Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 70540, 19 pages
doi:10.1155/2007/70540

Research Article

Virtual Reality System with Integrated Sound Field Simulation and Reproduction

Tobias Lentz,[1] Dirk Schröder,[1] Michael Vorländer,[1] and Ingo Assenmacher[2]

[1] *Institute of Technical Acoustics, RWTH Aachen University, Neustrasse 50, 52066 Aachen, Germany*
[2] *Virtual Reality Group, RWTH Aachen University, Seffenter Weg 23, 52074 Aachen, Germany*

Received 1 May 2006; Revised 2 January 2007; Accepted 3 January 2007

Recommended by Tapio Lokki

A real-time audio rendering system is introduced which combines a full room-specific simulation, dynamic crosstalk cancellation, and multitrack binaural synthesis for virtual acoustical imaging. The system is applicable for any room shape (normal, long, flat, coupled), independent of the a priori assumption of a diffuse sound field. This provides the possibility of simulating indoor or outdoor spatially distributed, freely movable sources and a moving listener in virtual environments. In addition to that, near-to-head sources can be simulated by using measured near-field HRTFs. The reproduction component consists of a headphone-free reproduction by dynamic crosstalk cancellation. The focus of the project is mainly on the integration and interaction of all involved subsystems. It is demonstrated that the system is capable of real-time room simulation and reproduction and, thus, can be used as a reliable platform for further research on VR applications.

1. INTRODUCTION

Virtual reality (VR) is an environment generated in the computer with which the user can operate and interact in real time. One characteristic of VR is a three-dimensional and multimodal interface between a computer and a human being. In the fields of science, engineering, and entertainment, these tools are well established in several applications. Visualization in VR is usually the technology of primary interest. Acoustics in VR (auralization, sonification) is not present to same extent and is often just added as an effect and without any plausible reference to the virtual scene. The method of auralization with real-time performance can be integrated into the technology of "virtual reality."

The process of generating the cues for the respective senses (3D image, 3D audio, etc.) is called "rendering." Apparently, simple scenes of interaction, for instance, when a person is leaving a room and closes a door, require complex models of room acoustics and sound insulation. Otherwise, it is likely that coloration, loudness, and timbre of sound within and between the rooms are not sufficiently represented. Another example is the interactive movement of a sounding object behind a barrier or inside an opening of a structure, so that the object is no longer visible but can be heard by diffraction.

1.1. Sound field modeling

The task of producing a realistic acoustic perception, localization, and identification is a big challenge. In contrast to the visual representation, acoustics deal with a frequency range involving three orders of magnitude (20 Hz to 20 kHz and wavelengths from about 20 m to 2 cm). Neither approximations of small wavelengths nor large wavelengths can be assumed with general validity. Different physical laws, that is, diffraction at low frequencies, scattering at high frequencies, and specular reflections have to be applied to generate a physically based sound field modeling. Hence, from the physical point of view (this means, not to mention the challenge of implementation), the question of modeling and simulation of an exact virtual sound is by orders of magnitude more difficult than the task to create visual images. This might be the reason for the delayed implementation of acoustic components in virtual environments.

At present, personal computers are just capable of simulating plausible acoustical effects in real time. To reach this goal, numerous approximations will still have to be made. The ultimate aim for the resulting sound is not to be physically absolutely correct, but perceptually plausible. Knowledge about human sound perception is, therefore, a very important prerequisite for evaluating auralized sounds.

Cognition of the environment itself, external events, and—very important—a feedback of one's own actions are supported by the hearing event. Especially in VR environments, the user's immersion into the computer-generated scenery is a very important aspect. In that sense, immersion can be defined as addressing all human sensory subsystems in a natural way. As recipients, humans evaluate the diverse characteristics of the total sound segregated into the individual objects. Furthermore, they evaluate the environment itself, its size, and the mean absorption (state of furniture or fitting). In the case of an acoustic scene in a room, which is probably typical for the majority of VR applications, a physically adequate representation of all these subjective impressions must, therefore, be simulated, auralized, and reproduced. Plausibility can, however, only be defined for specific environments. Therefore, a general approach of sound field modeling requires a physical basis and applicability in a wide range of rooms, buildings, or outdoor environments.

1.2. Reproduction

The aural component additionally enforces the user's immersive experience due to the comprehension of the environment through a spatial representation [1, 2]. Besides the sound field modeling itself, an adequate reproduction of the signals is very important. The goal is to transport all spatial cues contained in the signal in an aurally correct way to the ears of a listener. As mentioned above, coloration, loudness, and timbre are essential, but also the direction of a sound and its reflections are required for an at least plausible scene representation. The directional information in a spatial signal is very important to represent a room in its full complexity. In addition, this is supported by a dynamically adapted binaural rendering which enables the listener to move and turn within the generated virtual world.

