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

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

[1]International Digital Laboratory, WMG, University of Warwick, UK
[2]Dolby Laboratories, San Francisco, CA, USA
[3]Arup, London, UK
[4]Department of Electronics, University of York, UK

Abstract

In recent years research in the 3-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. However, 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 some findings which may be exploited in other fields. This STAR 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.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-dimensional Graphics and Realism—Virtual Reality I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—Physically based modeling I.6.8 [Simulation and Modeling]: Types of Simulation—Animation

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 STAR 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 [Cat, Nay93], advanced multimedia applications in Virtual Reality to enhance presence [CDG*93, MZP*95] and, more recently, immersive video games [MBT*07, RLC*07, GBW*09]. 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 [DM95] 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 [Bla97].

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 perfor-

† V.Hulusic@warwick.ac.uk
‡ Carlo.Harvey@warwick.ac.uk

mance of a rendering system. Auditory and visual limitations have been exploited in order to decrease the auditory [TGD04, MBT*07] or visual [CCL02, KK07, RFWB07, RBF08] 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 [MDCT05a, HWBR*10]. This can be used for improving the perception of a material quality [BSVDD10], Level-of-Detail (LOD) selection [GBW*09] or for increasing the spatial [MDCT05a, HAC08] and temporal [MDCT05b, HCD*09, HDAC10a, HDAC10b] quality of visuals by coupling it with the corresponding auditory stimulus.

While there have been surveys on acoustic rendering in the past [FJT02, MLC*09] in the field of computer graphics and on aspects of cross modality [SKS04] within the field of psychology, this is one of 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 [KvdP05], with the focus on multi-media applications rather then computer graphics.

2. Acoustics and the Human Auditory System

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

2.1. Sound

Since sound is an oscillation of pressure transmitted in a wave, modelling sound propagation is, 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 $343.2m \cdot s^{-1}$. 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: $c = \sqrt{\gamma \cdot \frac{p}{\rho}}$. As a reference sound travels at roughly 4.3 times faster in liquids and 15 times faster in non-porous solids. These travel delays are audible to humans and as lights travel time is typically ignored during light transport simulation, this cannot be the case when simulating acoustic transport. Delay and amplitude along travel paths must be encoded into the Impulse Response.

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 $343.26m \cdot s^{-1}$ 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 and this assertion remains until a source of distortions scale upon a plane is larger than that of the sound signals wavelength impingent 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 key to note that as 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

Time and Phase Dependance: 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 travelling 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 $R = \frac{S}{r}$, where $r = \sqrt{(R_x - S_x)^2 + (R_y - S_y)^2 + (R_z - S_z)^2}$. In addition to this attenuation, materials that are collided with 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. R 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).

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

More simple, yet acceptable, methods exist using a scalar across frequency octave bands (125, 250, 500, 1000, 2000, and 4000Hz). The absorption coefficient is the energy ratio between the absorbed and the incident energies. $R(\omega)$ is the pressure of the wave reflected from the surface at a given frequency and $\alpha(\omega)$ is the frequency dependant absorption coefficient on a scale of 0 to 1 calculated as $\alpha(\omega) = 1 - |R(\omega)|^2$. Such that an absorption coefficient of 0.9 at a frequency of 4kHz would reflect 90% of the pressure of the incoming wave into the exiting wave at 4kHz. Frequency dependant materials profiles can be created for various absorbers, either through industrial or independent measurements.

2.2. 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 channeled 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- lled 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 molecule density. Therefore, the ossicles, besides transmitting, also amplify the vibrations from the outer ear into the inner ear. The ossicles consist of three bones: hammer, anvil and stirrup. In order for the middle ear to function correctly, the air pressure must be equal to the atmospheric pressure in the ear canal. The mechanism for the pressure equalisation is provided by the Eustachian tube, the small canal connecting the middle ear and the throat.

The inner ear consists of few parts and two major functions: maintaining the balance and orientation in space; and frequency and intensity analysis. The first function is achieved through a specialised sensory system called semicircular canals. The other part of the inner ear, responsible for hearing, is the cochlea. The cochlea is spiral shaped and comprises of three chambers: vestibular canal, cochlear duct and tympanic canal. The fi rst and the last are connected at the end (a place called the apex). The vibrations from the middle ear are transmitted through the oval window, located at the base of the vestibular canal. At the base of the tympanic canal there is another tiny membrane, the round window, that compensates the pressure caused by the inward movement of the oval window. The cochlear duct is a separate chamber, containing a di erent type of liquid. It is separated from the tympanic canal by a basilar membrane. On top of the basilar membrane there is a structure named the Organ of Corti, which contains the receptors - hair cells - and transforms the fluid vibrations into neural impulses. More details can be found in [Moo82, Bre93, Yos00, Alt04, BS06].

PART ONE

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 [Kut91]. 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) as shown in Figure 1.

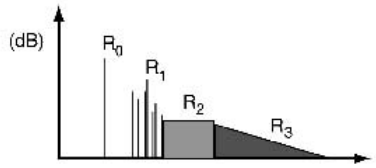

Figure 1: *Impulse response profile from a typical room.*

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 [Beg94, CM78, Har97] allowing a listener to discern some information about the shape and scale of the environment that the sound is reverberating within [Beg94, Har83, Nie93, Wag90]. 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. This is 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 [Ahn93, SHHT96].

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 $y_n = \Sigma i_k \cdot x_{n-k}$, where y is the output waveform, x_n are samples of the audio to be modelled and 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 [Kaj86]. 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 [BW99].

Sound scattering waves from source to a receiver introduce a multitude of different pathways: reflections, refractions, diffraction's 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.

3.3.1. Image Source

Allen et al.'s Image Source Method [AB79]: involved mirroring sound sources across all planes in an environment, constructing virtual sources as shown in Figure 2. 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 [Bor84]: 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 [AB79], a regular lattice of virtual sources ensued. Virtual source position is trivial to calculate in this format

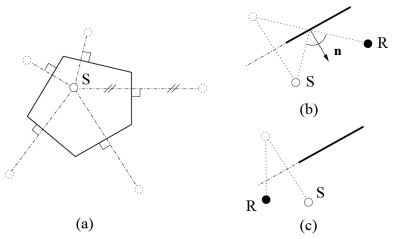

Figure 2: *Virtual source mirroring for the Image Source technique. Figure (a) shows a sound source (S) and its image sources of first order reflection for a pentagon. Figure (b) represents a valid image source for a receiver (R). Figure (c) represents an invalid virtual source because the path reflected between the virtual source and the receiver does not intersect the virtual source's reflector.*

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 the paper.

Savioja et al. [SHLV99]: introduced a novel 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 path reflections with directionality attributes; several key aims to preserve the integrity of the late reverberation information that are used as heuristics to guide the feedback reverberater in this technique are:

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 sources 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.

3.3.2. Finite Element and Boundary Element Methods (FEM and BEM)

Kludszuweit's Time Iterative Boundary Element Method (TIBEM) [Klu91]: 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 [KL95]: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 [Ihl98]: The wave equation is solved using a discrete set of linear equations on elements of subdivided space. At limit, Finite Element Techniques provides 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.

3.3.3. Digital Waveguide Mesh

Campos et al's Mesh [CH05]: 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 [CHD01].

Mullen et al's Waveguide Mesh Vocal Tract Model [MHM06]: 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 [MHM06]. 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 [MKMS07]: 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 intersections. 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 [CH05] have been implemented to mitigate this. Furthermore, frequency warping [SV01] has also been used to deal with this. Due to the dispersion the model is useful for frequencies below the update frequency.

3.3.4. Volumetric Methods

Funkhouser et al.'s Beam Tracing [FCE*98, FMC99]: 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. [HH84]. This removes the problems in sampling and aliasing that plague ray traced approaches as first discussed by Lehnert [Leh93].

Tsingos et al.'s extension based on the Uniform Theory of Diffraction (UTD) [TFNC01]: This builds upon the previous work by Funkhouser et al. [FCE*98] by incorporating the UTD into the model for propagation within the Beam Tracing architecture.

Laine et al.'s Accelerated Beam Tracing Algorithm [LSL*09]: 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.

3.3.5. Particle Based Methods

Kapralos et al.'s Sonel Mapping [KJM04]: 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.

The mechanical wave of sound propagation is approximated via ray tracing 1 or more sonel emitted for each sound source. The trace continues until the sonel encounters a surface. Information carried by each sonel is similar to traced photons (position, velocity: incident direction, energy, distance travelled and frequency). Each sonel represents the frequency distribution for one frequency band. Diffraction is handled by dilating edges of geometry the sonels hit by frequency dependant amount of $\frac{\lambda}{2}$ where λ is the wavelength for the frequency band. This creates a locus around and within the geometry the sonel has hit. These zones are categorised into diffraction zones within the locus of $\frac{\lambda}{2}$ and non-diffraction zones further inside of it dependant upon where the sonel hit. A sonel incident within the non-diffraction zone will either reflect specularly (perfect specular assumed) or diffusely guided by a Russian-roulette strategy. If diffusely the sonel emits across the hemisphere from the incident point. If diffracted the sonel is reflected over the hemisphere randomly about the diffraction point.

The echogram is then estimated from a sonel map generated from each incident hit point and a mix of distributed ray tracing. The second pass is then an acoustical visibility test from the receiver at which point the sampling strategy adopts different strategies for different reflections. For a diffuse reflection the technique uses the sonel map to provide an estimate of the energy leaving the hit point and reaching the receiver via density estimation methods. The energy

is attenuated based on medium, however no note was made about attenuation based on distance travelled. This energy is then added to the accumulated echogram. Specular reflections are handled in the same way as the first pass. Diffraction's of acoustical visibility rays use a modified version of the Huygens-Fresnel principle. Direct sound is computed via shadow rays from receiver to listener.

Using this two pass system for source and receiver means that one pass can be optimised out of recomputation should either source or listener move within the environment. This technique offers some advantages over standard deterministic approaches to sound simulation: the Russian roulette sampling strategy offers adaptability to increase the number of initial samples exiting a source at a trade off: computation time for accuracy. It also offers the ability to navigate arbitrarily lengthy paths. This is an advantage over employing traditional Monte-Carlo techniques because of the exponential running times and multiple new spawns at hit points.

Bertram et al.'s Phonon Tracing [BDM*05]: Inspired by the photorealism obtained by methods such as Photon Mapping [Jen96]; for a given source and listener position, this method computes an RIR based on particle ditributions 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 such that 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. [KJM04] in that it employs a two pass algorithm for emission of phonons and collection of phonon contributions for generation of the Impulse Response. Again operating within frequency bands each phonon is assumed to carry a collection of bands to save on computation cost. Collection of the emitted phonon samples from the map is done via a Gaussian strategy to generate smoother filters since more phonons contribute weighted by their shortest distance.

In addition to not supporting low frequency sound this technique does not consider the properties of air absorption on top of the materials absorbing energy. However in a system derived for high frequency sound scapes the scale of environments applicable to it would tend to be small enough for air absorption to be negligible. As such this is suitable for more complexly detailed environments whereas typically sound simulation environments tend to be modelled more coarsely due to the nature of the wavelength of sound not impacting so severely on reflections, even by corrugated materials.

3.3.6. Ray-Based Methods

Krokstad et al.'s Ray-Traced Acoustical Room Response [KSS68]: A Ray-Traced method, as first introduced to the computer graphics field in the form of ray casting [App68] and recursive ray tracing [Whi79], 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, 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 mis-counted diffraction paths.

To model the ideal Impulse Response all sound reflection paths should be discovered. This being a Monte Carlo approach to ray tracing it samples these paths to give a statistical approximation and whilst higher order reflections can be considered 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.

3.3.7. Volume Sampling

Rajkumar et al.'s Ray-Beam Tracing [RNFR96]: 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.

Lauterbach et al.'s Frustrum Tracing [LCM07]: 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.

