-head recordings through loudspeakers. *J. Audio Eng. Soc.*, 37(1/2):30–33.
Møller, H. (1992). Fundamentals of binaural technology. *Applied Acoustics*, 36:171–218.
Moore, B. C. (1982). *An Introduction to the Psychology of Hearing*. Academic Press, 2nd edition.
Moore, B. C. (1997). *An introduction to the psychology of hearing*. Academic Press, 4th edition.
Morein-Zamir, S., Soto-Faraco, S., and Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Brain Res Cogn Brain Res*, 17(1):154–163.
Moss, W., Yeh, H., Hong, J.-M., Lin, M. C., and Manocha, D. (2010). Sounding liquids: Automatic sound synthesis from fluid simulation. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*.
M.Rath, Avanzini, F., Bernardini, N., Borin, G., Fontana, F., Ottaviani, L., and Rochesso, D. (2003). An introductory catalog of computer-synthesized contact sounds, in real-time. *Proc. of the XIV Colloquium on Musical Informatics, Firenze, Italy*.
Mullen, J., Howard, D., and Murphy, D. (2006). Waveguide physical modeling of vocal tract acoustics: Improved formant bandwidth control from increased model dimensionality. In *IEEE Transactions on Speech and Audio Processing*, volume 3, pages 964–971.
Murphy, D., Kelloniemi, A., Mullen, J., and Shelley, S. (2007). Acoustic modeling using the digital waveguide mesh. *Signal Processing Magazine, IEEE*, 24(2):55–66.
Naylor, J. (1993). ODEON - another Hybrid Room Acoustical Model. *Applied Acoustics*, 38(1):131–143.
Nielsen, S. H. (1993). Auditory Distance Perception in Different Rooms. *J. Audio Eng. Soc.*, 41(10):755–770.
O'Brien, J. F., Cook, P. R., and Essl, G. (2001). Synthesizing sounds from physically based motion. In *In Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 529–536. ACM.
O'Brien, J. F., Shen, C., and Gatchalian, C. (2002). Synthesizing sounds from rigid-body simulations. In *In Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 175–181. ACM.
O'Grady, P., Pearlmutter, B., and Rickard, S. (2005). Survey of sparse and nonsparse methods in source separation. *Intl. Journal on Imaging Systems and Technology (IJIST), special issue on Blind source separation and deconvolution in imaging and image processing*.
Painter, E. M. and Spanias, A. S. (2000). Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4).
Parker, J. and Chan, S. (2003). Sound synthesis for the web, games, and virtual reality. *International Conference on Computer Graphics and Interactive Techniques*.
Pashler, H. (1999). *The psychology of attention*. The MIT Press.
Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. *Cuidado project report, Institute of Research and Musical Coordination (IRCAM)*.
Pellegrini, R. (2001a). Quality assessment of auditory virtual environments. *icad*, (ICAD2001).
Pellegrini, R. (2001b). *A virtual Listening Room as an application of auditory virtual Environment*. PhD thesis, Ruhr-Universitat, Bochum.
Perrott, D. R. and Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, 87(4):1728–1731.
Potard, G. and Burnett, I. (2004). Decorrelation techniques for the rendering of apparent source width in 3D audio displays. *Proc. of 7th Intl. Conf. on Digital Audio Effects (DAFX'04), Naples, Italy*.
Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc.*, 45(6):456–466.
Pulkki, V. (2006). Directional audio coding in spatial sound reproduction and stereo upmixing. In *28th International Conference: The Future of Audio Technology - Surround and Beyond*.
Pylyshyn, Z. W. (2006). *Seeing and Visualizing: It's not what you Think*. MIT Press.
Radke, R. and Rickard, S. (2002). Audio interpolation. In *the Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio (AES'02)*, pages 51–57.
Raghuvanshi, N., Galoppo, N., and Lin, M. (2008). Accelerated wave-based acoustics simulation. *SPM '08: Proceedings of the 2008 ACM symposium on Solid and physical modeling*, pages 91–102.
Raghuvanshi, N., Lauterbach, C., Chandak, A., Manocha, D., and Lin, M. C. (2007). Real-time sound synthesis and propagation for games. *Commun. ACM*, 50(7):66–73.
