CAMPFIRE

Acoustic Rendering for Virtual Environments

Preconference Proceedings

Snowbird, Utah, USA

26th - 29th May 2001

CAMPFIRES: Synergies for the Future

ACM SIGGRAPH and EUROGRAPHICS

CAMPFIRE: Acoustic Rendering for Virtual Environments
ACM SIGGRAPH AND EUROGRAPHICS
Snowbird, Utah, May 26-29, 2001

http://www.bell-labs.com/topic/conferences/campfire

Campfire on Acoustic Rendering for Virtual Environments

May 26th-29th 2001, Snowbird, Utah, USA

Sponsored by

ACM SIGGRAPH
Association for Computing Machinery
Special Interest Group on Graphics

and

EUROGRAPHICS
European Association for Computer Graphics

Co-chairs

Nicolas Tsingos
Bell Laboratories

Alan Chalmers
Bristol University

Motivation

Acoustic rendering aims at rendering an audible virtual sound field to a user immersed in a virtual world. With its roots in acoustics, it now has key applications in realistic virtual environment design, acoustic simulation, gaming and entertainment, art, architecture and telecommunications.

Combining rendered sound with 3D graphics enhances the sense of presence in the virtual environment. Audio cues through their reverberation properties can provide an enhanced impression of the spatial arrangement of objects within an environment.

Acoustic rendering shares many of its computational techniques and algorithms with computer graphics. This is not surprising, since both light and sound can be modeled as wave phenomena, and historical progress in wave physics tends to belong alternatively to optics or acoustics. In the world of digital simulations, computer graphics and virtual acoustics share the same geometrical tools, e.g. ray-tracing or cone tracing in worlds described by 3D primitives such as polygons.

By bringing together researchers from the computer graphics and virtual acoustics community, we can expect an extension of the historical multi-disciplinary interaction between classical optics and acoustics. We also want to take advantage of the multi-disciplinary nature of the campfires to explore a variety of applications of audio rendering technology outside the pure computer graphics, visualization and acoustics worlds.

In this Campfire, we thus invite researchers and professionals from the computer graphics, acoustics, audio signal processing, audio hardware design, virtual reality/simulation, psychology and architecture/art domains to share their expertise and ideas to help acoustic rendering and virtual worlds attain a new degree of efficiency and realism.

Nicolas Tsingos and Alan Chalmers

Acknowledgments

We would like to thank ACM SIGGRAPH and Eurographics for sponsoring the campfire events. We also express our gratitude to the following people for helping us with the contents, design and setup of the campfire web site: Carol O'Sullivan, Ann McNamara, John Dingliana, Agata Opalach and Aleta Lapone. Erica Johnson (ACM SIG Program Director), Norma Smith and the staff at Snowbird offered prompt and efficient help with the organization — thank you all.

Technical Program - Table of contents

Sunday, May 27 2001

Monday, May 28 2001

11:00 am – 12:00 pm
Panel III- The ultimate reproduction system

2:00 pm – 3:15 pm
Session IV – Audio for augmented reality

3:15 pm – 4:15 pm
Panel IV – Virtual audio for the mobility and orientation community

4:30 pm – 5:30 pm
Session V - Systems and applications (I)

5:30 pm – 6:15 pm
Panel V – Towards standard software/hardware tools for audio/visual rendering

Tuesday, May 29 2001

9:00 am – 10:15 am
Session VI – Systems and applications (II)

10:30 am – 11:30 am
Panel VI – Cross modal interactions and perceptual issues in virtual environments

Methods for Computing Sound Propagation Paths in Polyhedral Environments

Thomas Funkhouser (funk@cs.princeton.edu)
Princeton University
http://www.cs.princeton.edu/~funk

Computer-aided sound propagation prediction tools are important for design and simulation of three dimensional environments. For instance, modeling of the acoustical environment can be used to provide sound cues to aid understanding, navigation, and communication in interactive virtual environment applications, particularly if the models can be updated at interactive rates. For example, the voices of users sharing a virtual environment may be spatialized according to each user's avatar location.

A difficult challenge in acoustic modeling is computation of propagation paths from a sound's source position to a listener's receiving position. As sound may travel from source to receiver via a multitude of reflection, transmission, and diffraction paths, accurate simulation is extremely compute intensive. Prior approaches to acoustic simulation have used the image source method, whose computational complexity grows with O(n^r) (for n surfaces and r reflections), or ray tracing methods, which are prone to sampling error and require lots of computation to trace many rays.

We have been investigating data structures and algorithms to compute early propagation paths incorporating specular reflections, transmissions, and wedge diffractions in a large polygonal model fast enough to be used for interactive applications. Our approach is to precompute and store a spatial data structure that can be later used during an interactive session for evaluation of propagation paths. Briefly, our system executes as follows. During an off-line precomputation, we construct a spatial subdivision in which 3D space is partitioned into convex polyhedra (cells). Then, for each sound source, we trace beams through the spatial subdivision constructing a ``beam tree'' data structure encoding convex polyhedral regions of space possibly reached by different sequences of transmissions, specular reflections, and diffractions from the source. Then, during an interactive session, the beam trees are used to find propagation paths between pairs of sources and receivers quickly enough to enable impulse response updates for real-time auralization.

The most interesting features of this approach are that it scales well with increasing geometric complexity in densely-occluded environments and that it handles propagation paths with any combination of transmissions, specular reflections, and diffractions without aliasing. We have incorporated these data structures and algorithms into a system that supports real-time auralization and visualization of large virtual environments

Edge Diffraction and Surface Scattering in Auralization

R. R. Torres[1], M. Kleiner, U. P. Svensson[2], B.-I. Dalenbäck
Chalmers Room Acoustics Group, Chalmers University of Technology, SE-41296 Gothenburg, Sweden

[1] *Current address*: Program in Architectural Acoustics, Rensselaer Polytechnic Institute, School of Architecture and Building Science, 110 8th St., Troy, NY, 12180-3590, USA
[2] *Current address*: Gruppen for akustikk, Institutt for teleteknikk (NTNU), NO-7491 Trondheim, Norway

Abstract

In order for auralization to better replicate the binaural listening experience in a space, acoustic scattering in the room impulse response (RIR) must be more accurately calculated. Three models are discussed and used in room computations and auralizations: Lambert scattering, edge-diffraction models based on the Biot-Tolstoy-Medwin technique, and boss models for protuberances on reflecting surfaces. A psychoacoustic study with the Lambert "diffusion" model shows that the ear can clearly hear frequency-dependent changes in the scattering coefficient from 125 – 4000 Hz and that the perceived quality of the changes depends greatly on the temporal and spectral characteristics of the input signal. Edge-diffraction modeling offers a greater degree of accuracy compared to Lambert-diffusion, and a validated time-domain model is applied to calculate the RIR for a stage house, showing that reflected-diffracted combinations are significant and that even small spectral changes of up to 2 dB are audible as coloration differences in the room response. Finally, a boss model of scattering from baffled hemispheres is implemented, and binaural impulse responses are used in initial listening tests to investigate the audibility of the boss scattering. As an improvement upon Lambert-diffusion methods, a hybrid model is proposed incorporating BTM-based edge-diffraction and boss models.

Introduction

This paper briefly discusses several studies of edge diffraction and surface scattering as applied to auralization, i.e., the binaural replication of an acoustical environment. (A comprehensive definition of auralization, the acoustical analogue of visualization, was introduced by Kleiner in [1].) These studies provide investigations of the computation and perception of scattering in room simulations. The end aim is to improve the "aural accuracy" (and not simply the realism) of auralization.

For clarification it is helpful first to discuss similar but distinct terms for acoustical scattering phenomena. The term "scattering" generally refers to the redirection of sound when it interacts with a body and thus encompasses transmitted, reflected, and diffracted waves [2]. For non-specular components of this redirection of sound, and if the scattering area is periodic or statistically "rough" with respect to wavelength, one may also use the more descriptive term "surface scattering." "Edge diffraction" refers here to scattering from a wedge of a given angle, including planar "wedges" and interior corners. Finally, the term "diffusion" (sometimes equated with "diffuse reflection") is typically related to the redirection of a portion of the specular energy into non-specular directions (in its broad usage in room acoustics). Since phase is ignored, diffusion models cannot directly simulate edge-diffraction effects where, for example, the edge contributions interfere destructively with the specular reflection for certain source-receiver orientations and wedge angles. Surface "diffusion" can also be confused with diffusivity of the sound field, which is related but not equivalent to diffuse reflection. Despite these shortcomings, models of Lambert surface "diffusion" still offer a practical starting point for investigating perception of surface scattering, as discussed in the first study.

1. Perception of Lambert Surface "Diffusion"

This first study (see [3] for details) investigates the temporal-spectral perception of surface scattering as modeled by one of the most basic approximations, i.e., with a surface-diffusion coefficient δ that follows Lambert's law [4]. The study addresses the questions (a) "Is the ear sensitive to changes in δ in all frequency ranges?" and (b) "How are these changes perceived?" by employing listening tests to compare computed auralizations of a concert hall.

Binaural room impulse responses (BRIR) are computed with the program CATT-Acoustic, based on randomized cone-tracing [5]. The surface diffusion coefficient δ in the entire room is adjusted from 10% to 60% in each of three frequency regions within the 125 to 4000 Hz octave bands: "High" (2 and 4 kHz), "Mid" (500 and 1000 Hz), and "Low" (125 and 250 Hz). (The BRIR, however, cover the entire audio range.) The frequency dependence is described by a quasi-step function, which corresponds to increased surface scattering at an onset frequency given by each region. Moreover, the first BRIR pair is in the "High" frequency region, with diffusion compared at 10% and 60%, while constant in the lower regions at 1% (numerically extreme but representative of purely specular reflection). This quasi-step function then slides to the middle ("Mid") region where 10% and 60% diffusion is compared, with 1% diffusion below and 60% above. Finally, the difference in diffusion is compared in the "Low" region, the upper regions having 60% diffusion. In total, three pairs of BRIR are constructed (for the three frequency regions), each having one BRIR with 10% surface diffusion (signal "A") and another at 60% (signal "B"). The auralizations are based on a concert hall with hexagonal shape (8650 m^3) and reverberation times from 1.9 – 2.2 seconds. A rear-center position is used for this initial study and is expected to most strongly reveal the effects of varying δ, as the seat is near a reflecting surface and since the perceived comb-filter effect from the nearest wall is on-axis with the direct sound (and thus greatest). This yields an upper limit for this study, which can be complemented in future work with other reference points.

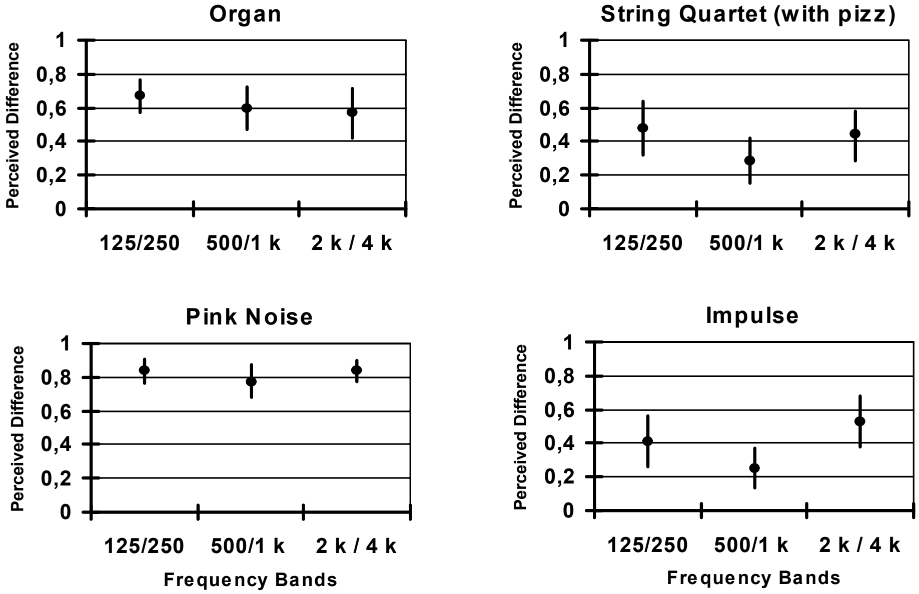


Figure 1. Average perceived difference (solid circles) and standard deviations (vertical lines), when the diffusion coefficient is varied in different regions. The average levels depend on the input signal. Rankings of frequency regions relative to each other (for a given input signal) varied depending on the listener.

The three BRIR pairs are convolved with anechoic recordings: two "sustained" (i.e., synthesized organ chord, and five seconds of pink noise), and two "impulsive" (string quartet with pizzicato, and the unconvolved BRIR). These signals are chosen to highlight time vs. frequency effects. Binaural pair comparisons are conducted with equalized headphones, and listeners rate the overall difference between A and B. In addition, the 15 listeners may specify whether they hear a difference in coloration and/or spaciousness and/or any other quality (described in a comment area). If a spaciousness difference is heard, the listener specifies whether "A" or "B" is more spacious.

Figure 1 shows the general perceived difference when δ is varied in different regions (average values in solid circles, with standard deviations in vertical lines), and there is a clear dependence on the input signal. For example, the general perceived differences with pink noise are greater than those for the impulse or string quartet. For some signals the differences are audible in all frequency regions, which shows that scattering must be treated with frequency dependence (and not simplistically represented by a single scattering coefficient, as in some algorithms).

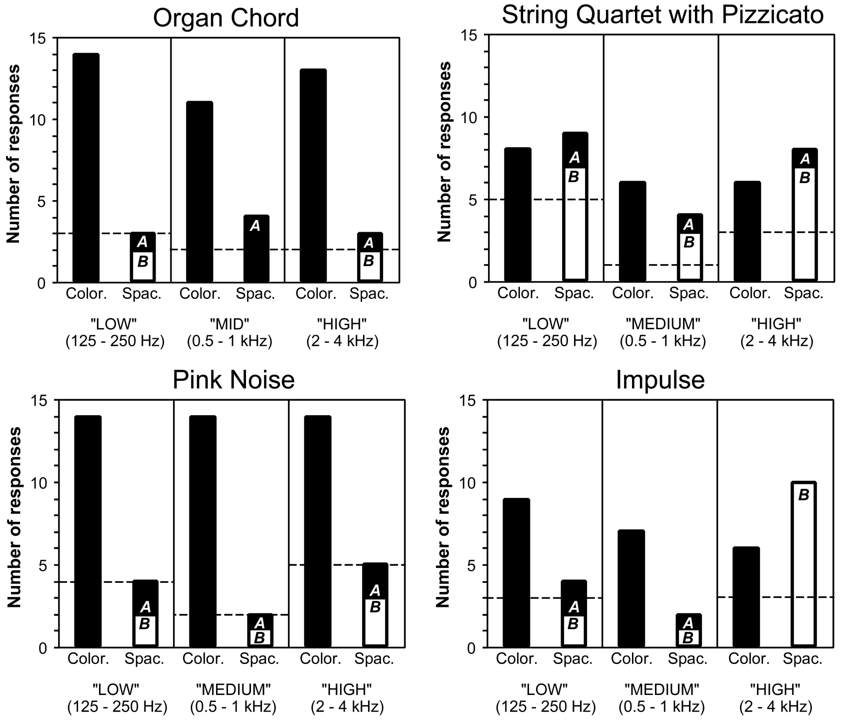


Figure 2. Listeners specified whether they heard differences in coloration, spaciousness, and/or other qualities. In the second vertical bar of each pair, the white portion shows how many thought signal "B" (with 60% diffusion) sounded more spacious. The dashed lines show how many people heard differences in *both* spaciousness and coloration.

Figure 2 shows the listeners' characterization of the differences. Each of the three frequency regions contains a pair of vertical bars. The left, black bar depicts how many people heard differences in coloration between signal "A" and signal "B." The right bar shows how many listeners heard differences in spaciousness; this bar is divided into those who thought either "A" or "B" was more spacious. The horizontal dashed line shows how many people heard differences in *both* spaciousness and coloration for a given pair. The results show that coloration differences are perceived more strongly for sustained signals (organ, pink noise), as compared with the impulsive signals. Also, for the sustained signals, very few heard only spaciousness differences (see dashed lines) without also hearing coloration differences. As discussed in [3], a significance level of $P < 0.5$ corresponds to approximately 11 of 15 responses. The optional listener comments [3] are also entirely consistent in identifying in which frequency ranges δ is varied. For example, when comparing the organ signals, listeners wrote for the “High” frequency region: “B less treble,” “B more bass.” This is reasonable, as the high-frequency specular component is reduced in signal “B” compared with signal “A” (signal “A” has $\delta = 10\%$, signal “B” $\delta = 60\%$). Similarly, for the “Low” region, listeners wrote: “A more bass,” “A less high frequency,” “B brighter.” Moreover, increasing low-frequency surface diffusion in signal B is perceived as either making signal B “brighter” *or endowing* signal A with “more bass.” Differences in spaciousness were audible but less obvious to the listeners, which may depend on room geometry, non-personal HRTFs, and other factors. Coloration differences also become less obvious for impulsive signals. Finally, reverberation does not seem to obscure perception of coloration changes (from varying the surface diffusion); this is shown for the sustained signals, where, despite reverberation, most test takers still heard differences in coloration that relate to the stationary part of the organ chord or pink noise auralizations.

In summary: For some signals, changes in the diffusion coefficient are clearly audible within a wide frequency region, indicating the necessity of frequency-dependent scattering models. The perception of these changes depends on the input signal. Listeners, though uninformed of the differences between high-or low-diffusion signals, still give consistent answers regarding perceived changes in coloration.

2. Experiments with Edge Diffraction

The second study [6] focuses on improving the accuracy of calculating the early part of the room impulse response by utilizing a validated time-domain edge-diffraction model (Svensson *et al.*’s analytical extensions to the Biot-Tolstoy-Medwin technique [7]. As discussed in [6], edge diffraction improves upon geometrical acoustics by (1) maintaining a continuous acoustic field around the edge and (2) correcting reflection strength from finite room surfaces. As derived in [7], the edge diffraction computations consist of dividing edges into sources with analytically derived strengths. One sample of the impulse response *h* is then (Eq. 35, [7]):

$$\Delta h_i \approx -\frac{\nu}{4\pi}\frac{\beta(\alpha,\gamma,\theta_S,\theta_R,)}{ml}\Delta z \qquad (1)$$

where Δz is the length of the source at z_i along the edge, ν is a “wedge index” describing the wedge’s concavity ($> \pi$) or convexity ($< \pi$), *m* and *l* are distances to the source *S* and receiver *R*, and β is an analytical edge-source directivity-function depending on the location of the source and receiver relative to a given edge source. This method can be used for finite wedges (with rigid or pressure-release boundary conditions), even if curvilinear.

In addition to “direct” diffraction paths (from the source to an edge to the receiver), the following combinations of “specular/diffractive” components are computed in [6]:

$$\begin{aligned} h_{sp-ed}(t;S\,|\,R) &= \sum_i \sum_j h_{diffr1}(t;S'_i\,|\,E_j\,|\,R) \\ h_{ed-sp}(t;S\,|\,R) &= \sum_i \sum_j h_{diffr1}(t;S\,|\,E_j\,|\,R'_i) \\ h_{sp-ed-sp}(t;S\,|\,R) &= \sum_i \sum_j \sum_k h_{diffr1}(t;S'_i\,|\,E_j\,|\,R'_k) \end{aligned} \qquad (2)\text{-}(4)$$

where, e.g., h_{sp-ed} is the impulse response for the *sp-ed* paths (“specular reflection to edge diffraction”) between the source *S* and the receiver *R*, h_{diffl} represents first-order edge diffraction from an image source

S_i' via a visible edge E_j to R, and t is time. (An image receiver is denoted by R'.) The summations are done over the indices that remain after edge-visibility checks. These combinations are significant components of the total computed diffraction, as shown by comparisons with measurements (of a simplified stage house) [6] in Fig. 3 and as shown in computations in [8]. Peak-by-peak comparison shows that the computed impulse response includes nearly all of the edge diffraction in the measured RIR. (The extra scattering from the source's bracing is not included in computations.)

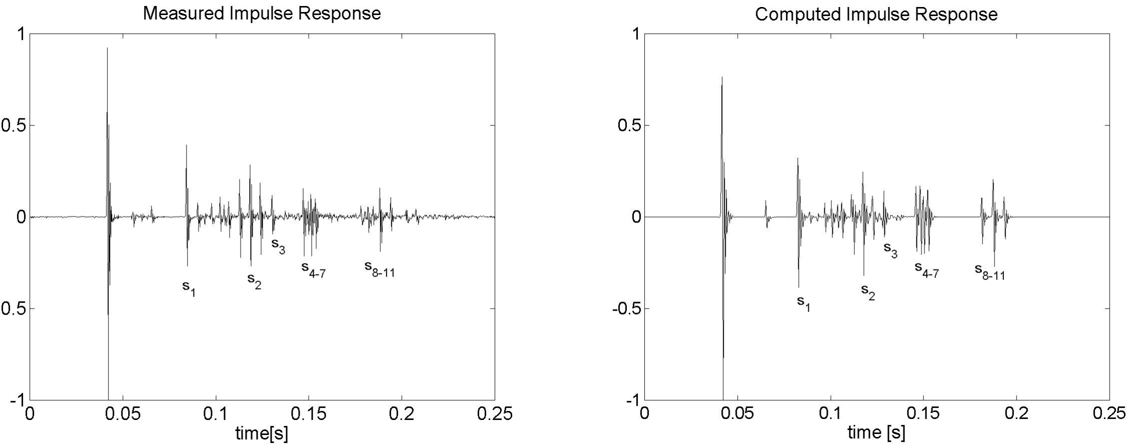


Figure 3. The left figure shows the measured impulse response in front of the stage house; the right figure, the computed response convolved with the spark source. The specular reflections are denoted by "*s*"; the non-specular arrivals (especially between s_1 and s_3) are edge diffractions.

For the large, smooth surfaces modeled in [6], the inclusion of diffraction (added to the geometrical acoustics solution) resulted in level differences of only 1-2 dB, below about 160 Hz. However, these small coloration changes are still clearly audible for input signals with low-frequency content, as demonstrated by double-blind ABX listening tests. In these monaural tests, impulse responses with different diffraction combinations are convolved with anechoic pink noise, organ music, speech, and a unit impulse. A significance level of 0.05 is used, and 18 listeners take the test. Results [6] show that the total computed diffraction is audible for the pink noise, organ, and impulse signals, which have richer low-frequency content than the speech signals. Here, the receiver is visible to the source, which makes the test for audibility more conservative than if the receiver is shadowed. The results also indicate that second-order diffraction is not significantly audible and can be neglected for similar source-receiver orientations. Although these particular results seem to apply only to low frequencies, realistic cases would actually have more diffracting wedges of smaller scales, such that diffraction effects would extend higher in frequency, when the wavelength is comparable with the dimensions of the reflecting facet [9]. The perception of edge diffraction in the early RIR could possibly be altered by reverberation, although the previous study [3] shows that changes in early coloration due to varying surface diffusion are still clearly audible even with 2 seconds reverberation and using a continuous input signal.

This edge diffraction model can also be used to calculate scattering from rough surfaces or objects modeled as a construction of wedges. Such application has often been called the "wedge assemblage method" [10]. One could use this method to model many surfaces in the low- to mid-frequency region, i.e., where the wavelength is greater than or comparable to the characteristic dimensions of a room's surfaces. Future work should also utilize binaural simulations. As a practical approximation for binaural modeling [6], one could use the least-time point on the wedge as a coordinate representing the entire edge relative to the listener. With this approximation each wedge then only requires one HRTF, whose angle corresponds to the least-time point.

3. Surface Scattering with “Boss” Models

Although scattering from many surfaces can be modeled with edge diffraction, a complementary method is desirable for objects not resembling constructions of wedges. Boss models of scattering from one or more “bosses” on a plane (i.e., protuberances such as hemispheres or semi-cylinders) can be applied to isolated scatterers as well as to periodic or statistical (random) scattering surfaces. One approach is a Green’s function method, where the total field is written as a sum of the incident field and the fields scattered from all the bosses. The scattering from an embedded sphere (i.e., a hemisphere on a plane) is depicted in Fig. 4. The total pressure at the receiver is the sum of the incident and reflected *specular* components, plus the incident and reflected *boss-scattered* components above the plane. With the latter three grouped together, the total pressure is the sum of the incident and total scattered components:

$$p_{tot} = p_i + p_{scatt} = p_i + \left(p_r^{sp} + p_i^{sc} + p_r^{sc}\right). \tag{5}$$

Moreover, the method combines the method of images and an analytical solution for scattering from a rigid sphere. This is an attractive way to treat widely spaced scatterers such as statues, which can be modeled at lower frequencies as arrangements of cylinders and spheres, or other canonical objects (e.g., prolate spheroids) for which scattering solutions are known. The main requirement is that the object is symmetrical about the reflecting plane.

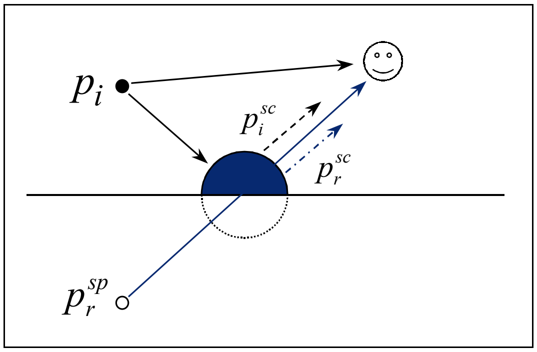


Figure 4. The total pressure at the receiver is the sum of the incident and specularly reflected components (p_i and p_r^{sp}) and the incident and reflected scattered components (p_i^{sc} and p_r^{sc}) above the reflecting plane.

The combined application of the above methods forms a proposed hybrid approach which may prove to be an acceptable balance between accuracy and utility in auralization: use of edge-diffraction methods at lower frequencies, boss models and the wedge-assemblage method to account for scattering from surfaces at mid-frequencies, and possibly a Lambert- or Tangent-Plane Approximation [11,12] for higher frequencies, with the restriction that edge diffraction near the boundaries is always taken into account, especially at grazing angles where the projected area is dominated by the edges.

Listening tests with binaural auralizations are in progress. (See [13].) The listening tests have two parts. The first concentrates on the binaural scattering components alone, to confirm that the frequency-dependence is audible and investigate how it is perceived. The subjects perform an ABX test to determine whether the differences between the boss sizes are aurally significant. They also rate the “general difference” on a scale and describe the character of the difference. Four signals are used: an organ chord, a string quartet, a chirp, and the impulse itself. The second part compares the early specular BRIR “with and without" bosses. The largest boss size is used (corresponding to 125 Hz), along with the four input signals listed above. The total number of cases is limited to avoid fatigue in the test listeners.

Initial ABX results show that the frequency-dependence is clearly audible in the scattering alone, as one should expect. The greatest perceived differences, with the lowest standard deviations, are observed for the pairs comparing the lowest and highest frequency ranges (i.e., largest and smallest boss radii). The second set of listening tests indicate a clear dependence on the input signal, with the greatest audibility of the boss scattering related to input signals more transient in nature.

Future work could include determining more physically-based values for scattering coefficients as input data into current auralization programs and studying whether the perceptual differences are significant between more accurate scattering models and more approximate approaches. The derivation of a time-domain boss model might also be of practical use in computation of room impulse responses.

References

1. M. Kleiner, B.-I. Dalenbäck, P. Svensson, "Auralization–an overview," J. Audio Eng. Soc. **41**, 861-875 (1993).
2. H. Medwin and C. S. Clay, *Fundamentals of Acoustical Oceanography* (Academic Press, 1998), p. 24.
3. R. R. Torres, M. Kleiner, B.-I. Dalenbäck, "Audibility of 'diffusion' in room acoustics auralization: an initial investigation," *Acustica/Acta Acustica* (Special Issue on Room Acoustics) **86**(6), 919-927 (2000). (Excerpts reprinted with permission, Deutscher Apotheker Verlag).
4. H. Kuttruff, *Room Acoustics* (Elsevier Applied Science, 1991), Third Ed., pp. 84-85, 110.
5. B.-I. Dalenbäck, "Verification of prediction based on Randomized Tail-Corrected cone-tracing and array modeling," Proc. of 137th Meeting of ASA, 2nd Convention of EAA (Forum Acusticum 99), Berlin (1999).
6. R. R. Torres, U. P. Svensson, M. Kleiner, "Computation of edge diffraction for more accurate room acoustics auralization" *J. Acoust. Soc. Am.* **109**(2), 600-610 (2001). (Excerpts reprinted with permission, Copyright 2001, Acoustical Society of America.)
7. U. P. Svensson, R. I. Fred, and J. Vanderkooy, "An analytic secondary source model of edge diffraction impulse responses,"J. Acoust. Soc. Am. **106**, 2331-2344 (1999).
8. U. P. Svensson, R. R. Torres, H. Medwin, "The color of early sound arrivals in an auditorium," *J. Acoust. Soc. Am.* **108**(5), 2648 (A) (2000).
9. R. R. Torres and M. Vorländer, "Scale-model MLS-measurements of scattering from overhead panel arrays" (submitted December 2000 to *Acustica/Acta Acustica*).
10. R. S. Keiffer and J. C. Novarini, "A time domain rough surface scattering model based on wedge diffraction: Application to low-frequency backscattering from two-dimensional sea surfaces," J. Acoust. Soc. Am. **107**, 27-39 (2000).
11. J.A. Ogilvy, *Theory of Wave Scattering from Random Rough Surfaces*, Adam Hilger (1991).
12. A.Voronovich, *Wave Scattering from Rough Surfaces*, Springer (1999).
13. R. R. Torres, *Studies of Edge Diffraction and Scattering: Applications to Room Acoustics and Auralization*, doctoral thesis, ISBN 91-7197-956-5, Chalmers Univ. of Tech., Sweden (2000), Paper IV.

Acoustic rendering beyond geometrical acoustics

Peter Svensson (svensson@tele.ntnu.no)
Acoustics group, Department of telecommunications, Norwegian University of Science and Technology
http://www.tele.ntnu.no/users/svensson

Most acoustic rendering systems of today are based on geometrical acoustics. This makes it possible to transfer tricks and techniques from computer graphics (e.g., ray tracing) and generally works well.
For some cases, however, these methods fail, and in particular at low frequencies and for sound diffracting around corners. This presentation will discuss various wave based techniques for predicting sound fields at low frequencies, and in particular methods based on edge diffraction and TLM (Transmission Line Modelling) or waveguide methods. Examples from auditoria and city street environments will be presented.

Digital Waveguide Mesh for Room Acoustic Modeling

Lauri Savioja and Tapio Lokki
Helsinki University of Technology
Telecommunications Software and Multimedia Laboratory
P.O.Box 5400, FIN-02015 HUT, FINLAND

Abstract

The two main approaches for room acoustic modeling are the wave-based and the ray-based techniques. In this paper we briefly overview the digital waveguide mesh method which is a wave-based model operating in the time domain. As a case study we show visualizations of edge diffraction modeled with the waveguide mesh technique. Some preliminary analysis of computational requirements for real-time auralizations are also presented, and the idea of frequency domain hybrid model is revisited.