3.3.8. GPU Accelerated Approaches

Jedrzejewski at al.'s application of ray based methods to programmable video hardware [JM06]: 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 [TDLD07]: 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 [RKM07]: 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.

A more comprehensive report and overview on the topic of using programmable graphics hardware for acoustics and audio rendering can be found in [Tsi09b].

3.3.9. 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 highlights which drawbacks associated with spatialisation techniques effect which in a succinct manner.

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 [0:40ms] 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 reverberation 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 [Pel01b]. 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 [Gar97, Roc02]. 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 [Sch62, Jot99, Gar97, Roc02].

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 [VKS06, Tsi09a, MP04].

4. Synthesising Virtual Sound Sources

Whilst section 3.3 covers algorithms for generation of sound filters to give a particular sound the prevailing acoustical properties of the propagating enviroment there is a need to generate virtual sound effects for other properties.

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

Technique	Speed	Accuracy	Comment
FEM/BEM	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.
Image Source Methods	Fast	Accurate	Only considers specular reflection paths, diffraction and material scattering is ignored. Drawbacks over low frequency bands.
Ray Tracing	Very Fast	Inaccurate*	Does not natively support diffraction effects. *Only accurate without work arounds for high frequency bands
Beam Tracing	Fast	Accurate	Scattering effects are not accounted for, geometric clipping techniques have always been a bottleneck.
Particle Methods	Slow-Fast	Accurate	Does not natively support diffraction.

Table 1: *Classification and drawbacks of various Sound Synthesis techniques*

sound events (i.e., sampling). One or several recordings, generally monophonic, can be combined to re-synthesised 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 [LWZ04, PC03, JB04, SAP98, AE03, DS03]. Given an example sound, the goal is to synthesise a similar and non-repetitive signal of arbitrary duration. A common approach is concatenative synthesis. They segment the example signal into a collection of short segments or "grains" and compute transitions probabilities for each pair of grains, thus creating a transition graph [LWZ04, Jeh05]. 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 multiscale wavelet analysis [DBJEY*02] or fit parametric models based on the statistics of the input signal [DCH, BJLW*99].

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 [OCE01a, DKP01, OSG02b, RL06, JBP06, BDT*08]. 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 [MAB*03]. 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 [DP98, vdDKP01, vdDPA*02, vdDKP04, OSG02a], 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 [vdDKP01]. In the context of rigid bodies, it is possible to first calculate the matrix of vibration modes using a 3D mesh [OSG02a]. For deformable objects, the synthesis requires more complex calculations; a basis of finite element, which prevents suitability for real time applications [OCE01b].

An alternative synthesis technique is a combined analysis of recordings and resynthesis. For example, one approach measures the acoustical response of real objects [vdDKP01]. 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 we can construct a 2D texture representing the impulse response of the object at different points on its surface. Analysis of recorded results allows extraction of parameters of the main modes of vibration 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 [DYN03] (wind, swish of a sword) or combustion noise and explosions [DYN04]. 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 [Doe04, Doe05]. 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 resonate frequency of a bubble in an infinite volume of water as $f = 3/r$ which leads to the equation for the formation of the sound of a bubble over time as: $Lambda(t) = A \cdot e^{-dt} \sin(2\pi ft)/\ \Lambda(t)$ is the impulse response at time t, e^{-dt} is a decay coefficient, f is Minneart's frequency. 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.

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) [RBF03], 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 [PB04]. In some case, it is possible to individually record the different spatial or directional components of the sound source using directional microphones [AWBW05, Mal01, Men02, ME04a] 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 [DYN03, DYN04, ZJ09, MYH*10].

5. Structured Audio Rendering and Perceptual Optimisations

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. Moreover, the solutions using rendering hardware [EAX04] support only a limited number of simultaneous sound sources, also called "channels". A large number of sound sources is necessary to render a realistic environment. Rendering of early propagation paths also requires rendering many secondary sources. In some applications, like video games, background music can also be rendered spatially using a set of specific 3D sound sources. A problem which is then tackled either via defning many sources either in software or by using dynamic mapping on a limited number of hardware channels. Rendering a scene with multiple sound sources has been researched extensively [Bre90, BvSJC05, BSK05]. A feature of these approaches is mapping the contents of signals to be spatialised for properties of the human listener. In practice, mastering the complexity of the 3D audio rendering process involves three main aspects: the relative importance of different sound sources in the scene, the complexity of the scenes space and complexity in signal processing.

5.1. Perceptual Aspects of Spatial Audio Rendering

Handling 3D audio simulation is a key factor for creating convincing interactive virtual environments. 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 [HB96, LVK02]. Our spatial auditory perception will be solicited for localising objects in direction and distance, discriminating between concurrent audio signals and analysing spatial characteristics of the environment (indoor vs. outdoor contexts, size and materials of the room). Typical situations encountered in interactive applications such as video games and simulators require processing of hundreds or thousands of sources, which is several times over the capabilities of common audio dedicated hardware. 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 [Pel01a, LHS01], 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. Several auditory perceptual properties may be exploited in order to simplify the rendering pipeline with limited impact on the overall perceived audio quality. The general approach is to structure the sound scene by (1) sorting the relative importance of its components, (2) distributing properly the computer resources on the different signal processing operations and (3) handling the spatial complexity of the scene. These techniques, derived from psycho-acoustics, perceptual audio-coding and auditory scene analysis introduce several concepts similar to those found in computer graphics: selective, progressive and scalable rendering (e.g., visibility/view-frustum culling and geometrical/shading level-of-detail).

5.2. Masking and Illusory Continuity

Selective audio processing approaches build upon prior work from the field of perceptual audio coding that exploits auditory masking. When a large number of sources are present in the environment, it is very unlikely that all will be audible due to masking occurring in the human auditory system [Moo97]. This masking mechanism has been successfully exploited in perceptual audio coding (PAC), such as the well known MPEG I Layer 3 (mp3) standard [PS00] and several efficient computational models have been developed in this field. In the context of interactive applications, this approach is thus also linked to the illusion of continuity phenomena [KT02a], 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.

5.3. Importance and Saliency of Sound Sources

Evaluating all possible solutions to the optimisation problem required for optimal rendering of a sound scene would be computationally intractable. An alternative is to use greedy approaches which first require estimating the relative importance of each sources in order to get a good starting point. A key aspect is also to be able to dynamically adapt to the content. Several metrics can be used for this purpose such as energy, loudness or the recently introduced saliency. Recent studies have compared some of these metrics showing that they might achieve different results depending on the nature of the signal (speech, music, ambient sound "textures"). Loudness has been found to be generally leading to better results while energy is a good compromise between complexity and quality.

5.4. Limitations of Spatial Hearing in Complex Soundscapes

Human spatial hearing limitations, as measured through perceivable distance and angular thresholds [Beg94] can be exploited for faster rendering independently of the subsequent signal processing operations. This is useful for applications where the reproduction format is not set in advance. Recent studies have also shown that our auditory localisation is strongly affected in multi-source environments. Localisation performances decrease with increasing number of competing sources [BSK05] showing various effects such as pushing effect (the source localisation is repelled from the masker) or pulling effects (the source localisation is attracted by the masker) which depend on the time and frequency overlapping between the concurrent sources [BvSJC05]. As a result, spatial simplification can probably be performed even more aggressively as the complexity of the scene, in particular the number of sound sources, grows.

5.5. Perceptual Importance of Sound Sources and Auditory Masking

The notion of sound source importance is fundamental to the structure and optimisations of processing techniques. It can guide different types of simplifications of the soundstage. Also, sorting by importance of sound sources is the most common technique used to compress a large number of sources into a smaller subset to define the most important sources for each audio frame. A fundamental question is then to define a good metric of importance. The metric most commonly used estimates the attenuation of different sound sources in the scene (eg, due to the distance, dimming, etc..), possibly combined with information on the duration of the sound source (a sound source which has completed most of it duration can be interrupted more easily). Finally, the user is free to adjust the importance values to give more weight to certain sounds. It is clear that in the event that the sounds are somewhat similar in terms of level or loudness, this approach can yield satisfactory results very efficiently. Nevertheless, in most cases it can lead to a suboptimal solution where perceptual quality will degrade significantly when the number n of playable sources simultaneously decreases. To mitigate these problems, we can draw on two findings. First, changes in sound energy over time in the same signal can be very important. In general, energy varies rapidly and discontinuously. Compared with the geometric criteria that it varies continuously and slowly as the source moves. Accordingly, these variations can be far more important than the attenuation of sources, most of which are in a limited area around the listener, and are attenuated in a similar way.

The combination of the instantaneous energy of the emitted signal in combination with the attenuation is therefore a good criteria to define the importance of a sound source. Recent work on the synthesis phase of sound using this principle supports this hypothesis [GLT05, Tsi05]. Properties of the signal can also be pre-calculated. MPEG7

and other similar standards and work in audio indexing databases [HSP99, Log00, Pee04] are descriptors that can be stored in a wide range of sound signals with a very limited impact on the memory required [TGD04]. Ultimately, this method remains very inefficient while adapting to the characteristics of signals to be processed. When several simultaneous sound sources are incurred, it is very unlikely that we perceive all of the sources separately. Indeed, complex auditory masking phenomena come into play as was the case in audio compression (with standards like MP3 [PS00] for example), various approaches have been developed to take advantage of these phenomena in order to optimize rendering sound synthesis by removing parts of the sound profile that will not be heard. Again, one can draw parallels with the approaches to elimination of hidden parts used to optimise rendering interactive 3D graphics. Lagrange and Van Den Doel [vdDPA*02, LM01, vdDKP04] for example, proposes using a model of an acoustic masking algorithm to speed modal synthesis methods by removing inaudible artefacts. Similarly in [TGD04] algorithms have been proposed to estimate effectively the audible sound sources within the a sound profile. This greedy algorithm starts by sorting sources by importance (In [TGD04] 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 [Ran01, KAG*02]. The algorithm then dynamically determines the number of audible sources. It has also been applied successfully to the optimisation calculations of reverberation by convolution with long impulse responses by cutting the filter into small blocks and considering each block as a separate sound source to be mixed [GLT05, Tsi05]. The measure of the importance of a sound source is not limited necessarily to energy properties within the sounds profile. Other criteria [ELD91, HC95] can also be used to quantify the relative importance of different sound sources from the environment to adapt the signal processing techniques.

5.6. Spatial Level of Detail and Sound Source Clustering: Auditory Impostors

Managing the complexity of the spatial scene is a very important aspect for rendering 3D audio. A large number of effects and processes depend on the spatial position of different sound sources in 3D space. However, our spatial perception of sound has its limitations (eg., frequency masking and temporal precision of sound localisation) [Moo97, Bla97, BvSJC05, BSK05]. Creating simplified representations of the sound stage has its benefits. This is especially the case if the number of simultaneous sound events is large, since we can only devote a limited set of resources to each event, or a subset of those resources [BvSJC05]. To this end, several approaches have been developed to create representations of a hierarchical soundstage. As such, they can be held hand in hand with level-of-detail algorithms and used to simplify the 3D geometry.

For the sake of compatibility with standard rendering approaches, impostor sounds can be constructed as a subset of point sources representing the scenes 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 [TGD04]. It is also necessary to determine a signal equivalent to the impostor noise, eg. 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 [Her99, SW04] or by adaptive clustering, k-means clustering algorithms [TGD04]. The adaptive clustering algorithms have several advantages: they can produce a number of target groups, they concentrate their resolution where it is necessary and can be controlled by a variety of error metrics. In particular, the importance of sound signals can be used to control the grouping of sources [TGD04].

Another similar example of such a technique is "Binaural Cue Coding (BCC)" [BF03, FB03, FM05], 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 [TGD04]. 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 [Bre90] 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 [MM95], 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 [FvDFH90].

5.7. Progressive Signal Representations and Processing Scalability

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 wavelet [DDS02], or to use a frequency representation in Fourier space [Tsi05]. Another family of approaches performs processing on signals directly compressing with the help of a perceptual codec (MPEG I Layer 3 (mp3) standard [PS00]), 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 [Tou00, TEP04]. 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.

