3D Reproduction of Room Acoustics using a Hybrid System of Combined Crosstalk Cancellation and Ambisonics Playback

S. Pelzer[1], B. Masiero[1], M. Vorländer[1]

[1] *Institute of Technical Acoustics, RWTH Aachen University, Germany, Email: spe@akustik.rwth-aachen.de*

Abstract

Room acoustic simulations have evolved to hybrid models with a deterministic and computationally expensive algorithm for a precise calculation of the early specular reflections and a stochastic and computationally efficient algorithm for the later part of the impulse response. Splitting the impulse response into an early and a late part is reasonable in a psychoacoustical sense, since the early part is responsible for the localization of sources, making the correct reproduction of its specific time-frequency structure important, while the later part is responsible for the sense of spaciousness, which mainly depends on the spectral, temporal and spatial envelope of the decay but not on the fine structure.

Nevertheless, in virtual reality systems the reproduction of the whole impulse response is done through the same reproduction system, even though there are systems more prone to coherent reproduction (important for the early arrivals of an impulse response) and others more prone to the reproduction of random-phase fields (the reverberant tail of an impulse response).

A hybrid reproduction approach is presented which uses an eight-channel loudspeaker system for both the playback of a binaural signal containing direct sound and early reflections via crosstalk cancellation, as well as the playback of a 2nd order Ambisonics signal containing the diffuse decay.

Introduction

Today's state-of-the-art room acoustic simulation tools [1][2][3] usually base on the image source (IS) model introduced by Allen and Berkley [4] in combination with acoustic ray tracing (RT) as proposed by Krokstad [5]. The IS method is well suited for a deterministic modeling of precise early reflections, while the stochastic RT technique is capable of effective modeling of much higher reflection orders, including important scattering effects. More accurate acoustic modeling algorithms, such as Finite- and Boundary-Element-Methods, suffer from high computational demands and are hardly applicable for normal to larger rooms or full broadband simulations, while geometrical acoustic modeling already achieved real-time capabilities [6].

Sound propagation in rooms and mixing time

Room impulse responses can generally be divided into an early part which is dominated by distinct strong early reflections and a late part that mainly consists of reflections which have been reflected and scattered several times, so that they thoroughly overlap due to increased reflection density over time and the broadening of the impulses with higher reflection orders.

Many attempts have been made to define the transition time between these two parts on a physical basis [7][8], but recent conclusions show that physical mixing does not explain diffusion and does not define the moment when a sound field turns diffuse [9]. Furthermore it must be asked how many real rooms exist which establish a perfectly diffuse reverberation?

A detailed comprehensive overview of physical predictors for the estimation of the transition time as well as their evaluation on a perceptual basis can be found in a recent publication by Lindau [10]. The investigated predictors comprised model-based ones (deriving the transition time from room parameters such as volume and mean free path length) as well as impulse response based ones (analyzing the time domain impulse response).

Shoebox shaped rooms usually have longer mixing times, due to their long unobstructed path length and regular shape. For these enclosures, Lindau found a transition time t_m, proportional to the mean free path length, with $t_m = 20\frac{V}{S} + 12\ [ms]$, *V* being the room volume and *S* the room's surface area. Absorption and reverberation time were not found to have significant influence.

Room acoustics simulation: Image sources and ray tracing

Going back to the IS model for prediction of early reflections, we find that the time range in an impulse response that is covered by a constant order of image sources is proportional to the mean free path length, just as the transition time itself, as proposed by Lindau. This concludes to the necessary image source order O_{IS} being a constant factor between mean free travel time $\bar{t} = 4\frac{V}{cS}$ (*c*: speed of sound) and transition time t_m:

$$O_{IS} \cdot \bar{t} = t_m \qquad (1)$$

For a simplified estimation of the image source order, the additional *12ms* in the transition time formula will be neglected in favor of a full additional order of image sources, which is a valid approximation for even small rooms with at least *4m* of mean free path length. Including this simplification, the necessary image sources order can be estimated independently of reverberation time, volume or absorption to $O_{IS,min}$, with:

$$O_{IS,min} = \frac{t_{m-12}}{\bar{t}} + 1 \approx 2.7 \qquad (2)$$

This means that for rooms, as selected by Lindau, which had shoebox shape and volumes in a wide range from *182m³* up to *8500m³*, each with varied mean absorption, a minimally

necessary IS order can be defined and this results to a number of three. After the third reflection, the sound field can be expected to be mostly mixing, uniform and isotropic, yielding a diffuse late reverberation. Similar observations were already made by Kuttruff and published in [11].