1 Introduction

In real-time acoustic modeling the main emphasis has been on geometrical acoustics. This is not sufficient for authentic auralization due to lack of wave-based phenomena such as diffraction and interference. During the recent years the computational capabilities in an ordinary PC have grown enormously. Due to this it is not obvious anymore that all the wave-based methods are out of reach. These methods can be grossly divided into two categories. The ones operating in the time domain such as the FDTD (finite difference time domain), and the ones operating in the frequency domain such as FEM (finite element method) and BEM (boundary element method). At this point, we see the time domain approach to be much more appropriate for real-time acoustic rendering.

In this paper we briefly review one wave-based method, the digital waveguide mesh method and discuss its suitability for auralization at low frequencies. We have been developing this method since 1994, first for analysis of low frequency behavior of closed spaces such as listening rooms and loudspeaker enclosures. Nowadays the computation power has increased so much that the method is suitable also for larger spaces. In addition, the algorithm has improved a lot since its early days, and the technique is applicable in real cases.

In the campfire, we would like to have discussion on the position of wave-based modeling in the field of acoustic rendering both in real-time and in non-real-time simulations.

2 Digital Waveguide Mesh

The digital waveguide mesh is an extension of the one-dimensional digital waveguide technique [1, 2, 3]. The mesh can be used for simulation of two- and three-dimensional wave propagation in musical instruments and acoustic spaces. Mathematically it is very close to the finite difference methods. The original rectangular digital waveguide mesh algorithm suffers from direction-dependent dispersion. Alternative geometries, such as the triangular mesh, have been proposed to improve the performance of the mesh [4, 5]. Another choice to overcome this problem is use of multidimensional interpolation [3]. These techniques enhance the direction dependency problem, but there still remains dispersion. This dispersion error can be compensated to a certain degree by frequency warping, but in the case of real-time simulations, this is not possible with current algorithms [6, 3].

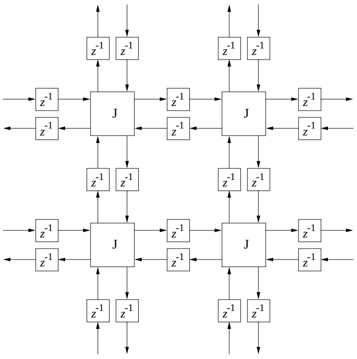


Figure 1: In the original 2-D digital waveguide mesh each node is connected to four neighbors with unit delays [2]. For 3-D meshes additional lines to upward and downward directions are required.

So far, the best solution for wave-based room acoustic modeling in the time domain, is the optimally interpolated three-dimensional digital waveguide mesh [7]. In this paper we still concentrate on the original 3D mesh.

2.1 Mesh Structure

A digital waveguide mesh is a regular array of discretely spaced 1-D digital waveguides arranged along each perpendicular dimension, interconnected at their intersections. A two-dimensional case is illustrated in Fig. 1. The resulting mesh of a 3-D space is a regular rectangular grid in which each node is connected to its six neighbors by unit delays [8, 1, 3].

The equations governing the mesh can be represented either by means of the nodes or by the means of the waveguides connecting the nodes. In this paper we apply the node approach. The difference equation for 3-D rectangular mesh is [8]

$$\begin{aligned} p(n+1,x,y,z) & \\ = \; & \tfrac{1}{3}[p(n,x+1,y,z)+p(n,x-1,y,z) \\ & +p(n,x,y+1,z)+p(n,x,y-1,z) \\ & +p(n,x,y,z+1)+p(n,x,y,z-1)] \\ & -p(n-1,x,y,z) \end{aligned} \tag{1}$$

where p represents the sound pressure at a junction at time step n, and x, y, and z are the coordinates of a node. This equation is equivalent to a difference equation derived from the Helmholtz equation by discretizing time and space. The update frequency of an N-dimensional mesh is:

$$f_s = \frac{c\sqrt{N}}{\Delta x} \approx \frac{588.9}{\Delta x} Hz \tag{2}$$

where c represents the speed of sound in the medium and Δx is the spatial sampling interval corresponding to the distance between two neighboring nodes. The approximate value stands for a typical room simulation ($c = 340m/s, N = 3$). That same frequency is also the sampling frequency of the resulting impulse response.

2.2 Computational Complexity

Let us consider the computational load by an example. If we study a room of size $5m \times 10m \times 3m$ with grid spacing of 0.2m we have $25 \times 50 \times 15 = 18750$ nodes. For each node six additions and one multiplication,

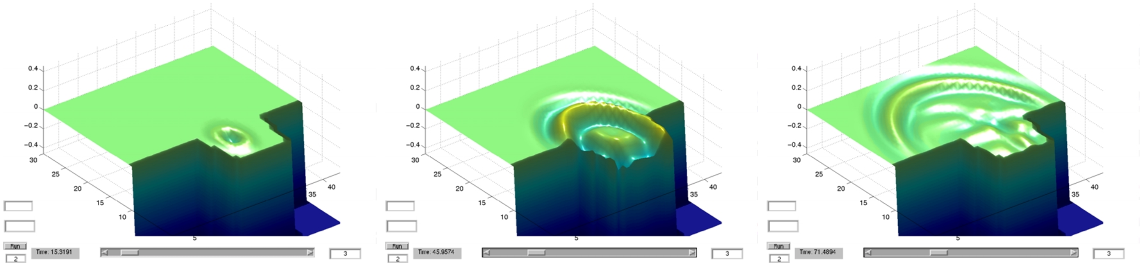


Figure 2: A set of visualized slices of the sound pressure level in a stage house illustrating diffraction.

altogether seven operations, are required. This means that each time sample takes 131 250 instructions. With 0.2m grid spacing the sampling frequency will be 3kHz thus resulting in 386 MIPS (millions of instructions per second). The valid frequency range for auralizations depends on the application, but at most it is one fourth of the sampling rate. In this case the auralizations up to 750 Hz could be achieved with the given computational load. The load is still quite heavy, but we believe that in the near future it is possible to apply the technique in real-time at the low end of the frequency band.

3 Applications

In the following we describe a couple of application areas for the digital waveguide mesh.

3.1 Traditional Room Acoustic Modeling

So far we have concentrated in finding modes in a given space with the method. One interesting study has been carried out dealing with diffraction. In Fig. 2 there are a couple of visualizations of a stage house applied in the diffraction study. The sound source is on the stage and there are several listeners in the hall. In these visualizations a cross-section showing the sound pressure level at a given height are shown. In the campfire we will show these visualizations as animations.

3.2 Auralization

We haven't made any auralizations yet with the method, but the first experiments will be done in the near future. To achieve realistic auralizations we need to make frequency domain hybrid renderer [9] in which the lowest end is calculated with the digital waveguide mesh and the upper end with our current DIVA system [10] which is based on the image-source method and artificial late reverberation.

4 Conclusions

In this paper we have discussed the digital waveguide mesh method. We believe that in the near future the room acoustic modeling techniques applied in auralization will include also some wave-based methods. Especially the ones operating in the time domain are interesting in this sense. Our goal is to develop a frequency domain hybrid in which we use both wave-based and ray-based methods to make realistic auralizations.

It would be nice to have some discussion in the campfire dealing with the future of the wave-based modeling methods.

References

[1] S. Van Duyne and J. O. Smith. Physical modeling with the 2-D digital waveguide mesh. In *Proc. Int. Computer Music Conf. (ICMC'93)*, pages 40–47, Tokyo, Japan, Sept. 1993.

[2] S. Van Duyne and J. O. Smith. The 2-D digital waveguide mesh. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 1993.

[3] L. Savioja and V. Välimäki. Reducing the dispersion error in the digital waveguide mesh using interpolation and frequency-warping techniques. *IEEE Trans. Speech and Audio Process.*, 8(2):184–194, March 2000.

[4] F. Fontana and D. Rocchesso. A new formulation of the 2D-waveguide mesh for percussion instruments. In *Proc. XI Colloquium on Musical Informatics*, pages 27–30, Bologna, Italy, 8-11 Nov. 1995.

[5] S. Van Duyne and J. O. Smith. The 3D tetrahedral digital waveguide mesh with musical applications. In *Proc. Int. Computer Music Conf. (ICMC'96)*, pages 9–16, Hong Kong, 19-24 Aug. 1996.

[6] L. Savioja and V. Välimäki. Reduction of the dispersion error in the triangular digital waveguide mesh using frequency warping. *IEEE Signal Processing Letters*, 6(3):58–60, March 1999.

[7] L. Savioja and V. Välimäki. Interpolated 3-D digital waveguide mesh with frequency warping. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, 15-19 May 2001.

[8] L. Savioja, T. Rinne, and T. Takala. Simulation of room acoustics with a 3-D finite difference mesh. In *Proc. Int. Computer Music Conf.*, pages 463–466, Aarhus, Denmark, 12-17 Sept. 1994.

[9] L. Savioja, J. Backman, A. Järvinen, and T. Takala. Waveguide mesh method for low-frequency simulation of room acoustics. In *Proc. 15th Int. Congr. Acoust. (ICA'95)*, volume 2, pages 637–640, Trondheim, Norway, June 1995.

[10] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. Creating interactive virtual acoustic environments. *J. Audio Eng. Soc.*, 47(9):675–705, Sept. 1999.

LOCALIZING SOUND IN ROOMS

Barbara Shinn-Cunningham

Department of Cognitive and Neural Systems and Biomedical Engineering
Boston University, 677 Beacon St., Boston, MA 02215
Email: shinn@cns.bu.edu
Ph: 617-353-5764
FAX: 617-353-7755

INTRODUCTION

Relatively little psychoacoustic work has examined how realistic echoes and reverberation affect spatial auditory perception. Within psychoacoustics, echoes and reverberation are generally thought to 1) cause little degradation in directional perception (as suggested by studies of the "precedence effect"; e.g., see Litovsky, Colburn, Yost & Guzman, 1999) and 2) improve distance perception (by some essentially unknown mechanism; e.g., see Mershon & King, 1975).

Head-related transfer functions (HRTFs) show how the signals that reach the two ears are related to the original source signal from a specific location in space (Wightman & Kistler, 1989a; Wightman & Kistler, 1989b; Wenzel, 1992; Carlile, 1996). HRTFs have been examined in detail in anechoic space as a function of source direction and, more recently, as a function of source distance (Brungart & Rabinowitz, 1999b). Typically, such HRTFs are relatively smooth (as a function of frequency) at low frequencies, with notches and peaks above about 6 kHz. The frequency locations of these notches and peaks depend on source elevation and are used by listeners to determine source elevation (e.g., see Wenzel, Arruda, Kistler & Wightman, 1993; Middlebrooks, 1997). Changes in the laterality of the source (relative to the median plane) cause changes in the interaural time difference (ITD) between the signals reaching the left and right ear, a cue known to mediate perception of source laterality (for a review, see Middlebrooks & Green, 1991). Changes in both source laterality and source distance cause changes in the interaural level difference (ILD, difference in the magnitude spectra of the left and right HRTFs; e.g., see Shinn-Cunningham, Santarelli & Kopčo, 2000b). Recent studies of anechoic localization show that ILDs convey some distance information to listeners when sources are near the head (Brungart & Durlach, 1999a).

Recent work in my laboratory addresses how echoes and reverberation influence localization in two ways: by 1) taking empirical measures of the sounds that reach a listener's ears in a room (and studying how these signals vary with source location and listener position) and 2) measuring human localization performance (in three dimensions) when listeners are presented with realistic reverberant signals. Results suggest that spatial perception is affected by room acoustics more than the literature might suggest; and that high-level factors, such as knowledge and experience, have a notable impact on how subjects interpret spatial cues in a reverberant space.

ACOUSTIC MEASURES

In order to understand how human perceivers perceive auditory source position in rooms, it is important to examine how echoes and reverberation affect the cues thought to underlie spatial perception. HRTFs were measured for a source and listener in a reverberant room (broadband T_{60} ~

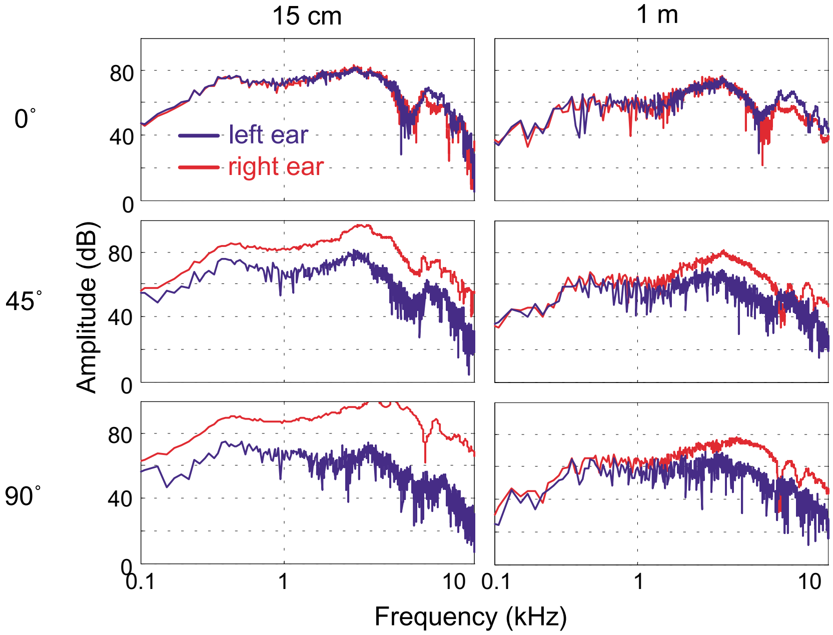


Figure 1: Magnitude spectrum (dB) of HRTFs in the center of a reverberant room as a function of source position relative to listener. Left and right columns show near (15 cm) and far (1 m) sources, respectively. Top, middle, and bottom rows show the lateral angle of the source relative to median plane (0˚, 45˚, and 90˚ to the right, respectively).

450 ms) using a maximum-length-sequence technique. Measurements were made for individual human listeners as well as a KEMAR manikin for sources at different positions (relative to the head) as well as different listener positions within the room (Brown, 2000).

Figure 1 shows the HRTF magnitude spectra for a source at various positions relative to KEMAR, who was positioned in the center of the room. In addition to randomly distorting the signal spectra reaching the ears, reverberation reduces the depth of any spectral notches. Because the notch depths are reduced by reverberation, one might expect that elevation perception is less robust in a real room than it is in anechoic space (see also Begault, 1992b). The effects are greatest at the ear farther from the source because it receives less direct energy (making the reverberant energy relatively stronger). For the source positions shown (to the right of the listener), the left ear signal (blue) is affected more than the right ear signal (red). The effect of reverberation increases with distance for both left and right ears because the direct sound level decreases; for the cases shown, the effect of reverberation is greater in the right column (source at 1 m) than the left column (source at 15 cm). Finally, source laterality affects the influence of reverberation as well; the effects increase at the left ear and decrease at the right ear as the source moves from 0° (top row) to 90° right (bottom row).

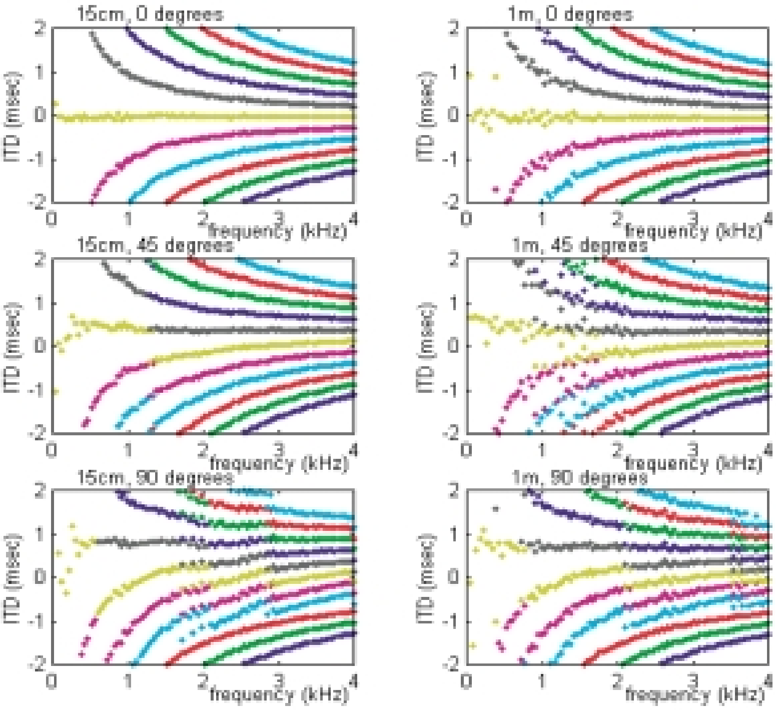


Figure 2: Interaural phase difference versus frequency for the same listener and source positions as in Figure 1.

Echoes and reverberation also distort interaural differences, and the amount of distortion grows with source distance and laterality. Figure 2 shows ITD as a function of frequency for the same source positions and listener position shown in Figure 1. At any single frequency, there is an essential ambiguity in the interaural time difference that corresponds to a phase difference (at that frequency) of 2 rad. The "true" interaural time delay is that value which yields approximately the same ITD at all frequencies. In anechoic space, similar calculations lead to an essentially flat line as a function of frequency (e.g., see blue symbols in Figure 4). However, as seen in Figure 2, the effect of reverberation is to introduce noise into the ITD as a function of frequency. Thus, one might expect judgments of source laterality to be affected by echoes and reverberation, although these effects may be small due to the precedence effect (e.g., see Litovsky et al., 1999). Similar results obtain when one examines interaural level differences (ILDs), although there is a tendency for echoes and reverberation to reduce the ILD magnitude in addition to generating frequency-to-frequency distortions.

Results show that the effects of echoes and reverberation depend on the location of the source relative to the listener. Of course, results also depend on the listener location in the room. For a listener located near a wall or other reflective surface, the influence of the resulting early-arriving, intense echo can cause large distortions in the magnitude spectra at the ears, the interaural phase differences, and the interaural level differences. These distortions are much more dramatic than those that occur when the listener is in the center of the room. In fact, early-arriving reflections cause comb-filtering effects characterized by deep notches and rapid phase shifts with frequency, both of which can lead to large distortions of spatial cues.

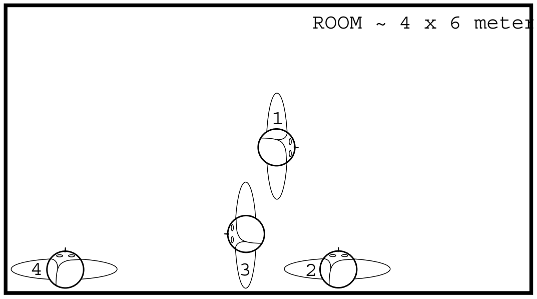


Figure 3: Four listener configurations for which HRTFs were measured in a reverberant room (not to scale).

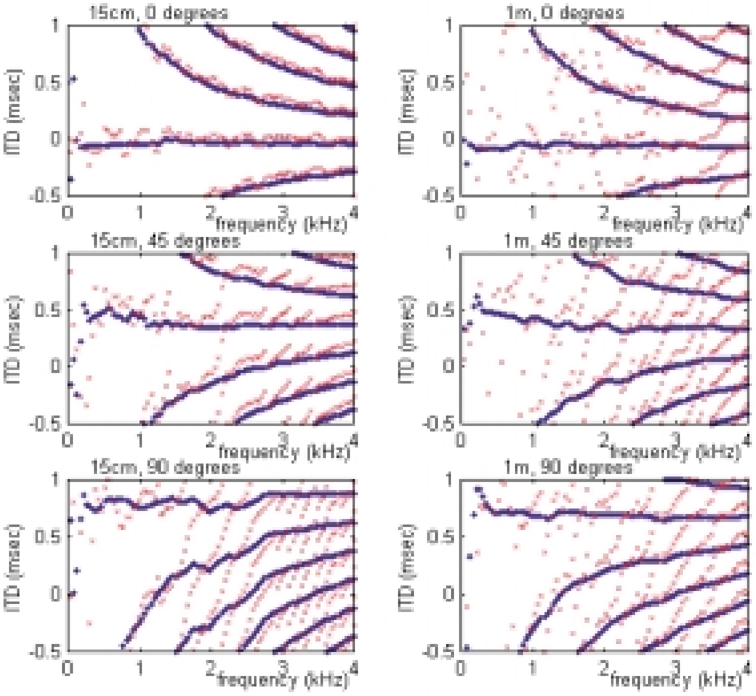


Figure 4: ITD versus frequency for the same source positions as in Figure 2. Blue symbols show anechoic and red symbols reverberant results for a listener located in the corner of the room.

In order to systematically evaluate HRTFs as a function of listener location and orientation, HRTFs were measured for four different configurations of the listener (Kopčo & Shinn-Cunningham, 2001). Figure 3 diagrams the listener positions/orientations for which HRTFs were measured (for the same six relative source positions shown in Figures 1 and 2). Results show that the "cleanest" results are obtained when a listener is in the center of the room (configuration 1 in Figure 3). Spatial cues becoming increasingly degraded as the listener approaches a wall (configuration 3), are even worse when the subject has his back to the wall (configuration 2), and are most distorted when the listener is located in the corner of the room (configuration 4).

Figure 4 demonstrates how much worse the acoustic distortion can be by showing ITD as a function of frequency for the same relative source positions as in Figure 2, but for listener configuration 4 (corner of the room; note that Figs 2 and 4 use different ITD scales). The blue symbols show the ITD that would arise for anechoic HRTFs; the red symbols show the corresponding ITD for the reverberant HRTFs. For a source near to and directly in front of the listener, echoes and reverberation only marginally affect ITD; however, for all other conditions, the ITD is dramatically distorted.

Taken as whole, acoustic measures suggest that directional localization performance should be degraded in a room compared to in anechoic space, and that this degradation should depend on where the listener is located in the room. Directional performance should be worst when a subject is located in the corner of the room and best when a listener is in the center of the room. In contrast, reverberation should provide source distance information. The degree to which distance perception varies with source and listener position may help in teasing out what aspects of reverberation provide distance information to the listener.

BEHAVIORAL MEASURES

Human localization performance was measured in the same room in which acoustic measures were made (Santarelli, Kopčo & Shinn-Cunningham, 1999; Santarelli, 2000; Santarelli, Kopčo & Shinn-Cunningham, 2000; Kopčo et al., 2001) using an experimental procedure essentially identical to that employed in a previous anechoic localization study (Brungart et al., 1999a). In the experiments, a human experimenter positioned a small speaker at a random location near the listener, whose eyes were closed, and a broadband signal was presented. The actual position of the speaker was measured using an electromagnetic tracker (Polhemus) mounted on the speaker, and the speaker was moved to a neutral position. The listener then opened his eyes and used a pointer to indicate the heard position of the source (in three-dimensional space). A second electromagnetic tracker, affixed to the end of the pointer, measured the response. At the beginning of the experiments, subjects were given an hour of practice on the task, just as in the previous anechoic study (Brungart et al., 1999a).

An initial experiment (Santarelli et al., 1999) confirmed that directional perception was degraded in the room compared to anechoic space, but that distance perception was vastly improved. However, in this initial experiment, two conditions were run. In both conditions, the listener was located in the center of the room. In the first condition, there were no objects near the listener. In the second condition, a 6' x 4' plywood board, covered in acrylic paint, was positioned just to the left of the listener. We anticipated that subjects' localization accuracy would be much worse in the second condition compared to the first, due to the presence of the board (and the concomitant early, intense reflections). Instead, we found that the listeners, all of whom performed the two conditions in the same order (first without the board, then with the board in place), were more accurate in localizing sources in the second condition in every spatial dimension. Further examination of the data showed that listener's accuracy improved over hours of practice in the first condition but was essentially unchanged during the second condition. No similar change were seen in the previous anechoic data (reanalyzed for these trends). These results imply that subjects "adapt" to a room over time, and that whatever the subjects learn transfers from one configuration (without a board) to another (with the board in place) that is very different, acoustically.

A follow-up study was recently conducted to explore how robust these effects are (Kopčo et al., 2001). We hypothesized that with practice in a room, subjects adapt and localization improves, and that this learning transfers from one listener configuration to another; i.e., that there is some "room specific" characteristics of reverberation common across all listener positions and orientations in the room. To examine these hypotheses, two groups of listeners performed a localization task similar to that in the initial experiment. Each listener performed four sessions of localization, each from one of the four configurations shown in Figure 3. The first group performed the sessions in the order indicated in Figure 3, starting in the center of the room (configuration 1) and ending in the corner of the room (configuration 4). The second group performed the sessions in the opposite order.

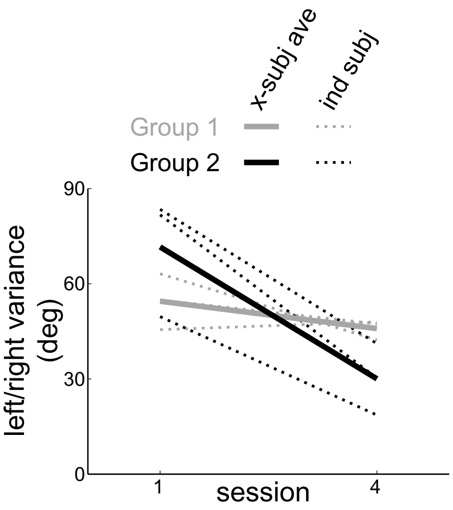


Figure 5: Response variability versus session. Solid lines show across-subject means. Dashed lines show individual subjects.

To the extent that room position affected localization accuracy, we hypothesized that performance would be best when listeners were in the center and worst when listeners were in the corner of the room. To the extent that practice in the room improved localization, performance should be better in the last session of the experiment and worst in the initial session, independent of room configuration order. If both factors influence localization accuracy, the second subject group should show the largest improvement from session one to session four, because both the acoustic and the learning effects would push the results in the same direction. In contrast, for the first subject group, who begin the experiment in the easiest acoustic setting (but without any prior experience in the room), the two effects would interact. In this case, insight into the relative importance of learning and room acoustics on localization performance could be gleaned by comparing results for the two groups.

Response variability in the left/right dimension is shown in Figure 5 for the two groups for the initial session (left) and the final session (right). Results show that the Group 2 subjects (for whom both learning and acoustic effects should cause performance to be best in session 4) show much larger changes in response variability between session 1 (the most acoustically-challenging, corner configuration) and session 4 (the room center configuration). The Group 1 subjects, who started in the easy room configuration and moved to the hardest room configuration, showed only a modest decrease in variability between sessions 1 and 4.

These results support the hypothesis that both learning and room acoustics influence localization accuracy. In addition, since the learning transfers across room configurations that are acoustically very different (and that lead to very different signals at the ears), the results suggest that with practice on the task, subjects learn some very general characteristic about the room reverberation that is similar for all room positions, independent of the exact structure of the echoes and reverberation interacting with the direct sound.

In another set of experiments (Shinn-Cunningham, Santarelli & Kopčo, 2000a), measured HRTFs were used to simulate anechoic and reverberant listening conditions under headphones. Subjects were asked to indicate the heard distance of the simulated sources for sources that were presented both binaurally and monaurally, for sources to the side (along the interaural axis) and to the front. Simulated source distances ranged from 15 cm to 1 m, the range in which ILD cues vary dramatically with distance for sources along the interaural axis. We expected to find that subjects could judge source distance accurately for binaural presentations of lateral sources because subjects in a real anechoic space have been shown to do relatively well on a similar task. Binaural and monaural presentations of reverberant simulations were used so that we could determine whether the reverberation cue for source distance arose from monaural effects (such as the direct-to-reverberant energy ratio; e.g., see Mershon & King, 1975; Bronkhorst et al., 1999) or binaural effects (such as the interaural decorrelation caused by reverberant energy, which is correlated with the direct-to-reverberant energy ratio).

Results of this study were compelling. In every anechoic condition, subject performance was near chance. In all reverberant conditions, subject performance was well above chance. Further, for lateral sources simulated with the reverberant HRTFs, monaural and binaural distance perception was essentially equal; binaural cues were irrelevant for the task. Interestingly, for medial sources, turning off one ear did affect distance judgments slightly, with subjects consistently overestimating the simulated source distance. However, we believe this bias arises because the simulated sources were heard in the wrong direction (i.e., along the interaural axis), where the pattern of reverberation varies differently with distance than it does for medial sources.

Results suggest that reverberation is an important distance cue. Even when sources are so close to the listener that there exist reliable ILD distance cues, these cues are ignored when listeners expect (are calibrated for) a reverberant listening environment. The cue for distance is probably correlated with the direct-to-reverberant energy ratio, although it is unlikely that the human auditory system can accurately compute such a ratio from the total signal reaching the ear. Further, the distance cue provided by reverberant energy is not a binaural cue, but a monaural cue; however, perceived direction (which is strongly influenced by binaural cues) affects perceived distance.

SUMMARY

Inclusion of realistic echoes and reverberation in virtual auditory environments will have a number of dramatic effects, including increasing the realism of the display (Begault, 1992b; Durlach, Rigapulos, Pang, Woods, Kulkarni, Colburn & Wenzel, 1992; Gilkey, Simpson & Weisenberger, 2001), improving distance perception (Shinn-Cunningham, 2000a), providing information about the room itself (Gilkey et al., 2001), and degrading directional accuracy, albeit slightly (Shinn-Cunningham, 2000b). Relatively little is known about which aspects of reverberation are most critical for each of these perceptual results. Further, it is likely that these different perceptual effects arise from different aspects of the reverberation. For instance, while our results hint that distance perception is driven more by monaural than binaural cues, impressions about room size depend on the amount of interaural decorrelation induced by echoes and reverberation, a binaural cue.

These results have a number of implications for the design of effective, efficient acoustic room simulators, pointing to the need to take into account how various aspects of reverberation influence perception. Further work is necessary to tease apart how reverberation influences various percepts important in virtual environments. More specifically, we must examine how accurately room reflection patterns must be simulated in a virtual environment to achieve accurate distance perception as well as realism (while some work addressed these issues, e.g., Begault, 1992a; Zahorik, Kistler & Wightman, 1994, much more work remains). The fact that, in a real reverberant room, listeners adapt their spatial percepts over time suggests that the human perceiver makes subtle perceptual calibrations in ways that we don't yet understand. In turn, this fact hints that listeners are perceptually sensitive to room acoustics in ways that must be explored and understood in order to develop room simulations that recreate what is important for the human perceiver.

ACKNOWLEDGEMENTS

Norbert Kopčo, Scott Santarelli, Tara Brown, and Scarlet Constant assisted in data collection and analysis. Portions of this work were supported by grants from the Air Force Office of Scientific Research, the Whitaker Foundation, and the Sloan Foundation.

REFERENCES

Begault, D. R. (1992a). "Binaural auralization and perceptual veridicalityy." Journal of the Audio Engineering Society, preprint 3421.

Begault, D. R. (1992b). "Perceptual effects of synthetic reverberation on three-dimensional audio systems." Journal of the Audio Engineering Society, **40**(11): 895-904.

Bronkhorst, A. W. and T. Houtgast (1999). "Auditory distance perception in rooms." Nature, **397**(11 February): 517-520.

Brown, T. J. (2000). Characterization of Acoustic Head-Related Transfer Functions for Nearby Sources. Electrical Engineering and Computer Science. Cambridge, MA, Massachusetts Institute of Technology.

Brungart, D. S. and N. I. Durlach (1999a). "Auditory localization of nearby sources II: Localization of a broadband source in the near field." Journal of the Acoustical Society of America, **106**(4): 1956-1968.

Brungart, D. S. and W. M. Rabinowitz (1999b). "Auditory localization of nearby sources I: Head-related transfer functions." Journal of the Acoustical Society of America, **106**(3): 1465-1479.

Carlile, S. (1996). Virtual Auditory Space: Generation and Applications. New York, RG Landes.