6. Rendering From Spatial Recordings

In this section we discuss methods to capture impulse response filters from real world sources [Kut91, Beg94, SHLV99]; 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 [GTL07, GT07]. 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) [MP05] and the extension, Directional Audio Coding (DirAC) [Pul06] 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.

6.1. Coincident Recordings and Directional Decompositions

Processing and compositing live multi-track recordings is of course a widely used method in motion-picture audio production [Yew03]. 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) [Stra, Strb], which historically evolved from stereo recording, can also be used for acquiring a sound-field suitable for restitution in typical cinemalike 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) [Ger85, MM95, DRP98, Lee98, Mer02] or generalized Fourier-Bessel functions [LBM03]. 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 [JLP99]. Most of them also require specific microphones [AW02, ME04b, Sou, LBM04], especially when higher-order decompositions must be captured.

6.2. 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 [AV02, Do04] using simpler omnidirectional microphones in the manner of lumigraphs or view interpolation in computer graphics [CW93, BBM*01, HAA97].

Radke and Rickard [RR02] proposed an approach aimed at interpolating in a physically consistent way the audio signal captured along a line joining two microphone. Their work relies on a time-frequency decomposition of the recordings derived from blind source separation [JRY00]. This approach has been extended by Gallo et al. to arbitrary numbers of microphones sparsely distributed throughout the environment to capture [GTL07].

Other approaches [AV02, Do04] densely sample the plenacoustic function and interpolate it directly. However, these approaches remain mostly theoretical due to the required spatial density of recordings.

6.3. 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 [Bre90]. This usually involves extracting spatial information about the

sound sources and segmenting out their respective content. Spatial feature extraction and restitution Some approaches extract spatial features such as binaural cues (interaural time-difference, interaural level difference, interaural correlation) in several frequency subbands of stereo or surround recordings. A major application of these techniques is efficient multi-channel audio compression [BF03, FB03] 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 [MP04] and analysis of room responses [Mer02] by extracting direction of arrival information from coincident or near-coincident microphone arrays [Pul06].

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 [VRR*03, OPR05]. 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 [Ric06]. Some work has shown that such sparse coding does exists at the cortex level for sensory coding [Lew02]. Several techniques have been proposed such as independent component analysis (ICA) [Com94, SAMM] or the DUET technique [JRY00, YR04] which can extract several sources from a stereophonic signal by building an inter-channel delay/amplitude histogram in Fourier frequency domain. In this aspect, it closely resembles the aforementioned binaural cue coding approach. However, most BSS approaches do not separate sources based on spatial cues, but directly solve for the different source signals assuming a priori mixing models which are often simple. Our context would be very challenging for such techniques which might require knowing the number of sources to extract in advance, or need more sensors than sources in order to explicitly separate the desired signals. In practice, 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 [Ave03].

7. Interfaces for Spatial Audio Reproduction (Auditory Displays)

The last step in the aurilisation pipeline is listening to the sound produced, however this cannot be done trivially, it is necessary to direct the sound to an auditory device designed to recreate the sound field simulated for the listener. A direct parallel to this for the visual graphics community is auto-stereoscopic displays. This section overviews a number of techniques and devices used to provide a listener with the auditory cues derived from the spatialisation simulation:

7.1. Binaural Techniques

These techniques use headphones directly at the ears of the listener (binaural) [JLW95, Møl89, Møl92]. 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.

7.2. 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 [Ste89, SE98]. 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 Inter-Aural Time Difference (ITD) and Inter-Aural Level Difference (ILD). This reproduction creates a phantom source image, which can be freely positioned (or "panned") along the line joining the two speakers.

7.3. 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 [CMRT10]. This type of configuration is also widely use 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 [DFMM99]. A commonly used approach for general 3D loudspeaker arrays is Vector-Based Amplitude Panning (VBAP) [Pul97] which extends stereo pair-wise panning techniques to triples of speakers in 3D space. Despite the existence of a number of empirical recording systems (e.g. Surround microphone trees [Stra]), extending traditional stereophonic recording techniques to multi-channel systems is difficult due to the limitation of the directional response of traditional microphone capsules. This lead to the development of more mathematically grounded sound field analysis theories and corresponding recording systems.

7.4. 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 [BdVV93]. 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 [LLBM04]. Holophony assumes that the sound field is modelled by an infinite number of secondary point sources continuously located on a closed surface (Kirchhoff-Helmoltz theorem) or infinite plane (Rayleigh theorem) which separates the domain of the sound emitting objects and the listening area. A simplified practical implementation has been developed by Berkhout and De Vries [BdVV93], who introduced a set of approximations and associated corrective terms to the original theory.

A related set of approaches, such as Ambisonics [Ger85, MM95, Lee98] 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 [Hob55, LLBM04]. 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 fied can be described independently from the reproduction system.

A recent overview of holophony/wave-field synthesis and Ambisonic techniques can be found in [Ahr10].

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. However, as the frequency up to the spatial aliasing frequency We refer the reader to [LLBM04, Ahr10] for in depth discussion of these effects.

7.5. Comparison and Integration in Virtual Reality Environments

Binaural reproduction techniques generally lead to high quality results but are limited to single-user scenarios. However, it is possible to extend them to multiple users e.g., using wireless headphones. Combined with head-tracking systems, binaural reproduction has been used successfully in numerous virtual or augmented reality applications. Binaural rendering can also be easily integrated into CAVE environments where loudspeakers installation is challenging.

Multi-speaker systems do not suffer from some limitations of binaural approaches, such individual differences between users and offer a very practical solution for multi-user scenarios. Stereophonic or multi-channel surround systems offer an effective reproduction in the horizontal plane and are standardized, widely available and cost-effective. However, their performance quickly degrades outside of standard and well calibrated configurations.

Systems based on wave-field synthesis or harmonic decompositions offer the best compromise between a large listening area and a good spatial reproduction. They also allow for rendering improved distance and auditory parallax effects. However, they require a large number of loudspeakers which makes their integration more difficult in virtual reality environments. As a result, wave-field synthesis systems are often limited to an “acoustical window” in front of the listening area, for instance supporting a projection display. Ambisonic techniques require loudspeakers fully surround the listening area since all loudspeakers, including rear-speakers, contribute to the rendering of a frontal sound source.

7.6. Latency and Synchronization with other Modalities

An important factor for immersive applications where sound reproduction complements other modalities is the global latency of the different rendering systems as well as their synchronization. Several studies have been conducted in order to estimate the impact of global latency of a binaural rendering systems with head-tracking on the sound localisation accuracy. Some work [Wen01, MAWM03] shows that a 500ms latency does not affect the localisation accuracy. Below 250ms, the latency is not significantly perceived by the subjects. Such large thresholds do not extend to the perception of synchronous multi-modal events augmented by auditory feedback (e.g., synthesis of contact sounds). In such scenarios, perceivable synchrony thresholds of about 20ms have been reported and could be as low as a few milliseconds [ALE03, ABAW03]. For a brief overview of these phenomena, we refer the reader to [LMCJ00].

8. 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 [TGD04, KdP05, Tsi07, VTJ07, VSF07].

Computer sound synthesis modeling of virtual environment is clearly in a very mature state with Image Source, Ray/Pyramid/Beam Tracing, FDTD, 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.

The 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 TWO

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, discuss perceptual limitations on attentional resources, and examine how they have been exploited separately and jointly for the benefit of computer graphics algorithms, Figure 3.

9. Human Sensory System (HSS)

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.

9.1. Vision

The Human Visual System (HVS) comprises three major parts: the eye, visual pathways and visual cortex. Each part has its own functionality and relies on the functionality of the other two. The light from an environment is received by the eye, transmitted through the visual pathways and processed in the visual cortex [Roo02, BS06, Kai]. When light passes through the cornea, it enters the lens through a small opening called the pupil. After passing the lens, it travels through the vitreous humour and finally reaches the retina at the back of the eye, which contains photoreceptor cells: rods and cones. The cones are responsible for colours and they are mostly concentrated in the fovea, a small region of the retina with the highest visual acuity. The rods, on the other hand, are mainly sensitive to light and benefit vision in low light conditions. They are concentrated around the fovea and their density decreases towards the periphery of the eye. The photoreceptors are connected to the ganglion cells, which transmit visual information from the retina to the visual cortex in the brain. Although an amazing sensory organ, HVS is limited and is able to process only certain amount of information at any point in time.

9.1.1. Limitations

HVS is sensitive to only a portion of the electromagnetic wavelength spectrum. This segment ranges from around 400nm to 700nm and it is 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 [Jam90, HB89]. The area of the foveal vision covers only 2 degrees 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 10). However, there are some other factors, such as spatial frequency, which might influence the perception [LM00]. 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 [BS06]) than the audition. However, a threshold of the temporal visual sensitivity is 26Hz [FN05], which is more than three times lower than for the audition. Nevertheless, we perceive visual stimuli as continuous thanks to the phenomenon called the flicker fusion. The reason for this is the persistence of vision, which is the ability of the retina to retain an image for a period of 1/20 to 1/5 a second after the exposure [Rog25].

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 [SD83, AA93, SPP00, Get07]. 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 [MMC09].

9.2. Audition

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. Unlike the eyes, that can be shut and block incoming light, our ears are constantly exposed to sound (see Section 2.2). A sound can differ in many properties, such as location, loudness, rhythm, complexity, duration, etc. (see Section 2.1). It is an important modality which helps us to learn about an environment and to identify surrounding objects and their features [Moo82, Yos00, Alt04, BS06].

9.2.1. 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 head-related transfer functions (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

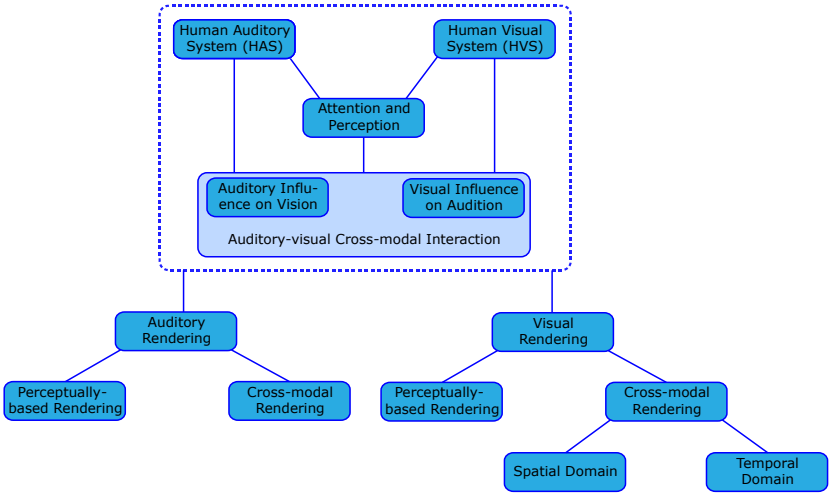

Figure 3: *Diagram showing the structure of the Part II.*

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° [PS90]. This makes hearing substantially weaker than vision in spatially related tasks. However, the temporal resolution of the HAS is rather high comparing to the visual, and according to Fujisaki et al. it is 89.3Hz [FN05].

10. 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. [MGKP08]. 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" [Jam92]. Research in psychology has considered the perception of individual senses separately [Bro58, BS06, Pyl06, Sch01], and across different modalities [DS98, Duf99, BdG04]. 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 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 the value and need. This was described in 1947 by Jerome Bruner and initiated a movement later named "new look in perception" [Pyl06]. 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

[Jam90, Jam92]. 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 [The91].

The 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 a salient objects or their salient features, or by a sudden motion [Yar67, IK01, Sch01]. 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 objects starts moving, our attention will 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 [KU85]. These maps combined form a saliency map.

10.1. 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. This phenomenon demonstrates the inability to detect salient objects in the centre of our gaze, when performing a task irrelevant to the distracting object [RLGM92, MR98]. 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 [SC99].