Scattered reflections in the early part und all reflections after the image sources cut-off time are then calculated using the ray tracing technique which builds temporal envelopes in certain frequency bands.

Spatial sound reproduction in 3D

Not only the simulation algorithms should be optimized and matched to the acoustic effects they are meant to predict, but also on the side of the reproduction of simulated sound fields there must be commitment to the psychoacoustical demands of the different stages of sound propagation in rooms.

Early reflections and especially the direct sound have to be reproduced with highest precision in terms of time and direction of arrival and frequency spectrum. Due to the precedence effect, the direct sound has a major influence on the localization of a source and the early reflections will affect the perceived source width. The reproduction system has to make sure that localization is as natural as possible, including exact compliance with frequency-dependent interaural level and time differences.

Many spatial reproduction techniques miss the point of three-dimensionality. A full 3D reproduction includes not only horizontally distributed sources, but also the incidence from elevated angles and near field effects for sources that are close to the head of the listener [12].

Even large and expensive wave field synthesis (WFS) systems usually miss out on height information. More commonly used and more affordable systems such as vector-base amplitude panning (VBAP) and Ambisonics can theoretically reproduce elevated sources, but there are only few implementations that support realistic distance perception. VBAP has no support for close-by sources and Ambisonics only in near-field compensated higher order setups (NFC-HOA) [13].

On the other hand, binaural technology might be the 3D technology that is the closest one to the way how the human ear perceives sound in nature. It is best implemented when using individual HRTF data. But even then, as a major disadvantage it is difficult to play back through a loudspeaker system. Using headphones on the other hand is not only problematic in terms of comfort and externalization, but also usually not able to impart the feeling of envelopment in diffuse sound fields. Additional problems such as the necessity to compensate for individual headphone transfer functions accrue.

A promising technology was found in the crosstalk cancellation (CTC) [14][15], also called *transaural* in some publications. It uses a regular loudspeaker system, with only two speakers required, and takes advantage of wave interference to achieve a sufficient channel separation between the left and right ear of the listener. The main drawback of this system is the requirement to accurately know the current position of the user, which is typically solved using a tracking system and continuous adaption of the CTC filters [16]. Thus, this technique is often found in virtual reality systems, when the user is already tracked for interaction or 3D visualization.

Reproduction technique	Advantages	Drawbacks
Binaural CTC	Precise and easy localization Good readability *Near field sources*	Poor realism and lack of immersion/ envelopment *Needs individual HRTF*
Ambisonics	Strong immersion and envelopment	Poor localization readability
Stereo Panning	Very precise localization	Lack of immersion/ envelopment

Table 1: Comparison of different reproduction techniques, as published by Guastavino [17]. Additional comments by this author are in *italics* and colored light green.

Guastavino et al. [17] compared different reproduction techniques (CTC, Ambisonics, Panning) and came to similar results as described above and summarized in Table 1 (with additional comments by this author). This makes obvious that just as for the simulation algorithms, also the reproduction side has to adapt to the particular purpose and orient itself at psychoacoustic effect and phenomena in room acoustics.

Hybrid reproduction systems

Favrot proposed to apply the idea of hybrid reproduction that is matched to the events in a RIR to room acoustics prediction models which can generate spatial IRs for existing or virtual halls [18]. He used a variable Ambisonics order for the early and late part of the RIR to benefit from reduced computation load for late reverberation and better localization of the direct sound.

Virtual room acoustics with hybrid 3D reproduction

In this present contribution a combined hybrid system is introduced that uses one common loudspeaker system to play a CTC and an Ambisonics signal at the same time. The binaural signal will ensure high detail of temporal and spectral features of the direct sound and early reflections, while the Ambisonics signal is used to produce a spacious and enveloping diffuse sound field. The poor localization abilities of Ambisonics are published in a variety of studies [17][19], and the poor immersion of binaural or transaural reproduction is documented as well [17]. Both observations clearly motivate the hybrid approach where binaural signals are used for the direct sound and early reflections and Ambisonics for the late decay.

Table 1 shows how the Pros and Cons of these two technologies are close to being perfectly complementary. Minor open questions such as the benefit and effort of binaural signals that use individual head-related transfer functions (HRTF) are recently investigated and the interested reader is kindly referred to a recent publication about very fast acquisition of individual HRTF [20] (measurement duration of roughly 7 minutes for 4000 directions).