Durlach, N. I., A. Rigapulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn and E. M. Wenzel (1992). "On the externalization of auditory images." Presence, **1**: 251-257.

Gilkey, R., B. D. Simpson and J. M. Weisenberger (2001). Creating auditory presence. Human Computer Interaction, International, New Orleans, LA.

Kopčo, N. and B. G. Shinn-Cunningham (2001). Effect of listener location on localization cues and localization performance in a reverberant room. 24th MeetingAssoc Res Otolaryng, St. Petersburg Beach, FL.

Litovsky, R. Y., H. S. Colburn, W. A. Yost and S. J. Guzman (1999). "The precedence effect." Journal of the Acoustical Society of America, **106**(4): 1633-1654.

Mershon, D. H. and L. E. King (1975). "Intensity and reverberation as factors in auditory perception of egocentric distance." Perception and Psychophysics, **18**: 409-415.

Middlebrooks, J. C. (1997). Spectral shape cues for sound localization. Binaural and Spatial Hearing in Real and Virtual Environments. R. Gilkey and T. Anderson. New York, Erlbaum**:** 77-98.

Middlebrooks, J. C. and D. M. Green (1991). "Sound localization by human listeners." Annual Review of Psychology, **42**: 135-159.

Santarelli, S. (2000). Auditory Localization of Nearby Sources in Anechoic and Reverberant Environments. Cognitive and Neural Systems. Boston, MA, Boston University.

Santarelli, S., N. Kopčo and B. G. Shinn-Cunningham (1999). Localization of near-field sources in a reverberant room. 22nd MeetingAssoc Res Otolaryng, St. Petersburg Beach, FL.

Santarelli, S., N. Kopčo and B. G. Shinn-Cunningham (2000). "Distance judgements of nearby sources in a reverberant room: Effects of stimulus envelope." Journal of the Acoustical Society of America, **107**(5).

Shinn-Cunningham, B. G. (2000a). Distance cues for virtual auditory space. Proceedings of the IEEE-PCM 2000, Sydney, Australia.

Shinn-Cunningham, B. G. (2000b). Learning reverberation: Implications for spatial auditory displays. International Conference on Auditory Displays, Atlanta, GA.

Shinn-Cunningham, B. G., S. Santarelli and N. Kopčo (2000a). Distance perception of nearby sources in reverberant and anechoic listening conditions: Binaural vs. monaural cues. 23rd MeetingAssoc Res Otolaryng, St. Petersburg Beach, FL.

Shinn-Cunningham, B. G., S. Santarelli and N. Kopčo (2000b). "Tori of confusion: Binaural localization cues for sources within reach of a listener." Journal of the Acoustical Society of America, **107**(3): 1627-1636.

Wenzel, E. M. (1992). "Localization in virtual acoustic displays." Presence, **1**(1): 80-107.

Wenzel, E. M., M. Arruda, D. J. Kistler and F. L. Wightman (1993). "Localization using nonindividualized head-related transfer functions." Journal of the Acoustical Society of America, **94**: 111-123.

Wightman, F. L. and D. J. Kistler (1989a). "Headphone simulation of free-field listening. I. Stimulus synthesis." Journal of the Acoustical Society of America, **85**: 858-867.

Wightman, F. L. and D. J. Kistler (1989b). "Headphone simulation of free-field listening. II. Psychophysical validation." Journal of the Acoustical Society of America, **85**: 868-878.

Zahorik, P., D. J. Kistler and F. L. Wightman (1994). Defining and redefining limits on human performance in auditory spatial displays. Second Intl Conf Aud Display, Santa Fe, NM, Santa Fe Institute.

Investigation of multisensory spatial hearing: from the sense of audition to multisensory interactions

Klaus A J Riederer
Helsinki University of Technology, Laboratory of Computational Engineering, Cognitive Science & Technology
Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing
P.O. Box 9400, FIN-02015 HUT, Finland, E-mail: Klaus.Riederer@hut.fi, URL: http://www.lce.hut.fi/~kar

Background of author

MSc EE Klaus Riederer has worked as a research scientist on audio, acoustics and specifically on human spatial hearing for more than five years at the Helsinki University of Technology (HUT). He has a strong background of 10 years in professional level, and non-professional level of 15 years in technical sciences. He has gained a wide-scaled theoretical and practical know-how in various fields of engineering, mechanics, craftsmanship and photography. The emphasis of his academic research is on various aspects of human spatial hearing; theory, practice and analysis of acoustical head-related transfer function (HRTF) measurements. He has measured ca. 160 subjects' HRTFs (from 252 sound incidents) with the high-quality automatized measurement system, he has devised. He is the only person of this basic research field in Finland. He has published various papers on these issues and currently he is devising perceptual experiments. A long-time close collaboration is running with his former affiliation Laboratory of Acoustics and Audio Signal Processing (HUT), Brain Research Unit at Low Temperature Laboratory (HUT) and Unides Design Ay. (Helsinki, Finland). With the latter, groundbreaking hardware for binaural technology has been engineered.

Abstract

The study of human multisensory, especially audio-visual perception, has recently obtained increased attention with the development of virtual reality systems, teleconferencing, computer games and home theatre systems. An immersive sound scape can be created by three-dimensional sound applying head-related transfer functions that "model" human spatial hearing. Already from practice one realizes that the (spatial) hearing sense is truly multi-modal, applying also other senses, such as motion (moving the head), vision, tactile sensing etc. However, due to the technical difficulties and other complexity, serious efforts to investigate the sensory interactions (especially concerning spatial hearing) are still lacking. Therefore, a deep understanding of our second most important sense — hearing — is most incomplete. The author's research focuses on the true spatial hearing including interactions between other sensory modalities, such as audio-visual, audio-motional, audio-visual-motional. This basic research has strong interdisciplinary connections in various fields of science and numerous application areas. Novel research paradigms will be addressed under various objectives.

Keywords: 3-D sound, audio-visual perception, basic research, heterosensory, homosensory, HRTF, multidisciplinary, multi-modal perception, spatial hearing, virtual reality

Desires of author

The author is anticipating extensive discussions on the current status of other scholars' research activities, interests and future plans concerning the various issues on multi-modal perception and applications thereof. Various viewpoints are to be considered, ranging from technical (e.g., audio-visual synchrony), methodological (e.g., perceptual metrics, analysis methods) to psychological (e.g., cognition) aspects.

The author's main interest is in multidisciplinary basic research, focused on a) investigating accurately the human spatial hearing (applying HRTFs) and b) studying sensory interactions between spatial hearing and other sensory modalities, such as audio-visual, audio-motional, audio-tactile, audio-visual-motional and audio-visual-motional-tactile. The ultimate aim is to understand better the enigma of the human being, her/his behaviour and capabilities. To accomplish this, the author is seeking after skilled scholars with overlapping interests for possible collaboration, believing that elaborate methods, strong will and intelligent elucidation will lead to able scientific results.

Introduction

A natural immersive sound scape can be accomplished by three-dimensional sound applying so-called head-related transfer functions (HRTFs) that comprise the homosensory cues of spatial hearing. In reality, spatial hearing applies also other sensory modalities, a fact that is demonstrated by the following examples of cross-modal induction in perception. These matters set the fundamental framework for the author's research and are hence discussed below.

Homosensory cues of spatial hearing

The *primary localization* cues, the *binaural interaural level* and *time differences* (ILDs and ITDs) were formulated to the public already in 1907 by Lord Rayleigh, after collection of his previous hypothesis (1877) and Giovanni Venturi's founding research almost a century earlier. Localization performances a) on the median plane and b) monaurally (see, e.g., Hebrank and Wright 1974ab) prove that ILD and ITD cannot be the only localization cues. In headphone listening sound is often perceived inside the head to the middle (*inside-the-head-locatedness*), which is neither explained by the duplex theory (Yost and Hafter 1987). The necessary *spectral filtering* caused by the body, torso, head and especially the *pinnae* is denoted as the *monaural localization cue.* Obviously, binaural localization yields more precise accuracy than monaural (e.g., Blauert 1997). These cues apply only one sensory modality (hearing) and are thus *homosensory.* They are embodied in the *head-related transfer functions* (*HRTFs*) that involve measured (or modeled) responses of a sound source in free field to a point in the ear canal (Blauert 1997, Møller 1992). Spatial hearing perception based on HRTFs has been widely investigated in recent years (see, e.g., Wightman et. al 2001 and Møller et. al 2001).

In literature, a number of different spatial hearing models applying monosensory (acoustic) cues have been proposed: *physical, psychoacoustical (behavioral)* and *functional.* All the models need to generalize over inter-personal differences (and inaccuracies due to various reasons) in human anatomy and perception. Furthermore, strong enough experimental data is required to support the hypothesis.

Schelhammer presented a hypothesis of the sound gathering effect of the pinna already in 1684. Now, after three hundred years of investigation, there still remains research work to solve the puzzle of the homosensory spatial hearing. At least from the neurophysiogical point of view, research is still in an infant level. The reason for this is also obvious: the introduction of the modern brain imaging techniques and binaural techniques have finally made possible non-invasive (cortical) measurements, in which human binaural neural processing of natural three-dimensional sounds can be investigated.

Multi-modal interactions in spatial hearing

If the location of the auditory event does not unequivocally correlate with the sound signals at the eardrums, also further supplemental sensory information is needed (Blauert 1997). In practice this means that at least in ambiguous cases humans incorporate inter-sensory information for determining the locations and distances of sound sources. The *bone-conduction* theories are homosensory, other theories of spatial hearing are heterosensory. The latter involve interaction between audition and other modalities. They are called *motional, visual, vestibular* and *tactile* theories.

The so-called *motional* (or *motoric*) *theories* describe relationships between the position of the auditory event and the variations to the ear input signals during head movements (Blauert 1997). They also characterize changes in other attributes, such as loudness and tone color, which the subject can utilize in sound localization. Humans (and animals) tend to move naturally their heads; often only a slight head movement will remarkably improve the localization accuracy.

It is also evident that *vision* has a powerful effect on spatial hearing, usually seeing the sound source improves the localization. Auditory localization is rather poor compared to vision under everyday conditions, typical errors of localization are 4-10° for the horizontal plane and much worse for elevations. In the sharpest detection area, i.e., the forward horizontal plane, a minimum angular separation between two (pure-tone) sound sources as small as 1° can be detected (Blauert 1997). According to Begault (1999), in multi-modal interaction experiments (e.g., audio-visual) the absolute accuracy of a particular modality should be regarded secondary. Instead, the evaluations have to focus on the overall "quality" of the perception, where positions produced by two modalities are judged relative to one another.

Vestibular theories present the organ of balance would have some direct effect on the spatial hearing, besides the obvious indirect influence in strong accelerations etc. These hypotheses, as well as *tactile* and *vibro-tactile* (e.g., "feeling the (live) music") *theories*, present that interaction between modalities can make a particular component of the modality in question either more or less noticeable. Formerly, these approaches have not been considered meaningful in regard to spatial hearing. However, this attitude has chanced as the research on multimodal interaction and perception has gained more interest. For example, force feedback is regarded an important technology development for virtual reality applications (Begault 1999).

Cross-modal induction in human perception

In certain conditions (temporal or spatial asynchrony or contradictory stimulae etc.) cross-modal induction is produced in human perception. This non-typical operation, e.g., modal override or fusion in sensory processing, presents useful starting points for the research of human perception and, e.g., cognitive systems. In the following, such cases of audio-visual perception are presented. These topics are the basic problem fields of *auditory scene analysis* (*ASA*), which is a method to understand brain and auditory system processing of complex sound environments.

In some cases vision overrides the auditory cues in (spatial) hearing; e.g., when watching television the sound seems to come from the screen, though it actually comes from the loudspeakers nearby. This so-called *ventriloquism effect* is defined as the spatially biased perception of the auditory stimulus from the same point as the visual stimulus (Shinn-Cunningham *et al.* 1997). Vision does not only reinforce the spatial auditory perception, especially in equivocal directions, but it also gives a great improvement in distance judgments of sound sources. However, vision can also diminish an auditory event. This can happen, e.g., in a concert hall that gives to the listener an inconsistent image between the auditory and visual space. Once the person closes her/his eyes, the music seems to sound better, because vision gives no disturbing information. Also, steering attention to one modality may improve one's performance, e.g., one can concentrate better to music with eyes closed. Furthermore, Paulsen and Ewertsen (1966) report a so-called *audio-visual reflex* as an involuntary turning of the eyeballs to front the direction of the sound source. This reflex requires at least some level of awareness of the direction of the sound source.

Many audio-visual psychophysical studies examine the influence of visual stimulation to auditory detection and vice versa. Welch and Warren (1986) present in their review that it is easier to detect auditory signal with the presence of a visual stimulus. In overall, the sensitivity to audio-visual events is increased under multi-modal conditions. The effect of *perceptual defense* states that presenting emotionally reserved auditory stimuli (i.e., words) raises the threshold of visual perception, and vice versa (Hardy and Legge 1968). These results have later been found controversial because of strong differences between subjects. In any case, the theory has proven to been beneficial in revealing pathological cases, such as various syndromes.

Audio-visual interaction on speech intelligibility is well-known from the *McGurk effect*, where conflicting audio-visual cues affect the intelligibility, and can create a fusion response that differs from both the auditory and visual stimulus (McGurk and McDonald 1976).

The *cocktail-party effect* is known as the ability to focus listening attention to on a single talker amidst a cacophony of conversations and background noise (Cherry 1953). The ability still exists when listening to high quality binaural recordings. Regardless of the wide-scale research, the underlying explanation for this effect is still not clear. Apparently, it is linked to human speech production system, auditory system and/or high-level perceptual and language processing systems.

Research paradigms

The basis for the author's research is the employment of the HRTF data that he has measured since the year 1997. He has set the following research paradigms that have not been solved in full by other scholars:

- What are the features that constitute the idiosyncratic features in individual HRTFs?
    - What are the roles of clothes, hair, hairstyle and headgear?
    - What is the effect of anatomical attributes, such as cranial and pinnae size?

  - To what extent are these idiosyncratic features of HRTFs necessary for reproducing perceptually accurate 3-D sound?
  - What are the errors and consequences in using non-individual HRTFs for perceptual experiments, compared to the use of individual HRTFs?
- Is it possible to find a (more) generic HRTF model (based on the previous)?
  - Would such a model be good enough for scientific experiments, i.e.,
    What are the errors and consequences in using such a generic HRTF model for perceptual experiments, compared to the use of individual HRTFs?
- What is the accuracy of auditory perception considering the whole 3-D space?
  - How does the perception change in the presence of distracters in other sensory modalities?
  - How does the perception change in the presence of supporting information from other modalities?
- What are the neurophysiological (cortical) findings related to the paradigms above, and what do they reveal about human multimodal information processing?

These paradigms will be addressed in the research objectives discussed below. The objectives are highly inter-related, and the weight is put to basic research — to understand how the human spatial hearing really works in its all complexity.

Objective I: Analysis of HRTF quality

The investigation of the *fundamental HRTF data quality* (Riederer 1998a, Riederer and Karjalainen 1998, Riederer 2000) is vital, because only this way the basis for the whole spatial hearing research can be confirmed. Repeatability investigations demonstrate the (high) *measurement system quality* (Riederer 1998b, Riederer 2000). Non-quantitative characteristics of HRTFs, based on the individual anatomy, are investigated by *structural HRTF analysis.*

Objective II: Quantitative analysis of HRTFs

Quantitative matters are strongly interlinked to Objective I, allowing direct utilization of methods and results between Objective I & II. The latter aims to the most enchanting research result on spatial hearing ever: a *generic HRTF model.* Such a model would give a *deeper universal understanding of spatial hearing*: to comprehend in detail (e.g., as a function of azimuth and elevation angle, *person-independently*) how the basic binaural cues are constituted. It is most obvious that there would not exist a *single pair of ears* ("*golden HRTFs*") but perhaps independent groups with common (idiosyncratic) features (e.g., "big heads/ears", "small heads/ears"). The Objective II concentrates on various issues around *HRTF classification* (Riederer 2000) and *distance-dependence HRTF analysis* (Riederer 1998a, Huopaniemi and Riederer 1998).

Objective III: Conversion methods

Objective III concentrates on digital signal processing issues, in order to make possible the empirical verification to the results of Objectives I and II, when applying binaural recordings and multichannel recordings (e.g., Dolby Digital program material) as test stimuli. The focus is to produce natural three-dimensional sound listening experiences by headphones for perceptual studies discussed in Objectives IV & V. This necessitates the implementation of *equalization* (free-field, diffuse field) and *conversion methods*, e.g., multichannel–binaural and individual binaural–generic binaural.

Objective IV: Psychophysiological studies

Careful psychophysiological experiments have been planned in order to address the research paradigms stated earlier. Differing from all the published research, a large amount of directions will be covered. This will reveal the true capability of human spatial hearing *covering the whole 3-D auditory space.* In order to make this possible, novel methods have been devised. Basically, virtual sources (sounds created via HRTFs, reproduced by headphones) are compared to the reference sources (loudspeakers at fixed positions). The subject is sitting on a rotating turntable; thus a great number of sound incidents are efficiently investigated. *Auditory, visual* and *motional, uni-, bi-,* and *tri-modal interactions* are to be investigated.

Objective V: Neurophysiological studies

Collaborative research (since 1997) continues with the Low Temperature Laboratory (LTL), Brain Research Unit. Individual and non-individual HRTFs will be applied for the investigation of the binaural neural processing with the 306-channel MEG instrument, utilizing the custom-made high quality tube headphones. The special interest is in comparing the metrics in the cortical representations of the three-dimensional auditory space in front-back, left-right and horizontal planes. Also multimodal experiments applying both EEG and MEG instruments will be performed.

Postscriptum

The theoretical and empirical results of the HRTF investigation by the author will be utilized in the audio-visual speech perception research at the Cognitive Science and Technology group, Laboratory of Computational Engineering (HUT). The author is also at the Finnish Graduate School of Electronics, Telecommunication and Automatization, and its financial support is greatly acknowledged.

The author's general research topics of the author involve experimental and theoretical analysis of hearing, vision, attention, motor control and haptic perception. The behavioral studies will touch many of the matters described in Introduction, including evaluation of reaction time, (a)synchrony, modal override and fusion, errors in perception etc. Also fundamental and applied exploration on virtual reality and virtual environments could are considered. Ultimately, a more comprehensive model of the human spatial hearing would be postulated.

References

Begault D. R., 1999. Auditory and non-auditory factors that potentially influence virtual acoustic imagery. *Proceedings of 16th Audio Engineering Society conference*, Rovaniemi, April 10-12, pp. 1-14.

Blauert J., 1997. *Spatial Hearing. The Psychophysics of Human Sound Localization.* Revised Edition, MIT Press, Cambridge, Massachusetts, USA, 494 p.

Cherry E. C., 1953. Some experiments on the recognition of speech with one and with two ears. *J. Acoust. Soc. Amer.*, vol. 25, pp. 975-979.

Hardy G. R. and Legge D., 1968. Cross-modal induction of changes in sensory thresholds. *Quartely Journal of Experimental Psychology*, vol. 20, pp. 20-29.

Hebrank, J., and Wright, D. 1974a. Are two ears necessary for localization of sound sources on the median plane? *J. Acoust. Soc. Am.*, vol. 56, no. 3, pp. 935–938.

Hebrank, J., and Wright, D. 1974b. Spectral cues used in the localization of sound sources on the median plane. *J. Acoust. Soc. Am.*, vol. 56, no. 6, pp. 1829–1834.

Huopaniemi J. and Riederer K. A. J., 1998. Measuring and modeling the effect of source distance in head-related transfer functions. In *ICA/ASA* 1998 conference, Seattle, USA, 20-26.6.1998.

McGurk H. and McDonald J., 1976. Hearing lips and seeing voices. *Nature*, 264, 746-748.

Møller H. et. al, 2001. Department of Acoustics, Aalborg University, Denmark. Publications list at http://acoustics.auc.dk/publications/pubframe.html

Møller, H. 1992. Fundamentals of binaural technology. *Applied Acoustics*, vol. 36, pp. 171–218.

Paulsen J. and Ewertsen H. W., 1966. Audio-visual reflex. *Acta Oto-laryngol. Suppl.*, vol. 224, pp. 217-221.

Rayleigh Lord (Strutt J. W.) (Ed.), 1907. On our perception of sound direction. *Philosophical magazine*, vol. 13, pp. 214–232. Cited in Blauert (1997).

Riederer K. A. J and Karjalainen M., 1998. DSP aspects of head-related transfer function measurements. In *IEEE Nordic Signal Processing Symposium, NORSIG'98*, Vigsø Holiday Resort, Denmark, 8-11.6.1998.

Riederer K. A. J, 1998a. *Head-related transfer function measurements.* Master's Thesis. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Espoo, Finland, 134 p.

Riederer K. A. J, 1998b. Repeatability analysis of head-related transfer function measurements, *105th Audio Engineering Society Convention*, San Francisco, Sept. 26-29. Preprint no. 4846, 62 p.

Riederer K. A. J, 2000. Computational Quality Assessment of HRTFs In *European Signal Processing Conference EUSIPCO (X)*, 5-8.9.2000, Tampere, Finland.

Schelhammer G. C., 1684. De auditu, liber unus. Lugduni Batavorum. Cited in Békésy G. von, 1960. *Experiments in hearing.* McGraw-Hill, New York, USA. Cited in Blauert (1997).

Shinn-Cunningham B., Lehnert H., Kramer G., Wenzel E. and Durlach N., 1994. Auditory displays, pp. 611-663. In Gilkey R. and Anderson T. (Eds.) (1997), *Binaural and Spatial Hearing in Real and Virtual Environments.* Lawrence Erlbaum Associates, New Jersey, USA, 795 p.

Welch R. B., and Warren D. H., 1986. Intersensory Interactions. In *Handbook of Perception and Human Performance,* K. R. Boff, L. Kaufman, and J. P. Thomas (Eds.), Ch. 25. New York, Wiley.

Wightman F. L. et. al, 2001. Hearing Development Research Laboratory, Waisman Center, University of Wisconsin. List of spatial hearing publications at http://www.waisman.wisc.edu/hdrl/index.html

Yost W. A and Hafter E. R., 1987. Lateralization. In Yost W. A. and Gourevitch G. (Eds.) *Directional Hearing.* Springer-Verlag, New York, pp. 49-84.

Perceptual and Statistical Models for Virtual Audio Environments

Jean-Marc Jot
Creative Advanced Technology Center.
1500 Green Hills road. Scotts Valley, CA 95067.
jmj@atc.creative.com

Abstract

Perceptual and statistical models provide alternatives to the physical/geometrical models for describing and rendering virtual audio environments. They are most relevant in applications where *plausibility* of the virtual world is sufficient (as opposed to physical accuracy), or in applications involving the creation of *imaginary* acoustic worlds (as opposed to realistic virtual worlds). Having reviewed the possible types of applications for virtual audio environments and their requirements, we compare three approaches to acoustic scene modeling (*geometrical*, *perceptual* and *statistical*) and describe an authoring tool proposed for environmental audio authoring in interactive video games. We conclude with general considerations and proposals regarding the standardization of a low-level environmental audio rendering API and of higher-level acoustic scene description models for interactive audio.

1. Overview and Classification of Targeted Applications

Virtual acoustic rendering technology has its origins in research carried out in the 1970's, which targeted two distinct applications:

Architectural acoustics. Schroeder et al. developed simulation methods based on geometrical acoustics (specular reflection model, ray tracing), to derive a computed echogram from the geometry and absorption properties of room boundaries and the position of the source and listener [1].

Computer music. Chowning developed a system for simulating dynamic movements of sounds in a virtual room, based solely on perceptual control parameters [2]. It allowed independent control of three positional parameters for each source: apparent distance to the listener, apparent direction of sound arrival, and Doppler shift. Chowning's system used an artificial reverberation algorithm (from Schroeder) to provide parametric control of the reverberation (intensity and decay time). Later, Moore proposed a more sophisticated system in which distance, Doppler and reverberation parameters were simulated and controlled via geometrical acoustic models as used for architectural acoustics [3].

Today, we can see a continuous spectrum of potential applications for 3D audio and virtual environment simulation technology, including: architectural acoustics, simulation, training, games, telepresence, multimedia installations, movie/video soundtracks, computer music. These various applications can be classified along an axis ranging from *"realistic"* virtual worlds to *"imaginary"* virtual worlds, and including *"augmented reality"* applications.

2. Requirements and Challenges

General requirements of virtual acoustic technology are listed below (some of these requirements may be optional depending on the specific application). Some additional specific requirements and challenges are then reviewed for two opposite application scenarios: realistic virtual worlds to imaginary virtual worlds.

General Requirements for Virtual Acoustic Systems

- *Interactivity* requires real-time rendering/mixing of multiple audio events/streams (sound sources).

- *Signal processing resources* will be limited in any practical system, and will vary across platforms (from specialized DSP hardware to "software fallback" using the general-purpose processor of a typical PC).

- The *positional representation* of sound events should be independent from the multi-channel playback format or the 3D audio positioning/panning technique used by the rendering system. It should therefore use absolute or head-relative 3D coordinates (in a cartesian or polar system).

- An *acoustic scene description* model must be provided, covering both positional and environmental parameters, for creating the virtual scene and driving the rendering or simulation system.

  We can distinguish:

  - *program-driven applications*, in which this description is provided in the form of an API (Application Programming Interface), such as DirectX, Java3D or OpenAL.

  - *description-driven applications*, in which the scene description takes the form of metadata (as in MPEG-4 Advanced Audio BIFS).

- *Authoring or design tools* must be provided for creating and building virtual acoustic worlds.

- *Standardization* of the scene description models is beneficial in order to enable:

  - *platform-independent playback* of content/scenes

  - *re-usability* of scene elements by authors and designers.

Realistic Virtual Acoustic Worlds

We define *realistic virtual worlds* as applications in which the acoustic environment is specified via a physical/geometrical and/or graphical representation of the world. The typical example is found in architectural acoustics, where the purpose of the simulation is the auditory evaluation of the acoustics of a room or building on the sole basis of architectural data. Other examples can be found in virtual reality applications or video games.

In terms of their requirements, realistic virtual worlds involve the following decisions.

- *Accuracy vs. plausibility*

  Accuracy is required when the exact simulation of an acoustic reality is the main goal. This is particularly true in architectural acoustics, whose purpose is an assessment of acoustical quality (the visual information, if presented, is merely used to support that assessment). In virtual reality applications, audio information will generally not be more important than visual information, and the latter will often preempt in the perceived scene when audio and visual cues are not consistent.

On the other hand, in certain virtual reality applications and particularly video games, the virtual audio environment is not a simulation of any existing environment, even though it may be specified geometrically and physically. Therefore, the requirement becomes *plausibility*, i. e. providing sufficiently valid and perceptible audio cues to support and complement the visual information, and contribute to the "suspension of disbelief".

- *Level of detail*

  The notion of "level of detail" is relevant in the context of acoustic rendering as in 3D computer graphics. In general, it will refer to simplifications that can be applied at the rendering stage in order to optimize the use of computing resources. When acoustic accuracy is required, any simplification in the rendering model must be applied with care on the basis of proven psycho-acoustical knowledge. Otherwise, a simplification may precisely discard an effect that is the object of the assessment. If plausibility is sufficient, simplifications that have a perceptible effect can be allowed, as long as they can only be noticed by comparing with the non-simplified model.

Imaginary Acoustic Worlds

By *imaginary acoustic worlds*, we refer to applications where the audio scene is specified explicitly and directly via a description model that does not restrict authoring or creation possibilities. The typical example is music (including interactive music or soundscapes in games). Accuracy is not an issue, since there is no physical reference to compare the virtual audio scene to. However, plausibility (or *naturalness*) is required, in the sense that no audio effects or artifacts should be noticed (other than are intended by the author).

The requirements for description and rendering models in imaginary acoustic worlds may include the following:

- Allow for the creation of virtual acoustic worlds that are not derived from physically and geometrically realistic representations. The audio scene may be impossible to recreate physically (although it may still need to be *plausible*).

- Allow for the composition of audio scenes prescribed purely by the intended auditory sensation (music).

- Retaining some perceptually salient geometry-based relationships between the audio experience and a graphical/geometrical representation (video games).

  Examples include:

  - the navigation of the listener between rooms having different reverberation properties;

  - muffling (occlusion/obstruction) of sounds by obstacles and partitions;

  - sources heard through openings to adjacent rooms;

  - effects related to the position of listener or sources relative to room walls (per-source control of early reflections);

  - acoustic reflections on distant obstacles (open environments).

3. Scene Description and Rendering Models

In this section, we briefly review and compare three existing scene representation models, which illustrate three different acoustical models: geometrical, perceptual and statistical. The first two are taken from AABIFS

(Advanced Audio Binary Interface For Scene description) in the MPEG-4 standard, version 2 [4]. The third example is EAX, an environmental audio API developed by Creative Labs [5] and partially covered by I3DL2, a vendor-neutral 3D audio rendering guideline published by the IA-SIG [6].

MPEG-4 AABIFS Physical Approach

The physical approach in MPEG-4 Advanced Audio BIFS is derived from the DIVA system developed at the Helsinki University of Technology [7]. The reflected sound is modeled as a combination of discrete individually spatialized reflections followed by an exponentially decaying reverberation tail.

The scene description model includes:

- a sound source model, providing:
    - a frequency-dependent, axissymmetric directivity model (a set of parametric filter models associated to a set of radiation angles);
    - an adjustable speed of sound parameter affecting Doppler effects and propagation delays;
- a room model (called "Acoustic Scene"), characterized by:
    - late reverberation parameters (decay time, level, delay);
    - a set of polygons (each having reflectivity and transmissivity filters defined in the same manner as the source directivity filters);
    - a rectangular bounding box (used for detecting the presence of sources or the listener in a room).

In the context of imaginary acoustic worlds, such as musical or "audio-only" applications, the inconvenient of this physical approach is that audio effects are dependent on geometrical and physical parameters [8]. It is not possible to override or modify the reflection and reverberation parameters or the muffling effects that are derived automatically according to the locations of the sound sources and the listener at runtime. These geometrical effects are eliminated if no acoustically active polygons are included in the scene. However, the acoustic response of the room is then reduced to a simple late reverberation tail whose parameterization is not sufficient to satisfy musical applications.

MPEG-4 AABIFS Perceptual Approach

The perceptual approach in MPEG-4 Advanced Audio BIFS is derived from the Spatialisateur system developed by IRCAM and France Telecom [8]. The room reverberation response is divided into three temporal sections: a group of directional early reflections (R_1) coming from an angular sector centered on the direction of the sound source, a group of diffuse early reflections (R_2) and the exponentially decaying late reverberation (R_3).

The scene description model includes:

- a sound source model (identical to the sound source model used in the physical approach);
- a set of environmental parameters associated to each individual sound source, comprising:
    - the time limits, the cross-over frequencies and the modal density defining the reverberation response model and its division into three temporal sections;
    - a set of nine *perceptual parameters*, which determine the energy levels in the three temporal sections and the late reverberation decay time (in three frequency bands);

  - a reference distance.

This environmental model enables detailed tuning of environmental effects for each source, without relying on geometrical or physical environment data. It addresses a music playback scenario in which the following simultaneous processes can be combined:

- a "reverberation preset" is applied to each individual sound source (the reverberation parameters can vary in time according to a predefined "score", or can, optionally, be manually adjusted by the user at playback time);
- the value of one of the nine perceptual parameters, called *source presence*, is automatically adjusted as the relative source-listener distance varies, according to the combination of:
  - trajectories predefined in the "score" (in absolute world coordinates),
  - the movements of the listener in the virtual world (optionally),
  - manual actions of the listener to modify source positions during playback (optionally).