1.3. System

In this contribution, we describe the physical algorithmic approach of sound field modeling and 3D sound reproduction of the VR systems installed at RWTH Aachen University (see Figure 1). The system is implemented in a first version. It is open to any extended physical sound field modeling in real time, and is independent of any particular visual VR display technology, for example, CAVE-like displays [3] or desktop-based solutions. Our 3D audio system named VirKopf has been implemented at the Institute of Technical Acoustics (ITA), RWTH Aachen University, as a distributed architecture. For any room acoustical simulation, VirKopf uses the software RAVEN (room acoustics for virtual environments) as a networked service (see Section 2.1). It is obvious that video and audio processing take a lot of computing resources for each subsystem, and by today's standards, it is unrealistic to do all processing on a single machine. For that reason, the audio system realizes the computation of video and audio data on dedicated machines that are interconnected by a network. This idea is obvious and has already been successfully implemented by [4] or [5]. There are even

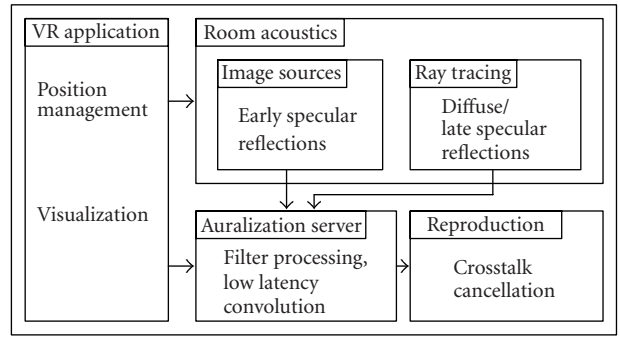


Figure 1: System components.

commercially available solutions, which have been realized by dedicated hardware that can be used via a network interface, for example, the Lake HURON machine [6]. Other examples of acoustic rendering components that are bound by a networked interface can be found in connection with the DIVA project [7, 8] or Funkhouser's beam tracing approach [9]. Other approaches such as [2] or [10] have not been implemented as a networked client-server architecture but rely on a special hardware setup.

The VirKopf system differs from these approaches in some respects. A major difference is the focus of the VirKopf system, offering the possibility of a binaural sound experience for a moving listener without any need for headphones in immersive VR environments. Secondly, it is not implemented on top of any constrained hardware requirements such as the presence of specific DSP technology for audio processing. The VirKopf system realizes a software-only approach and can be used on off-the-shelf custom PC hardware. In addition to that, the system does not depend on specially positioned loudspeakers or a large number of loudspeakers. Four loudspeakers are sufficient to create a surrounding acoustic virtual environment for a single user using the binaural approach.

2. ROOM ACOUSTICAL SIMULATION

Due to several reasons, which cannot be explained in all details here, geometrical acoustics is the most important model used for auralization in room acoustics [11]. Wave models would be more exact, but only the approximations of geometrical acoustics and the corresponding algorithms provide a chance to simulate room impulse responses in real-time application. In this interpretation, delay line models, radiosity, or others are considered as basically geometric as well since wave propagation is reduced to the time-domain approach of energy transition from wall to wall. In geometrical acoustics, deterministic and stochastic methods are available. All deterministic simulation models used today are based on the physical model of image sources [12, 13]. They differ in the way how sound paths are identified by using forward (ray) tracing or reverse construction. Variants of this type of algorithms are hybrid ray tracing, beam tracing, pyramid tracing, and so forth [14–20]. Impulse responses from image-like models

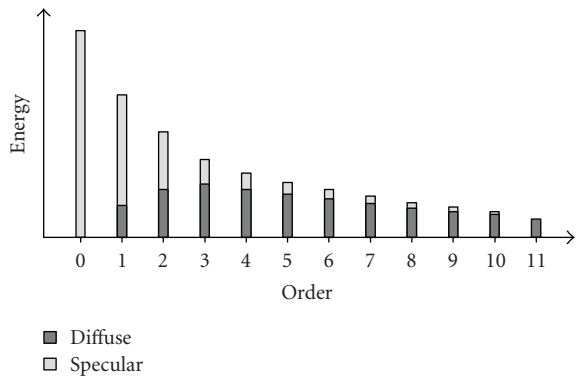


Figure 2: Conversion of specularly into diffusely reflected sound energy, illustrated by an example (after Kuttruff [23]).

consist of filtered Dirac pulses arranged accordingly to their delay and amplitude and are sampled with a certain temporal resolution. In intercomparisons of simulation programs [21, 22], it soon became clear that pure image source modeling would create too rough an approximation of physical sound fields in rooms since a very important aspect of room acoustics—surface and obstacle scattering—is neglected.

It can be shown that, from reflections of order two or three, scattering becomes a dominant effect in the temporal development of the room impulse response [23] even in rooms with rather smooth surfaces (see Figure 2). Fortunately, the particular directional distribution of scattered sound is irrelevant after the second or third reflection order and can well be assumed as Lambert scattering. However, in special cases of rooms with high absorption such as recording studios, where directional diffusion coefficients are relevant, different scattering models have to be used. Solutions for the problem of surface scattering are given by either stochastic ray tracing or radiosity [14, 18, 24–27]. Furthermore, the fact that image sources are a good approximation for perfectly reflecting or low absorption surfaces is often forgotten. The approximation of images, however, is valid in large rooms at least for large distances between the source, wall, and receiver [28]. Another effect of wave physics—diffraction—can be introduced into geometrical acoustics [29, 30], but so far the online simulation has been restricted to stationary sound sources. Major problems arise, however, when extending diffraction models to higher orders. Apart from outdoor applications, diffraction has not yet been implemented in the case of applications such as room acoustics. It should, however, be mentioned that numerous algorithmic details have already been published in the field of sound field rendering so far. New algorithmic schemes such as those presented by [31] have not yet been implemented. It should be kept in mind here that the two basic physical methods—deterministic sound images and stochastic scattering—should be taken into account in a sound field model with a certain performance of realistic physical behavior. Sound transmission as well as diffraction must be implemented in the cases of coupled rooms, in corridors, or cases where sound is transmitted through apertures.