Similar limitation 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 [Moo82]. This is the ability to pick out and listen to a single sound in a noisy environment. Another HAS limitation is the continuity illusion [WWP88, KT02b]. 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 [Pas99]. 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 [DS94, SJD00, SJ01], and the second that resources are individual, intra-modal, where each modality has its own attentional pool [AAR72, DMW97, BH98, AMB06, BA06]. 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 [MW77]. 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 [DS98]. The paper shows that shifts of covert attention in one modality induce the attentional shift in other modalities. Similar results can be found in [SM04].

10.1.1. 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 the mobile phone can disrupt the car driving performance, due to the attention diversion [SJ01]. Additionally, there is a difficulty in attending to different locations in the two modalities [DS94]. 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 [SJD00].

10.1.2. Intra-modal

On the other hand, Alais et al., in a study dealing with attentional resources for vision and audition [AMB06], 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 [AAR72, BH98, DMW97, BA06].

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 [LMBB03]. 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.

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 [GM59, Shi64]. 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 [HT66, WW80]. 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 [GGB05, Rec03, BA06] and that these cross-modal effects must be taken into consideration when the perception of distinct sensory modalities is investigated [SS01, SKS04].

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 [GM59, Shi64, WKN03, Rec03]. Initially, the experimental results did not show the reverse effect [GM59, Shi64]. 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 [WKN03], 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 [SKS00, SKS02]. Nevertheless, when a single beep is accompanied by multiple flashes, only one beep is perceived [SKS02].

An analogue phenomenon to the visual ventriloquism effect (see Section 11.2) is the temporal ventriloquism [MZSFK03, BA03, AB03, BBM09]. 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 [MZSFK03]. Aschersleben and Bertelson showed that the temporal ventriloquism works in the opposite direction, but to a much lesser extent [AB03].

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 [HT66, CWGJ75, VBG98, VdG04]. The effect was named by Howard Templeton after the illusion created by ventriloquists when producing the words without moving their lips [HT66]. 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 [VdG04].

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* [MM76]. 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 complement 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 stimulation. A study by Stein et al. demonstrated that a simultaneous auditory stimulus can increase the perceived visual intensity [SLWP96]. 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. showed that in a visual search task, a single synchronised auditory *pip*, regardless of its position, significantly decreases the search time [VdBOBT08]. Another study demonstrated that a single auditory click can change the meaning of the visual information [SSL97]. 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 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 [BA06]. The optimal estimate can be calculated following

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 $\hat{S}_A$ and $\hat{S}_V$ 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 [AB04]. 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 [CD09].

12. Perception and Cross-modal Interaction in Computer Graphics

In previous sections findings on related work in psychology has 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.

12.1.1. Perceptually-based Auditory Rendering

Perceptually-based approach has been used for auditory rendering enhancement. It utilises limitations described in Sections 9.2.1 and 10.1. Since our nervous system in 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 red fruit in a green bush or an emergency siren. The proposed auditory saliency map, based on the visual saliency model discussed below, consists of three features: intensity, frequency contrast and temporal contrast, combined into a single map [KPLL05]. 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. [MBT*07] in acoustic rendering, by integrating saliency values over frequency subbands. Although the approach showed certain limitations, Moeck et al. suggest using audio saliency for clustering stage.

In another study, Tsingos et al. proposed a perceptual rendering pipeline, in which spatial rendering of a complex auditory environment with hundreds of dynamic auditory sources can be significantly simplified using interactive sound masking and spatial LOD, without any perceivable difference [TGD04]. The pipeline consists of four stages: culling of the perceptually inaudible (masked) audio sources; clustering the remaining sources; generating equivalent signals for each cluster; sending pre-mixed audio signals and source positions for rendering. This approach allows for rendering hundreds of dynamic audio sources on a standard hardware, without significant perceptual difference in audio quality. For a complete overview on perceptually-based auralisation see [Tsi07].

12.1.2. Cross-modal Interactions in Auditory Rendering

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]. One of the first studies, conducted by Moeck et al. investigated sound source clustering [MBT*07]. 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 [GBW*09] based on [BSVDD10]. 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

Similar to cross-modal auditory rendering, presented in Section 12.1.2, the findings described in Sections 9.1.1 and 10.1 can be exploited and utilised for visual rendering in computer graphics, in order to speed up the rendering process.

12.2.1. Perceptually-based Visual Rendering

Perceptually-based rendering in computer graphics has focused on taking advantage of exogenous visual attention via saliency maps [YPG01], originally introduced by Itti and Koch [IKN98], and endogenous visual attention [CCW03].

Saliency maps [KU85, YPG01] are based on the exogenous visual attention. They were first introduced by Koch and Ulman [KU85]. A mathematical model, based on feature maps was later developed [IKN98]. Those feature maps, based on colour, intensity and orientations are then combined into single topographical saliency map. The model

was first used by Yee et al. [YPG01] and later by Chalmers et al. [CDS06] and Longhurst et al. [LDC06]. For adapting the concept of saliency for dynamic content, Yee et al. developed a spatiotemporal error tolerance map, named Aleph map [YPG01]. The map is generated for each frame of the animation, increasing the animation rendering speed in return. It uses the saliency maps with motion features, and spatiotemporal frequency in order to calculate the tolerable error threshold for the observed region. As opposed to saliency maps, task maps [CCW03] use endogenous visual attention model. Using this method, task related objects in the virtual scene are used for the task map creation. The map is used in rendering process so that only task related parts of the scene are rendered in high quality and the remainder in low quality, without perceptual degradation in visual quality. Sundstedt et al. developed a map, that combines those two approaches, using both exogenous and endogenous attention, called the importance map [SDL*05]. For a complete overview on perceptually-based rendering see [HL97, OHM*04, BCFW08].

12.2.2. Cross-modal Interactions in 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 [Sto98]. The author showed that high-quality audio further increases the perceptual quality of the high-quality video. Furthermore, high-quality video further decreases 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 infleunce 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 [MDCT05a, Mas06]. The authors tried to attract users' attention towards the sound emitting object using abrupt sounds. 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 might be used in conjunction with the Aleph map, described above [YPG01].

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 [HAC08]. 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 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. Such work is presented below.

Mastoropoulou et al. investigated how music can affect temporal visual perception [MC04, Mas06], 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 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 to small; animation clips lasted for 40 seconds, which is far 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 [MDCT05b]. 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 [HCD*09]. 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 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.

In subsequent studies, Hulusic et al. investigated the influence of the movement related sound effects on temporal visual perception [HDAC10a, HDAC10b]. The results indicate that introducing the sound effect of footsteps to walk-

Phenomenon	Used in
angular sensitivity [Jam90, HB89]	[YPG01, CCW03, MDCT05a, Mas06, CDS06, LDC06]
inattentional blindness [RLGM92, MR98, SC99]	[CCW03, MDCT05a, Mas06]
modality appropriateness hypothesis [HT66, WW80]	[MC04, Mas06, HCD*09, HDAC10a, HDAC10b]
auditory driving effect [GM59, Shi64, WKN03, Rec03]	[MC04, Mas06, HCD*09, HDAC10a, HDAC10b]
temporal ventriloquism [MZSFK03, BA03, AB03, BBM09]	[HCD*09, HDAC10a, HDAC10b]
illusory flash induced by sound [SKS00, SKS02]	[HCD*09, HDAC10a, HDAC10b]
stimuli weighting [BA06]	[CD09]
ventriloquism effect [HT66, CWGJ75, VBG98, VdG04]	[MBT*07]

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

ing 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

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 the algorithm development, it 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 10.1 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 [LMBB03], the McGurk effect [MM76], the audio effect on visual intensity [SLWP96] / colour perception, the effect of audio on visual search [VdBOBT08] and bouncing targets / circles [SSL97].

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 are 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 of the field.

14. Conclusions

Sound remains a fundamental component if virtual environments are to deliver high-fidelity experiences. This state-of-the-art-report 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" [WHT06]. As more co-ordinated multi-disciplinary efforts are made to provide physically accurate audio and visuals in virtual environments in real-time, this STAR should provide a valuable resource from which this future research can build.

References

[AA93] ANDERSON J., ANDERSON B.: The Myth of Persistence of Vision Revisited. *Journal of Film and Video 45*, 1 (1993), 3–12. 18

[AAR72] ALLPORT D. A., ANTONIS B., REYNOLDS P.: On the division of attention: a disproof of the single channel hypothesis. *Q J Exp Psychol 24*, 2 (May 1972), 225–235. 20

[AB79] ALLEN J., BERKLEY D.: Image Method For Efficiently Simulating Small Room Acoustics. *The Journal of the Acoustical Society of America 65*, 4 (1979), 943–950. 6

[AB03] ASCHERSLEBEN G., BERTELSON P.: Temporal ventriloquism: crossmodal interaction on the time dimension: 2. evidence from sensorimotor synchronization. *International Journal of Psychophysiology 50*, 1-2 (2003), 157 – 163. Current findings in multisensory research. 21, 24

[AB04] ALAIS D., BURR D.: The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology 14*, 3 (February 2004), 257–262. 22

[ABAW03] ADELSTEIN B. D., BEGAULT D. R., ANDERSON M. R., WENZEL E.: Sensitivity to haptic-audio asynchrony. *5th International Conference on Multimodal Interfaces, ACM, Vancouver, Canada* (2003), 73–76. 17

[AE03] ATHINEOS M., ELLIS D. P.: Sound texture modelling with linear prediction in both time and frequency domains. 11

[Ahn93] AHNERT W.: Ears auralization software. In *J. Audio Eng. Soc.* (november 1993), vol. 11, pp. 894–904. 6

[Ahr10] AHRENS J.: *The Single-layer Potential Approach Applied to Sound Field Synthesis Including Cases of Non-enclosing Distributions of Secondary Sources*. PhD thesis, Technische Universitat Berlin, 2010. 17

[ALE03] ADELSTEIN B. D., LEE T. G., ELLIS S. R.: Head tracking latency in virtual environments: Psychophysics and a model. *Proceedings of theHuman Factors and Ergonomics Society 47th Annual Meeting* (2003). 17

[Alt04] ALTEN S. R.: *Audio in Media*, 7th ed. Wadsworth Publishing, 2004. 5, 18

[AMB06] ALAIS D., MORRONE C., BURR D.: Separate attentional resources for vision and audition. *Proc Biol Sci 273*, 1592 (Jun 2006), 1339–1345. 20

[App68] APPEL A.: Some techniques for shading machine renderings of solids. In *AFIPS '68 (Spring): Proceedings of the April 30–May 2, 1968, spring joint computer conference* (1968), ACM, pp. 37–45. 9

[AV02] AJDLER T., VETTERLI M.: The plenacoustic function and its sampling. *Proc. of the 1st Benelux Workshop on Model-based processing and coding of audio (MPCA2002), Leuven, Belgium* (Nov. 2002). 15

[Ave03] AVENDANO C.: 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* (Oct. 2003). 16

[AW02] ABHAYAPALA T., WARD D.: Theory and design of high order sound field microphones using spherical microphone array. 15

[AWBW05] ALLMAN-WARD M., BALAAM M., WILLIAMS R.: Source decomposition for vehicle sound simulation. *available from www.mts.com/nvd/pdf/source_decomp4veh_soundsim.pdf* (2005). 12

[BA03] BERTELSON P., ASCHERSLEBEN G.: Temporal ventriloquism: crossmodal interaction on the time dimension: 1. evidence from auditory-visual temporal order judgment. *International Journal of Psychophysiology 50*, 1-2 (2003), 147 – 155. Current findings in multisensory research. 21, 24

[BA06] BURR D., ALAIS D.: Combining visual and auditory information. *Prog Brain Res 155* (2006), 243–258. 20, 21, 24

[BBM*01] BUEHLER C., BOSSE M., MCMILLAN L., GORTLER S., COHEN M.: Unstructured lumigraph rendering. *Proc. of ACM SIGGRAPH* (2001). 15