Raghuvanshi, N. and Lin, M. C. (2006). Interactive sound synthesis for large scale environments. In *I3D '06: Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 101–108. ACM.
Raghuvanshi, N., Narain, R., and Lin, M. (2009). Efficient and accurate sound propagation using adaptive rectangular decomposition. *IEEE Transactions on Visualization and Computer Graphics*, 15(5):789–801.
Raghuvanshi, N., Snyder, J., Mehra, R., Lin, M. C., and Govindaraju, N. (2010). Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. *ACM Trans Graph (proceedings of SIGGRAPH 2010)*, 29(3).
Rajkumar, A., Naylor, B. F., Feisullin, F., and Rogers, L. (1996). Predicting rf coverage in large environments using ray-beam tracing and partitioning tree represented geometry. *Wirel. Netw.*, 2(2):143–154.
Ramanarayanan, G., Bala, K., and Ferwerda, J. A. (2008). Perception of complex aggregates. In *SIGGRAPH '08: ACM SIGGRAPH 2008 papers*, pages 1–10, New York, NY, USA. ACM.
Ramanarayanan, G., Ferwerda, J., Walter, B., and Bala, K. (2007). Visual equivalence: towards a new standard for image fidelity. *ACM Trans. Graph.*, 26(3):76.
Rangachar, R. (2001). Analysis and improvement of the MPEG-1 audio layer III algorithm at low bit-rates. Master of science thesis, Arizona State University.
Rber, N., Kaminski, U., and Masuch, M. (2007). Ray acoustics using computer graphics technology. In *In Proc. 10th Intl. Conf. on Digital Audio Effects (DAFx07), Bordeaux*, pages 274–279.
Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of neurophysiology*, 89:1078–1093.
Rickard, S. (2006). Sparse sources are separated sources. *Proceedings of the 16th Annual European Signal Processing Conference, Florence, Italy*.
Rocchesso, D. (2002). Spatial effects. In Ed., U. Z., editor, *DAFX - Digital Audio Effects*, page Chapter 6. Wiley.
Rocchesso, D., Bresin, R., and Frenström, M. (2003). Sounding objects. *IEEE Multimedia*, 10(2):42–52.
Rock, I., Linnett, C. M., Grant, P., and Mack, A. (1992). Perception without attention: results of a new method. *Cognit Psychol*, 24(4):502–534.
Roget, P. M. (1825). Explanation of an Optical Deception in the Appearance of the Spokes of a Wheel Seen through Vertical Apertures. *Philosophical Transactions of the Royal Society of London (1776-1886)*, 115(-1):131–140.
Roorda, A. (2002). *Human Visual System - Image Formation*, volume 1, pages 539–557.
Saint-Arnaud, N. and Popat, K. (1998). Analysis and synthesis of sound textures.

Sakamoto, S., Nagamoto, H., Ushiyama, A., and Tachibana, H. (2008). Calculation of impulse responses and acoustic parameters in a hall by the finite-difference time-domain method. *Acoustical Science and Technology*, 29(4).
Sakamoto, S., Ushiyama, A., and Nagatomo, H. (2006). Numerical analysis of sound propagation in rooms using the finite difference time domain method. *The Journal of the Acoustical Society of America*, 120(5):3008.
Savioja, L., Huopaniemi, J., Huotilainen, T., and Takala, T. (1996). Real-time virtual audio reality. In *In Proc. ICMC 1996*, pages 107–110.
Savioja, L., Huopaniemi, J., Lokki, T., and Vn, R. (1999). Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47(9).
Savioja, L., Manocha, D., and Lin, M. C. (2010). Use of gpus in room acoustic modeling and auralization. *In: Proceedings of the international symposium on room acoustics*.
Savioja, L. and Valimaki, V. (2001). Interpolated 3-d digital waveguide mesh with frequency warping. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 01)*, volume 5, pages 3345–3348.
Sawada, H., Araki, S., Mukai, R., and Makino, S. (2006). Blind extraction of dominant target sources using ica and time-frequency masking. *IEEE Trans. Audio, Speech, and Language Processing*.
Scholl, B. J. (2001). Objects and attention: the state of the art. *Cognition*, 80(1-2):1–46.