2.1. Real-time capable implementation

Any room acoustical simulation should take into account the above-mentioned physical aspects of sounds in rooms. Typically, software is available for calculating room impulse responses of a static source and a listener's position within a few seconds or minutes. However, an unrestricted movement of the receiver and the sound sources within the geometrical and physical boundaries are basic demands for any interactive on-line auralization. Furthermore, any interaction with the scenery, for instance, opening a door to a neighboring room, and the on-line-update of the change of the rooms' modal structures should be provided by the simulation to produce a high believability of the virtual world [32].

At present, a room acoustical simulation software called RAVEN is being developed at our institute. The software aims at satisfying all above-mentioned criteria for a realistic simulation of the aural component, however, in respect of real-time capability. Special implementations offering the possibility of room acoustical simulation in real time will be described in the following sections. RAVEN is basically an upgrade and enhancement of the hybrid room acoustical simulation method by Vorländer [20], which was further extended by Heinz [25]. A very flexible and fast-to-access framework for processing an arbitrary number of rooms (see Section 2.2) has been incorporated to gain a high level of interactivity for the simulation and to achieve real-time capability for algorithms under certain constraints (see Section 5.2). Image sources are used for determining early reflections (see Section 2.3) in order to provide a most accurate localization of primary sound sources (precedence effect [33]) during the simulation. Scattering and reverberation are estimated on-line by means of an improved stochastic ray tracing method, which will be further described in Section 2.4.

2.2. Scene partitioning

The determination of the rooms' sound reflections requires an enormous number of intersection tests between rays and the rooms' geometry since geometrical acoustics methods treat sound waves as "light" rays. To apply these methods in real time, data structures are required for an efficient representation and determination of spatial relationships between sound rays and the room geometry.

These data structures organize geometry hierarchically in some n-dimensional space and are usually of recursive nature to accelerate remarkably queries of operations such as culling algorithms, intersection tests, or collision detections [34, 35].

Our auralization framework contains a preprocessing phase which transforms every single room geometry into a flexible data structure by using binary space partitioning (BSP) trees [36] for fast intersection tests during the simulation. Furthermore, the concept of scene graphs [37], which is basically a logical layer on top of the single room data structures, is used to make this framework applicable for an arbitrary number of rooms and to acquire a high level of interactivity for the room acoustical simulation.

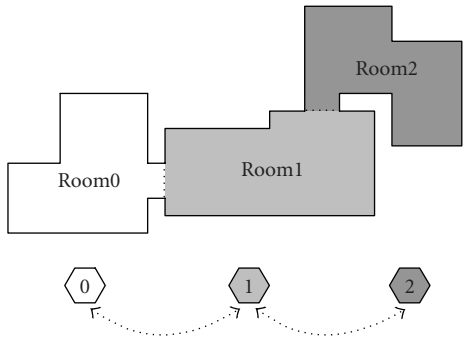


Figure 3: The scenery is split into three rooms, which are represented by the nodes of the scene graph (denoted through hexagons). The rooms are connected to their neighboring rooms by 2 portals (room0/room1 and room1/room2, denoted through the dotted lines).

2.2.1. Scene graph architecture

To achieve an efficient data handling for an arbitrary number of rooms, the concept of scene graphs has been used. A scene graph is a collection of nodes which are linked according to room adjacencies.

A node contains the logical and spatial representation of the corresponding subscene. Every node is linked to its neighbors by so-called portals, which represent entities connecting the respective rooms, for example, a door or a window (see Figure 3). It should be noted that the number of portals for a single node is not restricted, hence the scenery can be partitioned quite flexibly into subscenes. The great advantage of using portals is their binary nature as two states can occur. The state "active" connects two nodes defined by the portal, whereas the state "passive" cuts off the specific link. This provides a high level of interactivity for the room acoustical simulations as room neighborhoods can be changed on-line, for instance, doors may be opened or closed. In addition, information about portal states can be exploited to speed up any required tests during the on-line room acoustical simulation by neglecting rooms which are acoustically not of interest, for example, rooms that are out of bounds for the current receiver's position.

2.3. Image source method

The concept of the traditional image source (IS) method provides a quite flexible data structure, as, for instance, the online movement of primary sound sources and their corresponding image sources is supported and can be updated within milliseconds. Unfortunately, the method fails to simulate large sceneries as the computational costs are dominated by the exponential growth of image sources with an increasing number of rooms, that is, polygons and reflection order. Applying the IS method to an arbitrary number of rooms would result in an explosion of IS to be processed, which would make a simulation of a large virtual environment impossible within real-time constraints due to the extreme number of IS to be tested online on audibility.