[BBM09] BURR D., BANKS M., MORRONE M.: Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research 198* (2009), 49–57. 10.1007/s00221-009-1933-z. 21, 24

[BCFW08] BARTZ D., CUNNINGHAM D., FISCHER J., WALLRAVEN C.: The role of perception for computer graphics. In *Eurographics State-of-the-Art-Reports* (0 2008), pp. 65–86. 23

[BdG04] BERTELSON P., DE GELDER B.: *Crossmodal Space and Crossmodal Attention*. Oxford University Press, USA, May 2004, ch. The Psychology of Multimodal Perception. 19

[BDM*05] BERTRAM M., DEINES E., MOHRING J., JEGOROVS J., HAGEN H.: Phonon tracing for auralization and visualization of sound. *In Proceedings of IEEE Visualization* (2005), 151–158. 9

[BDT*08] BONNEEL N., DRETTAKIS G., TSINGOS N., VIAUD-DELMON I., JAMES D.: Fast modal sounds with scalable frequency-domain synthesis. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings) 27*, 3 (August 2008). 11

[BdVV93] BERKHOUT A., DE VRIES D., VOGEL P.: Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America 93*, 5 (May 1993), 2764–2778. 16, 17

[Beg94] BEGAULT D.: 3-D Sound for Virtual Reality and Multimedia. *Academic Press Professional* (1994). 6, 13, 15

[BF03] BAUMGARTE F., FALLER C.: Binaural cue coding - part i: Psychoacoustic fundamentals and design principles. *IEEE Trans. on Speech and Audio Proc 11*, 6 (2003). 14, 16

[BH98] BONNEL A. M., HAFTER E. R.: Divided attention between simultaneous auditory and visual signals. *Percept Psychophys 60*, 2 (Feb 1998), 179–190. 20

[BJLW*99] BAR-JOSEPH Z., LISCHINSKI D., WERMAN M., EL-YANNIV R., DUBNOV S.: Granular synthesis of sound textures using statistical learning. 11

[Bla97] BLAUERT J.: *Spatial Hearing : The Psychophysics of Human Sound Localization*. M.I.T. Press, Cambridge, MA, 1997. 3, 14

[Bor84] BORISH J.: Extension of the image model to arbitrary polyhedra. *The Journal of the Acoustical Society of America 75*, 6 (1984). 6

[Bre90] BREGMAN A.: *Auditory Scene Analysis, The perceptual organisation of sound*. The MIT Press, 1990. 12, 14, 15

[Bre93] BREGMAN A. S.: *Thinking in sound: the cognitive psychology of human audition*. Oxford University Press, 1993, ch. Auditory scene analysis: hearing in complex environments, pp. 10–36. 5

[Bro58] BROADBENT D. E.: *Perception and communication*. Oxford : Oxford University Press, 1958. 19

[BS06] BLAKE R., SEKULER R.: *Perception*, 5th ed. McGraw-Hill Higher Education, 2006. 5, 18, 19

[BSK05] BRUNGART D. S., SIMPSON B. D., KORDIK A. J.: Localization in the presence of multiple simultaneous sounds. *Acta Acustica united with Acustica 91* (May/June 2005), 471–479(9). 12, 13, 14

[BSVDD10] BONNEEL N., SUIED C., VIAUD-DELMON I., DRETTAKIS G.: Bimodal perception of audio-visual material properties for virtual environments. *ACM Transactions on Applied Perception 7*, 1 (Jan. 2010), 1–16. 4, 22

[BvSJC05] BEST V., VAN SCHAIK A., JIN C., CARLISLE S.: Auditory spatial perception with sources overlapping in frequency and time. *Acta Acustica united with Acustica 91* (May/June 2005), 421–428(8). 12, 13, 14

[BW99] BORN M., WOLF E.: *Principles of Optics*. 7th Edition, Pergamon Press, 1999. 6

[Cat] CATT-Acoustic, Gothenburg, Sweden. http://www.netg.se/catt. 3

[CCL02] CATER K., CHALMERS A., LEDDA P.: Selective quality rendering by exploiting human inattentional blindness: Looking but not seeing. In *Symposium on Virtual Reality Software and Technology 2002* (November 2002), ACM, pp. 17–24. 4

[CCW03] CATER K., CHALMERS A., WARD G.: Detail to attention: exploiting visual tasks for selective rendering. In *EGRW '03: Proceedings of the 14th Eurographics Workshop on Rendering Techniques* (Leuven, Belgium, 2003), Eurographics Association, pp. 270–280. 22, 23, 24

[CD09] CHALMERS A., DEBATTISTA K.: Level of realism for serious games. *Games and Virtual Worlds for Serious Applications, Conference in 0* (2009), 225–232. 22, 24

[CDG*93] CALVIN J., DICKENS A., GAINES B., METZGER P., MILLER D., OWEN D.: The Simnet Virtual World Architecture. *In Proceedings of the IEEE Virtual Reality Annual International Symposium* (Sep 1993), 450–455. 3

[CDS06] CHALMERS A., DEBATTISTA K., SANTOS L. P.: 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* (New York, NY, USA, 2006), ACM Press, pp. 9–18. 23, 24

[CH05] CAMPOS G., HOWARD D.: On the computational efficiency of different waveguide mesh topologies for room acoustic simulation. In *IEEE Transactions on Speech and Audio Processing* (2005), vol. 5, pp. 1063–1072. 8

[CHD01] CAMPOS G., HOWARD D., DOBSON S.: Acoustic reconstruction of ancients structures using three-dimensional digital waveguide mesh models. In *Computer Applications in Archaeology (CAA2001)* (2001), G. Burenhult B. I. S. . ., (Ed.), pp. 173–176. 8

[CM78] CREMER L., MÜLLER H.: Principles and Applications of Room Acoustics. *Applied Science 1* (1978). 6

[CMRT10] CHABANNE C., MCCALLUS M., ROBINSON C., TSINGOS N.: Surround sound with height in games using dolby pro logic iiz. In *Audio Engineering Society Convention 129* (11 2010). 16

[Com94] COMON P.: Independent component analysis: A new concept. *Signal Processing 36* (1994), 287–314. 16

[CRO] CROSSMOD project, cross-modal preceptual interaction and rendering. http://www-sop.inria.fr/reves/CrossmodPublic/index.php. 22

[CW93] CHEN S., WILLIAMS L.: View interpolation for image synthesis. vol. 27, pp. 279–288. 15

[CWGJ75] CHOE C. S., WELCH R. B., GILFORD R. M., JUOLA J. F.: The "ventriloquist effect": Visual dominance or response bias? *Perception and Psychophysics 18*, 1 (1975), 55–60. 21, 24

[DBJEY*02] DUBNOV S., BAR-JOSEPH Z., EL-YANNIV R., LISCHINSKI D., WERMAN M.: Synthesis of sound textures by learning and resampling of wavelet trees. 11

[DCH] DESAINTE-CATHERINE M., HANNA P.: Statistical approach for sound modeling. 11

[DDS02] DARLINGTON D., DAUDET L., SANDLER M.: Digital audio effects in the wavelet domain. In *Proceedings of COST-G6 Conference on Digital Audio Effects, DAFX2002, Hamburg, Germany* (Sept. 2002). 15

[DFMM99] DICKINS G., FLAX M., MCKEAG A., MCGRATH D.: Optimal 3D-speaker panning. *Proceedings of the AES 16th international conference, Spatial sound reproduction, Rovaniemi, Finland* (april 1999), 421–426. 16

[DKP01] DOEL K. V. D., KRY P., PAI D.: 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,* (2001), pp. 537–544. 11

[DM95] DURLACH N., MAVOR A.: *Virtual Reality Scientific and Technological Challenges*. Tech. rep., National Research Council Report, National Academy Press, 1995. 3

[DMW97] DUNCAN J., MARTENS S., WARD R.: Restricted attentional capacity within but not between sensory modalities. *Nature 387*, 6635 (Jun 1997), 808–810. 20

[Do04] DO M.: Toward sound-based synthesis: the far-field case. *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada* (May 2004). 15

[Doe04] DOEL K. V. D.: Physically-based models for liquid sounds. 12

[Doe05] DOEL K. V. D.: Physically based models for liquid sounds. In *ACM Trans. Appl. Percept.* (2005), vol. 2, pp. 534–546. 12

[DP98] DOEL K. V. D., PAI D.: The sound of physical shapes. *Presence 7*, 4 (1998), 382–395. 11

[DRP98] DANIEL J., RAULT J.-B., POLACK J.-D.: Ambisonic encoding of other audio formats for multiple listening conditions. *105th AES convention, preprint 4795* (Aug. 1998). 15

[DS94] DRIVER J., SPENCE C.: *Attention and Performance XV*. MIT Press, 1994, ch. Spatial Sy, pp. 311–331. 20

[DS98] DRIVER J., SPENCE C.: Crossmodal attention. *Curr Opin Neurobiol 8*, 2 (Apr 1998), 245–253. 19, 20

[DS03] DI-SCIPIO A.: Synthesis of environmental sound textures by iterated non linear functions. 11

[Duf99] DUFOUR A.: Importance of attentional mechanisms in audiovisual links. *Exp Brain Res 126*, 2 (May 1999), 215–222. 19

[DYN03] DOBASHI Y., YAMAMOTO T., NISHITA T.: Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics. *ACM Transactions on Graphics 22*, 3 (Aug. 2003), 732–740. (Proceedings of ACM SIGGRAPH 2003). 12

[DYN04] DOBASHI Y., YAMAMOTO T., NISHITA T.: Synthesizing sound from turbulent field using sound textures for interactive fluid simulation. *Computer Graphics Forum (Proc. EUROGRAPHICS 2004) 23*, 3 (2004), 539–546. 12

[EAX04] Eax. Ű environmental audio extensions 4.0, creative, 2004. 12

[ELD91] EDWORTHY J., LOXLEY S., DENNIS I.: Improving auditory warning design: Relationship between warning sound parameters and perceived urgency. In *Human Factors* (1991), vol. 33, pp. 205–231. 14

[FB03] FALLER C., BAUMGARTE F.: Binaural cue coding - part ii: Schemes and applications. *IEEE Trans. on Speech and Audio Proc 11*, 6 (2003). 14, 16

[FCE*98] FUNKHOUSER T., CARLBOM I., ELKO G., PINGALI G., SONDHI M., WEST J.: A beam tracing approach to acoustic modeling for interactive virtual environments. *ACM Computer Graphics, SIGGRAPH'98 Proceedings* (jul 1998), 21–32. 8

[FJT02] FUNKHOUSER T., JOT J., TSINGOS N.: Sounds good to me! computational sound for graphics, vr, and interactive systems. *SIGGRAPH 2002 Course Notes Number 45* (2002). 4

[FM05] FALLER C., MERIMAA J.: Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. 3075–3089. 14

[FMC99] FUNKHOUSER T., MIN P., CARLBOM I.: Real-time acoustic modeling for distributed virtual environments. *ACM Computer Graphics, SIGGRAPH'99 Proceedings* (aug 1999), 365–374. 8

[FN05] FUJISAKI W., NISHIDA S.: Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Exp Brain Res 166*, 3-4 (October 2005), 455–464. 18, 19

[FvDFH90] FOLEY J. D., VAN DAM A., FEINER S. K., HUGHES J.: *Computer graphics, principles and practice*. Addison Wesley, 1990. 14

[Gar97] GARDNER W.: Reverberation algorithms. In *Applications of Digital Signal Processing to Audio and Acoustics* (1997), Kahrs M., Brandenburg K., (Eds.), Kluwer Academic Publishers, pp. 85–131. 10

[GBW*09] GRELAUD D., BONNEEL N., WIMMER M., ASSELOT M., DRETTAKIS G.: 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* (New York, NY, USA, 2009), ACM, pp. 177–182. 3, 4, 22