Schroeder, M. (1962). Natural sounding artificial reverberation. *JAES*, 10(3):219–223.
Sekuler, R., Sekuler, A. B., and Lau, R. (1997). Sound alters visual motion perception. *Nature*, 385(6614):308.
Shams, L., Kamitani, Y., and Shimojo, S. (2000). What you see is what you hear. *Nature*, 408:788+.
Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. *Brain Res Cogn Brain Res*, 14(1):147–152.
Shams, L., Kamitani, Y., and Shimojo, S. (2004). Modulations of visual perception by sound. in the handbook of multisensory processes. pages 27–33.
Shimojo, S. and Shams, L. (2001). Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 11(4):505–509.
Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science*, 145:1328–1330.
Siltanen, S., Lokki, T., Kiminki, S., and Savioja, L. (2007). The room acoustic rendering equation. *J. Acoust. Soc. Am.*, 122(3):1624–1632.
Simons, D. and Chabris, C. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *perception*, 28:1059–1074.
Spence, C., Jane, R., and Driver, J. (2000). Cross-modal selective attention: on the difficulty of ignoring sounds at the locus of visual attention. *Perception & Psychophysics*, 62(2):410–424.
Spence, C. and McDonald, J. (2004). *The Handbook of Multisensory Processes*, chapter The Cross-, pages 3–25. MIT Press, Cambridge, MA.
Staal, H. E. and Donderi, D. C. (1983). The effect of sound on visual apparent movement. *The American Journal of Psychology*, 96(1):95–105.
Stein, B. E., London, N., Wilkinson, L. K., and Price, D. D. (1996). Enhancement of Perceived Visual Intensity by Auditory Stimuli: A Psychophysical Analysis. *J Cog Neurosci*, 8(6):497–506.
Steinman, R. M., Pizlo, Z., and Pizlo, F. J. (2000). Phi is not beta, and why wertheimers discovery launched the gestalt revolution. *Vision Research*, 40(17):2257–2264.
Storms, R. L. (1998). *Auditory-Visual Cross-Modal Perception Phenomena*. {PhD} thesis, Naval Postgraduate School.
Strar, W. and Wand, M. (2004). Multi-resolution sound rendering. *Symp. Point-Based Graphics*.
Strayer, D. L. and Johnston, W. A. (2001). Driven to distraction: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone. *Psychol. Sci.*, 12(6):462–466.
Streicher, R. The decca tree. web. http://mixonline.com/recording/applications/audio_decca_tree/.
Streicher, R. The decca tree – it's not just for stereo anymore. http://www.wesdooley.com/pdf/Surround_Sound_Decca_Tree-urtext.pdf.
Streicher, R. and Everest, F., editors (1998). *The new stereo soundbook, 2nd edition*. Audio Engineering Associate, Pasadena (CA), USA.
Theeuwes, J. (1991). Exogenous and endogenous control of attention: the effect of visual onsets and offsets. *Perception & psychophysics*, 49(1):83–90.
Touimi, A. B. (2000). A generic framework for filtering in subband domain. In *In Proceeding of IEEE 9th Workshop on Digital Signal Processing, Hunt, Texas, USA*.
Touimi, A. B., Emerit, M., and Pernaux, J.-M. (2004). Efficient method for multiple compressed audio streams spatialization. In *In Proceeding of ACM 3rd Intl. Conf. on Mobile and Ubiquitous multimedia*.
Tsingos, N. (2005). Scalable perceptual mixing and filtering of audio signals using an augmented spectral representation. *Proc. of 8th Intl. Conf. on Digital Audio Effects (DAFX'05), Madrid, Spain*.
Tsingos, N. (2007). Perceptually-based auralization. In *19th Intl. Congress on Acoustics*.
Tsingos, N. (2009a). Pre-computing geometry-based reverberation effects for games. *35th AES Conference on Audio for Games, London*.
Tsingos, N. (2009b). Using programmable graphics hardware for acoustics and audio rendering. In *Audio Engineering Society Convention 127*.
Tsingos, N., Dachsbacher, C., Lefebvre, S., and Dellepiane, M. (2007). Instant sound scattering. In *Rendering Techniques (Proceedings of the Eurographics Symposium on Rendering)*.