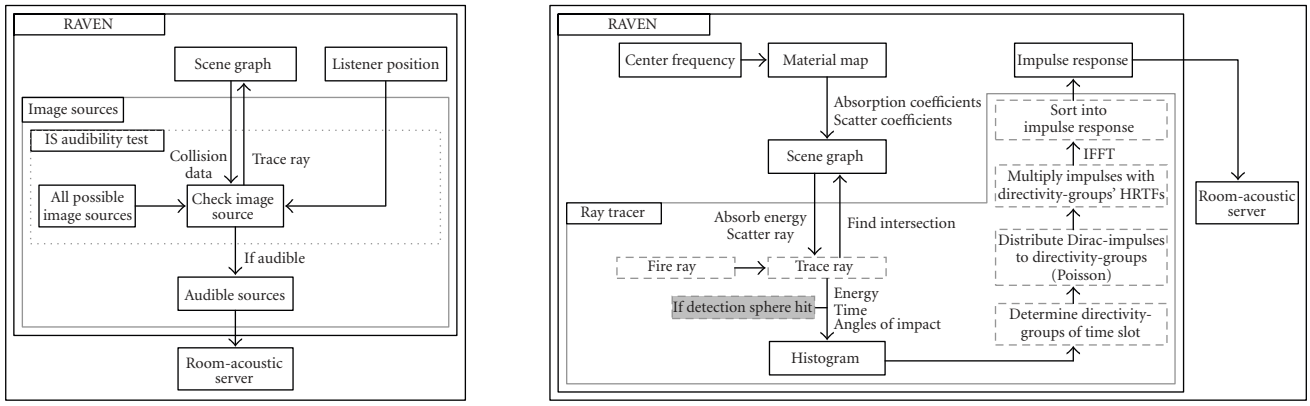
However, the scene graph data structure (see Section 2.2.1) provides the possibility of precomputing subsets of potentially audible IS according to the current portal configuration by sorting the entire set of IS dependent on the room(s) they originate from. This can easily be done by preprocessing the power set of the scene S, where S is a set of n rooms. The power set of S contains 2^n elements, and every subset, that is, family set of S refers to an n-bit number, where the mth bit refers to activity or inactivity of the mth room of S. Then, all ISs are sorted into the respective family sets of S by gathering information about the room IDs of the planes they have been mirrored on. Figure 5 shows exemplarily the power set P of a scenery S containing the three rooms $R2$, $R1$, $R0$, and the linked subsets of IS, that is, $P(S) = \{\{\text{Primary Source}\},\{\text{IS}(R0)\},\{\text{IS}(R1)\},\{\text{IS}(R1,R2)\},\{\text{IS}(R2)\},\{\text{IS}(R2,R0)\},\{\text{IS}(R2,R1)\},\{\text{IS}(R2,R1,R0)\}\}$.

During on-line auralization, a depth-first search [37] of the scene graph determines reachable room IDs for the current receiver's position. This excludes both rooms that are out of bounds and rooms that are blocked by portals. This set of room IDs is encoded by the power set P to set unreachable rooms invalid as they are acoustically not of interest. If in the case of this example room $R2$ gets unreachable for the current receiver's position, for example, someone closed the door, only IS family sets of P have to be processed for auralization that do not contain the room ID $R2$. As a consequence thereof, the number of IS family sets to be tested on audibility drops from eight to four, that is, $P(0)$, $P(1)$, $P(2)$, $P(3)$, which obviously leads to a significant reduction of computation time.

During simulation it will have to be checked whether every possible audible image source, which is determined as described above, is audible for the current receiver's position (see Figure 4(a)). Taking great advantage of the scene graph's underlying BSP-tree structures and an efficient tree traversing strategy [38], the required IS audibility test can be done very fast (performance issues are discussed in more detail in Section 5.2.1). If an image source is tested on audibility for the current receiver's position, all data being required for filter calculation (position, intersection points, and hit material) will be stored in the super-ordinated container "audible sources" (see Figure 4(a)).

2.4. Ray tracing

The computation of the diffuse sound field is based on the stochastic ray tracing algorithm proposed by Heinz [39]. For building the binaural impulse response from the ray tracing data, Heinz assumed that the reverberation is ideally diffuse. This assumption is, however, too rough, if the room geometry is extremely long or flat and if it contains objects like columns or privacy screens. Room acoustical defects such as (flutter) echos would remain undetected [40, 41]. For a more realistic room acoustical simulation, the algorithm has been changed in a way so that these effects are taken into account (see Figure 4(b)). This aspect is an innovation in real-time



(a) Image sources

(b) Ray tracing

Figure 4: (a) Image source audibility test, (b) estimation of scattering and reverberation.

ID	*R*2	*R*1	*R*0	IS subset
7	1	1	1	*R*2 *R*1 *R*0
6	1	1	0	*R*2 *R*1
5	1	0	1	*R*2 *R*0
4	1	0	0	*R*2
3	0	1	1	*R*1 *R*0
2	0	1	0	*R*1
1	0	0	1	*R*0
0	0	0	0	Primary source

Figure 5: IS/room-combination-power set $P(S)$ for a three-room situation. All IS are sorted into encapsulated containers depending on the room combination they have been generated from.

virtual acoustics, which is to be considered as an important extension of the perceptive dimension.

The BSP-based ray tracing simulation starts by emitting a finite number of particles from each sound source at random angles where each particle carries a source directivity dependent amount of energy. Every particle loses energy while propagating due to air absorption and occurring reflections on walls, either specular or diffuse, and other geometric objects inside the rooms, that is, a material dependent absorption of sound. The particle gets terminated as soon as the particle's energy is reduced under a predefined threshold. Before a time t_0, which represents the image source cut-off time, only particles are detected which have been reflected specular with a diffuse history in order to preserve a correct energy balance. After t_0, all possible permutations of reflection types are processed (e.g., diffuse, specular, diffuse, diffuse, etc.).

The ray tracing is performed for each frequency band due to frequency dependent absorption and scattering coefficients, which results in a three-dimensional data container called histogram. This histogram is considered as the temporal envelope of the energetic spatial impulse response. One single field of the histogram contains information about rays (their energy on arrival, time, and angles of impact) which hit the detection sphere during a time interval Δt for a discrete frequency interval f_b. At first, the mean energy for fields with different frequencies but the same time interval is calculated to obtain the short-time energy spectral density. This step is also used to create a ray directivity distribution over time for the respective rays: for each time slot, the detection sphere is divided into evenly distributed partitions, so-called directivity groups. If a ray hits the sphere, the ray's remaining energy on impact is added to the corresponding sphere's directivity group depending on its time and direction of arrival (see Figure 6).