[Ger85] GERZON M.: Ambisonics in multichannel broadcasting and video. *J. Audio Eng. Soc. 33*, 11 (1985), 859–871. 15, 17

[Get07] GETZMANN S.: The effect of brief auditory stimuli on visual apparent motion. *Perception 36*, 7 (2007), 1089–1103. 18

[GGB05] GUTTMAN S. E., GILROY L. A., BLAKE R.: Hearing what the eyes see: Auditory encoding of visual temporal sequences. *Psychological Science 16*, 3 (March 2005), 228–235. 21

[GLT05] GALLO E., LEMAITRE G., TSINGOS N.: Prioritising signals for selective real-time audio processing. In *Proceedings of Intl. Conf. on Auditory Display (ICAD) 2005, Limerick, Ireland* (July 2005). 13, 14

[GM59] GEBHARD J. W., MOWBRAY G. H.: On discriminating the rate of visual flicker and auditory flutter. *Am J Psychol 72* (Dec 1959), 521–529. 21, 24

[GT07] GALLO E., TSINGOS N.: Extracting and re-rendering structured auditory scenes from field recordings. *30th International Conference: Intelligent Audio Environments* (march 2007). 15

[GTL07] GALLO E., TSINGOS N., LEMAITRE G.: 3d-audio matting, postediting, and rerendering from field recordings. *EURASIP J. Appl. Signal Process.* (2007), 183–183. 15

[HAA97] HORRY Y., ANJYO K.-I., ARAI K.: 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* (New York, NY, USA, 1997), ACM Press/Addison-Wesley Publishing Co., pp. 225–232. 15

[HAC08] HULUSIC V., ARANHA M., CHALMERS A.: The influence of cross-modal interaction on perceived rendering quality thresholds. In *WSCG 2008 Full Papers Proceedings* (2008), Skala V., (Ed.), pp. 41–48. 4, 23

[Har83] HARTMANN W.: Localization of sound in rooms. *Journal of the Acoustical Society of America 74*, 5 (Nov. 1983), 1380–1391. 6

[Har97] HARTMANN W. M.: *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates, 1997. 6

[HB89] HUMPHREYS G. W., BRUCE V.: *Visual Cognition: Computational, Experimental and Neuropsychological Perspectives*. Lawrence Erlbaum Associates Ltd, East Sussex, BN3 2FA, UK, 1989. 18, 24

[HB96] HENDRIX C. M., BARFIELD W.: Presence within virtual environments as a function of visual display parameters. *Presence 5*, 3 (1996), 274–289. 12

[HC95] HAAS E. C., CASALI J. C.: Peceived urgency of and response time to multi-tone and frequency-modulated warning signals in broadband noise. *Ergonomics 38*, 11 (1995), 2313–2326. 14

[HCD*09] HULUSIC V., CZANNER G., DEBATTISTA K., SIKUDOVA E., DUBLA P., CHALMERS A.: Investigation of the beat rate effect on frame rate for animated content. In *Spring Conference on Computer Graphics 2009* (2009), Hauser H., (Ed.), Comenius University, Bratislava, pp. 167–174. 4, 23, 24

[HDAC10a] HULUSIC V., DEBATTISTA K., AGGARWAL V., CHALMERS A.: 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* (2010), IEEE Computer Society. 4, 23, 24

[HDAC10b] HULUSIC V., DEBATTISTA K., AGGARWAL V., CHALMERS A.: Maintaining frame rate perception in interactive environments byăexploiting audio-visual cross-modal interaction. *The Visual Computer* (2010), 1–10. 10.1007/s00371-010-0514-2. 4, 23, 24

[Her99] HERDER J.: Optimization of sound spatialization resource management through clustering. *The Journal of Three Dimensional Images, 3D-Forum Society 13*, 3 (Sept. 1999), 59–65. 14

[HH84] HECKBERT P. S., HANRAHAN P.: Beam tracing polygonal objects. *SIGGRAPH '84: Proceedings of the 11th annual conference on Computer graphics and interactive techniques* (1984), 119–127. 8

[HL97] HORVITZ E., LENGYEL J.: Perception, attention, and resources: A decision-theoretic approach to graphics rendering. In *1997, Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-97* (1997), pp. 238–249. 23

[Hob55] HOBSON E.: *The Theory of Spherical and Ellipsoidal Harmonics*. Chelsea Pub Co., 1955. 17

[HSP99] HERRERE P., SERRA X., PEETERS G.: Audio descriptors and descriptors schemes in the context of mpeg-7. *Procedings of International Computer Music Conference (ICMC99)* (1999). 14

[HT66] HOWARD I. P., TEMPLETON W. B.: *Human spatial orientation [by] I.P. Howard and W.B. Templeton*. Wiley, London, New York,, 1966. 21, 24

[HWBR*10] HARVEY C., WALKER S., BASHFORD-ROGERS T., DEBATTISTA K., CHALMERS A.: The Effect of Discretised and Fully Converged Spatialised Sound on Directional Attention and Distraction. Collomosse J., Grimstead I., (Eds.), Eurographics Association, pp. 191–198. 4

[Ihl98] IHLENBURG F.: *Finite Element Analysis of Acoustic Scattering*. SpringerVerlag, New York, 1998. 7

[IK01] ITTI L., KOCH C.: Computational modelling of visual attention. *Nat Rev Neurosci 2*, 3 (March 2001), 194–203. 20

[IKN98] ITTI L., KOCH C., NIEBUR E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell. 20*, 11 (1998), 1254–1259. 22

[Jam90] JAMES W.: *The principles of psychology*. Holt, New York, 1890. 18, 20, 24

[Jam92] JAMES W.: *Psychology*. McMilan and Co., 1892. 19, 20

[JB04] J.R.PARKER, BEHM B.: Generating audio textures by example. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSPŠ04)* (2004). 11

[JBP06] JAMES D. L., BARBIČ J., PAI D. K.: Precomputed acoustic transfer: Output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM Transactions on Graphics (SIGGRAPH 2006) 25*, 3 (Aug. 2006). 11

[Jeh05] JEHAN T.: *Creating Music by Listening*. PhD thesis, M.I.T., June 2005. 11

[Jen96] JENSEN H. W.: Global illumination using photon maps. *Proceedings of the eurographics workshop on Rendering techniques '96* (1996), 21–30. 9

[JLP99] JOT J.-M., LARCHER V., PERNAUX J.-M.: A comparative study of 3D audio encoding and rendering techniques. *Proceedings of the AES 16th international conference, Spatial sound reproduction, Rovaniemi, Finland* (april 1999). 15

[JLW95] JOT J., LARCHER V., WARUSFEL O.: Digital signal processing issues in the context of binaural and transaural stereophony. *Proc. 98th Audio Engineering Society Convention* (1995). 16

[JM06] JEDRZEJEWSKI M., MARASEK K.: Computation of room acoustics using programmable video hardware. In *Computer Vision and Graphics* (2006), vol. 32 of *Computational Imaging and Vision*, Springer Netherlands, pp. 587–592. 9

[Jot99] JOT J.-M.: Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. *Multimedia Systems 7*, 1 (1999), 55–69. 10

[JRY00] JOURJINE A., RICKARD S., YILMAZ O.: Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSPŠ00)* (2000). 15, 16

[KAG*02] KURNIAWATI E., ABSAR J., GEORGE S., LAU C. T., PREMKUMAR B.: 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* (2002). 14

[Kai] KAISER P.: The joy of visual perception (web book). 18

[Kaj86] KAJIYA J. T.: The Rendering Equation. *SIGGRAPH Comput. Graph. 20*, 4 (1986), 143–150. 6

[KdP05] KOHLRAUSCH A., DER PAR S. V.: Audio-visual interaction in the context of multimedia applications. In *Communication Acoustics* (2005), Springer Verlag, pp. 109–138. 17

[KJM04] KAPRALOS B., JENKIN M., MILIOS E.: Acoustic modeling utilizing an acoustic version of phonon mapping. *In Proc. of IEEE Workshop on HAVE* (2004). 8, 9

[KK07] KOZLOWSKI O., KAUTZ J.: Is accurate occlusion of glossy reflections necessary? In *APGV '07: Proceedings of the 4th symposium on Applied perception in graphics and visualization* (New York, NY, USA, 2007), ACM, pp. 91–98. 4

[KL95] KOPUZ S., LALOR N.: Analysis of interior acoustic fields using the finite element method and the boundary element method. *Applied Acoustics 45* (1995), 193–210. 7

[Klu91] KLUDSZUWEIT A.: 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* (1991), 17–27. 7

[KPLL05] KAYSER C., PETKOV C. I., LIPPERT M., LOGOTHETIS N. K.: Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology 15*, 21 (November 2005), 1943–1947. 22

[KSS68] KROKSTAD A., STRØM S., SØRSDAL S.: Calculating the acoustical room response by the use of a ray tracing technique. *J. Sound Vib. 8* (July 1968), 118–125. 9

[KT02a] KELLY M., TEW. A.: The continuity illusion in virtual auditory space. *Proc. of the 112th AES Conv., Munich, Germany* (may 2002). 13

[KT02b] KELLY M. C., TEW A. I.: The continuity illusion in virtual auditory space. In *in proc. of AES 112th Convention* (Munich, Germany, May 2002). 20

[KU85] KOCH, C., ULLMAN, S.: Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology 4*, 4 (1985), 219–227. 20, 22

[Kut91] KUTTRUFF H.: *Room Acoustics (3rd edition)*. Elseiver Applied Science, 1991. 6, 15

[KvdP05] KOHLRAUSCH A., VAN DE PAR S.: *Communication Acoustics*. Springer, 2005, ch. AudioŰVisual Interaction in the Context of Multi-Media Applications, pp. 109–138. 4

[LBM03] LABORIE A., BRUNO R., MONTOYA S.: A new comprehensive approach of surround sound recording. *114th convention of the Audio Engineering Society, preprint 5717* (2003). 15

[LBM04] LABORIE A., BRUNO R., MONTOYA S.: High spatial resolution multichannel recording. *Proc. 116th convention of the Audio Engineering Society, preprint 6116* (2004). 15

[LCM07] LAUTERBACH C., CHANDAK A., MANOCHA D.: Interactive sound propagation in dynamic scenes using frustum tracing. *IEEE Trans. on Visualization and Computer Graphics 13* (2007), 1672–1679. 9

[LDC06] LONGHURST P., DEBATTISTA K., CHALMERS A.: 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* (New York, NY, USA, 2006), AFRIGRAPH '06, ACM, pp. 21–29. 23, 24

[Lee98] LEESE M. J.: Ambisonic surround sound FAQ (version 2.8), 1998. http://members.tripod.com/martin_leese/Ambisonic/. 15, 17

[Leh93] LEHNERT H.: Systematic errors of the ray-tracing algorithm. *J. Applied Acoustics 38* (1993), 207–221. 8

[Lew02] LEWICKI M.: Efficient coding of natural sounds. *Nature Neuroscience 5*, 4 (2002), 356–363. 16

[LHS01] LOKKI T., HIIPALLA T., SAVIOJA L.: A framework for evaluating virtual acoustic environments. *AES 110th convention, Berlin, preprint 5317* (2001). 13

[LLBM04] LARCHER V., LABORIE A., BRUNO R., MONTOYA S.: Techniques de spatialisation des sons. In *Informatique Musicale - du signal au signe musical* (2004), Pachet F., Briot J.-P., (Eds.), Hermès science. 17

[LM00] LOSCHKY L. C., MCCONKIE G. W.: User performance with gaze contingent multiresolutional displays. In *ETRA '00: Proceedings of the 2000 symposium on Eye tracking research & applications* (New York, NY, USA, 2000), ACM, pp. 97–103. 18

[LM01] LAGRANGE M., MARCHAND S.: 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* (2001). 14

[LMBB03] LARSEN A., MCILHAGGA W., BAERT J., BUNDESEN C.: Seeing or hearing? perceptual independence, modality confusions, and crossmodal congruity effects with focused and divided attention. *Percept Psychophys 65*, 4 (May 2003), 568–574. 20, 24