However, because this model does not take into account any world geometry information, it provides no means for automatically applying muffling effects or adjusting reflection and reverberation parameters according to the positions of sources and the listener relative to walls and obstacles.

A Statistical Model: EAX

Unlike the two previous models, EAX is a primarily a low-level environmental rendering API (implemented in the form of extensions to existing 3D positional audio APIs: OpenAL and Microsoft's DirectSound) [5]. However, it also includes optional higher-level functions, which are based on a statistical model.

By exposing low-level rendering parameter in the scene description, one ensures that the model is not biased towards a particular type of interactive audio application, and can address realistic virtual worlds as well as imaginary acoustic scenes. The only limitations lie in the rendering model itself, which must be complete enough to cover all the perceptually relevant effects.

Currently (EAX 3.0), the low-level rendering parameters include:

- *Low-level reverberation parameters*. The room response model is decomposed as a group of early reflections followed by an exponentially decaying reverberation tail and can be parameterized as follows:
  - Basic reverberation parameters (EAX 2.0, I3DL2): initial delay and level of the reflections and of the reverberation, high-frequency attenuation, decay time at medium and high frequencies, modal density and echo density (or "diffusion").
  - Advanced reverberation parameters (EAX 3.0): low-frequency level and decay time, directional panning of the reflections and of the reverberation, periodic echo with adjustable salience and period, periodic pitch modulation with adjustable salience and period (for special effects).
- *Low-level source parameters*. For each source, the level of the direct-path sound and of the reflected sound can be controlled separately at low and high frequencies.

In order to facilitate the task of application developers and sound designers, the API provides several optional higher-level functions, all based on a statistical model of acoustic propagation in rooms.

- *Distance and directivity models*. The direct-path attenuation the and reflected-path attenuation at low and high frequencies can be automatically adjusted according to the directivity of the source (at low and high frequencies), the source-listener distance, the reverberation decay time and the air absorption coefficient [9]. These adjustments combine additively (on a decibel scale) with the low-level source parameters. As in the perceptual approach described earlier, the reverberation parameters define a "reverberation preset" corresponding to a reference source-listener distance.

- *Muffling effects (obstruction, occlusion, exclusion)*. These effects provide an attenuation and an increasing low-pass effect simultaneously via a single command, and also combine additively on a decibel scale with the above source parameters. *Occlusion* affects both the direct-path and the reflected path sound, whereas *obstruction* and *exclusion* affect only the direct path and only the reflected path.

- *Environment size control*. This function provides a simultaneous adjustment of all the low-level reverberation parameters in order to simulate a relative scaling of the room dimensions.

Future extensions of the EAX API, necessary for covering the main effects required in a wide range of applications, include the following functions.

- *Multiple environments*. This extension requires multiple artificial reverberators running in parallel and allows each source to feed one or several of them. The simulation of acoustical coupling between rooms also implies that an output signal from one reverberator can be fed into other reverberators.

- *Per-source control of early reflections*. This extension involves means for a source signal to feed several virtual sound sources with different delays and directions. It also includes an efficient processing architecture allowing each sound source to provide a multi-channel feed to a reverberator, so that the level, delay and direction of early reflections can be controlled separately and independent from late reverberation parameters.

4. Content Creation / Authoring: A Case Study

In this section, we introduce EAGLE (an authoring tool for creating virtual acoustic environments in video games), together with a high-level geometrical API and audio engine, EAX Manager (which enables runtime mapping from the geometrical world representation to the low-level audio rendering API parameters) [5].

EAGLE (Environmental Audio Graphical Librarian Editor) is designed to address the needs of sound designers in interactive audio applications and facilitates their collaboration with the application programmers. It provides the following functions.

- Sound design:
    - design environment models (reverberation presets);
    - design source models (reference distance, volume balance, directivity parameters, filters, Doppler effects);
    - design obstacle models (transmission properties used to drive occlusion effects);
    - tune diffraction model (used to drive obstruction effects).
- Mapping audio parameters to world geometry:
    - import a 3D world map provided by the application programmer;
    - partition the world into "environments" (rooms) and associate a reverberation preset to each environment;

- associate obstacle models to room walls;
- identify diffracting obstacles (used to drive obstruction effects).

EAGLE allows the sound designer to save all the above data in a single file, which is included by the programmer in the compilation of the game application. By linking the EAX Manager library with the application, the programmer can use a high-level geometrical API to interrogate a runtime geometrical engine. This engine returns recommended values for environment parameters, and for the occlusion and obstruction parameters settings associated to a given source, according to the position of the listener and sources relative to walls and obstacles.

The EAX Manager geometrical and physical engine performs the following functions:

- identify the environments in which sources and listeners are located;
- retrieve the appropriate obstacle model to control occlusion effects when source and listener are not located in the same room;
- detect diffracting obstacles located between a source and the listener and compute obstruction parameter settings according to the diffraction path.

Conclusion

Standardizing a low-level rendering API

In order to ensure cross-platform playback of interactive virtual audio environments (“write once, run everywhere”), there is a clear benefit in defining a standard low-level rendering model that is “agnostic” in terms of the type of application (from realistic worlds to imaginary worlds), and scalable in terms of the computational resources and audio playback system required at the rendering stage. An example of such a model is the OpenAL API, completed with environmental audio extensions [5].

Higher-level scene description models

In order to facilitate and promote the creation and re-usability of content and applications, there is value in developing standard higher-level scene description models associated with runtime rendering engines for mapping high-level descriptions to low-level API parameters. Different types of applications may call for different higher-level models, while being satisfied by a common low-level rendering API. Two examples of high-level models are the *physical* and *perceptual* parameter sets specified in the MPEG-4 v.2 Advanced Audio BIFS standard for environmental audio spatialization.

In addition to these two approaches, we have described a high-level *statistical* scene representation model whose purpose is to provide plausible reproduction of the key perceptually relevant environmental audio effects, and allow for tuning and exaggerating these effects (source distance and directivity, Doppler effects, occlusion effects…). Like the perceptual approach, the statistical approach circumvents any direct reference to the geometry and acoustical properties of walls or obstacles, and thus enables the creation of imaginary or musical soundscapes without physically based constraints. Like the geometrical approach, the statistical approach is based solely on general physical laws of room acoustics [9] and thus provides a plausible high-level acoustic model in a wide range of situations.

The low-level rendering API can be enhanced with optional higher-level environmental parameters exposing statistical or perceptual models (to allow for the automatic rendering of distance effects, for instance). This is beneficial as long as the effect of these parameters is designed to be *additive* with respect to the effect of the low-level rendering parameters. Geometrical environment parameters such as wall and obstacle positions should be left out of the low-level rendering API layer, because a geometrical acoustic propagation model cannot be combined in a additive manner with low-level rendering parameters. A geometrical model will typically *override* the low-level parameter values and is therefore best exposed via a higher-level API layer that is used only in applications where the audio scene must be driven by a physical and geometrical representation of the virtual world.

Description-driven applications vs. program-driven applications

Although description-driven applications do not differ from program-driven applications in the requirements they place on the low-level rendering API, they have different needs with regard to high-level scene description models.

In program-driven applications, such as video games, the main benefit offered by a higher-level scene description API is a reduction of the programming effort for the application developer, obtained by exposing complex operations via simple commands. For instance, the intensity and spectral parameters of the direct path, reflections and reverberation will be automatically adjusted according to the source-listener distance and several other parameters.

However, in description-driven (or metadata-driven) virtual audio environments, such as enabled by the MPEG-4 standard, some high-level environmental effects *must* be supported. They are required in order to enable user interaction within the scene at run time, where audio effects that depends on the position of the listener in the virtual world are rendered. This includes, at a minimum, a distance model to account for source-to-listener distances within a room. Furthermore, in order to account for the muffling effects of intervening obstacles or dynamic changes of reverberation or reflection parameters according to the navigation of the listener in the virtual world, a geometrical and physical scene description is necessary (such as the physical approach in the MPEG-4 description model).

References

[1] Schroeder, M. R. (1973). Computer models for concert hall acoustics. *American Journal of Physics*, Vol. 41, pp. 461-471.

[2] Chowning, J. (1971). The simulation of moving sound sources. *J. Audio Eng. Soc.*, Vol. 19, no. 1.

[3] Moore, F. R. (1983). A general model for spatial processing of sounds. *Computer Music J.*, Vol. 7, no. 6.

[4] Scheirer & al. (1999). *Multimedia Systems*, Vol. 7, no. 1.

[5] Creative Labs (2001). SDK Downloads (EAX, OpenAL, EAGLE). http://developer.creative.com/

[6] IA-SIG (1999). 3D Audio Rendering Guidelines, Level 2 (I3DL2). http://www.iasig.org/

[7] Savioja & al. (1999). *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675-705.

[8] Jot (1999). *Multimedia Systems*, Vol. 7, no. 1.

[9] Jot, J.-M., Cerveau, L., Warusfel., O. (1997). Analysis and synthesis of room reverberation based on a statistical time-frequency model. *Proc. 103rd Conv. of the Audio Eng. Soc* (preprint no. 4629).

Acoustic rendering with wave field synthesis

Marinus M. Boone
Lab. of Acoustical Imaging and Sound Control
Lorentzweg 1, 2628 CJ Delft, The Netherlands
e-mail: rinus@akst.tn.tudelft.nl

Abstract

Wave Field Synthesis is a method of sound reproduction, based on the precise construction of the desired wave field by secondary sources, implemented as arrays of loudspeakers. The method is derived from acoustic wave field theory, and implemented in a practical approach based on physical as well as perceptual laws. An overview is given of the theory, the implementation as a Laboratory Demonstration System and some interesting applications of the method. Presently, the method is further developed in an IST-project of the European Commission, called CARROUSO.

1 Introduction

Looking back in history, we see that the development of spatial sound reproduction started already at an early stage [1]. Originally, the stereophonic reproduction principle was not restricted to two channels. However, it was found that the effect of adding more than two channels did not produce so much better results that it would justify the additional technical and economical efforts. This was especially the case at a time when it was very difficult and expensive to develop a medium for the simultaneous recording of many channels. Besides that, the spatial effects that could be obtained with only two channels made it straightforward that spatial sound recording and reproduction focused on the well known two-channel stereophony.

In the 70's of the last century, efforts were taken to enlarge the spatial impact with the so-called quadraphony. However, the results that were obtained did not convince the public sufficiently and the development was stopped.

More recently, a new surround standard has been adopted, known as the 5-channel surround system. This system has been mainly developed for cinema use, but finds also application in video home theaters and audio-only reproduction.

Results that have been obtained with these systems range from excellent to poor, depending on the recorded material and the way of reproduction. This is strongly related to the physical properties of the reproduced sound field and the psychoacoustic effects that can be reached with a limited number of reproduction channels. The main difficulty with these reproduction principles is, that they strongly rely on the psychoacoustic effect of so-called phantom sources, i.e. one hears a sound source between two loudspeakers at a position, depending on amplitude and time differences between the loudspeaker signals. The apparent position of these phantom-sources is strongly dependent on the position of the listener. As such, these principles are not well suited for reproduction in larger listening areas. It is well-known that for two-channel stereophony one often has only one good listening position in a room. The wrong perceptual position of phantom sources is especially distracting in combination with visual reproduction. Reproduction from the middle front is therefore stabilized with a special dialog channel. When a signal is reproduced by one single loudspeaker, the location of the sound is stable for the whole audience. The reason for that is, that the directions of the primary wave fronts are correct for all listening positions. With phantom reproduction this is not so and a spatially correct reproduction for a large listening area can never be obtained in this way.

The idea that the correct curvatures of the wave fronts should be reproduced forms the backbone of Wave Field Synthesis (WFS). The idea goes back to the so-called Huygens principle,

stating that a wave front can be thought to be originating from many secondary sources. WFS is also based on secondary sources. In 1953, Snow published an overview of stereophonic techniques [2] and discussed the acoustic curtain as being the ideal stereophonic reproduction technique. Here we already see a resemblance with Wave Field Synthesis.
In the late eighties, WFS was introduced by Berkhout; see e.g. Berkhout [3] and Berkhout et al. [4]. The intuitive acoustic curtain concept is replaced here by a well funded wave theory.
The method attracted interest from several research institutes and industries. This resulted in an IST (Informatiion Society Technologies) project of the European Commission, called CARROUSO. The key objective of the project CARROUSO (for Creating, Assessing and Rendering in Real-time Of high-quality aUdio-viSual envirOnments in MPEG-4 context) is to provide a new technology enabling to transfer a sound field, generated at a certain real or virtual space, to another, remotely located, space. This will be possible with full interactive control of perceptually relevant temporal, spatial and perceptual properties of the sound space, especially in combination with the transmission of visual data. CARROUSO will merge a flexible and powerful coding technology such as the new MPEG-4 standard, allowing object-oriented and interactive sound manipulation. In this projectthe Wave Field Synthesis rendering technique plays an essential role, which makes it possible to produce a virtual sonic space, not its impression.

In this paper we will first give an overview of the underlying theory of WFS. Next we will describe the Laboratory Demonstration System, that we developed, based on this theory. An important part of this paper is also the discussion of techniques that can be used in practice to make recordings and obtain special reproduction effects with WFS for different applications.

2 Theory

It is known from general linear acoustic theory that an arbitrary sound field within a closed (fictive) volume can be generated with a distribution of monopole and dipole sources on the surface of this volume. The only restriction is that there are no acoustic sources within this volume. This can be expressed with the so-called Kirchhoff-Helmholtz integral, given by [5]:

$$P(\mathbf{r},\omega)=\frac{1}{4\pi}\iint_S\left[P(\mathbf{r}_S,\omega)\frac{\partial}{\partial n}\left(\frac{e^{-jk|\mathbf{r}-\mathbf{r}_S|}}{|\mathbf{r}-\mathbf{r}_S|}\right)-\frac{\partial P(\mathbf{r}_S,\omega)}{\partial n}\frac{e^{-jk|\mathbf{r}-\mathbf{r}_S|}}{|\mathbf{r}-\mathbf{r}_S|}\right]dS \tag{1}$$

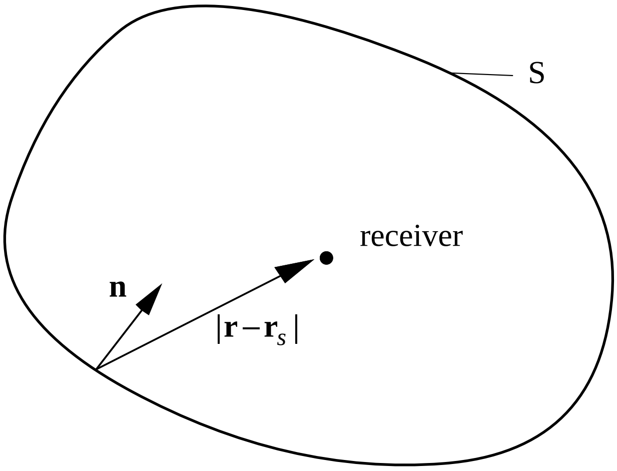


Figure 1: Geometry for the Kirchhoff-Helmholtz integral formulation of Eq. (1).

The geometry is shown in figure 1. *S* is the surface of the volume, **r** is the coordinate vector of an observation point, $\mathbf{r}_s$ is the coordinate vector of the integrand functions on *S*. The sound pressure in the Fourier domain is given by $P(\mathbf{r},\omega)$ and *k* is the wave number *ω/c*.
In this expression the first term represents a distribution of dipoles that have a source strength, given by the sound pressure of the sound field at the surface and the second term represents a distribution of monopoles that have a source strength given by the normal velocity of the sound field (which is proportional to $\partial P/\partial n$).
This theoretical result predicts that we can recreate an arbitrary sound field by making a recording of $p(t)$ and $\mathbf{v}_n(t)$ over some surface *S* during the actual musical performance and then reproduce the recording over a similar surface in a reproduction room with the help of a large number of monopole and dipole sources, fed by the recorded signals. Care must be taken that a sufficient number of reproduction sources is used to omit so-called spatial aliasing. For an exact reproduction of all propagating waves, thereby neglecting near field effects, the spacing of the loudspeakers must be less than half of the shortest wavelength of the reproduced sound. It will be evident without further proof that such a registration and reproduction system is far from realistic.
For practical purposes, this method has been adapted to make use of linear loudspeaker arrays surrounding the listening area, rather than planes of loudspeakers. This has several consequences:

1. With a straight line solution only monopoles are needed, no dipoles;
2. Reproduction is only correct for wave field components in the horizontal plane;
3. Because of the line reproduction the amplitudes of the reproduction are not correct over the whole listening area.

It can be shown [6] that for linear arrays the input signals of the loudspeakers are given by

$$E_i(\omega) = K\sqrt{jk}V_n(\mathbf{r}_i,\omega) \tag{2}$$

where $V_n(\mathbf{r}_i,\omega)$ equals the normal component of the particle velocity, virtually at the loudspeaker position $\mathbf{r}_i$, *k* is the wave number and *K* is a constant depending on the loudspeaker sensitivity, the distance between the loudspeakers and the desired sound pressure of the reproduction. In case of loudspeakers with a flat frequency response, *K* is frequency independent.
In practice the secondary loudspeakers will not behave as true monopoles. However, the synthesis operator can be corrected for that [7]. It was also found by experiment that spatial aliasing due to the finite loudspeaker spacing is not critical for high frequencies. In practice, a loudspeaker spacing of 0.125 m gives perceptually correct results. The wave fronts are then physically correct up to 1360 Hz.

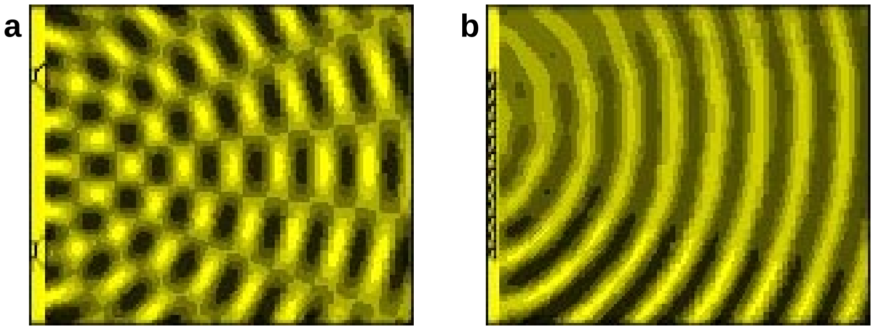


Figure 2: a) monochromatic source signal reproduced by two loudspeakers; b) monochromatic signal of a point source, reproduced by a loudspeaker array according to the WFS concept.

In figure 2 it is shown how, in contrast to solutions with conventional solutions with a few individual loudspeakers, in WFS the array loudspeakers together create a spatial replica of the original sound field.

Figure 3 shows that these virtual sources can be placed at will behind or in front of these loudspeaker arrays. For the latter situation, there is an equivalence with a focused image in optics. The concept of focused virtual sources is very important, because it enables a way of sound reproduction that cannot be reached with conventional means.

Notice that the virtual sources can also be placed at such a large distance, that plane waves result in the listening area.

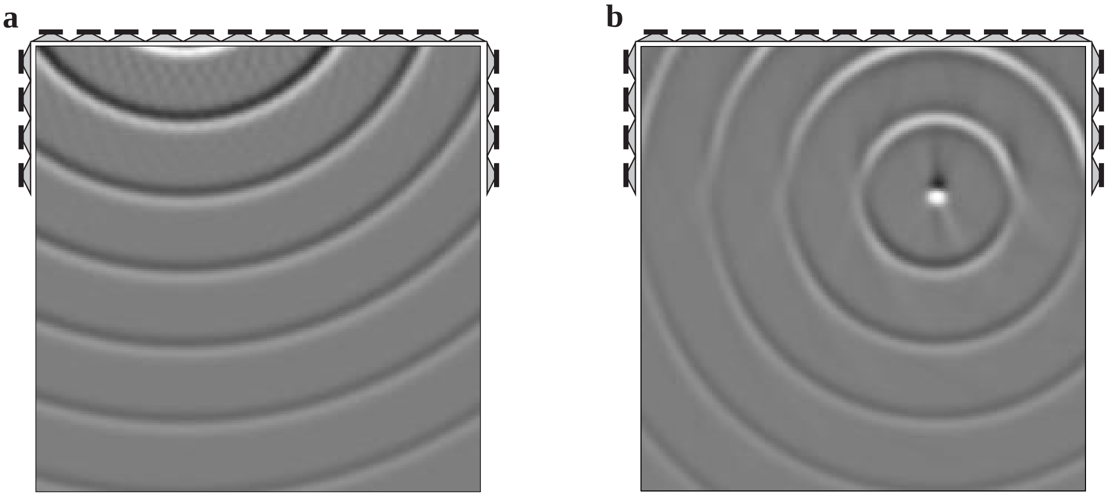


Figure 3: With WFS not only the sound of sources outside the listeners area (a), but also within that area (b) can be reproduced.

3 Laboratory Demonstration system

At Delft University, a WFS system has been implemented as a Laboratory Demonstration System, as shown in figures 4 and 5. The real time signal processing is carried out with a DSP-system that was specially built for this purpose. It consists of 8 Texas Instruments TMS320C32 processors. At the front-end there are 8 analog/digital audio inputs and at the rear-end 128 digital/analog outputs. The outputs feed a total of 160 loudspeakers that are fitted in front of the 4 walls of the studio, such that an effective listeners area of approximately 15 m^2 exists. Recently, video equipment has been added for research on WFS in multimedia applications such as video conference systems, cinemas and home theatres.

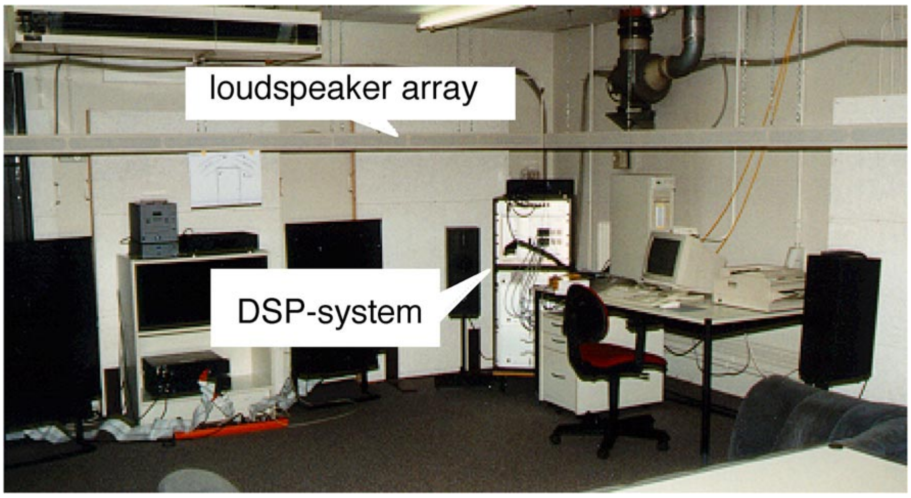


Figure 4: Survey of the WFS Laboratory Demonstration System at Delft University.

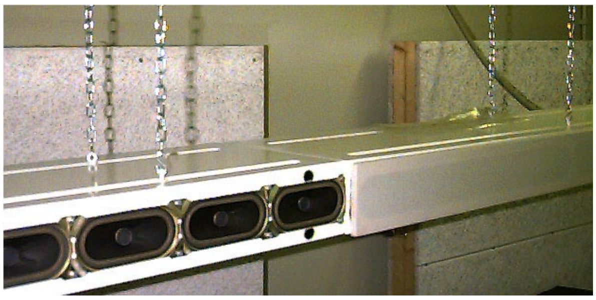


Figure 5: Detail of a loudspeaker array.

Another WFS system has been built, with 16 inputs and 96 outputs, where the outputs are fed to 192 loudspeakers. These loudspeakers are optimized to be used for direct sound enhancement (to be discussed later) in large auditoria.
The basic principle of the implemented facilities is that each input signal can be processed in such a way, that the outputs of this particular signal to the loudspeaker arrays form together the sound field of a virtual source somewhere in space.

4 Recording and reproduction techniques

In developing practical recording and reproduction techniques, experience has been gained with several methods and the quality was evaluated objectively and subjectively [8], [9]. It was found that it makes sense to distinguish between the direct sound, the early reflections and the reverberation. For that reason it is necessary to use a recording technique that can also distinguish between these components of the sound field. From a principle point of view, the best way of recording can be obtained with microphone arrays [10]. For practical reasons we decided to develop methods that require much less microphones.

4.1 Recording and reproduction of the direct sound

An approach which was found to be very practical is by recording the direct sound field of the different sound sources with spot microphones. These signals can then directly be processed as so-called virtual sources, which can be placed at each desired position. It should be realized

that the illusion of a source in front of the array is only obtained in the listening area of the convex waves and not at the other side. Hence, the usable listening area is reduced.
When the number of sound sources is large, such as in orchestral music, it will not be possible to record each sound source separately. It will then be needed to record groups of instruments as one so-called notional source, that is reproduced as a distinct virtual source. Alternatively, the whole or part of the orchestra is recorded with a conventional stereo recording technique with a left and right signal. With the WFS system these signals can be reproduced as plane waves from different directions. In that way a kind of mix can be obtained between true virtual sources (for soloists) and an adapted stereophonic reproduction technique (for the orchestra).

4.2 Recording and reproduction of early reflections and reverberation

When the original recording is made in for instance a concert hall with good acoustics, we also want to reproduce the reflections and reverberation from that hall. We have gained experience with a procedure where we make recordings of the early reflections and the reverberation with directional microphones pointing to different directions. We experimented with different set-ups, including cardioid microphones, but also a SoundField microphone, from which a number of different directional patterns can be output simultaneously. These recordings are reproduced as plane waves from the corresponding directions [9]. These techniques are also applicable to some extent to conventional surround sound recording and playback.

4.3 Artificial generation of reflections and reverberation

This method is appropriate when the original reflections and reverberation cannot be used, as is the case in many studio situations. In the same way that reflections and reverberation are added in conventional reproduction techniques with artificial reverberation devices, we can do the same with wave field synthesis. The best results are obtained by generating the early reflections as virtual sources, using image source theory for the reflections. The later reflections will effectively act as plane waves from different directions, making up a diffuse reverberant sound field. It was found by research at our laboratory that reverberation can be reproduced with WFS as plane waves from different directions in a very satisfactory way [14], [12]. Optimal results are obtained with reproduction from 10 different directions, but even quite satisfactory results are obtained with only 4 directions.

4.4 Compatible reproduction

It is also possible to reproduce conventional recordings with the WFS-system. This has the benefit that the reproduction is less dependent on the listener position than with normal loudspeaker set-ups. For instance 2-channel stereophonic material can quite well be reproduced with two plane waves at angles of ± 30 degrees with the front direction. If a listener moves from the middle to the left, the image still shifts to the left, because the left signal will reach the listener’s ears earlier. There are several ways to compensate for that. One way is to add a mono-mix at a virtual central source position that is reproduced a bit earlier in time at a low level. This stabilizes the stereo image for a large listening area. This procedure is also often applied with normal stereophony in theaters by adding a central speaker above the stage that is fed with the advanced mono signal. Another approach is to implement a directivity in the reproduced waves such that the travel time differences are compensated by intensity differences [11]. However, we found that this technique is limited in practice to only a small listening area.
Modern surround material is mixed according to one of the discrete multi-channel standards. For instance the 5.1 discrete surround method can very well be reproduced in WFS by 4 plane

waves and the dialog channel as a virtual source in the front center. A benefit is that the left-right and front-back imaging is much more stable for different listening positions with WFS than with discrete loudspeakers [12].

5 Applications

From the experience that has been gained at our laboratory we are now able to show a number of applications for which WFS can be of use to create a better spatial sound field than is possible with conventional means.

5.1 Direct sound enhancement in theaters

By reinforcing the voices of speakers or singers in shows, musicals, opera's etc., WFS offers a correct localization for all listeners in the audience.

5.2 Cinema's

The WFS-principle is ideal for cinema applications, to obtain high quality spatial sound over a large listening area. The system must be able to work under different conditions, such that conventional stereo or surround material can be enhanced as discussed in section 4, but the best results will be obtained with special cinema productions, where the sound processing is optimized for WFS. Special attraction for the audience will be reached with dynamically moving sound sources, for which a special processing algorithm has been developed [13]. The moving sound illusion, which is based on a dynamic temporal interpolation technique, includes the Doppler effects that accompany fast moving sound sources.

5.3 Home theaters

The sweet-spot drawback of conventional stereophony and surround sound is fully eliminated by WFS if special recorded audio programs are made available. Special loudspeaker arrays, preferably with integrated digital signal routing, converters and amplifiers will be needed to find acceptance by the consumer market. Combination with wide screen television makes it possible to have a true WFS cinema at the home.

5.4 Virtual reality theaters

For this application the WFS method can give even more impact than for the cinema. Especially 3D video projection, where the visual sources enter the audience area, will be enhanced when combined with sound sources, focused at the same position. For these applications special productions, optimized for WFS reproduction, should be developed.

5.5 Simulators

The sound quality of for instance flight simulators can be greatly enhanced with a WFS sound simulator. Because of the complexity of modern airplanes, vehicles and large vessels, high quality simulation of the real environment, including the sound field, is getting more and more important.

5.6 Auralization

The presently available auralization systems that are used to judge the quality of (future) concert halls and theaters, are based on binaural reproduction with the aid of HRTF's (Head Related Transfer Functions). Binaural reproduction with headphones gives very often so-called IHL (In Head Localization). A much better illusion can be obtained with the WFS principle, by making use of the methods discussed in section 4, especially in 4.3. For the auralization of existing halls a method has been developed where first multi-channel impulse responses of the hall are acquired along two perpendicular arrays. The measurements comprise pressure as well as particle velocity measurements. From these impulse responses directional impulse responses are calculated that apply to the loudspeaker positions of a WFS-system. These loudspeakers are fed with these impulse responses, convolved with dry recorded audio signals to produce a very realistic acoustic impression of the hall. For a more thorough description of the method, the underlying theory and the related signal processing, the reader is referred to [14].

5.7 Teleconference systems

One of the formats used with teleconferencing is based on the use of large video walls to connect two meeting rooms in a virtual way. In these applications it is known that as soon as people start to talk together it is very difficult to concentrate on the desired speech signal. It is well known that binaural cues help to concentrate on the desired signal (Cocktail Party effect). Therefore, a better spatial quality of the sound is also of great importance here. Besides that, WFS can generate a natural virtual acoustics, giving the illusion of “being there”. For teleconference systems it may not be necessary to surround the whole listening area with loudspeakers, if only the spatial impression from a projection screen needs to be correct.

6 New developments

Presently, research is carried out for new loudspeaker technology that is optimized for application in WFS-systems. A promising development is based on the concept of distributed Mode Loudspeakers (DML’s). Another development is related to the signal processing, for which special DSP-configurations are needed. Such DSP’s can very well be integrated into the housing of the loudspeaker arrays. It is expected that the first commercial applications will be found in special surround sound applications in large screen teleconferencing and large exhibitions. Next, applications in cinema’s can be expected and eventually the method might find its way into domestic applications such as home theaters. Of course the development in the CARROUSO project that started at the beginning of the year 2001, will play an important role, especially by making special WFS demonstrators available.