This energy distribution is used to determine a ray probability for each directivity group and each time interval Δt. Then a Poisson process with a rate equal to the rate of reflections for the given room and the given time interval is created. Each impulse of the process is allotted to the respective directivity group depending on the determined ray probability distribution. In a final step, each directivity group which was hit by a Poisson impulse cluster is multiplied by its respective HRTF, superposed to a binaural signal, and weighted by the square root of the energy spectral density. After that, the signal is transformed into time domain. This is done for every time step of the histogram and put together to the complete binaural impulse response. The ray tracing algorithm is managed by the room acoustics server to provide the possibility of a dynamic update depth for determining the diffuse sound field component (see Section 3). Since this contribution focuses on the implementation and performance of the complete system, no further details are presented here. A detailed description of the fast implementation and test results can be found in [42].

3. FILTER PROCESSING

For a dynamic auralization where the listener is allowed to move, turn, and interact with the presented scenery and

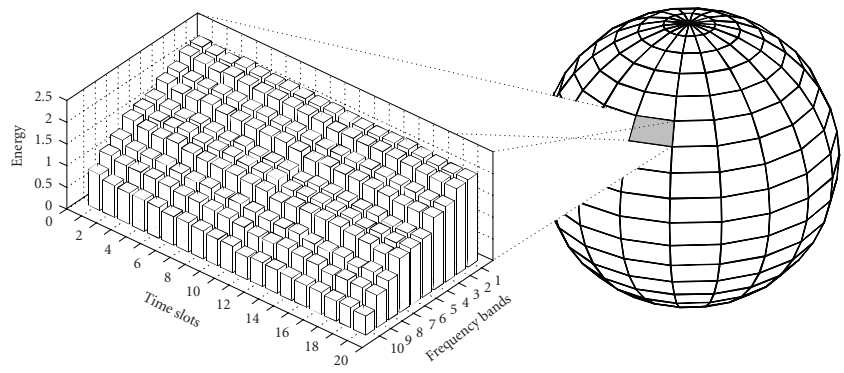


Figure 6: Histogram example of a single directivity group.

where the sources can also be moved, the room impulse response has to be updated very fast. This becomes also more important in combination with congruent video images. Thus, the filter processing is a crucial part of the real-time process [8]. The whole filter construction is separated into two parts. The most important section of a binaural room impulse response is the first part containing the direct sound and the early reflections of the room. These early reflections are represented by the calculated image sources and have to be updated at a rate which has to be sufficient for the binaural processing. For this reason, the operation interface between the room acoustics server and the auralization server is the list of the currently audible sources. The second part of the room impulse response is calculated on the room acoustics server (or cluster) to minimize the time required by the network transfer because the amount of data required to calculate the room impulse response is significantly higher than the resulting filter itself.

3.1. Image sources

Every single fraction of the complete impulse response, either the direct sound or the sound reflected by one or more walls, runs through several filter elements as shown in Figure 7. Elements such as directivity, wall, and air absorption are filters in a logarithmic frequency representation with a third octave band scale with 31 values from 20 Hz to 20 kHz. These filters contain no phase information so that only a single multiplication is needed. The drawback of using a logarithmic representation is the necessity of interpolation to multiply the resulting filter with the HRTF. But this is still not as computationally expensive as using a linear representation for all elements, particularly if more wall filters have to be considered for the specific reflection.

So far, the wall absorption filters are independent of the angle of sound incidence, which is a common assumption for room acoustical models. It can be extended to consider angle-dependent data if necessary. Reflections calculated by using the image source model will be attenuated by the factor of the energy which is distributed by the diffuse reflections. The diffuse reflections will be handled by the ray tracing algorithm, (see Section 3.2).

Another important influence on the sound in a room, especially a large hall, is the directivity of the source. This is even more important for a dynamic auralization where not only the listener is allowed to move and interact with the scenery but where the sources can also move or turn. The naturalness of the whole generated sound scene is improved by every dynamic aspect being taken into account. The program accepts external directivity databases of any spatial resolution, and the internal database has a spatial resolution of 5 degrees for azimuth and elevation angles. This database contains the directivity of a singer and several natural instruments. Furthermore, it is possible to generate a directivity manually. The air absorption filter is only distance dependent and is applied also to the direct sound, which is essential for far distances between the listener and source.

At the end of every filter pass, which represents, up to now, a mono signal, an HRTF has to be used to generate a binaural head-related signal which contains all directional information. All HRTFs used by the VirKopf system were measured with the artificial head of the ITA for the full sphere due to the asymmetrical pinnae and head geometry. Nonsymmetrical pinnae lead to positive effects on the perceived externalization of the generated virtual sources [43]. A strong impulse component such as the direct sound carries the most important spatial information of a source in a room. In order to provide a better resolution, even at low frequencies, an HRTF of a higher resolution is used for the direct sound. The FIR filter length is chosen to be 512 taps. Due to the fact that the filter processing is done in the frequency domain, the filter is represented by 257 complex frequency domain values corresponding to a linear resolution of 86 Hz.