[LMCJ00] LEVITIN D. J., MACLEAN K., CHU L. Y., JENSEN E. R.: The perception of cross-modal simultaneity. *International Journal of Computing and Anticipatory Systems* (2000), 323–329. 17

[Log00] LOGAN B.: Mel frequency cepstral coefficients for music modelling. *Proceedings of the International Symposium on Music Information Retrieval (Music IR 2000)* (October 2000). 14

[LSL*09] LAINE S., SILTANEN S., LOKKI T., , SAVIOJA L.: Accelerated beam tracing algorithm. *Applied Acoustics 70* (2009), 172–181. 8

[LVK02] LARSSON P., VÄSTFJÄLL D., KLEINER M.: 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* (June 2002), 31–38. 12

[LWZ04] LU L., WENYIN L., ZHANG H.-J.: Audio textures: Theory and applications. *IEEE Transactions on Speech and Audio Processing 12*, 2 (2004), 156–167. 11

[MAB*03] M.RATH, AVANZINI F., BERNARDINI N., BORIN G., FONTANA F., OTTAVIANI L., ROCHESSO D.: An introductory catalog of computer-synthesized contact sounds, in real-time. *Proc. of the XIV Colloquium on Musical Informatics, Firenze, Italy* (July 2003). 11

[Mal01] MALHAM D.: 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* (June 2001). 12

[Mas06] MASTOROPOULOU G.: *The Effect of Audio on the Visual Perception of High-Fidelity Animated 3D Computer Graphics*. PhD in Computer science, University of Bristol, 2006. 23, 24

[MAWM03] MILLER J. D., ANDERSON M. R., WENZEL E. M., MACLEAN B. U.: Latency measurement of a real-time virtual acoustic environment rendering system. *In Proceedings of the International Conference on Auditory Display (ICAD 2003), Boston, MA* (2003). 17

[MBT*07] MOECK T., BONNEEL N., TSINGOS N., DRETTAKIS G., VIAUD-DELMON I., ALLOZA D.: Progressive perceptual audio rendering of complex scenes. In *I3D '07: Proceedings of the 2007 symposium on Interactive 3D graphics and games* (New York, NY, USA, 2007), ACM, pp. 189–196. 3, 4, 22, 24

[MC04] MASTOROPOULOU G., CHALMERS A.: 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)* (Washington, DC, USA, 2004), IEEE Computer Society, pp. 128–134. 23, 24

[MDCT05a] MASTOROPOULOU G., DEBATTISTA K., CHALMERS A., TROSCIANKO T.: 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* (New York, NY, USA, 2005), ACM Press, pp. 363–369. 4, 23, 24

[MDCT05b] MASTOROPOULOU G., DEBATTISTA K., CHALMERS A., TROSCIANKO T.: 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* (New York, NY, USA, 2005), ACM Press, pp. 9–15. 4, 23

[ME04a] MEYER J., ELKO G.: 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* (2004). 12

[ME04b] MEYER J., ELKO G.: 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* (2004). 15

[Men02] MENZIES D.: W-Panning and O-format, tools for object spatialization. 12

[Mer02] MERIMAA J.: Applications of a 3D microphone array. *112th AES convention, preprint 5501* (May 2002). 15, 16

[MGKP08] MARAGOS P., GROS P., KATSAMANIS A., PAPANDREOU G.: *Cross-Modal Integration for Performance Improving in Multimedia: A Review*. Springer-Verlag, 2008. 19

[MHM06] MULLEN J., HOWARD D., MURPHY D.: Waveguide physical modeling of vocal tract acoustics: Improved formant bandwidth control from increased model dimensionality. In *IEEE Transactions on Speech and Audio Processing* (2006), vol. 3, pp. 964–971. 8

[MKMS07] MURPHY D., KELLONIEMI A., MULLEN J., SHELLEY S.: Acoustic modeling using the digital waveguide mesh. *Signal Processing Magazine, IEEE 24*, 2 (2007), 55–66. 8

[MLC*09] MANOCHA D., LIN M., CALAMIA P., SAVIOJA L., TSINGOS N.: Interactive sound rendering. *SIGGRAPH2009 Course Notes* (Aug 2009). 4

[MM76] MCGURK H., MACDONALD J.: Hearing lips and seeing voices. *Nature 264*, 5588 (December 1976), 746–748. 21, 24

[MM95] MALHAM D., MYATT A.: 3d sound spatialization using ambisonic techniques. *Computer Music Journal 19*, 4 (1995), 58–70. 14, 15, 17

[MMC09] MACKNIK S., MARTINEZ-CONDE S.: *Encyclopedia of Perception*. SAGE Press, 2009, ch. Vision: te, pp. 1060–1062. 18

[Møl89] MØLLER H.: Reproduction of artificial-head recordings through loudspeakers. *J. Audio Eng. Soc. 37*, 1/2 (jan/feb 1989), 30–33. 16

[Møl92] MØLLER H.: Fundamentals of binaural technology. *Applied Acoustics 36* (1992), 171–218. 16

[Moo82] MOORE B. C.: *An Introduction to the Psychology of Hearing*, 2nd ed. Academic Press, 1982. 5, 18, 20

[Moo97] MOORE B. C.: *An introduction to the psychology of hearing*. Academic Press, 4th edition, 1997. 13, 14

[MP04] MERIMAA J., PULKKI V.: Spatial impulse response rendering. *Proc. of the 7th Intl. Conf. on Digital Audio Effects (DAFX'04), Naples, Italy* (Oct 2004). 10, 16

[MP05] MERIMAA J., PULKKI V.: Spatial impulse response rendering I: Analysis and synthesis. *J. Audio Eng. Soc. 53* (Dec 2005), 1115–1127. 15

[MR98] MACK A., ROCK I.: *Inattentional Blindness*. The MIT Press, 1998. 20, 24

[MW77] MASSARO D. W., WARNER D. S.: Dividing attention between auditory and visual perception. *Perception & Psychophysics 21*, 6 (1977), 569–574. 20

[MYH*10] MOSS W., YEH H., HONG J.-M., LIN M. C., MANOCHA D.: Sounding liquids: Automatic sound synthesis from fluid simulation. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)* (2010). 12

[MZP*95] MACEDONIA M. R., ZYDA M. J., PRATT D. R., BRUTZMAN D. P., BARHAM P. T.: Exploiting Reality with Multicast Groups. *IEEE Computer Graphics and Applications 15* (Sep 1995), 38–45. 3

[MZSFK03] MOREIN-ZAMIR S., SOTO-FARACO S., KINGSTONE A.: Auditory capture of vision: examining temporal ventriloquism. *Brain Res Cogn Brain Res 17*, 1 (Jun 2003), 154–163. 21, 24

[Nay93] NAYLOR J.: ODEON - another Hybrid Room Acoustical Model. *Applied Acoustics 38*, 1 (1993), 131–143. 3

[Nie93] NIELSEN S. H.: Auditory Distance Perception in Different Rooms. *J. Audio Eng. Soc. 41*, 10 (Oct 1993), 755–770. 6

[OCE01a] O'BRIEN J. F., COOK P. R., ESSL G.: Synthesizing sounds from physically based motion. In *In Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (2001), ACM, pp. 529–536. 11

[OCE01b] O'BRIEN J. F., COOK P. R., ESSL G.: Synthesizing sounds from physically based motion. *ACM Computer Graphics, SIGGRAPH'01 Proceedings* (Aug. 2001), 545–552. 11

[OHM*04] O'SULLIVAN C., HOWLETT S., MCDONNELL R., MORVAN Y., O'CONOR K.: Perceptually adaptive graphics. In *Eurographics State-of-the-Art-Reports* (2004). 23

[OPR05] O'GRADY P., PEARLMUTTER B., RICKARD S.: Survey of sparse and non-sparse 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* (2005). 16

[OSG02a] O'BRIEN J., SHEN C., GATCHALIAN C.: Synthesizing sounds from rigid-body simulations. *Proc. of the ACM SIGGRAPH Symposium on Computer Animation, San Antonio, Texas* (July 2002), 175–182. 11

[OSG02b] O'BRIEN J. F., SHEN C., GATCHALIAN C.: Synthesizing sounds from rigid-body simulations. In *In Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation* (2002), ACM, pp. 175–181. 11

[Pas99] PASHLER H.: *The psychology of attention*. The MIT Press, 1999. 20

[PB04] POTARD G., BURNETT I.: 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* (Oct. 2004). 12

[PC03] PARKER J., CHAN S.: Sound synthesis for the web, games, and virtual reality. *International Conference on Computer Graphics and Interactive Techniques* (2003). 11

[Pee04] PEETERS G.: 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)* (2004). 14

[Pel01a] PELLEGRINI R.: Quality assessment of auditory virtual environments. 13

[Pel01b] PELLEGRINI R.: *A virtual Listening Room as an application of auditory virtual Environment*. PhD thesis, Ruhr-Universitat, Bochum, 2001. 10

[PS90] PERROTT D. R., SABERI K.: Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America 87*, 4 (1990), 1728–1731. 19

[PS00] PAINTER E. M., SPANIAS A. S.: Perceptual coding of digital audio. *Proceedings of the IEEE 88*, 4 (april 2000). 13, 14, 15

[Pul97] PULKKI V.: Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc. 45*, 6 (june 1997), 456–466. 16

[Pul06] PULKKI V.: Directional audio coding in spatial sound reproduction and stereo upmixing. In *28th International Conference: The Future of Audio Technology - Surround and Beyond* (june 2006). 15, 16

[Pyl06] PYLYSHYN Z. W.: *Seeing and Visualizing: It's not what you Think*. MIT Press, March 2006. 19

[Ran01] RANGACHAR R.: *Analysis and Improvement of the MPEG-1 Audio Layer III Algorithm at Low Bit-Rates*. Master of science thesis, Arizona State University, Dec. 2001. 14

[RBF03] ROCCHESSO D., BRESIN R., FRENSTRÖM M.: Sounding objects. *IEEE Multimedia 10*, 2 (Apr. 2003), 42–52. 12

[RBF08] RAMANARAYANAN G., BALA K., FERWERDA J. A.: Perception of complex aggregates. In *SIGGRAPH '08: ACM SIGGRAPH 2008 papers* (New York, NY, USA, 2008), ACM, pp. 1–10. 4

[Rec03] RECANZONE G. H.: Auditory influences on visual temporal rate perception. *Journal of neurophysiology 89* (Feb 2003), 1078–1093. 21, 24

[RFWB07] RAMANARAYANAN G., FERWERDA J., WALTER B., BALA K.: Visual equivalence: towards a new standard for image fidelity. *ACM Trans. Graph. 26*, 3 (2007), 76. 4

[Ric06] RICKARD S.: Sparse sources are separated sources. *Proceedings of the 16th Annual European Signal Processing Conference, Florence, Italy* (2006). 16

[RKM07] RÖBER N., KAMINSKI U., MASUCH M.: Ray acoustics using computer graphics technology. In *In Proc. 10th Intl. Conf. on Digital Audio Effects (DAFxŠ07), Bordeaux* (2007), pp. 274–279. 9

[RL06] RAGHUVANSHI N., LIN M. C.: Interactive sound synthesis for large scale environments. In *I3D '06: Proceedings of the 2006 symposium on Interactive 3D graphics and games* (2006), ACM, pp. 101–108. 11

[RLC*07] RAGHUVANSHI N., LAUTERBACH C., CHANDAK A., MANOCHA D., LIN M. C.: Real-time sound synthesis and propagation for games. *Commun. ACM 50*, 7 (2007), 66–73. 3

[RLGM92] ROCK I., LINNETT C. M., GRANT P., MACK A.: Perception without attention: results of a new method. *Cognit Psychol 24*, 4 (October 1992), 502–534. 20, 24