The earlier introduced transition time is perfectly qualified to define the crossover between the two reproduction systems, with the same motivation as for the simulation. Therefore the CTC is used to reproduce the direct sound and specular reflections up the order of 3. Further reflection paths and all scattered reflections are fed into the Ambisonics engine.

The presented idea is not meant to replace any cinema or public address system, due to the fact that the CTC is a single-user experience. It is more aimed at sophisticated room acoustics simulation and reproduction in virtual acoustics applications, such as virtual concert hall prototyping or fully immersive virtual environments.

In the proposed hybrid system Ambisonics is only used for late reflections, therefore the usual implementation of plane wave sources is sufficient and near-field compensation [13] is not applied.

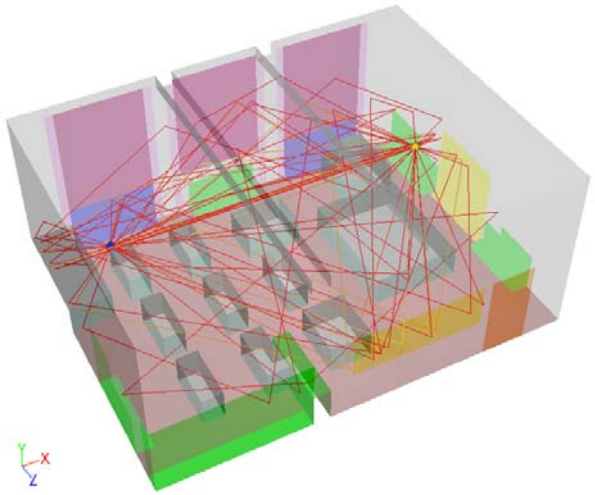

Figure 1: Classroom CAD model that is simulated with the image source model. Reflections up to order 3 are shown.

Room impulse response filter synthesis

After the design of a CAD room model and its parametrization, including material properties, source positions/directivities and receiver positions/HRTF, the IS model will return the positions and spectra of audible image sources and the ray tracer returns spatially discretized time-frequency energy histograms. To auralize the virtual scene, this information can now be translated into actual impulse responses. As proposed, the early reflections part is rendered into a binaural IR, while the scattered and late reflections are used to build an Ambisonics B-format IR.

Binaural synthesis

To generate a binaural filter, each audible image source is attenuated according to the distance law for spherical sources. The absorption coefficients of all walls in the reflection path are combined to a spectral filter which is then convolved with the source directivity. The last step of this filter chain adds the spatial information by including the HRTF data for the correct sight angle of the image source.

For sources closer than *2m* appropriate HRTF data is used that is measured in the near field (*0.2, 0.3, 0.4, 0.5, 0.75, 1.0, 2.0 meters*) [16]. If no such near field data is available, a range extrapolation should be applied, as proposed by Pollow [21].

To enable comparisons between different 3D reproduction systems, the binaural IR can also be extended to comprise all late reflections, so that the full room impulse response is ready for playback trough a CTC system.

Ambisonics B-format synthesis

The late reverberation is predicted using a ray tracer. Thus, the simulation result is a data structure that contains the amount of energy that is arriving from a certain direction at a certain time in a certain frequency band, as shown in Figure 2. The temporal, spectral and spatial domains are discretized, usually in accordance with the number of rays for the desired resolutions.

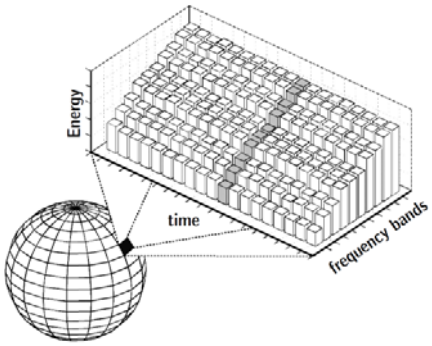

Figure 2: Acoustic ray tracing results in a spatial data structure with time-frequency information of the energy of incident rays for each detection sphere (after [22]).

Virtual Environment with hybrid reproduction

A very interesting application for the hybrid 3D sound system is found in virtual environments. Both 3D visualization as well as 3D sound reproduction need each other and benefit from multi-modal stimulation. If 3D sound is combined with appropriate graphics representation, the immersion is massively increased. Visual and audio cues are combined by the human sensory system, and the listening experience, including multi-modal tasks such as source localization, are presented in a manner close to the natural experience in real life.