Furthermore, the database does not only contain HRTFs measured at one specific distance but, also near-field HRTFs. This provides the possibility of simulating near-to-head sources in a natural way. Tests showed that the increasing interaural level difference (ILD) becomes audible at a distance of 1.5 m or closer to the head. This test was performed in the semianechoic chamber of the ITA, examining the ranges

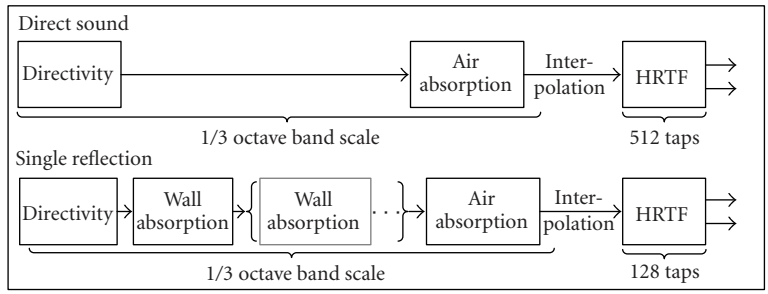


Figure 7: Filter elements for direct sound and reflections.

where different near-field HRTFs have to be applied. The listeners were asked to compare signals from simulated HRTFs with those from correspondingly measured HRTFs on two criteria, namely, the perceived location of the source and any coloration of the signals. The simulated HRTFs were prepared from far-field HRTFs (measured at a distance of two meters) with a simple-level correction applied likewise to both channels. All of the nine listeners reported differences with regard to lateral sound incidences in the case of distances being closer than 1.5 m. No difference with regard to frontal sound incidences was reported in the case of distances being closer than 0.6 m. These results are very similar to the results obtained by research carried out in other labs, for example, [44]. Hence, HRTFs were measured at distances of 0.2 m, 0.3 m, 0.4 m, 0.5 m, 0.75 m, 1.0 m, 1.5 m, and 2.0 m. The spatial resolution of the databases is 1 degree for azimuth and 5 degrees for elevation angles for both the direct sound and the reflections.

The FIR filter length of 128 taps used for the contribution of image sources is lower than for the direct sound, but is still higher than the limits to be found in literature. Investigations regarding the effects of a reduced filter length on localization can be found in [45]. As for the direct sound, the filter processing is done in the frequency domain with the corresponding filter representation of 65 complex values. Using 128 FIR coefficients leads to the same localization results, but brings about a considerable reduction of the processing time (see Table 3). This was tested as well in internal listening experiences but is also congruent to the findings of other labs, that is, [46]. The spatial representation of image sources is realized by using HRTFs measured in 2.0 m. In this case, this does not mean any simplification because the room acoustical simulation using image sources is not valid anyway at distances close (a few wavelengths) to a wall. A more detailed investigation relating to that topic can be found in [28, 47].

3.2. *Ray tracing*

As mentioned above, the calculation of the binaural impulse response of the ray tracing process is done on the ray tracing server in order to reduce the amount of data which has to be transferred via the network. To keep the filters up-to-date according to the importance of the filter segment, which is related to the time alignment, the auralization process can send interrupt commands to the simulation server. If a source or the listener is moving too fast to finish the calculation of the filter within an adequate time slot, the running ray tracing process will be stopped. This means that the update depth of the filter depends on the movements of the listener or the sources. In order to achieve an interruptible ray tracing process, it is necessary to divide the whole filter length into several parts. When a ray reaches the specified time stamp, the data necessary to restart the ray at this position will be saved and the next ray is calculated. After finishing the calculation of all rays, the filter will be processed up to the time the ray tracing updated the information in the histogram (this can also be a parallel process, if provided by the hardware). At this time, it is also possible to send the first updated filter section to the auralization server, which means that it is possible to take the earlier part of the changed impulse response into account before the complete ray tracing is finished. At this point, the ray tracing process will decide on the interrupt flag whether the calculation is restarted at the beginning of the filter or at the last time stamp. For slight or slow movements of the head or of the sources, the ray tracing process has enough time to run through a complete calculation cycle containing all filter time segments. This also leads to the fact that the level of the simulation's accuracy rises with the duration the listener stands at approximately the same position and the sources do not move.

4. REPRODUCTION SYSTEM

The primary reproduction system of the room acoustical modeling described in this paper is a setup mounted in the CAVE-like environment, which is a five-sided projection system of a rectangular shape, installed at RWTH Aachen University. The special shape enables the use of the full resolution of 1600 by 1200 pixels of the LCD projectors on the walls and the floor as well as a 360 degree horizontal view. The dimensions of the projection volume are $3.60 \times 2.70 \times 2.70\ \mathrm{m}^3$ yielding a total projection screen area of $26.24\ \mathrm{m}^2$. Additionally, the use of passive stereo via circular polarization allows lightweight glasses. Head and interaction device tracking is realized by an optical tracking system. The setup of this display

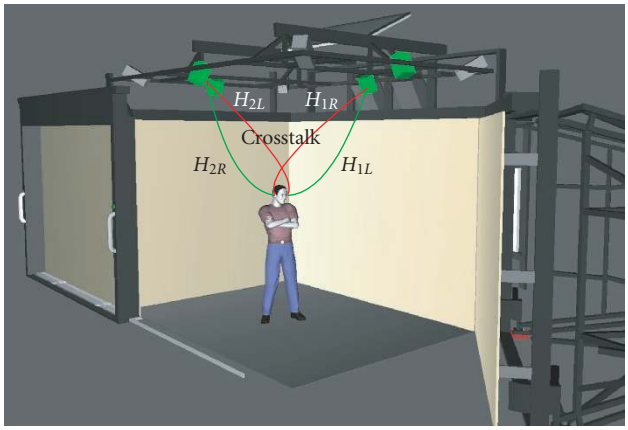


Figure 8: The CAVE-like environment at RWTH Aachen University. Four loudspeakers are mounted on the top rack of the system. The door, shown on the left, and a moveable wall, shown on the right, can be closed to allow a 360-degree view with no roof projection.