[RNFR96] RAJKUMAR A., NAYLOR B. F., FEISULLIN F., ROGERS L.: Predicting rf coverage in large environments using ray-beam tracing and partitioning tree represented geometry. *Wirel. Netw. 2*, 2 (1996), 143–154. 9

[Roc02] ROCCHESSO D.: Spatial effects. In *DAFX - Digital Audio Effects* (2002), Ed. U. Z., (Ed.), Wiley, p. Chapter 6. 10

[Rog25] ROGET P. M.: 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 (Jan. 1825), 131–140. 18

[Roo02] ROORDA A.: *Human Visual System - Image Formation*, vol. 1. 2002, pp. 539–557. 18

[RR02] RADKE R., RICKARD S.: Audio interpolation. In *the Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio (AES'02)* (2002), pp. 51–57. 15

[SAMM] SAWADA H., ARAKI S., MUKAI R., MAKINO S.: Blind extraction of dominant target sources using ica and time-frequency masking. *IEEE Trans. Audio, Speech, and Language Processing*. accepted for future publication. 16

[SAP98] SAINT-ARNAUD N., POPAT K.: Analysis and synthesis of sound textures. 11

[SC99] SIMONS D., CHABRIS C.: Gorillas in our midst: sustained inattentional blindness for dynamic events. *perception 28* (1999), 1059–1074. 20, 24

[Sch62] SCHROEDER M.: Natural sounding artificial reverberation. 219–223. 10

[Sch01] SCHOLL B. J.: Objects and attention: the state of the art. *Cognition 80*, 1-2 (June 2001), 1–46. 19, 20

[SD83] STAAL H. E., DONDERI D. C.: The effect of sound on visual apparent movement. *The American Journal of Psychology 96*, 1 (1983), 95–105. 18

[SDL*05] SUNDSTEDT V., DEBATTISTA K., LONGHURST P., CHALMERS A., TROSCIANKO T.: Visual attention for efficient high-fidelity graphics. In *SCCG '05: Proceedings of the 21st spring conference on Computer graphics* (New York, NY, USA, 2005), ACM Press, pp. 169–175. 23

[SE98] STREICHER R., EVEREST F. (Eds.): *The new stereo soundbook, 2nd edition*. Audio Engineering Associate, Pasadena (CA), USA, 1998. 16

[SHHT96] SAVIOJA L., HUOPANIEMI J., HUOTILAINEN T., TAKALA T.: Real-time virtual audio reality. In *In Proc. ICMC 1996* (august 1996), pp. 107–110. 6

[Shi64] SHIPLEY T.: Auditory flutter-driving of visual flicker. *Science 145* (Sep 1964), 1328–1330. 21, 24

[SHLV99] SAVIOJA L., HUOPANIEMI J., LOKKI T., VÄÄNÄNEN R.: Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society 47*, 9 (Sep 1999). 7, 15

[SJ01] STRAYER D. L., JOHNSTON W. A.: Driven to distraction: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone. *Psychol. Sci. 12*, 6 (2001), 462–466. 20

[SJD00] SPENCE C., JANE R., DRIVER J.: Cross-modal selective attention: on the difficulty of ignoring sounds at the locus of visual attention. *Perception & Psychophysics 62*, 2 (2000), 410–424. 20

[SKS00] SHAMS L., KAMITANI Y., SHIMOJO S.: What you see is what you hear. *Nature 408* (2000), 788+. 21, 24

[SKS02] SHAMS L., KAMITANI Y., SHIMOJO S.: Visual illusion induced by sound. *Brain Res Cogn Brain Res 14*, 1 (Jun 2002), 147–152. 21, 24

[SKS04] SHAMS L., KAMITANI Y., SHIMOJO S.: Modulations of visual perception by sound. in the handbook of multisensory processes. 27–33. 4, 21

[SLWP96] STEIN B. E., LONDON N., WILKINSON L. K., PRICE D. D.: Enhancement of Perceived Visual Intensity by Auditory Stimuli: A Psychophysical Analysis. *J Cog Neurosci 8*, 6 (1996), 497–506. 21, 24

[SM04] SPENCE C., MCDONALD J.: *The Handbook of Multisensory Processes*. MIT Press, Cambridge, MA, 2004, ch. The Cross-, pp. 3–25. 20

[Sou] SOUNDFIELD:. http://www.soundfield.com. 15

[SPP00] STEINMAN R. M., PIZLO Z., PIZLO F. J.: Phi is not beta, and why wertheimers discovery launched the gestalt revolution. *Vision Research 40*, 17 (2000), 2257–2264. 18

[SS01] SHIMOJO S., SHAMS L.: Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology 11*, 4 (August 2001), 505–509. 21

[SSL97] SEKULER R., SEKULER A. B., LAU R.: Sound alters visual motion perception. *Nature 385*, 6614 (January 1997), 308. 21, 24

[Ste89] Stereophonic Techniques - *An anthology of reprinted articles on stereophonic techniques*. Audio Engineering Society, 1989. 16

[Sto98] STORMS R. L.: *Auditory-Visual Cross-Modal Perception Phenomena*. {PhD} thesis, Naval Postgraduate School, 1998. 23

[Stra] STREICHER R.: The decca tree. http://mixonline.com/recording/applications/audio_decca_tree/. 15, 16

[Strb] STREICHER R.: The decca tree – it's not just for stereo anymore. http://www.wesdooley.com/pdf/Surround_Sound_Decca_Tree-urtext.pdf. 15

[SV01] SAVIOJA L., VALIMAKI V.: Interpolated 3-d digital waveguide mesh with frequency warping. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)* (2001), vol. 5, pp. 3345–3348. 8

[SW04] STRASSER W., WAND M.: Multi-resolution sound rendering. *Symp. Point-Based Graphics* (2004). 14

[TDLD07] TSINGOS N., DACHSBACHER C., LEFEBVRE S., DELLEPIANE M.: Instant sound scattering. In *Rendering Techniques (Proceedings of the Eurographics Symposium on Rendering)* (2007). 9

[TEP04] TOUIMI A. B., EMERIT M., PERNAUX J.-M.: Efficient method for multiple compressed audio streams spatialization. In *In Proceeding of ACM 3rd Intl. Conf. on Mobile and Ubiquitous multimedia* (2004). 15

[TFNC01] TSINGOS N., FUNKHOUSER T., NGAN A., CARLBOM I.: Modeling acoustics in virtual environments using the uniform theory of diffraction. *In Proc. of ACM SIGGRAPH* (2001), 545–552. 8

[TGD04] TSINGOS N., GALLO E., DRETTAKIS G.: Perceptual audio rendering of complex virtual environments. *ACM Trans. Graph. 23*, 3 (2004), 249–258. 4, 14, 17, 22

[The91] THEEUWES J.: Exogenous and endogenous control of attention: the effect of visual onsets and offsets. *Perception & psychophysics 49*, 1 (1991), 83–90. 20

[Tou00] TOUIMI A. B.: A generic framework for filtering in subband domain. In *In Proceeding of IEEE 9th Workshop on Digital Signal Processing, Hunt, Texas, USA* (October 200). 15

[Tsi05] TSINGOS N.: 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* (Sept. 2005). 13, 14, 15

[Tsi07] TSINGOS N.: Perceptually-based auralization. In *19th Intl. Congress on Acoustics* (sep 2007). 17, 22

[Tsi09a] TSINGOS N.: Pre-computing geometry-based reverberation effects for games. *35th AES Conference on Audio for Games, London* (2009). 10

[Tsi09b] TSINGOS N.: Using programmable graphics hardware for acoustics and audio rendering. In *Audio Engineering Society Convention 127* (10 2009). 10

[VBG98] VROOMEN J., BERTELSON P., GELDER B. D.: A visual influence in the discrimination of auditory location. 21, 24

[VdBOBT08] VAN DER BURG E., OLIVERS C. N., BRONKHORST A. W., THEEUWES J.: Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of experimental psychology. Human perception and performance 34*, 5 (October 2008), 1053–1065. 21, 24

[vdDKP01] VAN DEN DOEL K., KRY P. G., PAI D. K.: Foleyautomatic: Physically based sound effects for interactive simulation and animation. *ACM Computer Graphics, SIGGRAPH'01 Proceedings* (Aug. 2001), 545–552. 11

[vdDKP04] VAN DEN DOEL K., KNOTT D., PAI D. K.: Interactive simulation of complex audio-visual scenes. *Presence: Teleoperators and Virtual Environments 13*, 1 (2004). 11, 14

[vdDPA*02] VAN DEN DOEL K., PAI D. K., ADAM T., KORTCHMAR L., PICHORA-FULLER K.: Measurements of perceptual quality of contact sound models. In *Proceedings of the International Conference on Auditory Display (ICAD 2002), Kyoto, Japan* (2002), pp. 345–349. 11, 14

[VdG04] VROOMEN J., DE GELDER B.: Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon. in the handbook of multisensory processes. 140–150. 21, 24

[VKS06] VICKERS E., KRISHNAN P., SADANANDAM R.: Frequency domain artificial reverberation using spectral magnitude decay. *Proceedings of the 121th AES convention, Preprint 6926* (Oct 2006). 10

[VRR*03] VINCENT E., RODET X., RÖBEL A., FÉVOTTE C., CARPENTIER E. L., GRIBONVAL R., BENAROYA L., BIMBOT F.: 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* (Apr. 2003). 16

[VSF07] VÄLJAMÄE A., SOTO-FARACO S.: Audio-visual interactions in dynamic scenes: implications for multisensory compression. In *Invited paper at 9th International Congress on Acoustics Ű ICAŠ07* (2007). 17

[VTJ07] VÄLJAMÄE A., TAJADURA-JIMÉNEZ A.: Perceptual optimization of audio-visual media: Moved by sound. In *Narration and Spectatorship in Moving Images* (2007), Cambridge Scholars Press. 17

[Wag90] WAGENAARS W. M.: Localization of Sound in a Room with Reflecting Walls. *J. Audio Eng. Soc. 38*, 3 (Mar 1990). 6

[Wen01] WENZEL E.: Effect of increasing system latency on localization of virtual ssounds with short and long duration. *Proceeding of ICAD 2001, Espoo, Finland* (august 2001). 17

[Whi79] WHITTED T.: An improved illumination model for shaded display. In *SIGGRAPH '79: Proceedings of the 6th annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 1979), ACM, p. 14. 9

[WHT06] WARD J., HUCKSTEP B., TSAKANIKO E.: Sound-colour synaesthesia: to what extent does it use cross-modal mechanisms common to us all? In *Cortex* (2006), vol. 42, pp. 264–280. 24

[WKN03] WADA Y., KITAGAWA N., NOGUCHI K.: Audio-visual integration in temporal perception. *Int J Psychophysiol 50*, 1-2 (October 2003), 117–124. 21, 24

[WW80] WELCH R. B., WARREN D. H.: Immediate perceptual response to intersensory discrepancy. *Psychological bulletin 88*, 3 (November 1980), 638–667. 21, 24

[WWP88] WARREN R. M., WRIGHTSON J. M., PURETZ J.: Illusory continuity of tonal and infratonal periodic sounds. *The Journal of the Acoustical Society of America 84*, 4 (1988), 1338–1342. 20

[Yar67] YARBUS A. L.: *Eye Movements and Vision*. Plenum Press, New York, NY, 1967. 20

[Yew03] YEWDALL D.: *Practical Art of Motion Picture Sound (2nd edition)*. Focal Press, 2003. 15

[Yos00] YOST W. A.: *Fundamentals of hearing : an introduction*, 4th ed. Academic Press., 2000. 5, 18

[YPG01] YEE H., PATTANAIK S., GREENBERG D. P.: Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Trans. Graph. 20*, 1 (2001), 39–65. 22, 23, 24

[YR04] YILMAZ O., RICKARD S.: Blind separation of speech mixtures via time frequency masking. *IEEE Transactions on Signal Processing 52*, 7 (2004), 1830–1847. 16

[ZJ09] ZHENG C., JAMES D. L.: Harmonic fluids. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)* (2009). 12