Real-time implementation for virtual reality

Virtual Reality (VR) systems usually feature a tracking system for stereoscopic displays, so that the CTC can be driven dynamically. The late reverberation that uses much longer computation time due to the ray tracer only need few updates and only if the user moves already several meters or from one room to another. Ambisonics playback can stay stationary most of the time, without losing spatial cues. The B-format IR construction needs more channels (only 2 for binaural, but $(N+1)^2$ for the N^{th} Ambisonics order), but there is no HRTF data required and no costly convolutions in the filter synthesis. Decoding B-format to the available loudspeaker setup is only a simple matrix multiplication, as described earlier, so that this method is well suited for use in real-time applications.

The Institute of Technical Acoustics at RWTH Aachen University implemented the hybrid reproduction technique in their VR laboratories [6][16][23]. The system features a 3D stereoscopic display (using polarization), an electromagnetic motion tracker as well as fully dynamical crosstalk cancelled 3D audio, currently using four loudspeakers. In total, eight loudspeakers are available arranged in a cube configuration and thus ready for Ambisonics playback of 1st or 2nd order signals. A real-time spatial audio mixing console that was developed at ITA allows comparison of CTC, Ambisonics and VBAP. The real-time room acoustics simulation (RAVEN [6]) that was also used for this contribution is to be connected to this reproduction system for fully immersive and interactive acoustic and visual walk-throughs. Therefore a small compute cluster with 4 Quadcore PCs is available; two of them prepared for high performance GPU operations.

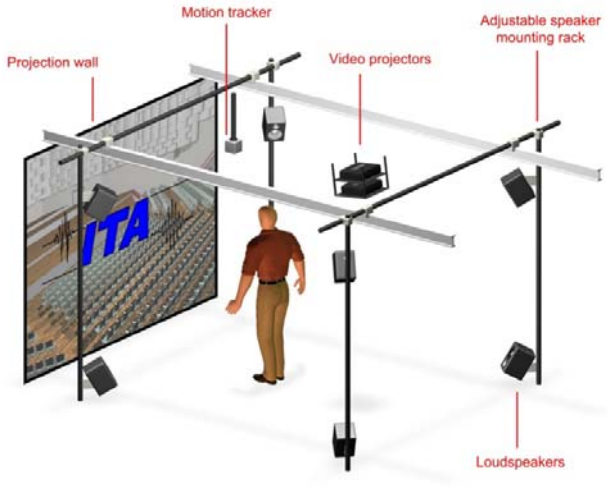

Figure 3: Virtual reality laboratory of the Institute of Technical Acoustics at RWTH Aachen University. The system features a 3D projection wall and a sound reproduction system with 8 K&H O100 loudspeakers in variable configurations. For the hybrid CTC/Ambisonics playback the loudspeakers stay at the corners of a cube. The user's head position can be detected using electromagnetic tracking for virtual reality applications.

Conclusion and outlook

This contribution proposes a realistic and natural sounding high quality auralization of sound sources in enclosures. Direct sound of a virtual source is conveyed to the listener using a precise reconstruction of the appropriate sound pressure time signals at his eardrums including all important cues that are interpreted by the human auditory system for the localization and identification of sound sources just like in nature. This is achieved by using binaural technology including near-field effects for close sources and employment of individual head-related transfer functions. For a loudspeaker based reproduction of this binaural signal, the crosstalk cancellation technique is applied.

Binaural technique (including crosstalk cancellation) is applied to auralize all early reflections of the virtual room the user is located in. These are calculated using the image source model. It was shown that after three surface reflections the sound-field can be expected to mix so that isotropic diffuse reflections dominate. These late reflections are modeled using a ray tracer which returns information about the time, spectrum and direction of energy that arrives at the listener. These results are transformed into a spatial Ambisonics B-format impulse response which contains all scattered and late reflections.

This 2nd order Ambisonics reverberation signal is decoded to 8 channels for playback on a cubical loudspeaker array. The same loudspeakers are used to play the crosstalk cancellation for direct sound and all early reflections. Multiple sources can be auralized by simple superposition.

The proposed hybrid approach was very convincing in preliminary listening tests. The subjects reported a precise localization of sound sources joined with the feeling of engulfment and envelopment. Known problems such as inside-head localization did not occur. Especially in halls with long reverberation the immersion was on a very high level and the listening experience felt natural.

The simulation and reproduction including all convolutions and dynamically adapting crosstalk cancellation is already fully real-time capable. Due to the modular concept of simulation stages, convolution and reproduction and the availability of a small compute cluster, the parts are already in the process of joining wires. When finished, the entire system will enable the motion tracked user to freely move through virtual scenes that are auralized in real-time.