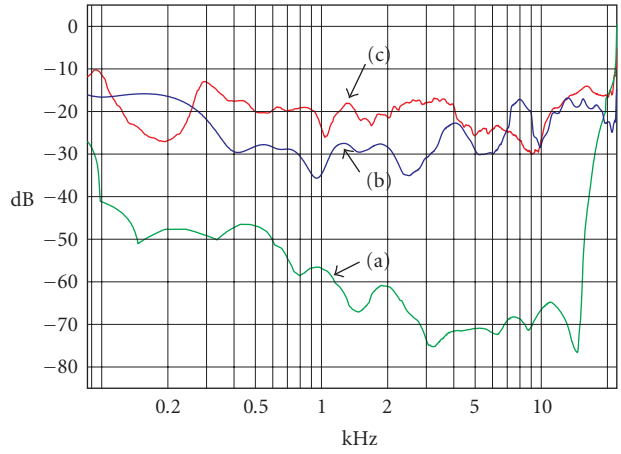


Figure 9: Measurement of the accessible channel separation using a filter length of 1024 taps. (a) = calculated, (b) = static solution, (c) = dynamic system.

system is an improved implementation of the system [48] that was developed with the clear aim to minimize attachments and encumbrances in order to improve user acceptance. In that sense, much of the credibility that CAVE-like environments earned in recent years has to be attributed to the fact that they try to be absolutely nonintrusive VR systems. As a consequence, a loudspeaker-based acoustical reproduction system seems to be the most desired solution for acoustical imaging in CAVE-like environments. Users should be able to step into the virtual scenery without too much preparation or calibration but still be immersed in a believable environment. For that reason, our CAVE-like environment depicted above was extended with a binaural reproduction system using loudspeakers.

4.1. Virtual headphone

To reproduce the binaural signal at the ears with a sufficient channel separation without using headphones, a crosstalk cancellation (CTC) system is needed [49–51]. Doing the CTC work in an environment where the user should be able to walk around and turn his head requires a dynamic CTC system which is able to adapt during the listener's movements [52, 53]. The dynamic solution overrides the sweet spot limitation of a normal static crosstalk cancellation. Figure 8 shows the four transfer paths from the loudspeakers to the ears of the listener (H_{1L} = transfer function loudspeaker 1 to left ear). A correct binaural reproduction means that the complete transfer function from the left input to the left ear (reference point is the entrance of the ear canal) including the transfer function H_{1L} is meant to become a flat spectrum. The same is intended for the right transfer path, accordingly. The crosstalk indicated by H_{1R} and H_{2L} has to be canceled by the system.

Since the user of a virtual environment is already tracked to generate the correct stereoscopic video images, it is possible to calculate the CTC filter online for the current position and orientation of the user. The calculation at runtime enhances the flexibility of the VirKopf system regarding the validity area and the flexibility of the loudspeaker setup which can hardly be achieved with preprocessed filters. Thus, a database containing "all" possible HRTFs is required. The VirKopf system uses a database with a spatial resolution of one degree for both azimuth (φ) and elevation (ϑ). The HRTFs were measured at a frequency range of 100 Hz–20 kHz, allowing a cancellation in the same frequency range. It should be mentioned that a cancellation at higher frequencies is more error prone to misalignments of the loudspeakers and also to individual differences of the pinna. This is also shown by curve (c) in Figure 9. The distance between the loudspeaker and the head affects the time delay and the level of the signal. Using a database with HRTFs measured at a certain distance, these two parameters must be adjusted by modifying the filter group delay and the level according to the spherical wave attenuation for the actual distance.

To provide a full head rotation of the user, a two loudspeaker setup will not be sufficient as the dynamic cancellation will only work in between the angle spanned by the loudspeakers. Thus, a dual CTC algorithm with a four-speaker setup has been developed, which is further described in [54]. With four loudspeakers, eight combinations of a normal two-channel CTC system are possible and a proper cancellation can be achieved for every orientation of the listener. An angle dependent fading is used to change the active speakers in between the overlapping validity areas of two configurations.

Each time the head-tracker information is updated in the system, the deviation of the head to the position and orientation compared to the information given which caused the preceding filter change is calculated. Every degree of freedom is weighted with its own factor and then summed up. Thus, the threshold can be parameterized in six degrees of

freedom, positional values $(\Delta x, \Delta y, \Delta z)$, and rotational values $(\Delta\varphi, \Delta\vartheta, \Delta\rho)$. A filter update will be performed when the weighted sum is above 1. The lateral movement and the head rotation in the horizontal plane are most critical so $\Delta x = \Delta y = 1$ cm and $\Delta\varphi = 1.0$ degree are chosen to dominate the filter update. The threshold always refers to the value where the limit was exceeded the last time. The resulting hysteresis prevents a permanent switching between two filters as it may occur when a fixed spacing determines the boundaries between two filters and the tracking data jitter slightly.

One of the fundamental requirements of the sound output device is that the channels work absolutely synchronously. Otherwise, the calculated crosstalk paths do not fit with the given condition. On this account, the special audio protocol ASIO designed by Steinberg for professional audio recording was chosen to address the output device [55].