7 Conclusions

In this paper it has been shown that, starting with the physical laws of acoustic wave fields, in combination with perceptual knowledge and experience, the principle of Wave Field Synthesis has been worked out for different kinds of audio applications. These applications have in common that a high quality spatial impression is required that is valid over a large listening area. It has been shown by the use of demonstrations with a Laboratory Demonstration system how WFS can be applied for applications ranging from compatible surround playback to auralization. A next step will be the implementation of WFS in affordable systems for different applications. From then on, practical applications are only limited by our imagination.

8 References

1 J. C. Steinberg and W. B. Snow, “Auditory Perspective—Physical Factors,” AIEE, vol. 53, pp. 12–15, 1934.

2 W. B. Snow, “Basic Principles of Stereophonic Sound,” Journal of the SMPTE, vol. 61, pp. 567 – 589, 1953.

3 A. J. Berkhout, “A Holographic Approach to Acoustic Control,” J. Audio Eng. Soc., vol. 36, pp. 977-995, 1988.

4 A. J. Berkhout, D. de Vries, and P. Vogel, “Acoustic Control by Wave Field Synthesis,” J. Acoust. Soc. Am., vol. 93, pp. 2764–2778, 1993.

5 A. J. Berkhout, Applied Seismic Wave Theory. Amsterdam: Elsevier, 1987.

6 M. M. Boone and E. N. G. Verheijen, “Multi-channel sound reproduction based on wave field synthesis,” presented at 95th Convention of the AES, New York, 1993, Preprint 3719.

7 D. de Vries, “Sound reinforcement by wavefield synthesis: adaptation of the synthesis operator to the loudspeaker directivity characteristics”, J. Audio Eng. Soc., vol. 44, p. 1120, 1996.

8 M. M. Boone and E. N. G. Verheijen, “Qualification of Sound Generated by Wave Field Synthesis For Audio Reproduction,” presented at 102nd Convention of the AES, Munich, 1997, Preprint 4457.

9 W. P. J. de Bruijn, T. Piccolo, and M. M. Boone, “Sound recording techniques for wave field synthesis and other multichannel souns sytems,” presented at 104th Convention of the AES, Amsterdam, 1998.

10 A. J. Berkhout, “A wavefield approach to multichannel sound,” presented at 104th Convention of the AES, Amsterdam, 1998.

11 R. M. Aarts, “Enlarging the Sweet Spot for Stereophony by Time/Intensity Trading,” presented at 94th Convention of the AES, 1993, Preprint 3473.

12 M.M. Boone, W.P.J. de Bruijn and U. Horbach, “Virtual Surround Speakers with Wave Field Synthesis”, 106th Convention of the AES, 8-11 May 1999, Munich, preprint 4928, 12 pages.

13 E. N. G. Verheijen, Sound Reproduction by Wave Field Synthesis, PhD Thesis, TU Delft, 1998.

14 A. J. Berkhout, D. de Vries and J.J.Sonke (1997): Array technology for acoustic wave field analysis in enclosures, J.Acoust.Soc.Am., vol.102, pp. 2757-2770.

Acoustic rendering using loudspeaker arrays

Ulrich Horbach (ulrich.horbach@studer.ch)
Studer Prof. Audio AG
http://www.studer.ch

Studer Prof. Audio, Switzerland, is developing and offering 3D-audio immersive rendering technologies since a couple of years. Among these are a binaural processor with multichannel input, using measured room impulse responses and a headtracker [1], and software plug-ins for our flagship digital mixing console *D950*, containing room simulation tools and distance pan-pots [2].

Recently, we have started a European project, together with 10 partners (among others Univ. of Delft, Univ. of Erlangen, IRCAM Paris, France Telecom, Fraunhofer Gesellschaft), called *Carrouso* [3], which I would like to introduce at the “campfire“. The project aims at building a complete reproduction chain for 3D audio environments, based on array technologies and MPEG-4 encoded information channels. Soundfields are captured and rendered by arrays of microphones and loudspeakers, respectively. Goal is to achieve a sweet-spot independend, extended listening area, with sound sources positioned within a room (for example between the listener and the loudspeaker array). The rendering algorithm is a combination of well-known wave field synthesis methods [4], and inverse multichannel filtering [5], in order to control the acoustics of the listening room. We will perform a couple of controlled listening tests, in order to optimize the system in terms of cost-performance, and to gain more knowledge about the psychoacoustics of spatial hearing in this context.

For the transmission path, we pursue different options. In the first option, the recorded multichannel signals will be transmitted directly, after having passed a data-reduction step. Second, the acoustics of the recording room will be pre-measured by a microphone array and sent once over the channel, a so-called data-based rendering method [6]. We are also working on a third option, a processor which generates the appropriate multichannel impulse response sets in real-time, controlled by perceptual parameters.

Possible applications are wide-spread: from multimedia workstations (8-16 channels), small installations in private households (up to 100 channels), to large-scale systems for electronic cinemas or theatres (several hundred channels).

In the campfire, I would like to learn more about possible applications, other concepts and solutions, and come in contact with leading experts in the field.

References

[1] U. Horbach, A. Karamustafaoglu, R. Pellegrini, P. Mackensen, G. Theile: Design and Applications of a Data-based Auralisation System for Surround Sound. 106. AES Convention, Munich 1999, preprint 4976.

[2] Ulrich Horbach, Attila Karamustafaoglu, Etienne Corteel, Renato Pellegrini: Implementation of an Auralization Scheme in a Digital Mixing Console using Perceptual Parameters, 108. AES Convention, preprint 5099, Paris 2000.

[3] http://emt.iis.fhg.de/projects/carrouso/index.html

[4] Ulrich Horbach, Marinus M. Boone: Future transmission and rendering formats for multichannel sound. Proc. of the 16. AES Conference, Rovaniemi, Finland, April 1999.

[5] U. Horbach, A. Karamustafaoglu, R. Rabenstein, G. Runze, P. Steffen: Numerical Simulation of Wave Fields Created by Loudspeaker Arrays. 107. AES Conv., New York 1999, preprint 5021.

[6] Ulrich Horbach, Attila Karamustafaoglu, Marinus M. Boone: Practical Implementation of a Data-based Wave Field Reproduction System. 108. AES Convention, preprint 5098, Paris 2000.

Highly efficient methods for binaural 3D audio rendering

Jiashu Chen (jiashuchen@agere.com)
Agere Systems
http://www.agere.com

Head-related transfer functions (HRTFs) have been used in synthesizing 3D audio to place virtual sound source in desired locations for binaural presentation. It has been proven that HRTF filtering to the source, coupled with conventional binaural cues of ITD and IID, can greatly improve the externalization and positioning in 3D space, compared with simple panning in stereophony. However audio processing involved with HRTF filtering demands more computing power. In synthesizing complex acoustic scenes, in which multiple independent sources and multiple reflections are considered, the computing resource required can be overwhelming. In this presentation a filter bank is introduced to provide a very efficient algorithm to address this problem. This filter bank, derived from SFER model of measured HRTFs, replaces multiple HRTFs in implementing multiple source 3D positioning. Supporting computing architecture that handles ITD, distance introduced delay, and memory management are also discussed. Demonstrations will be given at the presentation.

Individualization of head-related transfer functions by spectral warping

Véronique Larcher (veronique.larcher@genesis.ac)
GENESIS
http://www.genesis.ac

Head-Related Transfer Functions (HRTF) vary upon individuals due to the specificity of one's head/ear/torso shapes.The use of non individual HRTF for binaural synthesis entails several perceptual artefacts, namely the increase of Inside-the-head localization and the rate of front-back confusion. They can be significantly reduced by scaling in frequency the non individual HRTF to best fit the listener's ones. This method, that was proposed by Middlebrooks ([1], [2]), consists in translating one head's spectra on the log-frequency axis to match the features of the second head's HRTF.

In this presentation, an extension of Middlebrooks's global scaling approach is described. In order to automatically set the scaling factors, correlation to the dimension of morphological features is studied. It is also shown that a satisfying estimate can be obtained that relies on a limited number of HRTF. They correspond to the "principal directions" that were given by a statistical analysis of HRTF as decribed in [3]. Finally, practical issues raised by a real-time implementation of the frequency scaling are discussed. The first issue is the choice for a satisfying "scalable" head, i.e. for the head that shows the best morphing properties to be scaled into the others. Several implementations of the digital filters that enable the proposed spectral warping method are also proposed and compared.

Individual adaptation of binaural synthesis has proved to enhance localization accuracy. So do head-tracking and early reflections rendering. It is poorly known however which of these factors should be taken into account first and which are second order factors only. Vision cues seem to solve most of the ambiguities raised by auditory localization. Does it make the preceeding factors useless in the audio-visual case ? Using these different tricks, how accurate/plausible can (needs) the sound rendering be achieved ?

This leads to the evaluation of immersive systems quality. Several questionnaires exist to investigate the "sense of presence" conveyed by a given installation. However, no real agreement exist on a test procedure to rate these systems. Should we target precise localization of sound sources ? Localization stability ? Envelopment ? Naturalness ? Other criteria ?

These are issues we would like to discuss during the campfire.

[1] J. Middlebrooks. Individual differences in external-ear transfer functions reduced by scaling in frequency. J. Acoust. Soc. Am., 106(3):1480-1492, 1999.
[2] J. Middlebrooks. Virtual localization improved by scaling non-individualized external-ear functions in frequency. J. Acoust. Soc. Am., 106(3):1493-1509, 1999.
[3] V. Larcher, J.-M. Jot, J. Guyard, and O. Warusfel. Study and comparison of efficient methods for 3D audio spatialization based on linear decomposition of HRTF data. Presented at the 108th convention of the Audio Eng. Soc. in Paris. Preprint #5097(E1), 2000.

ACN 051 198 273

Sonic Landscapes by Lake Technology

Prototype Virtual Augmented Audio Reality System

Authors: Nigel Helyer, Patrick Flanagan, Stephen Bennett, David McGrath.

- Sonic Landscapes is an advanced research and development project conducted between Artist Nigel Helyer and Lake Technology Limited, with technical assistance from the University of New South Wales Satellite Navigation and Positioning Group. The project is part funded by the Australian federal government (via the New Media Fund of the Australia Council for the Arts). Sonic Landscapes aims to demonstrate the concept of augmented acoustic reality (AAR), in which a three dimensional acoustic simulation is overlaid on real physical space.

- Traditionally Virtual Reality has been dominated by visual considerations. This project attempts to extend the concept of virtual space to the auditory domain. The goal is to demonstrate a system that allows a user to wander at will in physical space and simultaneously experience a three dimensional acoustic simulation that is overlaid on the environment. The acoustic simulation consists of a number of interactive sound objects which correspond to the location of objects in the 'real world' and which appear to respond to the users movements and position, thus creating the impression that they are real and inhabit physical space.

- The concept of Virtual Audio Reality is a novel way to 're-think' cyberspace, by escaping the 'perspective' constraints established by visual screen based systems. Sonic

Landscapes the user moves through physical terrain to navigate a virtual soundscape which is composed within a digitally mapped space.

- The 'Sonic Landscapes' technology has been prototyped as part of a soundscape installation created by Dr Nigel Helyer. The installation overlays an acoustic simulation onto the Gothic graveyard in Newtown, Sydney. This site is an ideal context as it combines multiple layers of physical/sculptural interest with a high level of content - as well as providing a quiet pedestrian environment. The prototype system integrated GPS positioning, headtracking, and spatial audio in a portable backpack.

- The applications of this technology range from dedicated museum and tourist guide systems to multipurpose mobile audio interfaces. The prototype is used to demonstrate the feasibility and usefulness of an audio only augmented reality system.

The LISTEN Vision

Gerhard Eckel (eckel@gmd.de)
GMD - German National Research Center for Information Technology
http://viswiz.gmd.de/~eckel

We report about the LISTEN project, a research project funded by the European Commission in the context of the Information Society Technology (IST) program. LISTEN, which started in January 2001, will provide users with intuitive access to personalized and situated audio information spaces while they naturally explore everyday environments. A new form of multi-sensory content is proposed to enhance the sensual impact of a broad spectrum of applications ranging from art installations to entertainment events. This is achieved by augmenting the physical environment through a dynamic soundscape, which users experience over motion-tracked wireless headphones. Immersive audio-augmented environments are created by combining high-definition spatial audio rendering technology with advanced user modeling methods. These allow for adapting the content to the users' individual spatial behavior. The project will produce several prototypes and a virtual-reality-based authoring tool. Technological innovations will be validated under laboratory conditions whilst the prototypes will be evaluated in public exhibitions.

Objectives

Intuitive access to information in everyday environments is becoming a central concern of new information society technologies. An important question is how established and well functioning everyday environments can be enhanced rather than replaced by virtual environments. Augmented or enhanced reality technologies address this issue but have concentrated so far on the visual sense and have mainly been used in industrial applications. Auditory augmentation of visually dominated everyday environments (such as exhibition spaces) is a new and very promising approach in creating user-friendly information systems, which are accessible to everybody. The complementarity between the visual and auditory sense is the basis for a new type of multi-sensory content, which will become feasible thanks to anticipated advances in auditory rendering, wireless tracking, and communication techniques in the context of this project.

LISTEN proposes a new type of information system for intuitive navigation of visually dominated exhibition spaces. Visitors are immersed in a dynamic virtual auditory scene that consistently augments the real space they are exploring. They wear motion-tracked wireless headphones for 3D spatial reproduction of the virtual auditory scene. A sophisticated auditory rendering process takes into account the current position and orientation of the visitor's head in order to seamlessly integrate the virtual scene with the real one. Speech, music and sound effects are dynamically arranged to form an individualized and situated soundscape offering exhibit-related information as well as creating context-specific atmospheres. The dynamic composition of the soundscape is personalized through each visitor's spatial behavior, the history of the visit, and interests or preferences either expressed explicitly by the visitor or inferred from the visitor's behavior. The proposed system is targeted at all kinds of exhibition applications ranging from art exhibitions to industrial fairs. Curators, artists, composers and sound designers will assist in the design of the system and help to shape this new form of multi-sensory content.

The evaluation of immersive audio-augmented environments will be carried out with virtual and physical prototypes. The virtual prototypes will be realized with an audio-visual surround-view display system using state-of-the-art virtual environment technology. The physical prototypes will be installed at 2 different sites. Out of the many possible applications of immersive audio-augmented environments, the validation will concentrate on museum applications. These are considered to be the most demanding in terms of perceptual quality, openness, flexibility and user-friendliness. The virtual prototypes will be used to develop different scenarios. Advanced audio guides will be developed showing the potential of the new form of content for pedagogical applications. In order to push the system to its limits, artistic applications will be realized as well in form of virtual prototypes. These prototypes will then be used to attract an internationally recognized artist who will be commissioned to realize the content for the main physical prototype. This physical prototype will be installed at a distinguished museum of modern art, and will be made accessible during several months in a public exhibition.

Innovation

Innovation in the LISTEN project is located on two levels: the conceptual/artistic and the scientific/technological one. On the conceptual level, LISTEN proposes a new form of interaction between information and people. The strength of the approach lies in its simplicity and intuitiveness experienced by the end users: people just put on discreet wireless headphones and explore physical space by walking about. The space they explore becomes the actual interface to the information, which is presented in a virtual auditory space consistently augmenting the real space. In order to achieve the necessary consistency of augmentation, the LISTEN project proposes a set of innovations on the scientific/technological level, which are necessary to implement the conceptual vision of the project. Apart from the innovations in the areas of wireless motion tracking, binaural rendering, user modeling and adaptation, virtual prototyping and authoring, the main technological innovation of the project lies in the integration of new generation technologies.

Concept

The idea of individual auditory augmentation of exhibition spaces is almost half a century old by now. Early audio guides used taped explanations about the artwork displayed in an exhibition. Visitors were wearing headphones and had to carry the playback device. They were also constrained to follow a particular path through the exhibition. This path was defined by the sequence of explanations stored on the tape, which allowed for linear access only. Apart from controlling the playback level, pausing the playback and eventually rewinding the tape, visitors could not interact with the presentation. Nowadays, audio guides use random access audio storage technologies (e.g. RAM, CD, MD, or CDROM based) allowing the visitors to enter exhibit specific codes to recall corresponding audio presentations. This simple and pragmatic solution is used by nearly all audio guide services currently offered by almost all big museums. Other types of audio guides were developed when wireless headphone technology became available in the eighties. Induction and infrared-based techniques were used to create zones around exhibits where visitors could hear audio clips repeated in loops. The main drawback of this presentation technique is that visitors will typically not arrive at the start of a loop and therefore hear the end before the beginning. The timing of the audio clips is not individually controllable but shared by all visitors. The main advantage of the wireless technology is that the users don’t need to carry any playback device and that they can interact with the information by naturally walking through the exhibition spaces. They create their individual soundscape by freely moving from one zone to the another – and it is this feature which is of central importance to the LISTEN project.

Audio Guide

With LISTEN we generalize the audio guide concept by conceiving an adaptive and personalized spatial audio information system. This generalization is motivated by the conviction that the time of auditory user interfaces, especially in form of audio-augmented environments, has finally arrived. What prevented the realization of immersive audio-augmented environments in the past, was the lack of the most important requirement for advanced auditory interfaces: the availability of a spatial audio technology refined enough to make full use of the human sense of spatial hearing. Such technology (i.e. affordable wireless wide-area high-definition motion tracking combined with advanced binaural rendering and wireless digital audio transmission) only becomes feasible now and will be developed further in the context of this project. With this new technology all the features known from traditional audio guide systems can be emulated and - more importantly - a revolutionary set of new features for the design of interactive soundscapes is created.

The key idea of the LISTEN concept is to place the notion of space – of visual, auditory and imaginary space and their relationships – at the center of the design. By moving through real space, users automatically navigate an attached acoustic information space designed as a complement or extension of the real space. Virtual acoustic landmarks will play an equally important role than the visual ones for the orientation of the users in this augmented environment. Acoustic labels can be attached to visual objects. The particularities of auditory and visual memory can be combined to create new forms of non-linear audio-visual narratives. Objects can acoustically address the visitors when he or she passed them, thus providing exhibit-related information and calling for attention. Objects not in the field of vision can gain the attention of visitors through localized acoustic cues. Spatial regions can be provided with particular acoustic ambiences creating atmospheres and contexts for the visual perception of objects. Music and sound effects can be used to create an individualized sound track along the freely chosen path through an exhibition. The concrete visual space may be overlaid with an abstract auditory space, which proposes an alternative spatial structure. This could be realized by permeable "acoustic walls", which invisibly separate zones in a visually continuous space. Along these borders room acoustic signatures could change, thus creating different acoustic spaces in one visual space. Spatial perception and navigation is one of the best-developed abilities of human beings and is therefore one of the most solid grounds an intuitive human-machine interface could be based on.

Immersive Audio-Augmented Environment

With the immersive audio-augmented environment, LISTEN defines a new format of interactive audio content. Rather than a predetermined, pre-recorded audio program, listeners are offered a personalized audio environment, based on their interaction with the real space. The enhanced audio format can provide deepening layers of information, giving increasing levels of involvement. It will allow the visitors to find their own level of engagement with an exhibition. The depth of experience may vary giving each person the chance to find his or her own level or area of comfort and interest. These adaptive features of the LISTEN system are based on advanced user modeling methods, which allow extracting certain preferences from the user’s spatial behavior. User modeling also allows avoiding redundancies in the presentation of audio information. The user model will keep track of each user’s visit history and adapt the presented information with respect to what the user has already experienced. This will avoid repetition of information where it is not explicitly desired by the user and communicated to the system with a simple remote control unit. By these means, LISTEN provides enhanced, interactive sound tailored to the interests and experiences of the individual visitor and to a variety of exhibit types. These will range from art exhibitions in museums to gallery installations and from scientific conferences to industrial fairs or marketing events.

Technology

Innovation on the scientific/technological level is concentrated in 4 areas: (1) motion tracking, (2) auditory rendering, (3) user and world modeling, and (4) authoring and simulation. Significant advances of the state of the art in the 4 areas are necessary to realize the LISTEN concept. Apart from the innovation in the individual areas, LISTEN proposes an integration of new generation technologies in an original way. Only the combination of large-area high-definition wireless motion tracking with advanced binaural rendering and wireless high-quality digital audio transmission provides for the degree of auditory immersion necessary to create convincing audio-augmented environments. Only the combination of novel user modeling techniques with advanced virtual-reality-based world modeling, authoring and simulation techniques can provide the basis for producing and experiencing a new form of content: the immersive audio-augmented environment. The project consortium has developed a detailed research agenda for the next 3 years in order to meet the objectives defined in the project. As the project only started a few weeks ago and is currently involved with the details of the system design, there are of course no detailed technological results to be reported so far.

Partners

The LISTEN consortium is composed of 5 experienced partners, which complement each other perfectly in order to achieve the objectives of the project. Three important research institutes (GMD - http://www.gmd.de, IRCAM - http://www.ircam.fr, IEMW - htp://www.iemw.tuwien.ac.at) bring their scientific and technological expertise and resources to the consortium. The partner representing the end users and content authors (Kunstmuseum Bonn - http://www.bonn.de/kunstmuseum) is from the artistic/cultural domain and will guarantee for a high quality of the public prototypes. The industrial partner (AKG - http://www.akg-acoustics.com) is a world-leader audio technology company who will ensure that the project results correspond to real-world needs and meet industrial-strength standards of quality and usability. The consortium is completed by a group of artists, composers and independent consultants (the most prominent of which is Larry Sider, director of the School of Sound - http://www.schoolofsound.co.uk) who helped to shape the project and will continue to actively support it through its lifetime.

Active sound field control for virtual reality

Peter Svensson (svensson@tele.ntnu.no)
Acoustics group, Department of telecommunications, Norwegian University of Science and Technology
http://www.tele.ntnu.no/users/svensson

The generation of a virtual sound field is straightforward in anechoic environments and several different techniques are available. However, in all other types of rooms, the interaction between the generated and the existing, passive sound field in the room must be considered. In this presentation the special large-scale case of electroacoustic systems for variable room acoustics in auditoria is explored. Results from experiments are presented for aspects such as feedback problems, the audibility of the virtual vs. the natural acoustic field, and limitations for how large changes that are possible. The need for the development of metrics for evaluating the quality of virtual sound fields and implications for smaller scale virtual or augmented reality cases is discussed.

Utilizing Audio in Immersive Visualization

Matti Gröhn
Helsinki University of Technology
Telecommunications Software and Multimedia Laboratory
PO Box 5400, FIN-02015 HUT, FINLAND
Matti Gröhn: E-mail: Matti.Grohn@hut.fi

ABSTRACT

This is the position paper covering my research area for Acoustic Rendering Campfire. My research is part of the work done in DIVA group. The main objective of my research is to find out the best possible ways to use spatial audio in typical tasks in immersive visualization (orientation, localization, navigation, and data representation). In the long run the goal is more efficient utilization of the spatial audio in immersive visualization application areas. In this position paper I have included an abridged version of the ten page paper[1] presented in January at Photonics West 2001.

1. BACKGROUND OF THE AUTHOR

I have ten year experience in scientific visualization. My master thesis covered data sonification, and I have explored that field from year 1992 for example actively participating ICAD (http://www.icad.org) conferences, and chairing the next one (http://www.acoustics.hut.fi/icad2001).

2. MY EXPECTATIONS ABOUT THE CAMPFIRE AND ISSUES I LIKE TO DISCUSS

In this Campfire I like to learn what is current situation in this field. Especially my interest areas are audio/visual interaction in virtual worlds, combined audio/visual perception, and virtual reality audio/visual system integration and applications.

I am looking forward to have discussions covering this complex area of cross-modal interaction of dynamic auditory and visual stimuli in virtual environment.

3. BACKGROUND AND MOTIVATION OF MY RESEARCH

Immersive visualization generally takes place in virtual environments, which provide an integrated system of 3D auditory and 3D visual display. The usage of 3D sound in virtual environments is a quite well established area,[2] and it is used to emphasize the sense of presence. This is normally achieved using recorded or simulated real world sounds to create virtual audio environment. The aim of my research is to find out new efficient ways to use audio in immersive visualization.

In the immersive scientific visualization the structures and objects might not have obvious up and down directions or any other orientation or wayfinding cues. For example, large molecules or large multidimensional datasets could be very complex and after few rotations and movements it is easy to loose orientation or location of the origin. In complex immersive visualization tasks audio can be utilized as a navigational aid or as a data representation method (sonification).

The main objective of my research is to find out the best possible ways to use audio in different tasks. In the long run the goal is more efficient utilization of the spatial audio in immersive visualization application areas. I have chosen the performance based approach for evaluation of usefulness of audio, and I run my experiments in our virtual room.[3] Due to a existence of DIVA group our virtual room has state of the art 3D-audio system[4–6]

4. RELATED RESEARCH

4.1. Sonification

Sound has been used in data analysis for several years. This is called sonification. Sonification brings out different aspects of data than visualization. It can reduce the visual overload. Ears are better than eyes for finding time dependent changes and identifying periodic patterns. With multiple data streams some of the streams can be shown in visual display and some can be presented on auditory display.

Since the beginning of the 80's the research has gone forward, and after the first International Conference of Auditory Display conference 1992,[7] the researchers have regularly done new experiments and shared the information. The ICAD organization* has published 'Sonification Report'.[8] This report was prepared at the request of National Science Foundation and the purpose is to provide an overview of sonification research, including the current status of the field and a proposed research agenda.

4.2. Spatial sound

Spatial sound is a wide research area. There are many subareas like virtual acoustics[9] and spatial sound reproduction,[10] which are quite well explored. These and some other areas are well covered also by Begault.[2] There is a recent study on perceptual issues on spatial reproduction systems,[11] though it is concentrated on virtual home theater systems.

Auditory localization of 3D sound sources have been tested in several experiments.[12–15] Unfortunately most of them have been done using static sound sources. The cross-modal perception of auditory and visual stimuli is explored mostly with animals.[16] The intersensory interaction of visual and auditory stimuli have also been explored.[17] Typically most of these tests have been done in static test situations. So far, little research have been done in the area of the cognitive aspects of simultaneous visual and auditory stimuli in dynamic environments.[18] My research is concentrated on this complex area of cross-modal interaction of dynamic auditory and visual stimuli in virtual environment.

5. DEFINITION OF THE TASKS

In my research I concentrate on role of the audio in four typical tasks (defined in Table 1) in immersive visualization: orientation, localization, navigation, and data analysis (sonification).

Task	Definition
Orientation	User awareness about the front-back, up-down, and left-right directions.
Localization	User ability to define direction and distance of the target
Navigation	User ability to move from starting point to target
Sonification	Use of nonspeech audio to convey information

Table 1. Definition of tasks

5.1. Orientation

In this research orientation is defined as a task, in which user is aware about the front-back, up-down, and left-right directions.

The orientation can be represented in such a way, that each direction has its own characteristic timbre and the sound source is located in that direction. While user rotates the global geometry the sound sources indicating orientation move as well. Applying this method the user hears all the time which way at the moment is for example the original front-back direction. In informal tests (done in horizontal plane) the method has been successful.

*http://www.icad.org

5.2. Localization

In this paper localization is defined as a task, in which user defines the direction and distance of the source (could be auditory, visual or combined).

In data analysis auditory beacons[7] or some other auditory stimuli are applied to localize the most interesting features of the data. For example, while a researcher is exploring a large protein, he can 'highlight' the most important amino acids with auditory beacons. In a dynamic representation it is important that the user is able to follow the location of the moving sound source.

5.3. Navigation

Navigation (also known as wayfinding) is a task which utilizes both localization and orientation information. In this research navigation is defined as a task, in which user goes from one specified position (starting point) to another specified position (target).

Typically in a complex visualization the visualized objects may occlude the target. If the target is presented using sound, it can be located even when it is not visible. For example while exploring large protein the user has a awareness of locations of the most important amino acids, and could easily move near them even when he doesn't at first see them behind the other chemical structures. Our first experiment described in paper[19] showed that navigation is possible with auditory cues.

5.4. Sonification

Sonification is a large and complex research area. In the 'Sonification Report'[8] the sonification is defined as the use of nonspeech audio to convey information. More specifically, sonification is transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation.

In my research I concentrate on spatial sonification.[7,20] In spatial sonification 3D audio techniques are used to place the sound sources in their locations. Spatial sonification enables separation and localization of the most interesting or critical values during the data exploration. For example, the spatial sonification of distance information is applicable as an aid in a molecule docking task. The listening point could be put inside the molecule and the user will hear the critical distance information from accurate direction.

6. RESEARCH METHOD

I have chosen the performance based approach for evaluation of usefulness of audio. The tasks described in section 5 are evaluated with several user tests. In these tests the subjects accomplish a well defined subset of these tasks. During the test I measure user performance data like speed and accuracy. Additionally I collect subjective evaluations, which will make it possible to find the least annoying variable combinations.

6.1. Test variables

Many different variables should be taken into account, while defining the most useful way to use auditory information. In my research I have considered to explore the effects of 3D-panning method, visual display, interaction, visual cues, viewing and listening position, and timbre. It is not reasonable to test all these variables and all different tasks in one large test (If I just test the orientation, navigation and localization and use three different audio stimuli, the amount of different test combinations is still over four hundred different subtasks for each subject). It is much more convenient to start (as I have already done) with smaller subset of the variables. After evaluation the best variables will be used as fixed variables in next tests.

6.1.1. 3D-panning method

Spatial sound can be reproduced using either headphones or speakers. Headphone playback is considered optimal for reproducing spatial sound because it allows the greatest degree of control over the location of the spatial source.[18] Head related transfer functions (HRTF)[10] are the most common method for headphone reproduction.

Multichannel sound reproduction using multiple loudspeakers is a more convenient solution for spatial sound in virtual rooms. With multichannel reproduction we avoid the need for individualized HRTFs and head tracking. In our system we apply vector base amplitude panning (VBAP), which is a simple mathematical way to calculate the

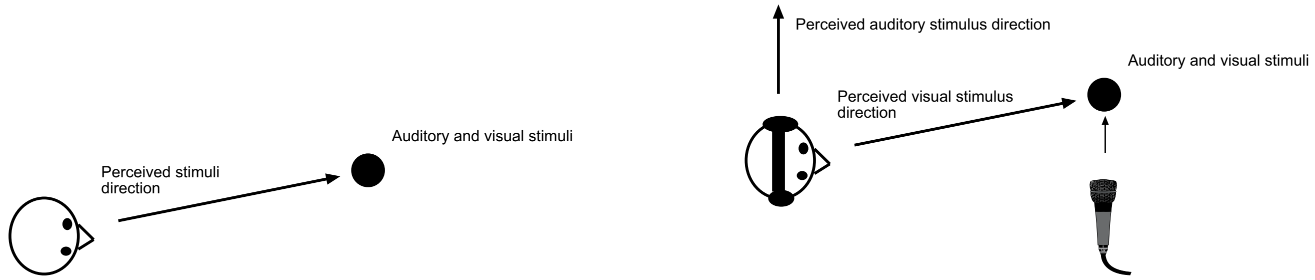

Perceived stimuli direction
Auditory and visual stimuli
Perceived auditory stimulus direction
Perceived visual stimulus direction
Auditory and visual stimuli

Figure 1. To the left is a normal situation where auditory and visual stimuli from the same source are perceived in the same direction. To the right the listening position is separated from the viewing position and the perceived directions of auditory and visual stimuli differ from each other.

gain coefficients for the loudspeakers.[21] VBAP also allows an arbitrary loudspeaker placement which is good feature in virtual rooms where mirrors and projectors hinder the fixed loudspeakers positioning.