References

[1] G. M. Naylor. ODEON - another hybrid room acoustical model. Applied Acoustics, 38:131, 1993

[2] CATT-Acoustic, www.catt.se

[3] Ahnert, W.; Feistel, R.: EARS Auralization Software. J. Audio Eng. Soc. Vol. 41 (11), 897-904, 1993

[4] Allen, J. B. and Berkley, D. A.: Image Method for Efficiently Simulating Small-Room Acoustics. Journal of the Acoustical Society of America, 65:943, 1979

[5] Krokstad, A.; Strom, S.; Sorsdal, S.: Calculating the acoustical room response by the use of a ray tracing technique. Journal of Sound and Vibration 8 (1968), Nr. 1, S. 118–125

[6] Schröder, D.: Physically based real-time auralization of interactive virtual environments. PhD thesis, RWTH Aachen University, 2011

[7] Reilly, A.; McGrath, D.; Dalenbäck, B.-I.: Using Auralisation for Creating Animated 3-D Sound Fields Across Multiple Speakers. Proc. of the 99th AES Conv., New York, 1995, preprint no. 4127

[8] Meesawat, K; Hammershøi, D.: The time when the reverberant tail in binaural room impulse response begins. Proc. of the 115th AES Conv., New York, 2003, preprint no. 5859

[9] Polack, J.-D.: Is mixing the source of diffusion? J. Acoust. Soc. Am. Volume 129, Issue 4, pp. 2502-2502, 2011

[10] Lindau, A.: Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses. Proceedings of the AES 128th Convention, London, UK, 2010 May 22–25

[11] Kuttruff, H.: Room Acoustics. 4th ed., New York: Routledge Chapman & Hall, 2000

[12] Masiero, B.; Vorländer, M.: Spatial Audio Reproduction Methods for Virtual Reality. Proceeding of the 42º congreso español de acústica, Caceres, Spain, 2011

[13] Daniel, J.: Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format. AES 23rd International Conference, Copenhagen, Denmark, 2003 May 23-25

[14] Bauer, B. B.: Stereophonic Earphones and Binaural Loudspeakers. Journal of the Audio Engineering Society, vol. 9, no. 2, pp. 148-151, 1961

[15] Atal, B. S.; Hill, M.; Schroeder, M. R.: Apparent Sound Source Translator. U.S. Patent 3,236,949

[16] Lentz, T.: Binaural technology for virtual reality. PhD thesis, RWTH Aachen University, 2011

[17] Guastavino, C. et al. :Spatial audio quality evaluation: comparing transaural, ambisonics and stereo. Proceedings of the 13th International Conference on Auditory Display, Montréal, Canada, June 26-29, 2007

[18] Favrot, S.; Buchholz, J. M.: LoRA: A Loudspeaker-Based Room Auralization System. ACTA ACUSTICA UNITED WITH ACUSTICA Vol. 96 (2010) 364 – 375

[19] Pulkki, V.: Evaluating Spatial Sound with Binaural Auditory Model. Proceedings of the International Computer Music Conference, pp. 73–76, Havana, Cuba, Sep. 2001.

[20] Masiero, B.; Pollow, M.; Fels, J.: Design of a Fast Broadband Individual Head-Related Transfer Function Measurement System. Acustica, Hirzel, 2011, Vol. 97, pp. 136-136

[21] Pollow, M.: Applying extrapolation and interpolation methods to measured and simulated HRTF data using spherical harmonic decomposition. Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics May 6-7, Paris, France, 2010

[22] Lentz, T., Schröder, D., Vorländer, M., Assenmacher, I.: Virtual Reality System with Integrated Sound Field Simulation and Reproduction. EURASIP Journal on Applied Signal Processing, Special Issue on Spatial Sound and Virtual Acoustics, 2007

[23] Schröder, D. et al.: Virtual Reality System at RWTH Aachen University. Proceedings of the International Symposium on Room Acoustics (ISRA), Melbourne, 2011

[24] Solvang, A.: Spectral impairment for two-dimensional higher order Ambisonics. Journal of the Audio Engineering Society 56 267–279, 2008

[25] Daniel, J.: Representation de champs acoustiques, application a la transmission et a la reproduction de scenes sonores complexes dans un contexte multimedia (in french). PhD thesis, 1996-2000 Universite Paris 6, 2000