Tsingos, N., Funkhouser, T., Ngan, A., and Carlbom, I. (2001). Modeling acoustics in virtual environments using the uniform theory of diffraction. *In Proc. of ACM SIGGRAPH*, pages 545–552.
Tsingos, N., Gallo, E., and Drettakis, G. (2004). Perceptual audio rendering of complex virtual environments. *ACM Trans. Graph.*, 23(3):249–258.
van den Doel, K., Knott, D., and Pai, D. K. (2004). Interactive simulation of complex audio-visual scenes. *Presence: Teleoperators and Virtual Environments*, 13(1).
van den Doel, K., Kry, P. G., and Pai, D. K. (2001). Foleyautomatic: Physically based sound effects for interactive simulation and animation. *ACM Computer Graphics, SIGGRAPH'01 Proceedings*, pages 545–552.
van den Doel, K., Pai, D. K., Adam, T., Kortchmar, L., and Pichora-Fuller, K. (2002). Measurements of perceptual quality of contact sound models. In *Proceedings of the International Conference on Auditory Display (ICAD 2002), Kyoto, Japan*, pages 345–349.
Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of experimental psychology. Human perception and performance*, 34(5):1053–1065.
Vickers, E., Krishnan, P., and Sadanandam, R. (2006). Frequency domain artificial reverberation using spectral magnitude decay. *Proceedings of the 121th AES convention, Preprint 6926*.
Vincent, E., Rodet, X., Röbel, A., Févotte, C., Carpentier, E. L., Gribonval, R., Benaroya, L., and Bimbot, F. (2003). A tentative typology of audio source separation tasks. *Proc. of the 4th Intl. Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan*.
Vam A. and Soto-Faraco, S. (2007). Audio-visual interactions in dynamic scenes: implications for multisensory compression. In *Invited paper at 9th International Congress on Acoustics ICA07*.
Vam A. and Tajadura-Jimz, A. (2007). Perceptual optimization of audio-visual media: Moved by sound. In *Narration and Spectatorship in Moving Images*. Cambridge Scholars Press.
Vroomen, J., Bertelson, P., and Gelder, B. d. (1998). A visual influence in the discrimination of auditory location.
Vroomen, J. and de Gelder, B. (2004). Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon. in the handbook of multisensory processes. pages 140–150.
Wada, Y., Kitagawa, N., and Noguchi, K. (2003). Audio-visual integration in temporal perception. *Int J Psychophysiol*, 50(1-2):117–124.
Wagenaars, W. M. (1990). Localization of Sound in a Room with Reflecting Walls. *J. Audio Eng. Soc.*, 38(3).
Ward, J., Huckstep, B., and Tsakaniko, E. (2006). Sound-colour synaesthesia: to what extent does it use cross-modal mechanisms common to us all? In *Cortex*, volume 42, pages 264–280.
Warren, R. M., Wrightson, J. M., and Puretz, J. (1988). Illusory continuity of tonal and infratonal periodic sounds. *The Journal of the Acoustical Society of America*, 84(4):1338–1342.
Welch, R. B. and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological bulletin*, 88(3):638–667.
Whitted, T. (1979). An improved illumination model for shaded display. In *SIGGRAPH '79: Proceedings of the 6th annual conference on Computer graphics and interactive techniques*, page 14, New York, NY, USA. ACM.
Yarbus, A. L. (1967). *Eye Movements and Vision*. Plenum Press, New York, NY.
Yee, H., Pattanaik, S., and Greenberg, D. P. (2001). Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Trans. Graph.*, 20(1):39–65.
Yewdall, D. (2003). *Practical Art of Motion Picture Sound (2nd edition)*. Focal Press.
Yilmaz, O. and Rickard, S. (2004). Blind separation of speech mixtures via time frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847.
Zheng, C. and James, D. L. (2009). Harmonic fluids. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*.
Zheng, C. and James, D. L. (2010). Rigid-body fracture sound with precomputed soundbanks. *ACM Transactions on Graphics (SIGGRAPH 2011)*, 29(3):69:1–69:13.
Zheng, C. and James, D. L. (2011). Toward high-quality modal contact sound. *ACM Transactions on Graphics (SIGGRAPH 2011)*, 30(4).