Another used multichannel spatial sound reproduction method is Ambisonics[22,23] which is suitable to create background soundscape because a recording and coding methods are available for Ambisonics. Due to it's known limitations (small "sweet spot") it is not suitable for our purposes.

6.1.2. Interaction methods and devices

In the immersive visualization the keyboard and mouse combination is not convenient solution. There are many alternative solutions for the interaction in virtual environments like data gloves, wands, and trackers.

6.1.3. Visual cues

I am interested in the real usage scenarios, where user has always visual and auditory information available. Typically it is considered, that when visual and auditory cues conflict, sounds are localized to the position of the visual stimuli. This is known as the "ventriloquism effect".[16] However, at least one study[24] suggests that visual dominance is not unilateral across azimuth positions, and there exist positions where auditory information provides more accurate localization information

6.1.4. Viewing and listening position

It is natural to think that the viewing and listening points are the same. That is the normal situation in everyday life. I have a hypothesis, that in some situations, it will be useful to separate viewing and listening points (figure 1).

Separate viewing and listening position is analogous with the situation where sound recorder is using a remote microphone. I have preliminarily studied these in.[20]

7. CONCLUSIONS AND FUTURE WORK

In this paper I defined four different task in which the spatial sound can be utilized in immersive visualization (orientation, localization, navigation, and sonification). The test setting with multiple variables was presented.

There is lot of work to be done in testing the tasks, and different combinations of test variables.

REFERENCES

1. M. Gröhn, T. Lokki, L. Savioja, and T. Takala, "Some aspects of role of audio in immersive visualization," *Proc. SPIE* **4302**, (San Jose, California), Jan 2001.
2. D. Begault, *3D Sound for Virtual Reality and Multimedia*, Academic Press, Cambridge, MA,, 1994.
3. J. Jalkanen, *Building a spatially immersive display - HUTCAVE. Licenciate Thesis*, Helsinki University of Technology, Espoo, Finland, 2000.
4. L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *Journal of the Audio Engineering Society* **47**, pp. 675–705, Sept. 1999.
5. J. Hiipakka, T. Ilmonen, T. Lokki, and L. Savioja, "Sound signal processing for a virtual room," *Proc. X European Signal Processing Conference (EUSIPCO 2000)* , (Tampere, Finland), Sep 2000.
6. J. Hiipakka, T. Ilmonen, T. Lokki, M. Gröhn, and L. Savioja, "Implementation issues of 3d audio in a virtual room," *Proc. SPIE* **4297B**, (San Jose, California), Jan 2001.
7. G. Kramer, *Auditory Display: Sonification, audification and auditory interfaces.*, Addison-Wesley, Reading, MA., 1994.
8. G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, J. Neuhoff, R. Bargar, S. Barrass, J. Berger, G. Evreinov, M. Gröhn, S. Handel, H. Kaper, H. Levkowitz, S. Lodha, B. Shinn-Cunningham, M. Simoni, W. Tecumseh Fitch, and S. Tipei, *Sonification Report: Status of the Field and Research Agenda.*, ICAD, 1999.
9. L. Savioja, *Modeling Techniques for Virtual Acoustics.* PhD thesis, Helsinki University of Technology, Telecommunications Software and Multimedia Laboratory, report TML-A3, 1999.
10. J. Huopaniemi, *Virtual acoustics and 3-D sound in multimedia signal processing.* PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, report 53, 1999.
11. N. Zacharov, *Perceptual Studies on Spatial Sound Reprodution Systems.* PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, report 57, 2000.
12. F. Wightman and D. Kistler, "Localization of virtual sound sources synthesized from model HRTFs," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'91)*, (New Paltz, NY), 1991.
13. E. Wenzel, "Localization in virtual acoustic displays," *Presence: Teleoperators and Virtual Environments* **1**(1), pp. 80–107, 1992.
14. E. Wenzel, M. Arruda, D. Kistler, and S. Foster, "Localization using non-individualized head-related transfer functions," **94**, pp. 111–123, 1993.
15. J. Blauert, *Spatial Hearing, The psychophysics of human sound localization.*, The MIT Press, Cambridge, MA,, 1997.
16. B. Stein and M. Meredith, *Merging the Senses,*, The MIT Press, Cambridge, MA, 1993.
17. R. Welch and D. Warren, *Intersensory interactions*, Wiley, New York, 1986.
18. D. Begault, "Auditory and non-auditory factors that potentially influence virtual acoustic imagery," in *Proc. AES 16th Int. Conf. on Spatial Sound Reproduction*, pp. 13–26, (Rovaniemi, Finland), April 10-12 1999.
19. T. Lokki, M. Gröhn, L. Savioja, and T. Takala, "A case study of auditory navigation in virtual acoustic environments," *Proc. ICAD 2000* , (Atlanta GA,), Apr 2000.
20. M. Gröhn and T. Takala, "Magicmikes: method for spatial sonification.," *Proc. SPIE* **2410**, (San Jose, California), Jan 1995.
21. V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society* **45**(6), pp. 456–466, 1997.
22. M. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society* **21**(1/2), pp. 2–10, 1973.
23. D. Malham and A. Myatt, "3-d sound spatialization using ambisonics techniques," *Computer Music Journal* **19**(4), pp. 58–70, 1995.
24. D. Perrot, "Auditory and visual localization: two modalities, one world," *Audio Engineering Society 12th International Conference: The perception of reproduced sound* , pp. 221–231, (Copenhagen), 1993.

The DIVA Auralization System

Tapio Lokki and Lauri Savioja
Helsinki University of Technology
Telecommunications Software and Multimedia Laboratory
P.O.Box 5400, FIN-02015 HUT, FINLAND

Abstract

This is the position paper of DIVA sound group for Acoustic Rendering Campfire. In this paper we briefly overview the DIVA auralization system and also tell our exceptations related to this Campfire.

1 Introduction

The basic principles and ideas of physics-based rendering were given already at 1983 by Moore [1]. However, one of the first complete sound rendering systems that creates natural sounding rendering, was implemented in Helsinki University of Technology (HUT) [2] and it is called Digital Interactive Virtual Acoustics (DIVA). The sound rendering part of the DIVA system has been reported in more detail by Savioja *et al.* [3]. In the Acoustic Rendering for Virtual Environments Campfire we will overview the auralization part and present the recent improvements to the system.

In VAE creation our ultimate goal has been to develop a sound rendering application that can be used in acoustic design. In its ideal form an authentic reproduction of a real environment would be indistinguishable from the real environment without any exception [4]. Of course, this is not possible because of simplifications that have to be made in room acoustic modeling. We try to do plausible, perceptually authentic, auralization that can be used as a reliable tool in room acoustics design. We know very well that the rendering with same level of naturalness can be implemented with more efficient algorithms with perceptual modeling. However, our starting point has been to take the room geometry and all the physics-based data that we can get and with this information to realize natural sounding auralization. To achieve this ambitious goal, we utilize the DIVA auralization system which is a physics-based room acoustic modeling and auralization system that does rendering in time domain. The applied auralization method enables dynamic rendering (also in real time) with normal PC (without any additional DSP-cards), running Linux operating system.

2 Room Acoustic Modeling and Auralization in DIVA System

In room acoustic simulation our goal is to create a totally artificial, but still plausible, virtual auditory environment. In other words, no measured room impulse responses are used in sound rendering. This means that the sound source characteristics, sound propagation in a room as well as the listener have to be modeled.

In DIVA auralization the modeling of room acoustics are divided into three parts: the modeling of direct sound, early reflections, and late reverberation. The direct sound and early reflections are modeled with the image-source method and late reverberation with an efficient recursive algorithm. With the image-source method the following parameters for each reflection, at each time instant, are calculated:

- orientation (azimuth and elevation angles) of sound source
- distance from listener
- material filter parameters
- azimuth and elevation angle with respect to the listener

These parameters are used in the auralization process that is implemented as the signal processing structure presented in Fig. 2. The signal processing blocks contain the following filters:

- $S_d(z)$ is a diffuse field filter of a sound source.
- $F_d(z)$ is a diffuse field filter of HRTFs.
- $T_{0...N}(z)$ contain the sound source directivity filter, distance dependent gain, air absorption filter and material filter (not for direct sound).
- $F_{0...N}(z)$ contain directional filtering realized with separated ITD and minimum-phase filters for HRTFs.
- *R* is a late reverberation unit.

Each of these blocks is discussed in mode detail in the Campfire.

3 Evaluation of DIVA Auralization System

We are currently evaluating our auralization system with the evaluation framework [5]. In this framework the evaluation of quality of auralization is done by comparing real-head recordings and auralized room acoustic simulation. In Campfire we will also present the first results of the evaluation.

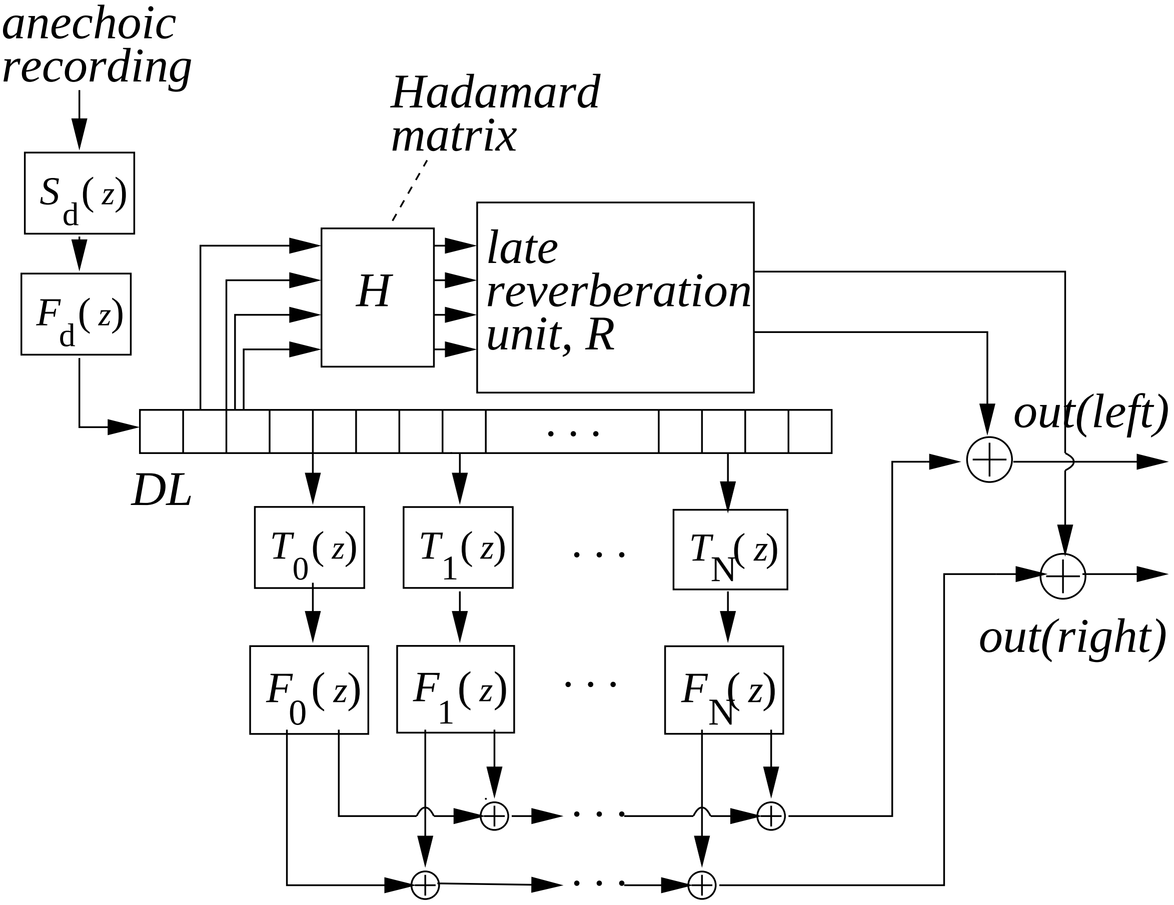


Figure 1: The DIVA auralization signal processing structure. From the long delay line DL the sound signal is picked to filter blocks $T_{0...N}(z)$ according to the distance of the image source from the listener.

4 Expectations about the Campfire

In this Campfire we like to learn what is the current state-of-the-art in sound rendering. We also want to find out new ideas about the applications of acoustic rendering. Naturally, we are also interested in the ideas how we can evaluate the quality of the virtual acoustic environments.

References

[1] F. R. Moore. A general model for spatial processing of sounds. *Computer Music J.*, 7(3):6–15, 1983 Fall.

[2] T. Takala, R. Hänninen, V. Välimäki, L. Savioja, J. Huopaniemi, T. Huotilainen, and M. Karjalainen. An integrated system for virtual audio reality. In *the 100th Audio Engineering Society (AES) Convention*, Copenhagen, Denmark, May 11-14 1996. preprint no. 4229.

[3] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. Creating interactive virtual acoustic environments. *J. Audio Eng. Soc.*, 47(9):675–705, Sept. 1999.

[4] R.S. Pellegrini. Quality assessment of auditory virtual environments. In *Proceedings of Internoise 2000*, volume 6, pages 3477–3483, Nice, France, Aug. 27-30 2000.

[5] T. Lokki, J. Hiipakka, and L. Savioja. A framework for evaluating virtual acoustic environments. In *the 110th Audio Engineering Society (AES) Convention*, Amsterdam, the Netherlands, May 12-15 2001. Accepted for publication.

Design of the Sound System at the PDC VR-CUBE at the Royal Institute of Technology.

Peter Lundén
The Interactive Institute
Emotional and Intellectual Interfaces
peter.lunden@interactiveinstitute.se

Gert Svensson
Royal Institute of Technology
Center for Parallel Computers (PDC)
gert@pdc.kth.se

Johan Vävare
ADL Konsult AB/
Asia Magic Advanced
Technology LTD
johan@asia-magic.co.th

This project is a co-operation between PDC, The Interactive Institute and ADL Konsult AB. The goal was to improve the sound system of the PDC VR-CUBE at the Royal Institute of Technology and to bring the quality of the aural experience to the same level as the visual experience. The VR-CUBE at PDC is the first six wall Cube ever built. The fact that it is closed on all six sides' lead to difficulties in the design of the sound system that has not been considered before.

Room acoustics

There where several acoustical problems to solve. The first problem was the acoustics of the hall that is housing the Cube. To make it possible to perceive the qualities of the aural virtual environments the acoustics of the hall had to be designed in such a way, that the aural images emanating from the loudspeaker system were not distorted by reverberation or unwanted reflections from the room. As the hall was quite big, and constructed with a combination of concrete and brick walls, the reverberation time of the hall was far too long for this purpose. This problem was solved by covering the main part of the walls in the hall with mineral wool absorbers and a black velvet fabric. By using black fabric the problem of light from the Cube reflecting on the wall surfaces was also reduced. The fabric also reduces high frequency sound reflections quite well, but to control the lower frequencies mineral wool absorbers were necessary. They were hung in long 'stripes' from the balcony, thus creating a suitable distance (about 1.5 m) to the walls for low frequency absorption. After the acoustic treatment of the hall the reverberation time was measured to be less than 0.4 ms, with a low frequency absorption efficient at least down to 40-50 Hz.

Loudspeaker system

The second acoustical problem was the placement of the loudspeakers and the acoustic properties of the Cube itself. TAN Projection Technology, which is the manufacture of the Cube, suggested the loudspeakers should be placed outside of the Cube behind the screens. They claimed that the frequency dependent loss in the high frequency range of the sound caused by the transmission of the sound through the screens could be compensated for although the plastic film the screens were made of is quite thick. A simple theoretical study of the problem gave at hand that the high frequency loss was too large to compensate for. The surface weight of the screens suggested that they were acoustically equivalent to a lowpass filter with a knee-point at less then 200 Hz, which gives a damping of more then 35 dB in the high end of the spectrum. The high damping was also confirmed by measurements of the acoustical transmittance of the screen material that we made. The bad acoustical transmittance of the screen is not only causing problems of getting sounds in to the Cube but also to get them out of there. That might not seam as a very relevant problem, but a deeper thought into the problem reveals a severe problem. As a large part of the acoustical energy will stay in the Cube then the room acoustics of the Cube will be very evident inside it, which will distort the impression of the room acoustics of the aural image. The only solution to both of the problems would be to replace the material of the screens with a material with better transmittance. It is very difficult to find a screen material that has both acceptable optical and acoustical qualities at the same time. We thought this would take to long time to wait for so we decided to go for another solution that will solve one of the problems. The solution to get sounds into the Cube was to place small high frequency loudspeakers inside of the Cube, one in each corner combined with full range speakers outside and a sub bass system on the floor about 3 meters below the cube. By using a DSP processor, the different speaker signals were properly time aligned and some filtering was applied in order to get a flat frequency response. Measurements done after the sound system installation showed, as expected, a significant reflection about 15 ms (1.5 times the cube size) in the high frequency register, coming from the screens. Except for this acoustic artefact, the speaker system turned out to be very accurate in terms of frequency and phase response.

3D sound software

Snd3D (Lundén, 2000) is the software that is used to control and feed the loudspeaker system. Snd3D is developed by the Interactive Institute and is a software based real-time 3D sound system based on ambisonics (Gerzon, 1980) and reproduction over loudspeakers aiming at VR and interactive applications. The system is able to perform real-time simulation of direction, distances and movements as well as the acoustic environment. The system is implemented in PD (Puckette, 1996), a graphical programming environment for audio and MIDI processing. One of the novelties with Snd3D is its ability to represent environmental sounds.

Future work

PDC has recently developed an open source navigation tool Navier for virtual environments. Navier is based on the Performer library and CAVElib. Today Navier only handles the visual aspect of the virtual environment. The plan for the future is to integrate Navier and Snd3D for the aural simulation. Special attention will be paid on developing methods for defining and simulating environmental sound fields, which cannot be defined as singular spherical sound sources - like traffic sounds, ocean waves, background noise etc.

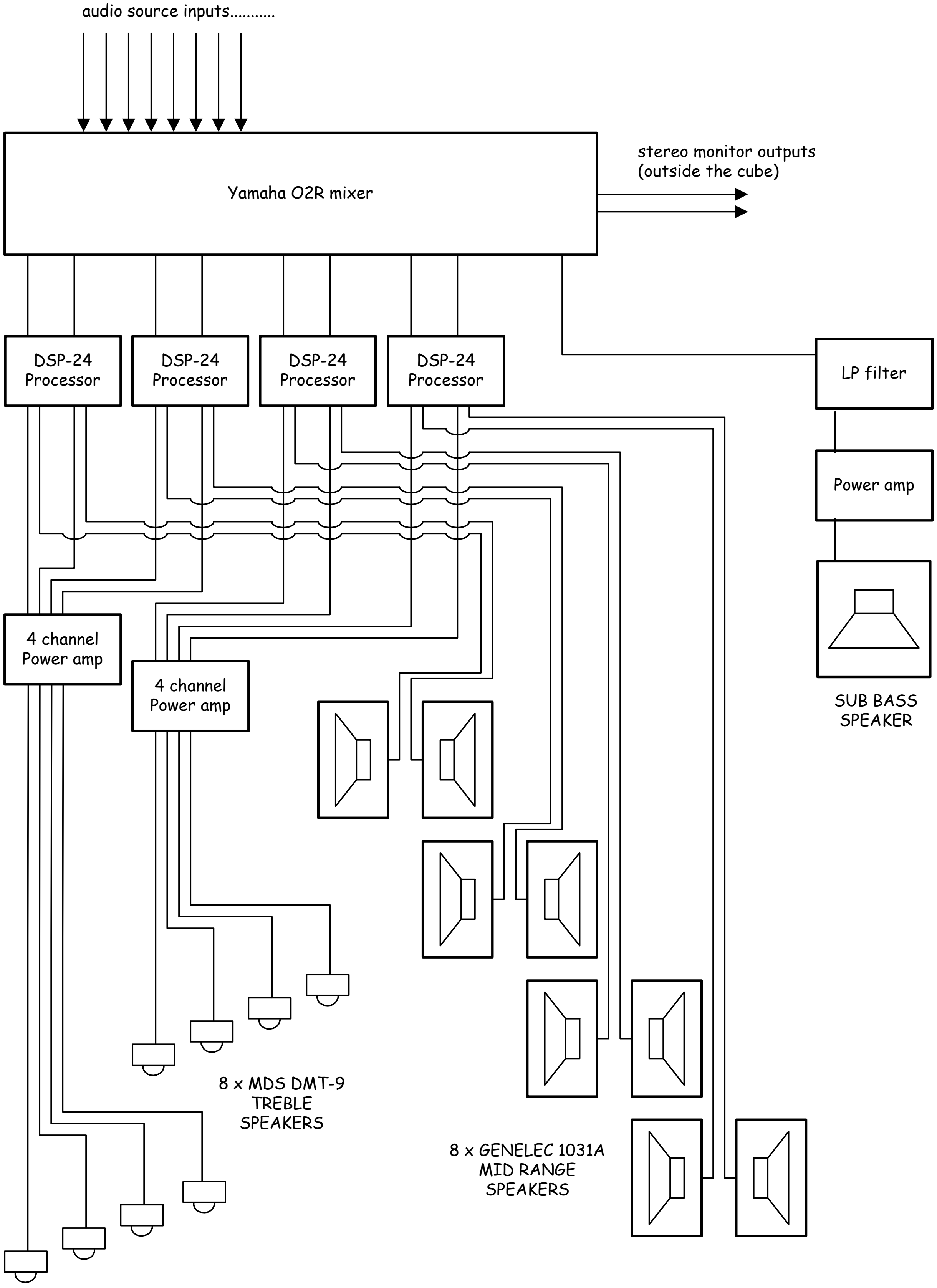
There is also work done to develop a system for playback of pre-recorded ‘paths’. This will include a system for audio synchronisation which combines pre-defined timeline synchronisation with event controlled real time interactive audio playback. There are thoughts and plans to extending the audio system to handle real time room acoustic simulation - both for pre-recorded audio material, modelling existing sound environments as well as simulating virtual spaces where sounds emitted from the user inside the cube will be part of the virtual acoustics.

References

Gerzon, M A. 1980. ”Practical Periphony: The Reproduction of Full-Sphere Sound”. Draft of lecture presented at Audio Engineering Socitey 65th Convention. Preprint 1571.

Lundén, P 2000, “ Snd3D; a 3D sound system for VR and interactive applications”. in Proc. Int. Computer Music Conference, pp 300-303 Berlin 27 Aug-1 Sep 2000. The International Computer Music Association, San Francisco, California, USA.

Puckette, M. 1996. "Pure Data." Proceedings, International Computer Music Conference. San Francisco: International Computer Music Association, pp. 269-272.’



PDC VR Cube
Audio system layout

ADL
KONSULT
AKUSTIK
DATA • LJUD

VETENSKAP OCH KONST
KUNGL
TEKNISKA
HÖGSKOLAN

Evaluation of Acoustic Comfort in Buildings

Julien Maillard (j.maillard@cstb.fr)
Jacques Martin (j.martin@cstb.fr)
Centre Scientifique et Technique du Bâtiment (CSTB)
http://www.cstb.fr

Most of the existing tools for the prediction and evaluation of acoustic comfort in buildings give the user quantitative results. Sound insulation is generally expressed in dB across the frequency spectrum in third octave bands.

While this approach is appropriate when comparing the acoustic performances of various building elements, it can be hard to use for the non-specialist. Also quantitative results do not allow the evaluation of complex situations in terms of masking effects for example. An interactive tool is proposed here for the evaluation of acoustic comfort in buildings from a perceptive point of view. The user evaluates the levels of isolation by listening to familiar sound scenes inside the building. These scenes are synthesized according to the construction materials and room configuration.

This paper presents the algorithm implemented in the audio restitution of sound scenes inside buildings. The algorithm is now part of the ACOUBAT© software developped at CSTB. The approach uses a set of calibrated source signals (appliances, conversations, street noise, etc...). Based on the construction materials and room configuration, ACOUBAT estimates the sound attenuation levels along the transmission paths between the rooms (inside noise) and the outside (exterior noise). These calculations follow the European Norm EN 12354 [1] for the prediction of sound insulation in buildings. Attenuation filters are constructed based on the third octave band transmission data associated with each source. The transmission filter output signals are then fed to the auralization and the artificial reverberation modules whose parameters depend on the source locations and the room reverberation time. The resulting auralized sound gives a realistic binaural impression of the sound scene including relative loudness and coloration. One of the key feature of the proposed algorithm is the ability to modify the sound scene configuration in real time. In other words, the user can change the construction materials, source positions, and room size interactively. This is an improvement over another similar tool which has been developped recently in Germany [2].

References

[1] EN 12354 part 1: Building acoustics - Estimation of acoustic performance of buildings from the performance of products.

[2] M. Vorländer, R. Thaden. Auralisation of airborne sound insulation in buildings. Acta Acustica, Vol. 86, 70-76 (2000).

Virtual Environment System for Interactive Acoustic Assess of Building

Ying Zhang, Terrence Fernando, Ting-Kai Wang
Center for Virtual Environment, Information System Institute, University of Salford,
Greater Manchester, M5 4WT, UK
y.zhang1@pgr.salford.ac.uk, T.fernando@salford.ac.uk

Abstract

The subjective analysis and evaluation of acoustic properties prior to the actual construction of a facility is very important aspect during the design stage. Virtual Reality technology, specially visualization and auralization, provides a useful method to interactively implement this task. This position paper presents our creating Virtual Environment-based system for interactive acoustic evaluation of buildings and integration of combined audio/visual rendering in real time. Some issues and challenges what we are of interest and will face are presented as well.

1 Introduction

The evaluation of acoustic properties is an important aspect during the design of buildings. Typically a building contains many spaces, each with its own requirements for acoustical quality and for isolation from noises and vibrations. For example, the quality of sound (e.g. speech intelligibility, sound from music instruments, etc.) is extremely important for buildings such as concert hall, theater and studio etc. Depending on the use of the building, the designers need to utilize different measures such as different materials, layout and shapes to control unwanted sounds and enhance wanted sounds.

The acoustical design for a particular building, or for a space within the building, must be compatible with other construction requirements and must be incorporated in the architectural drawings and specifications that form the basis for the awarding of a construction contract. The design of a building is generally marked by many complex, often conflicting goals and constraints. The designer has to make a series of compromises in which benefits for one discipline must be weighted against another. The designer should know to what extent acoustical goals may be compromised without significantly affecting the usefulness of the building. For example, how important is total control of exterior noise or speech privacy between spaces? Or, should the interior be designed to enhance or suppress projection of sound? Which kind of materials should be chosen within the project budget and acoustic requirement? In many renovations, budgetary, aesthetic, or physical impediments limit modifications, compounding the difficulties confronting the designer. There have been many occasions where the construction teams had to modify some buildings in order to achieve expected acoustic properties, resulting high cost and delay in completing the building.

The developments in DSP, acoustic rendering, auralization and computer power have made the acoustic simulation and prediction possible. The advent of computer simulation and visualization techniques for acoustic design and analysis has yielded a variety of approaches for modeling acoustic performance (**Borish 1984; Heinz 1993; Lewers 1993; Naylor 1993; Bose Corp. 1998**). Many researches in this area to date have mainly addressed the accuracy and speed of simulation algorithms. However, software tools currently available for architectural acoustic design and simulation (i.e. CATT-acoustic, Odeon etc) use 2D user interfaces to communicate acoustic properties to the designers. The input files are used to describe parameters such as the size and surface material of the rooms, source positions/properties and receiver positions/directions etc. The prediction outputs are text format files, graphs, charts and tables (**http://www.netg.se/~catt**) (**Naylor 1993**). These systems have some limitations. Firstly current systems can not efficiently employ the CAD models produced by the architectural designers during building design stage. The designer is burdened with specifying acoustic spaces and characteristics that is laborious and time consuming. Secondly, clients can not understand the graphs, data, charts and tables, even visualized output to gain a better understanding of the acoustic properties of the proposed design. Thirdly, design changes are tedious and time consuming. When the designer modifies one parameter of the geometrical models, they have to change several input files for the acoustic simulation separately in order to re-evaluate the acoustic properties of the design. This hinders the interactive and real-time acoustic analysis and prediction of the evolving design by the clients and the designers.

This paper introduces the framework of our project which is developing an interactive decision support system based on virtual environment that is suitable for evaluating and predicting the sound properties of the buildings during the design stage. The working process of the system is shown in Fig.1.1. New 3D CAD building geometric design and its acoustic requirements are imported into the system. This system uses the geometric data and relevant acoustic models generated by the system itself to implement visualization and auralization via VR facilities. The user can move freely in the immersive Virtual Environment (VE) to hear and see the aural/visual effect of the spaces within the building. If it does not reach the requirements, the user can modify the parameters and repeat above steps till satisfied with it and get the final design. In this way, we can implement interactive simulation and design.

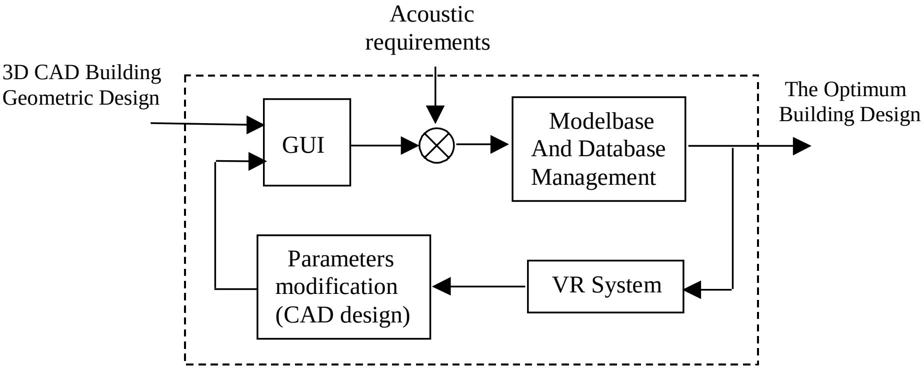

Acoustic requirements
3D CAD Building Geometric Design
GUI
Modelbase And Database Management
The Optimum Building Design
Parameters modification (CAD design)
VR System

Fig. 1.1 The process of decision-making

The system is composed of the Model-base Management System (MBMS), Database Management System (DBMS), Commercial building CAD tools, GUI and VE. Fig. 1.2 shows the structure of the system. MBMS tracks all of the possible models and schedules the relevant visual models and acoustic models. DBMS administrates the alternative visual/acoustic parameters of materials. CAD tools implement 3D geometric models design. Principles and methods of MBMS and DBMS can be found in many papers. The core part is VE system explained in later section.

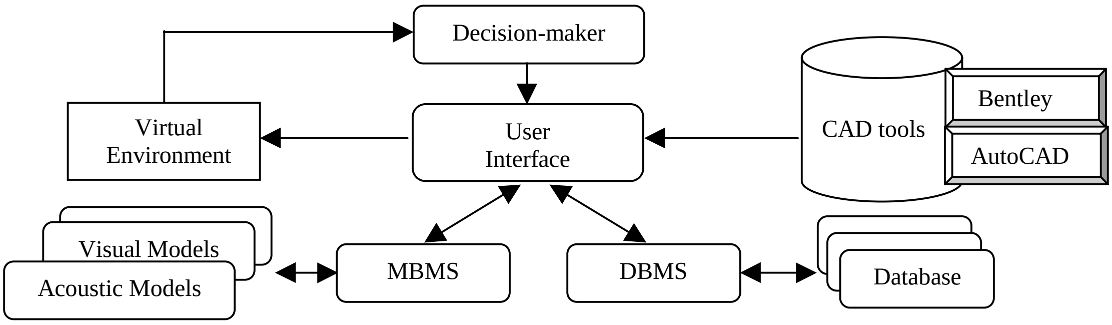


Fig. 1.2 Components of the System

This system will provide an immersive virtual environment in which the clients can be immersed within the proposed design and interact with the space to make design changes and experience the effect of that changes in acoustic properties of the building. The designer will be able to make changes to materials and space or introduce sound sources with various sound characteristics and immediately experience the acoustic properties of the building. As for a concert, auditorium, theater and studio, the listener will be able to place himself or herself at different position to feel the sound effect for different sound sources. This will give the clients a more intuitive feel for the acoustic effect of the corresponding enclosure space whilst by providing useful information on how to design and layout the dwelling, public, studying and working spaces for optimal comfort and work efficiency during the design of the buildings. For a renovation project, it may assist in optimizing an existing configuration with modifiable materials and components. The specific objectives of this research are as follows:

1) *Automatic generation of acoustic models and visual models from the CAD design produced by the designers.* Emerging CAD standards such as IFCs (Industrial Foundation Classes) will be used as the CAD input representation for generating the acoustic models and visual models. The research challenge in this objective is to generate appropriate 3D spaces from the CAD description and to automatically generate acoustic characteristics for each room in the building.
2) *Integration of immersive CAVE technologies, tracking technologies, audio server, visual models and acoustic models to provide an interactive design environment to analyze the acoustic properties of the proposed design.* A distributed environment will be designed and implemented to achieve this objective.
3) *Develop an intuitive design interface for the architects to analyze the acoustic properties of the evolving Design.* This work will develop an intuitive interface for the architects to change the geometric properties of the design using existing constraint-based modeling facilities, change material properties of the building components, define sound sources with various acoustic characteristics etc.
4) *Evaluate and enhance the system performance.* This work will involve localizing the acoustic simulation depending on the position of the user, conducting experiment to identify the affect of resolution of the RIRs and HRTFs on human auditory perception.
5) *Evaluate the System using an industrial case study and practicing architects.*

2 Hardware configuration of the VE system

The audiovisual virtual environment system is a network-based distributed VR system centered on CAVE which consists of a cube with display screen faces surrounding a viewer, Silicon Graphics Infinite Reality Engine, Huron PCI audio workstation, and real time interaction parts such as head trackers that allow updating the position of the viewerś head in relation to position of the virtual sound source, loudspeakers or headphones [Fig. 1.3].

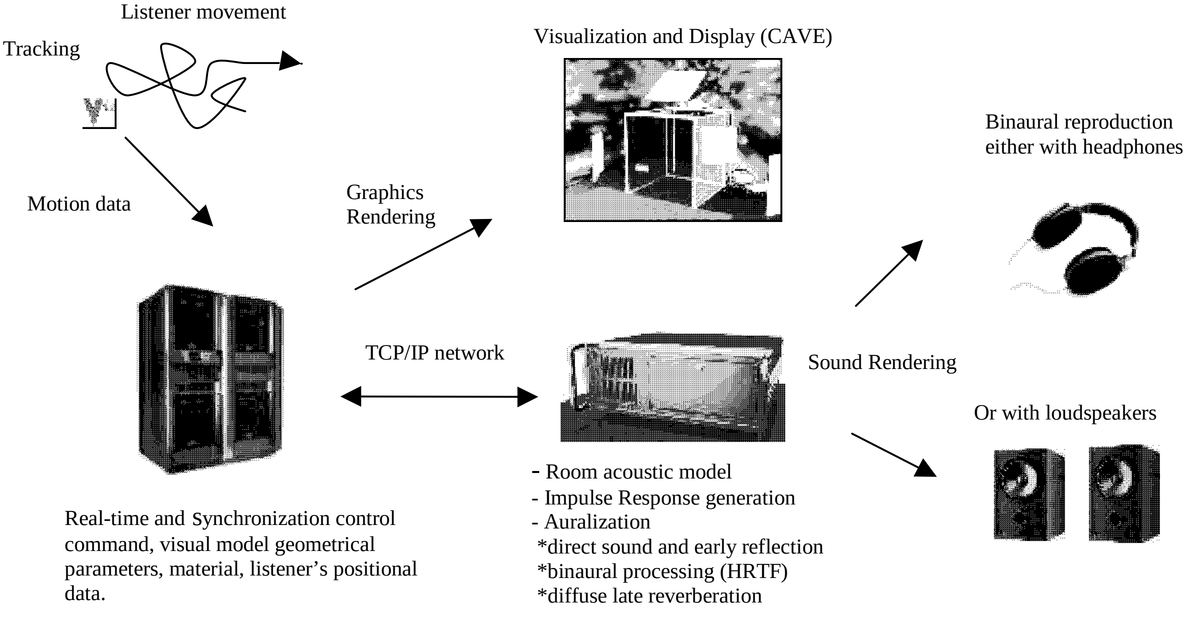

Listener movement
Tracking
Motion data
Visualization and Display (CAVE)
Graphics Rendering
Binaural reproduction either with headphones
TCP/IP network
Sound Rendering
Or with loudspeakers
Real-time and Synchronization control command, visual model geometrical parameters, material, listener’s positional data.
- Room acoustic model
- Impulse Response generation
- Auralization
*direct sound and early reflection
*binaural processing (HRTF)
*diffuse late reverberation

Figure 1.3 Audiovisual Virtual Environment System Architecture

3 Software Architecture

The software consists of Interface/Configuration Manager, World Manager, Input Manager, Viewer, Sound Manager and Database (**Fernando et al 2000**)[Fig.1.4]. The Interface/Configuration Manager gets the CAD design, parameters modification/system configuration from a file or GUI. It tracks all master processes to allow a run time configuration of the different modules.

The World Manager is responsible for administrating the overall system and coordinates the visualization, user inputs and sound management via network. The World Manager fetches the user input and passes its data (e.g. the positional information of the user) to the World Manager. The World Manager then uses these new data to update the scene graph and transfers to the Sound Manager. The Sound Manager sends the relevant commands to the audio workstation via network to switch the impulse response of the relevant position smoothly.

The Viewer renders the scene to the selected displays in the appropriate mode. Only the Viewer knows the display type (e.g. CAVE, Reality Room, Workbench, Monitor or other display). The Configuration Manager passes this information to the Viewer and according to the selected configuration the number of view cameras are created and the correct display type is initialized.

The Input Manager associates an input device to a virtual object, and it establishes the data flow between the user input and the object held by the World Manager. Typical objects are the head and hand(s), where the last one can be associated with two different input sources: a tracker and a glove.

The Sound Manager gets the traversed positional data of the listener (viewer) or sound sources and modified parameters from the user interface, and then use the relevant commands to control the audio workstation to generate the relevant sound via network configured for TCP/IP protocol.

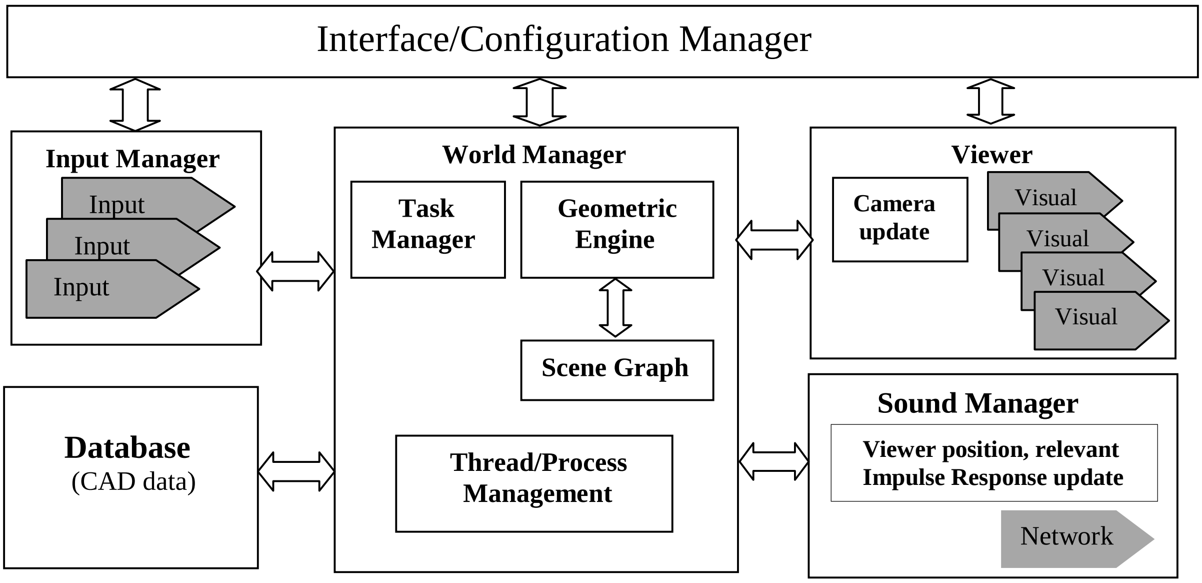


Figure 1.4 Software Architecture

4 Potential Benefits of this Research

This research will have several benefits. Firstly, the designer of the building can verify the compatibility with the clients' requirements during design stage so as to improve the decision level of the investor. Secondly, For the clients, they will be able to experience the visual and aural properties of the buildings during the design. For the designer, this system will provide a valuable interactive simulation and visualization environment for evaluating the different CAD designs. It will allow the designers to optimize the quality of the building from various perspectives. Furthermore, it acts as an avenue of research for those interested in improving the effectiveness of virtual reality experience.

The system and technology can be used to make acoustic evaluation and simulation for different buildings, spaces and environment according to different acoustical requirements.

1) Dwelling
   The most important acoustic requirements for dwellings, such as private homes, hotels and hospital wards etc, are control of intruding noise, so as to live in a comfort environment.

2) Work/Study
   Acoustic requirements for work or study facilities depend on the tasks to be accommodated. For open-plan offices, the background level should be low enough for telephone use yet high enough for some degree of speech privacy between workstations. For long open-plan spaces such as library reading rooms and museum galleries require non-projection of sound, that is, localization of activity sounds and control of reverberation.
3) Meeting/Hospitality
   Meeting spaces may vary from fixed-seating, special-purpose facilities such as lecture and demonstration rooms to flat-floor, divisible rooms that are easily adapted to many types of events. The size may vary from a classroom for 25-30 people to a convention facility seating over a thousand people, with amplified sound and projection systems.
4) Performance
   The primary acoustical requirement for performance space is a low background noise level, free from distracting noise intrusions, to allow the audience to hear stage whispers and pianissimos.
5) Worship
   Most worship requirements can be satisfied fairly well by designing the space to meet music requirements (i.e. long reverberation time) and by designing the sound amplification system to ensure intelligibility of speech.
6) Industrial/Transport
   Industrial machinery may create high noise levels, so sound-absorbing materials on walls and ceiling are necessary. Either air or ground transport systems may create serious noise problems for neighboring areas. Initial urban planning studies should include evaluation of compatibility and may entail preparation of an Environmental Impact Report.
7) Arena/Stadium
   Enclosed arena and stadiums are too large for unamplified speech or music such as orchestral or choral concerts. A basic requirement is sound absorbing wall and ceiling material for control of reverberation crowd noise and echoes (long-delayed sound reflections).

5 Concerned issues

The synchronism between visual and auditory information and the accuracy needed for simulation. The audio/visual perception, system performance and real time integration of the audio/visual rendering subsystems into virtual environment.

[1] Fernando, T., Marcelino, L., (2000). "Interactive Assembly Modeling within a CAVE Environment." 16-18th Feb. 2000. Portuguese Chapter of EuropGraphics. pp43-49.

[2] Vince. J., (1998). Essential Virtual Reality fast. Springer-Verlag London Ltd.

[3] Borish, J. (1984). "Extension of the Image Model to Arbitrary Polyhedra." *Journal of the Acoustical Society of America*, vol.75, no.6, pp.1827-1836.

[4] Heinz, R. (1993). "Binaural Room Simulation Based on an Image Source Model with Diffuse Sound Scattering of Walls and to Prediction the Reverberant Tail." *Applied Acoustics*, vol.38, pp. 148-159.

[5] Lewers, T. (1993). "A Combined Beam Tracing and Radiant Exchange Computer Model for Room Acoustics." *Applied Acoustics* vol.38, pp.161-178.

[6] Kleiner, M., Dalenback, B.-I.and Svensson, P.(1993). "Auralization—An Overview." J. Audio Eng. Soc., Vol. 41. No.11.

[7] http://www.netg.se/~catt CATT-Acoustic Software.

[8] Funkhouser, T., Carlbom, I., Elko,G., Pingali, G.,Sondhi, M. and West, J. (1998). " A Beam Tracing Approach to Acoustic Modeling for Interactive Virtual Environment." Computer Graphics (ACM SIGGRAPH'98 Proc.) pp. 21-23.

[9] Naylor, G.M. (1993). "Odeon-Another Hybrid Room Acoustical Model." *Applied Acoustics* vol.38 pp.131-143.

[10]Bose Corporation (1998). Auditioner Audio Demonstrator Technology in Depth. http://www.bose.com/technologies/prediction_simulation/html/principle.html.

[11]Manual of Huron digital Audio Convolution Workstation. Lake Corp.

[12]Manual of CAVElib.

Campfire: Acoustic Rendering and Virtual Environments
Snowbird, Utah May 2001

Robert Essert

Position Paper

Expertise

I have been acoustics designer of concert hall and theatre buildings for over 20 years. At the same time, I have been developing associated technologies to serve the consultancy, especially modelling, auralisation and measurement. My interests lie in the application of sound rendering to music and theatre projects and also in the basic modelling algorithms. My expertise includes a deep understanding of what is perceptually significant in concert hall acoustics, both subjectively and physically. My rendering focus for several years was therefore on sound quality and on practical useability. More recently I have become interested in (interactive) Virtual Environments and have been looking at tradeoffs between speed and detail. I have designed and specified the immersive audio system for the new VR Cave (SGI/Trimension ReACTor) at University College London, basically from available components. I have also assembled an in-house ambisonic sound system for Arup Acoustics' use on architectural project work. I have developed acoustical modelling and auralisation software from the ground up, and have experience as a user of commercial packages.

What I would like to learn (and discuss)

- Generally what people are doing and want to be doing about perceptually adaptive sound rendering (I hope we can cross over to the sister campfire on perceptually adaptive graphics)
- Specifically, about tradeoffs between listener movement and precision of rendering required
- Tradeoffs between source movement and precision of rendering required
- Pre-processing of various components of the sound field in virtual environments, such as:
  a) fixed source and environment, moving listener → maximum pre-processing possible
  b) moving source, fixed environment → initial pre-processing at the creation of the environment
  c) changeable environment → pre-process adaptive during the exploration

- Applications of spatial auditory displays including mapping of data to spatial aspects of sound:
  - experiencing multidimensional data sets as sound environments
  - tools for the blind
  - combination with visualisation tools

What I would like to discuss (and learn)

- Which physical phenomena should be/are modelled in various source/environment/listener situations? Are we modelling "what we can" or what is most relevant?
  Different situations might be categorised as some combination/balance of such variables as:
  - listener motion, source motion
  - listener concentration on sound – is sound the primary concern, as in a concert hall, or is it subservient?
  - frequency characteristics and amplitude envelopes of the source(s)
  - number of sources – few or many – and their spatial distribution
  - is the priority to be quality of information transmission or refinement of aesthetic judgement?

- What are the perceptual attributes of each situation? Static situations are complex enough. Dynamic situations (moving source/receiver) shift the audibility thresholds and scales of other subjective attributes.

- What can we hear that we can't model? In concert hall acoustics the live experience is still quite a bit more rich and subtle than can be modelled, either in real time or off line.

- What do we model that we can't hear? Others probably know more about this than I, but I'd bet that many algorithms spend more time than necessary on inaudible detail of the sound field. Those models need to be based on better knowledge of:
  - the subjective impressions of listeners,
  - the relationships between the subjective and objective, and
  - how those relationships change with other variables.

  In this last, we find the importance of adaptive systems.

- On such possibility is consolidation of directional information – what degree of precision is necessary at various points in an impulse response, at various states of listener and source movement, for different spatial extent/complexity of source? Can the level of detail be adapted in real time?

Position Paper

Esham Fouad (fouads@bellatlantic.net)
VRSonic Inc.

Areas of Expertise

Sound APIs, Sound for VE, Real-time Sound Rendering Techniques, Real-time distributed systems.

Topics for discussion

- New real-time acoustic simulation
- Integration of sound in VE systems
- Sound control techniques
- Why HRTFs don't work

Researchers have just recently begun concentrating on the problems of integrating sound in Virtual Environment (VE) systems. Research efforts to-date have concentrated mainly on techniques for localizing sounds; giving the listener the impression of a sound emanating from a particular direction. While this is an important problem, it is certainly not the whole picture. Three basic problem areas need to be addressed by a VE sound system:

- **Modeling the sonic environment.** Modeling abstractions describe the static and dynamic properties of the elements comprising a sonic environment.
- **Real-time sound generation**. Sound generation is the problem of modeling sounds at their source and of evaluating that representation of a sound at fixed intervals in order to produce an audio sample stream in real-time.
- **Real-time sound rendering**. Rendering sounds entails two problems: localization and simulating the environmental effects. Localization is the process of recreating spatial auditory cues so that sounds appear to emanate from a particular direction in 3D space. Calculating environmental effects requires that sound waves be traced from source to listener taking into account reflection, diffraction, and attenuation.

While very useful tools have been developed by researchers for creating Virtual Sonic Environments (VSE) [1, 3, 5, 6, 7, 10], we believe that a number of important technological problems remain to be addressed:

- Current systems do not model the true three-dimensional characteristics of sound sources.
- Current technology has not taken into account variations in auditory cues and sound properties when sounds are near by. We refer to these phenomenon as near field effects.
- Current localization techniques do not adequately localize both nearby and distant sounds.

Localization techniques currently utilize empirical approaches that fall into two general categories: One is the recreation of the Head Related Transfer Functions (HRTF) cues through filter convolution, and the other is the recreation of the sound field using free field loudspeakers.

HRTFs are modeled using finite impulse response filters (FIR filters). These filters are generated by actually recording the sound reaching the eardrum using a set of probe microphones placed in a listener’s ears. A set of noise pulses are generated from locations surrounding the head, and recorded inside the ear. In order to eliminate the effect of reverberant sounds reaching the listener’s ears, the recording takes place inside an anechoic chamber. The spectral, intensity and phase change in the recorded sound represents the effect of the HRTFs on the original sound. These changes are captured using a set of FIR filters corresponding to the locations of the sound sources around the listener’s head during the recording process. During playback, sounds are localized to a certain location by finding the corresponding filter or by interpolating the coefficients of four neighboring filters in order to obtain a filter for that location. The resultant filter is convolved with the sound signal and, when heard over headphones, gives the impression of a directional sound sources.

A major problem with this approach is that HRTFs are generated for the response of a particular person's head and ears. When recreated for other persons the effect may be suboptimal causing front back reversals. HRTF based systems often do not externalize sounds such that they appear to occur inside the listener's head. Even when sound appear to emanate from outside the listener’s head, they do not appear distant. Also localization along the median plane is generally not very good.

Another approach to localization was introduced in [11]. In this approach a set of loudspeakers are located around the listener. In order to simulate the effect of angular location and distance, the amplitude of the sound emanating from each speaker is scaled such that the resultant sound appears to be emanating from a particular direction. Speaker based systems cannot place sounds inside the space delineated by the speaker array, so that close in sounds cannot be modeled. Also speaker based systems give only a weak impression of a moving sound source [8].

Near field effects can be described as variations in the effectiveness of auditory cues and our response to those cues when sounds are near by or when the listener is near a reflecting surface. Such effects can have a dramatically effect how sounds are perceived. For example, research presented in [2] suggests that assumptions concerning localization with HRTFs do not necessarily hold up when sounds are close to the listener (ie < 1 meter). HRTFs measurements change when the sound source is near field suggesting that current HRTF based localization techniques are incorrect for those sounds. Other research [9] suggests that Interaural Level Differences (ILD) becomes a dominant distance as well as direction cue when sounds are near field. Finally, changes in the reverberant characteristics of enclosures when listeners are near a reflecting surface can effect the spectral characteristics of the sounds reaching the ear nearest the reflecting surface. Such effects can be readily observed when one places an ear very close to a wall in a room for example.

Current spatial sound systems do not generally take such cues into effect and thus do not correctly model near field effects.

Finally, sound sources in the real world exhibit varying characteristics based on the direction from which they are heard. A helicopter, for example, sounds markedly different when heard from the front than it does when heard from the rear. The sound in those two locations differs not only in intensity but also in spectrum. Current approaches to spatial sound generation, generally do not model those varying characteristics, or at best model them as variations in the sound’s intensity along a two dimensional plane bisecting the sound source. The equivalent to this in computer graphics would be to have two-dimensional models placed in a three dimensional space, a technique usually referred to in computer graphics as 2 ½ D modeling. While work has been done in [4] to accurately model the three-dimensional characteristics of musical instruments. More work need to be done to find effective ways to measure, represent and reproduce the three dimensional characteristics of sound sources for use in VE systems.

References

1. Bargar, R., Choi, I., Sumit, D., Goudeseune, C. *Model-based Interactive Sound for an Immersive Virtual Environment*, Proceedings of the International Computer Music Conference '94, 471-474, 1994.
2. Brungart, D.S., Durlach, N. I., *Auditory Localization of Nearby Sources II: Localization of a Broadband Source in the Near Field*. Journal of the Acoustical Society of America, 1999. !06(4), 1956-1968.
3. Burgess, D.A., Verlinden, J.C., *An Architecture for Spatial Audio Servers,* GVU Technical

Report, GIT-GVU-93-34.

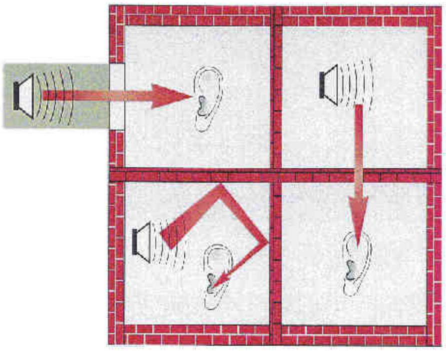
4. Cook, P.R., Essl, G., Tzanetakis, G., Trueman, D., *Multi-speaker Display Systems for Virtual Reality and Spatial Audio Projection,* proceedings of the International Conference on Auditory Displays, ICAD '98, Glasgow, 1998.
5. Das, S., DeFanti, T., Sandin, D. *An Organization for High-Level Interactive Control of Sound,* Proceedings of the International Conference on Auditory Displays, ICAD '94, 203-216, November 1994.
6. Hamman, M., Goudeseune, C. *Mapping Data and Audio Using an Event-Driven Audio Server for Personal Computers*, Proceedings of the International Conference on Auditory Displays, ICAD '97, 83-87, November 1997.
7. Huopaniemi, J., Savioja, L., Takala, T. *DIVA Virtual Audio Reality System*, Proceedings of the Third International Conference on Auditory Displays*, ICAD '96*, 111-116, November 1996.
8. R. Moore, *Elements of Computer Music*, Prentice Hall, Englewood Cliffs, New Jersey, 1990.
9. Shinn-Cunningham, B., *Learning Reverberation: Considerations for Spatial Auditory Displays*, Proceedings of the International Conference on Auditory Displays ICAD 2000, 126-134, April 2000.
10. Storms, R.L. *NPSNET-3D Sound Server: An Effective User of the Auditory Channel,* Master's Thesis, Naval Postgraduate School, Monterey California, September 1995.
11. J. M. Chowning. *The Simulation of Moving Sound Sources*, Journal of the Audio Engineering Soc., 1970.

Building Acoustical Simulation and Auralization

Klaus Naßhan (nasshan@ibp.fhg.de)
German Physical Society
http://www.ibp.fhg.de/

Introduction

One subject in building acoustics is the calculation of the propagation of sound from one room to another. The sending "room" can be situated also outside a building (see figure). The sound can be airborne sound (eg. speech, music, traffic noise) or impact sound (such as footsteps, hammer blows). In general a building acoustical situation is described by a single rating number (eg. weighted sound reduction index). This rating does neither include the spectral propagation of sound nor the characteristics of the sound source and the different paths of propagation. Experience and auralization experiments showed that such a single value does not correctly describe the auditory perception.



room and building acoustical situations

Physical problems of propagation

Let us take a look at the propagation of airborne sound from one room to an adjacent room. A source produces an airborne sound field in the sending room, which induces a sound field in the structure. In the receiving room the structure borne sound is radiated and received as airborne sound. The dominant energetic part in this propagation chain is at low frequencies. In this low frequency range sending and receiving room and the surrounding walls exhibit modal sound fields. Human hearing is not very sensitive at these low frequencies. The questions are: How exact must the virtual model be and how to validate it, even if the laboratory measurements are uncertain?

The sound fields within the structure and the rooms are diffuse at high frequencies. The received sound is very soft, but still noticeable for human hearing. The question is how to construct a fast filter for this?

At all frequencies the sound sources of the receiving room are area radiators. The general question is how to include these sources into room acoustical simulation and auralization efficiently?

My Experience

Implementation of a real-time auralization program for right parallelepipeds [1,2]. This program uses the mirror source model to extract the 128 first reflections and to calculate the reverberation time. These parameters are sent to an audioprocessor for auralization.

A program to estimate the sound transmission loss between adjacent rooms[3,4] was developed to compare estimates of sound reduction according to the drafted European standard EN 12354 with messurements of sound reduction in German houses.

Implementation of a building acoustical real-time auralization [5,6]. This program uses hardware functions of a sound card to auralize the sound reductions of different windows. In order to do this, it was necessary to pre-process the sound sources and to fit the sound reduction to the capabilities of the digital filters of the sound card.

Combination of a simple room-acoustical auralization and a building-acoustical auralization [7] and combination of a building acoustical simulation with auralization [8]. For both applications the results of one program were pipelined to another.

What will I learn?

How to link the perceptual approach with physical parameters?
State of the art in digital audio processing, fast filtering and convolution.
Combining ray-tracing, image sources and acoustical radiosity.
Do we need an Open Acoustical Language (OpenAL)?

References

[1] Naßhan, K., Akustische Virtuelle Realität mit ARS und AquA, Fortschritte der Akustik, DAGA 98, Zürich 1998

[2] Naßhan, K., Auralisation quaderförmiger Arbeitsräume. IBP-Mitteilung 24 (1997), Nr. 327

[3] Naßhan, K., Demonstrationsprogramm zur neuartigen Berechnung der Luft- und Trittschallübertragung zwischen benachbarten Räumen, IBP-Mitteilung 24 (1997), Nr. 326

[4] Naßhan, K., Fischer, H. M.: Vergleich und Bearbeitung von Vorschlägen europäischer Berechnungsvorschriften zur Bestimmung der akustischen Eigenschaften von Gebäuden aus Bauteileigenschaften, Berichte aus dem Fraunhofer-Institut für Bauphysik B-BA 3/1995

[5] Naßhan, K., Bauakustische Auralisation in Echtzeit, Fortschritte der Akustik, DAGA 2000, Oldenburg 2000

[6] Naßhan, K., Bauakustische Auralisation in Echtzeit, IBP-Mitteilung 26 (1999) Nr. 348

[7] Naßhan, K., Auralisationsprogramm zur Demonstration bau- und raumakustischer Wirkungen von Bauteilen, IBP-Mitteilung 27 (2000) Nr. 365

[8] Naßhan, K., Maysenhölder W., Mit Auralisation und rechnerischen Prognoseverfahren zur optimalen Schalldämmung, accepted for publication in Bauphysik

Position paper

Pedro Novo (novo@ika.ruhr-uni-bochum.de)
Institut of Communication Acoustics, University of Bochum,Germany
http://www.ika.ruhr-uni-bochum.de/

My present work is concentrated on real time acoustic simulation techniques using geometrical acoustics. I am currently involved in the audio aspects of a project that aims at creating an audio/visual virtual environment that simulates emergency situations. This virtual environment will be used as a training tool for people whose work involve managing emergency situations.

I am interested in the discussion of sound sources modelling, geometrical simplifications of the simulated environment, models for sound propagation in cities and perceptual based techniques. Besides I am interested in learning about how combined audio/visual perception can be used to improve current acoustic rendering systems and about audio/visual system integration.

Position paper

Roger D. Petersen (petersen@svpal.org)
Chairperson, Technology Committee (CCB)
California Council of the Blind/American Council of the Blind
http://www.acb.org/ccb

I come to this campfire as a blind person who navigates in an acoustical environment, with some hearing loss too I might add, an advocate for technology to be placed in the hands of blind people to "level the playing field" insofar as possible, and a former psychologist who had some training in psychoacoustics in a former life many years ago. I hope I can bring to bear the knowledge that now exists in the field of psychoacoustics upon the understanding of how acoustic information gathering can be optimized for blind people.

In a series of studies at Cornell University during World War II, Karl Dallenbach and his colleagues showed that a blind person's ability to perceive objects around him/her is auditory in nature, i.e., hearing is necessary and sufficient to observe the phenomenon, even though people often perceive the object as impinging on their faces rather than hearing it. Methodology was very primitive at that time and it might be useful to revisit those studies knowing what we now know.

Also, numerous investigators have proposed providing acoustic information artificially to the blind traveler's ears. However, we blind travelers must be shown that the information thus provided is superior and substitutable for the information we can get naturally. Otherwise we won't let you cover up our ears.

Thus, we can ask the questions:

Is there a way to optimize the information we can gather naturally, either by providing special hearing aids or a special sound transmitter for echo location?

Is there a way to provide information artificially to the auditory system of sufficient quality that we don't need to listen to the natural acoustic environment?

Or alternatively, can information be provided to my auditory system by artificial devices without covering my ears or masking the natural sound environment?

Sigurd Saue

Voxelvision AS * work: +47 73 87 36 97
Fjordgt. 56-58 * home: +47 74 85 70 67
P.O.Box 838 * fax: +47 73 87 36 99
N-7408 TRONDHEIM * email: sigurd.saue@voxelvision.com
NORWAY * URL: http://www.voxelvision.com

Background

M.Sc. in Acoustics with focus on music technology. Currently finishing a Ph.D. on sonification of spatial data. Several years of programming applications related to music research and sound installations.

Current research and development

My company, Voxelvision AS, develops 3D visualization software for the oil industry on the Windows NT platform. As a software designer my specific task is to introduce sound tools and to investigate possible applications of acoustic rendering of scientific data. I'm also engaged in a research project at Hydro Cave, adding sonification to their virtual seismic interpretation environment. On the side I still develop control applications for large, multichannel sound installations.

Position

My current research has three main objectives:

1. To establish a model for exploration of spatial data sets in auditory displays based on ecological acoustics.
2. To implement the main principles of the model in a platform and application independent sonification framework.
3. To explore the temporality of sound through different parameterizations of audible time.

I loosely define a spatial data set as (possibly multivariate) data that are defined on spatial coordinates of dimension 3 or less, but allowing time as a fourth dimension. Typical examples are exploratory seismic data and medical ultrasound or MR images. They constitute more or less immediate representations of real world structures.

The basic premise behind all my work is that sound is a temporal medium. Spatial relations are most effectively presented in a visual display. This is not a limitation in technology, but in perceptual ability. However, through a process of *temporalization* space could be mapped into time and subsequently rendered in an auditory display.

The model

How can we grasp 3D spatial relations through a 1D temporal medium? As perceiving subjects we do that all the time in our everyday world. A natural scene representation could exploit this innate capability. Ecological perception has already found relevance in auditory display design through Gaver's work on everyday listening[1]. His approach was however limited to the sound sources and less concerned with the active process of perception. We explore our world through movement, shifts of attention and decision making. An effective auditory display must model the interaction as well as the soundscape.

Many of the characteristic elements of virtuality (egocentric point-of-view, immersion, user-centered interaction, etc.) are applicable to auditory displays even if the accompanying visual display represents a flat, outside view. I therefore propose a general "sound walk" model for sonification of spatial data[2]: An active listener walks through the data set, listening to sounds locally and globally, and making new decisions along the way. The soundscape is organized around three kinds of objects: The listener's path, external sources and an area. The latter is a configurable limitation of the data in order to reduce perceptual complexity. It could be absolute (no sound outside) or a damping wall. The relevant sounds of the model are:

	NO. OF OBJECTS	SOUNDS	SPATIAL SOUND	POSITIONS
Listener	Single or multiple (comparisons)	MoveSounds, ExamineSounds	Near, dry	Listener only
Sources	Multiple	SourceSounds	Localized	Source and listener
Area	Single	AmbientSounds	Distant	Area only

The implementation

My goal has been to develop a general framework independent of the specific application, built as dynamically linked libraries. It connects to the application data through two abstract base classes, the data set (representing data values) and the data object (representing data structure/geometry). In order to integrate the sonification framework, the data set and object methods must be implemented in the host application.

The framework itself offers methods for mapping data to sound, for positioning and movement in the data set, for configuration and administration of interaction and soundscape. Platform specific code related to low level sound control, timing and thread synchronization is concentrated in a few modules, allowing simple preprocessor versioning at compile time.

The actual sonification is organized in two structures, a static mapping from data to sound, and a dynamic mapping from space to time. The static structure connects application data through one or more data transformations to normalized instrument parameters. The parameters does not have to be single-valued. Each instrument may carry information from several simultaneous data values (such as different seismic attributes). The instrument interface is very general and should give plenty of room for internal improvements, both perceptually and technologically.

The dynamic structure incorporates the main ideas from the model above. It connects data objects from the application to updateable position objects in the framework. Typical data objects are the entire data set, a surface or a path through data. The most important position is the listener, and in addition there could be spatially independent source positions and the surrounding area. Each data object is rendered through a set of players. They contain the static structure and update its elements according to various temporal strategies. All players relate to a possibly changing position. Source players relate to two positions, the listener and the source, transferring relative position to their instruments. All positions and all players may be updated independent of each other, mimicking the temporal richness of a natural scene.

Temporalization

The process of mapping spatially distributed data into the temporality of sound could be described in two steps: First we define a one-dimensional ordering of elements that has a well-defined mapping into sound (implicit time), and then we play through that sequence with given time intervals between elements (explicit time). This process might be instantaneous, as for the moving probe. In this case the user traverses data space interactively, defining both the sequence and the time interval between elements with the movements of an input device. Alternatively the trajectory through data is predefined or previously recorded, and then played back at constant speed. In the first approach the explication of time is movement driven, in the second it is constant.

More advanced temporal strategies permit the explication of time to be data driven. Both in my sonification and installation work I'm searching for perceptually meaningful ways to parameterize audible time. Among the promising parameters are density and metric regularity. The latter represents deviations from a regular meter and is implemented as frequency modulated rhythmic curves. Temporal density represents a direct control over the time interval between events. The perceptual effect depends strongly on time scale. On a fast, timbral level the density is perceived as changes in sound envelope. On an intermediate, rhythmic level it corresponds to changes in pulsing speeds. And on a slower event level it represents the frequency of occurrence of separate events, from overlaps to isolated sounds, a measure of soundscape activity.

Data driven time might be relevant for source sounds and ambient sounds. They add life and variation to potentially stationary objects. The source sounds primary function is to support orientation in large data sets. They should be spatially distributed. This will help the listener to perceptually organize the auditory scene, and to draw his visual attention to significant events in the data set. However, the application should not rely on the listener to hear the exact location.

I would like to discuss what role acoustic rendering could play in displays of scientific data. How can we exploit the capabilities of the different sense modalities most effectively? Visual and auditory perception work very well in parallel, but we must be aware of what information is best presented to which sense. Loosely speaking, vision deals with space and audition deals with time.

[1] Gaver, W.W. What in the world do we hear? An ecological approach to auditory source perception.*Ecological Psychology* 5, 1 (1993): 1-29

[2] Saue, S. A model for interaction in exploratory sonification displays. In*Proceedings of the International Conference on Auditory Display, ICAD 2000* (Georgia Institute of Technology, Atlanta, Georgia, USA, April 2-5, 2000).

Position paper

Robert Wall (robert.s.wall@vanderbilt.edu)
Vanderbilt University

Our group is working on a multifaceted project dealing with how people with visual impairments use sound to understand and navigate space. One portion of the project is attempting to devise a 3D representation of sound so that specific acoustic situations can be created in a laboratory setting. We are initially interested in the representation of vehicle trajectories but would like to be able to expand to build representation of entire intersections with complex traffic patterns and audible pedestrian signals. This is, perhaps, easy enough, but we want to do so using only two speakers. Eventually, we would like to be able to re-create the perception of 3D moving sound images,including above and behind a person, in a reverberant environment using a standard computer's speakers.

To work toward this goal, we are using two Tucker Davis Technologies set ups. In one lab we have a Power Dac and related system II components. In another lab attached to an anechoic chamber, we have an RP2 and related system III components. Each set of TDT hardware will be operated via MATLAB programming.

Actual representation of the 3D images is achieved by passing a signal through a set of filters using HRTFs. Currently we have managed to obtain very good externalization of stationary and moving sound phantoms in the horizontal plane with the sound moving 360 degrees around the listener. Intensity changes
will be used to alter the distance impression.Members of the team working on this project have backgrounds in psychoacoustics, developmental psychology, and special education. A varied lot with little firm grounding in 3D sound imaging or VR. This is why I was thrilled to hear of the Campfire. I hope to learn some of the basics that my self study has overlooked. I also want to find out about other hardware and software set ups in use for 3D acoustic VR and see how they compare to what we have devised. We have occasionally heard about devices on the market that purport to do what we are trying to accomplish but so far none of these devices have panned out. Usually, the on-the-fly adaptation and DSP of the filters is bypassed or truncated somehow in these systems so that the result is not as robust as we would like. The system III components from Tucker Davis should allow us to create what we want in the laboratory. Perhaps some swapping of MATLAB info would also be possible at the Campfire.

Jérôme Daniel
France Télécom R&D
DIH/IPS/ISI
Technopole Anticipa
2, Avenue Pierre Marzin
22307 Lannion Cedex
France
Phone: (+33) 2 96 05 27 96
Fax: (+33) 2 96 05 35 30
jerome.daniel@rd.francetelecom.fr
3D audio related web pages: http://gyronymo.free.fr

Position paper for the Campfire on Acoustic Rendering for Virtual Environments

Introduction: paper overview and additional resources

The area of expertise presented below in the first few sections issues substantially from my PhD thesis work, prepared at the Rennes Labs of France-Telecom R&D and recently defended (September 2000). It deals mainly with the reproduction techniques and the sound field representation that they are associated to, with the aim being to apply them to the 3D browsing in virtual environments. Among them, the ambisonic approach is more specifically developed: almost all of the aspects of the traditional first order systems are generalized to any higher order, for horizontal and full 3D reproduction configurations, and the usually referred psychoacoustic theories, based on the velocity and energy vectors, are thoroughly justified and interpreted.

For further information, the thesis document and the defense presentation (in french) are downloadable on my web pages (http://gyronymo.free.fr/audio3D/download_Thesis_PwPt.html) with english comments on each chapter, and an additional page (in english) gives commented sound and visual illustrations of higher order ambisonic rendering (see and hear: http://gyronymo.free.fr/audio3D/the_experimenter_corner.html). French and english abstracts are also available *via* http://gyronymo.free.fr/audio3D/accueil.html#lecture_audio3D.

The first section (3 pages) describes the ambisonic approach characteristics, the recent progress toward higher orders, and the expectations regarding its future.

The second section (3 other pages) opens a more general discussion on the reproduction techniques. The main classes of sound imaging strategies over loudspeakers (Amplitude Panning and Ambisonics, Transaural and Extended Transaural, WFS or Holophony) can be compared on the basis of acoustical considerations about the synthesized sound field. As a function of the chosen strategy and for given loudspeaker configurations, different compromises are achieved regarding the listening constraints, the satisfaction of natural localization mechanisms, the sound image accuracy, and the preservation of spatial qualities.

The third and last section (last page) briefly exposes current interests related to my recent activity in the 3D sound team of France Telecom R&D. Whereas the previous sections handle the reproduction, this last one deals with the content creation of virtual sound environments in a large sense, including the modeling of acoustical interactions (room effect, obstruction, etc…).

For these first two sections, the reproduction techniques are considered for their ability to reproduce the effect of each elementary event (wave front) of a pre-composed sound field, and in the end, to reproduce its macroscopic effect ("how preserved the global spatial qualities can be expected to be?").

First and higher order ambisonics

Brief overview

Ambisonics is worth being considered as a sound field *representation*, as a *sound imaging technique*, and as a whole reproduction *system*.

The ambisonic approach is based on *spherical harmonic decomposition of the acoustic field*, centered on the listener viewpoint. It has been known for a long time as a first order restricted form, which processes a minimal, **directional** sound field **encoding** through four components (B-format): W (pressure) and X, Y, Z (pressure gradient), offering easy sound field manipulations, such as rotations (see figures at http://gyronymo.free.fr/audio3D/accueil.html#choixsujet_audio3D). Ambisonic field can be encoded either acoustically, using a dedicated microphone, or synthetically, as a function of the directions of virtual sources and their associated reflections.

A **decoder** can be defined for various panoramic (2D) or periphonic (3D) loudspeaker rigs: it consists in matrixing ambisonic channels to feed the loudspeakers, in order to reproduce the original sound field at the listener place, or at least its perceptive effect. *Three primitive decoding solutions* had been defined for the first order systems to optimize the directional rendering in terms of the listening conditions: the LF-optimized (referred to as *"basic"*, later) and HF-optimized (*"max* r_E*"*) solutions, given by M.A.Gerzon for an ideal, centered listening, and *"in-phase"* decoding proposed by D.G.Malham for a collective, off-centered listening. Rendering can extend to headphones or a pair of loudspeakers *via* binaural techniques (virtual loudspeakers).

By considering in addition **higher order spherical harmonic components**, *the directional resolution* of the encoded sound scene *increases*. Quantitatively, the extended B-format consists of $K=(M+1)^2$ channels for a full 3D, M^th^ order representation, or only $K=2M+1$ channels for an horizontal restricted representation. The rendering requires more loudspeakers than ambisonic channels.

As a sound field representation based technique, Ambisonics is thus characterized by a **very appreciable versatility:**

- "Variable geometry" rendering (various loudspeaker configurations, plus possible headphone presentation)
- Ability to sound field transformations (rotations and perspectives deformations)
- "Variable resolution" sound field representation (scalability) used as a function of the transmission or/and the rendering capabilities
- "Variable listening area" decoding adaptability

As a system, Ambisonics has a quite **simple and low-cost implementation**, and offers **processing conveniences**:

All steps of the system are *simple linear operations* (substantially *matrix* operations, excepted the decoding for a binaural presentation), which are applied to the input or intermediary signals. These are: directional encoding of the sound field; optional sound field manipulations; optional mix of natural or synthetic sound fields; decoding (with optionally a low-cost *shelf-filtering*).

Note that the *decoding cost* doesn't depend on the original sound scene complexity (number of sources, reflections, etc.).

For a binaural presentation, decoding involves typically as many transfer functions as ambisonic channels. When dealing with many virtual sources, it can be interesting to use Ambisonics as an intermediate compact representation, in order to *factorize* the positional processing and to save CPU.

(Note: some emerging techniques dedicated to binaural synthesis do that with a better efficiency). Head-tracking can be handled by simply rotating the whole ambisonic field just before decoding.

As a rendering technique over loudspeakers, Ambisonics ensures **good predictability and homogeneity** of the rendered spatial qualities.

The encoding and decoding of each sound source (or phantom image) is equivalent to an *amplitude pan-pot*, thus the localization effect at the centered position can be predicted by the *velocity and energy vectors* ***V*** and ***E*** (ref thesis or any ambisonic related document). Since the decoder ensures that these vectors are compliant with the expected direction, *the directional information is preserved* (or controlled with virtual sources).

A *homogeneous* rendering is provided along all the directions; while ensuring the *loudspeaker "dematerialization"* (by avoiding to perceive them as individual sources). It also satisfies the naturalness of dynamic localization mechanisms (ITD and ILD variations due to head rotations, especially in low frequencies).

<u>1st order system limitations:</u> compared with the original "real" sound field experience, first order ambisonic rendering suffers from a lack of lateralization, which is felt as an elevation effect or as a *loss of image precision*. From a macroscopic point of view, considering a complex, reverberant field, the lack of lateral separation *may be* perceived as a partial loss of Spatial Impressions (S.I.) and envelopment (accompanied with a coloration effect).

<u>Using higher order harmonics</u>, which needs also more loudspeakers, allows to better benefit from the number of loudspeakers and their angular density (i.e. to use them more selectively). That way, the sound image robustness, its precision and the listening area are increased, and the spatial qualities better preserved thanks to a better lateral separation.

<u>Recent progress: theoretical developments and understanding</u>

Previous studies (Bamford95, Poletti96) have opened the way to the extension of ambisonic rendering to higher order, though offering partial view and extension of the approach. These have been completed by further studies (Daniel98, Nicol99, Furse&Malham99, Daniel00, etc.). In the following, I present the contributions issuing from my thesis work.

<u>Technical and mathematical aspects</u> [Ref chapter 3 of the thesis, plus defense presentation]

Most aspects of the traditional first order systems have been formally generalized to any higher order (for both 2D and 3D systems): the encoding, the decoding (major part of the work), and more partially the sound field transformations (rotations) and higher order microphone design.

For the **generic solving of the decoding problem**, underlying mathematics have been elucidated, in particular the directional sampling of the spherical harmonic basis (related to loudspeaker directions). Its regularity properties imply that the decoding matrix has a simple form, and that the local and global propagation properties (***V*** and ***E***) of the truncated sound field decomposition are preserved at the rendering. These concepts are also used for the design of higher order ambisonic microphones.

The primitive decoding solutions previously mentioned are generalized to higher orders into three families. They can be used separately or juxtaposed (per frequency band) to define an optimal decoder:

- The *"basic"* one optimizes the local centered reconstruction of the wave field (*i.e.* its extent regarding the wave length). It has to be used on a low frequency band, which narrows as the listening area extends.
- The *"max r_E"* one optimizes the "global propagation" ("global energy flow" ***E***), typically by "concentrating" the loudspeaker energy in the direction of the virtual source. It has to be used on the high frequency complementary band.
- The *"in-phase"* one minimizes directional artifacts and fluctuations when the listening area extends up to the loudspeaker perimeter.

Rendering prediction and characterization: "psychoacoustic" localization theories

Velocity and energy vectors (***V*** and ***E***, defined as the mean of the loudspeaker directions weighted by respectively the amplitude or the energy of their feedings) have been introduced by Gerzon (also referring to Makita) as representing the low and high frequency localization effect, and used as "psychoacoustic criteria" for the decoder optimization. It appeared necessary to clarify the foundations of these theories, in order to better characterize and interpret the expected spatial effect from these vectors.

For this purpose, ***V*** and ***E*** are first defined as characterizing respectively the local and the "global" sound propagation, then prediction laws of interaural difference are shown and their perceptive implications are interpreted as a function of head motions [sections 1.5, 2.2, 2.4 of the thesis]. The macroscopic interpretation (Spatial Impressions with a complex field) is also discussed.

An intrinsic link is shown between ambisonic representation (and its order *M*) and the potential properties of the rendered field (local reconstruction extent and "quality" of the global propagation), and as a consequence, the potential perceived spatial qualities (localization accuracy, image robustness, spatial impressions…).

Objective evaluations [Chapter 4] of localization cues (Spectra, ITD, ILD) issuing from the rendering confirm the contribution of higher orders and are correlated with the velocity and energy vector predictions. They are now supported by some additional sound demos (though with rather unrealistic examples: http://gyronymo.free.fr/audio3D/the_experimenter_corner.html).

Formal listening validations would have to be carried out. Moreover, generalized systems are still young or even not completely implemented. Their uses in interactive applications (within a complete spatialization environment, including room effect synthesis) still have to be more extensively experienced too.

The next future of higher order ambisonics

Extended ambisonic formats have certainly a future, but fast no past yet… How will they be used and found to be useful? The question involves many aspects.

- *A versatile use*: music or ambient sound recording; transmitting a room or space effect through *3D Impulse Responses*, mixing different sound scenes and factorizing positional processing, even for binaural presentation…
- *Rendering* high order ambisonics requires quite *a lot of loudspeakers*… as other rendering techniques like *Wave Field Synthesis* (see later) do. Thus adapted loudspeaker configurations are not a dream.
- *Implementation of extended B-format as an extension to the WAV-format* is being discussed.
- *A common destiny* of extended B-format: shared by Ambisonics and the binaural B-format strategy (Ref Jot, Larcher…)!
- *Sound field pickup*: higher order ambisonic microphones are in study. Their issue can be expected as a great step for the usefulness of ambisonic approach.
- There's a pool of ambisonics' defenders, still ready to promote such developments.

Opening a discussion: Ambisonics among other sound imaging techniques

The following discussion is based only on acoustical considerations about the synthesized sound field (and their perceptive implications), without worrying about system aspects like the transmission or the computation costs. The purpose is to highlight the potential of each strategy in terms of the sound image accuracy, the spatial quality preservation, the listening constraints and the satisfaction of natural localization mechanisms, all this, as a function of the number of loudspeakers involved. Sometimes, paradoxes will appear between the aim at satisfying natural hearing mechanisms, and the listening constraints. (Ref thesis + PowerPoint presentation: slide "principes de création d'image sonore" and the following ones).

Main classes of sound imaging strategies over loudspeakers

One can distinguish between at least three main classes of sound imaging strategies over loudspeakers:

- Using **amplitude differences** between loudspeaker signals (for each sound image), the loudspeakers being placed at the same distance from the center: *pair-wise pan-pot* (or reproduction issuing from *MS or XY stereophonic recordings*) and *Ambisonics*. Thus, **the contributing waves converge synchronously** at the center (thus **one focused point**), resulting (without the listener diffraction) in a **local, synthetic wave front** that has **uniform propagation properties** (apparent local direction and speed, characterized by the velocity vector) over the full frequency band (or over the bandwidths where amplitude ratio are constants), **extending** from the center **in proportion to the wavelength**.
- **Focusing on the field reconstruction at both ears** (thus **two focused points**, with account to the head diffraction): *Transaural or Stereo-Dipole, Double and Extended Transaural.*
- *Holophony (acoustic equivalent to Holography) /Wave Field Synthesis (WFS)*: **reconstructing the wave field over an area from its value on the area boundary** (Kirschhof Integral). Involving in practice a "sampled boundary", *i.e.* a **finite, discrete microphone/loudspeaker array**, reconstruction is quite **homogeneous over the whole area** for each frequency, but **spatial aliasing** occurs in a high frequency domain as a function of the spacing between loudspeakers.

A fourth class is omitted here – phantom source imaging using **time differences between loudspeakers** (issuing from *spaced microphones techniques*) – because it provides quite unpredictable (and wandering) sound images, though a better lateral decorrelation and enhanced spatial impressions, compared with reproduction issuing from coincident microphone techniques. (Note that it could be considered as a *very* degenerated case of holophonic methods.)

In the following, **we don't consider adaptive systems** (like head tracking cross-talk cancellation**).**

Comparison of systems will be made firstly with a *limited number of loudspeakers* (two speaker pan-pot, low order ambisonics, *versus* transaural and extended transaural) and a *single listener*, and secondly with *many loudspeakers* (high order ambisonics *versus* WFS/Holophony), with an *extended listening area or moving listeners*.

Preliminary: Some very general and evident laws

For the rendering of each elementary wave front, *interference figures*, which can be observed in the frequency domain, are created by combination of the contributing waves coming from loudspeakers.

*"In all cases, the interference figures have a size or a spatial periodicity that is typically **wavelength proportional**"*. This means that listening cues control becomes less stable or achievable as one considers a higher frequency domain, whereas things are quite easy with low frequencies, *i.e.* with wavelengths that are long enough regarding the listener scale. By the way, **all rendering techniques process similarly for (very) low frequencies**, and for a given loudspeaker configuration.

"It's as much difficult to reproduce the effect of a wave front (or a sound source), as its direction (or location) is far from the real, contributing sound sources (loudspeakers)".

- *"Difficult"* means "hard to achieve with stability and accuracy, or on a large area, or on a large frequency band". More technically, it needs more energy and implies the simultaneous participation of antagonist loudspeakers (thus a highly variant interference figure).
- *"The effect"* is in the end the perceptive effect, regarding static and dynamic listening mechanisms (localization cues and their variations by head rotation), or from an acoustic point of view, the sound field in the neighborhood of ears.

"The number of rendering control degrees is limited by the number of loudspeakers." The control degrees (or parameters) are typically the focused points (*e.g.* the ears, or the center) for the sound field reconstruction, and the axis along which variations are considered.

Limited number of loudspeakers, individual, centered listening

(Amplitude Pan-pot and Low Order Ambisonics *versus* Extended Transaural)

With only two frontal loudspeakers (traditional stereo *versus* transaural or stereo-dipole)

What is lacking: *traditional stereo* can control the direction (only frontal) of a synthetic wave front, but not its apparent propagation speed (not the natural sound celerity), while cross-talk cancellation is achieved only for given ear positions with *transaural*. As a consequence, variations of localization cues (especially ITD) by *slight head rotation* cannot be natural. This can be perceptively interpreted as: either an elevation effect (“under-lateralization”) for images between loudspeakers; or a directional move (“over-lateralization”) for images outside the loudspeaker span (only with transaural).

Sound scene extent and image accuracy: because of the cross-talk, traditional stereo offers only a smeared localization effect (predicted by the energy vector ***E*** in HF), especially for central images, and confines the sound scene within the frontal loudspeaker interval; transaural offers theoretically a full 3D sound scene with strong phantom images, but back-front reversals occur, probably because of contradictory cues variations by head rotation.

What are the freedom degrees: in both cases: moves are not critical in the median plane of the loudspeakers, including the front-back axis.

Image stability is critical with lateral head movements, depending on the lateral extent of the interference figure and its variance (amplitude). This lateral extent increases as the frequency decreases and as the loudspeakers narrow.

Compromise regarding the loudspeaker angle: in traditional stereo, loudspeakers are placed at +-30° as a compromise between a "not too confined" sound scene and a "not too poor" central imaging; applying the transaural approach to a +-5° speaker positioning (*stereo-dipole*) greatly enlarges the interference figure, thus the phantom image stability. [Footnote: Jerry Bauck "hybrid" system with two frontal pairs (substantially: transaural for low frequencies, stereo-dipole for higher frequencies).]

With four loudspeakers (1st order horizontal ambisonics *versus* double-transaural strategies)

What's improved: *With ambisonics*, sound scene extends to the full surround, while allowing synthetic wave fronts to have a natural propagation speed (thus correct dynamic lateralization in LF), but HF localization cues (ITD, ILD, spectral cues) still being smeared. *With double-transaural, i.e.* a transaural process distributed over a frontal and a back speaker pairs (ref Olivier Warusfel, IRCAM, or J-M Jot for binaural B-format rendering), back-front reversals do not longer occur, and it is even possible with slight refinement (proposed in my thesis) to provide natural ITD variations with slight (yaw) head rotations (at least with LF).

New constraints! Because both frontal and back loudspeaker pairs participate (especially for *lateral* virtual sources), an interference effect appears along the front-back axis, and the ear signal reconstruction is no longer stable considering front-back moves. A second positional constraint is added.

Paradox and critical situation for the "double-transaural": The "double transaural" and especially the "double-stereo-dipole" (speakers at +-5° and 180+-5°) are expected to be very unfavorable to *lateral* virtual sources, forcing to a very strict ear positioning along the front-back axis with regard to small wavelengths (HF). This is in contradiction with the aim to allow slight head rotations and to satisfy dynamic localization mechanisms.

The problem stands in the fact that these are *minimal layouts regarding the number of parameters* to be controlled (here: four). There's *no such problem with Ambisonics* (centered focused point), for which *cross-talk is anyway involved* in sound image illusion and localization for any head orientation, though it smears HF cues and cannot provide images as precise and strong as Transaural ideally does.

[Note that the comparison could extend to a "minimal" 3D (*e.g.* cubic) configuration: a new positional constraint (along the vertical axis) is added in this case.]

Increasing the number of loudspeakers: This paradox is progressively removed when loudspeakers are added without increasing the number of control parameters (*i.e.* without adding new dimensions). Comparatively, with more loudspeakers and higher order ambisonics, HF cues are less and less smeared while always featuring a natural dynamic localization.

It is likely that both kinds of rendering would converge, but Ambisonics is much easier to implement than Extended Transaural.

Many loudspeakers for an extended area (High Order Ambisonics *versus* WFS)

A concise comparison is given is the following table.

Rendering properties	**High Order Ambisonics**	**Wave Field Synthesis**
Sound field reconstruction (as the order increases)	Radial expansion (*kr*), wavelength proportional	Spectral expansion (*f*), uniform over the area
Loc. characterization outside the reconstruction domain	Energy vector ***E*** (HF/off-centered)	No prediction (above the spatial aliasing frequency)
Reference viewpoint	Unique (centered listener) and extrapolated	Global
Sound image projection (converging point of perceived directions from all listening positions)	Over the loudspeaker array (like usual visual image projections) (see comment *)	Beyond the array, with respect to the original source distance (like holographic images)

(*) To be more exact, high order ambisonics is able to reproduce the effect of sound sources *beyond* the loudspeaker array, but *only within the reconstruction domain*: it only requires compensating the near field effect of the loudspeakers.

The last two lines of the table introduce *the question of the audio-visual coherency*, since a true holographic visual rendering is not achieved. Such a coherency seems to be better achieved with High Order Ambisonics, which tends to act as the usual, visual projection (with the image corresponding to *one* viewpoint). Despite of absolute directional distortions perceived at off-center positions, perspective information is preserved through the relation between direct and reverberated sound. However, the level distortion caused by loudspeaker proximity can be a problem and its effect should be further evaluated.

Conclusion: Systems have been compared on the basis of objective arguments. It would be worth confronting these expectations to practical, audible experiences!

Bibliography

Refer to the work of: Gerzon, Malham, Bamford, Poletti, Daniel, Nicol for Ambisonics; Larcher, Jot, Warusfel for binaural B-format and double-transaural; Nicol and Delft University for WFS/Holophony.

Current activity and special interests: acoustic modeling and content creation tools

As a complement to the question of the sound field reproduction (or sound imaging) treated just above, my current interests are rather concerned with:

1. The content production of virtual sound environments;
2. The efficient integration of advanced technologies using existing hardware.

The first point, beyond the *ergonomics* of Human-Computer Interfaces of content creation tools, involves several aspects: *refinement of virtual acoustic modeling* for a more immersing and interactive rendering (room effect and coupling, occlusion and obstruction); its *translation* into parameters of standardized description formats; the extension of *description formats*.

The second purpose deals with technical questions such as the description formats, the plat-form variability, and the repartition of processing tasks between hardware and software. Regarding current API features, an additional question is *the control or the choice of the sound imaging technique* at the final stage of the rendering (which doesn't seem to be proposed yet).

It is hoped that the emphasized aspects will be further discussed during the Campfire.

Campfire participants

Rinus Boone
Lab. of Acoustical Imaging and
Sound Control,
Delft Univ., The Netherlands
rinus@akst.tn.tudelft.nl

Alan Chalmers
Dpt. of Computer Science,
Bristol University, UK
alan@cs.bris.ac.uk

Jiashu Chen
Agere systems, USA
jiashuchen@agere.com

Jerome Daniel
France Telecom, R&D, France
jerome.daniel@rd.francetelecom.fr

Robert Essert
ARUP, UK
Robert.Essert@arup.com

Terrence Fernando
Center for Virtual
Environments,
Univ. of Salford, UK
T.Fernando@salford.ac.uk

Patrick Flanagan
Lake DSP, Australia
P.Flanagan@lake.com.au

Hesham Fouad
VRSonic Inc., USA
fouads@bellatlantic.net

Thomas Funkhouser
Computer Science Dpt.,
Princeton University, USA
funk@cs.princeton.edu

Joachim Gossman
GMD National Research Center
for Information Technology,
Germany
joachim.gossmann@gmd.de

Matti Gröhn
Helsinki University of Technology,
Finland
mgrohn@csc.fi

Murray Hodgson
Dpt. of Mechanical Eng.,
Univ. of British Columbia, Canada
hodgson@mech.ubc.ca

Ulrich Horbach
Studer Professional Audio AG,
Switzerland
ulrich.horbach@studer.ch

Marty Johnson
Vibration and acoustics lab,
Virginia tech, USA
martyj@vt.edu

Jean-Marc Jot
Creative, USA
jmj@atc.creative.com

Véronique Larcher
GENESIS, France
veronique.larcher@genesis.ac

Tapio Lokki
Helsinki University of Technology,
Finland
Tapio.Lokki@hut.fi

Peter Lunden
The Interactive Institute,
Emotional and Intellectual
Interfaces, Sweden
ludde@kkh.se

Julien Maillard
CSTB, France
j.maillard@cstb.fr

Klaus Naßhan
German Physical Society,
Fraunhofer-Institut für Bauphysik,
Germany
nasshan@ibp.fhg.de

Pedro Novo
Institut of Communication
Acoustics,
University of Bochum,
Germany
novo@ika.ruhr-uni-bochum.de

Roger Petersen
Technology Committee chair,
California Council of the Blind,
USA
petersen@svpal.org

Klaus Riederer
Helsinki University of Technology,
Finland
kar@cc.hut.fi

Sigurd Saue
Voxelvision AS, Norway
sigurd.saue@voxelvision.com

Barbara Shinn-Cunningham
Hearing Research Center,
Boston University, USA
shinn@cns.bu.edu

Peter Svensson
Dpt. of telecom.,
Norwegian Univ. of Science and
Technology, Norway
svensson@tele.ntnu.no

Rendell Torres
Dpt. of Applied Acoustics,
Chalmers University of
Technology, Sweden
rendell@ta.chalmers.se

Nicolas Tsingos
Bell Laboratories/Lucent
Technologies, USA
tsingos@lucent.com

Johan Vävare
ADL Konsult AB, Sweden
Asia Magic Advanced Technology
Ltd., Thailand
johan@asia-magic.co.th

Robert Wall
Vanderbilt Univ., USA
robert.s.wall@vanderbilt.edu

Tomas Weber
Center for Parallel Computers,
Royal Institute of Technology,
Sweden
snus@pdc.kth.se

Ying Zhang,
Center for Virtual Environments,
Univ. of Salford, UK
y.zhang1@pgr.salford.ac.uk

- NOTES -

- NOTES -

- NOTES -

- NOTES -