To classify the performance that could be reached theoretically by the dynamic system, measurements of a static system were made to have a realistic reference for the achieved channel separation. Under absolute ideal circumstances, the HRTFs used to calculate the crosstalk cancellation filters are the same as during reproduction (individual HRTFs of the listener). In a first test, the crosstalk cancellation filters were processed with HRTFs of an artificial head in a fixed position. The windowing to a certain filter length and the smoothing give rise to a limitation of the channel separation. The internal filter calculation length is chosen to 2048 taps in order to take into account the time offsets caused by the distance to the speakers. The HRTFs were smoothed with a bandwidth of 1/6 octave to reduce the small dips which may cause problems by inverting the filters. After the calculation, the filter set is truncated to the final filter length of 1024 taps, the same length that the dynamic system works with. However, the time alignment among the single filters is not affected by the truncation. The calculated channel separation using this (truncated) filter set and the smoothed HRTFs as reference is plotted in Figure 9 curve (a). Thereafter, the achieved channel separation was measured at the ears of the artificial head, which had not been moved since the HRTF measurement (Figure 9 curve (b)).

In comparison to the ideal reference cases, Figure 9 curve (c) shows the achieved channel separation of the dynamic CTC system. The main difference between the static and the dynamic system is the set of HRTFs used for filter calculation. The dynamic system has to choose the appropriate HRTF from a database and has to adjust the delay and the level depending on the position data. All these adjustments cause minor deviations from the ideal HRTF measured directly at this point. For this reason, the channel separation of the dynamic system is not as high as the one that can be achieved by a system with direct HRTF measurement.

The theory of crosstalk cancellation is based on the assumption of a reproduction in an anechoic environment. However, the projection walls of CAVE-like environments consist of solid material causing reflections that decrease the performance of the CTC system. Listening tests with our system show [56] that the subjective localization performance is still remarkably good. Also tests of other labs [57, 58] and different CTC systems indicate a better subjective performance than it would be expected from measurements. One aspect validating this phenomenon is the precedence effect by which sound localization is primarily determined by the first arriving wavefront; the other aspect is the head movement which gives the user the ability to approve the perceived direction of incidence. A more detailed investigation on the performance of our binaural rendering and reproduction system can be found in [59].

The latency of the audio reproduction system is the time elapsed between the update of a new position and orientation of the listener, and the point in time at which the output signal is generated with the recalculated filters. The output block length of the convolution (overlap save) is 256 taps as well as the chosen buffer length of the sound output device, resulting in a time between two buffer switches of 5.8 milliseconds at 44.1 kHz sampling rate for the rendering of a single block. The calculation of a new CTC filter set (1024 taps) takes 3.5 milliseconds on our test system. In a worst case scenario, the filter calculation just finishes after the sound output device fetched the next block, so it takes the time playing this block until the updated filter becomes active at the output. That would cause a latency of one block. In such a case, the overall latency accumulates to 9.3 milliseconds.

4.2. Low-latency convolution

A part of the complete dynamic auralization system requiring a high amount of processing power is the convolution of the audio signal. A pure FIR filtering would cause no additional latency except for the delay of the first impulse of the filter, but it also causes the highest amount of processing power. Impulse responses of more than 100 000 taps or more cannot be processed in real time on a PC system using FIR filters in the time domain. The block convolution is a method that reduces the computational cost to a minimum, but the latency increases in proportion to the filter length. The only way to minimize the latency of the convolution is a special conditioning of the complete impulse response in filter blocks. Basically, we use an algorithm which works in the frequency domain with small block sizes at the beginning of the filter and increasing sizes to the end of the filter. More general details about these convolution techniques can be found in [60]. However, our algorithm does not operate on the commonly used segmentation which doubles the block length every other block. Our system provides a special block size conditioning with regard to the specific PC hardware properties as, for instance, cache size or special processing structures such as SIMD (single instruction multiple data). Hence, the optimal convolution adds a time delay of only the first block to the latency of the system, so that it is recommended to use a block length as small as possible. The amount of processing power is not linear to the overall filter length and also constrained by the chosen start block length. Due to this, measurements were done to determine the processor load of different modes of operation (see Table 1).

Table 1: CPU load of the low-latency convolution algorithm.

Impulse response length	Number of sources							
	3	10	15	20	3	10	15	20
	(Latency 256 taps)				(Latency 512 taps)			
0.5 s	9%	30%	50%	76%	8%	22%	30%	50%
1.0 s	14%	40%	66%	—	11%	33%	53%	80%
2.0 s	15%	50%	74%	—	14%	42%	71%	—
3.0 s	18%	62%	—	—	16%	53%	—	—
5.0 s	20%	68%	—	—	18%	59%	—	—
10.0 s	24%	—	—	—	20%	68%	—	—

5. SYSTEM INTEGRATION

The VirKopf system constitutes the binaural synthesis and reproduction system, the visual-acoustic coupling, and it is connected to the RAVEN system for room acoustical simulations. The complete system's layout with all components is shown in Figure 10. As such it describes the distributed system which is used for auralization in the CAVE-like environment at RWTH Aachen University, where user interaction is tracked by six cameras. As a visual VR machine, a dual Pentium 4 machine with 3 GHz CPU speed and 2 GB of RAM is used (cluster master). The host for the audio VR subsystem is a dual Opteron machine with 2 GHz CPU speed and 1 GB of RAM. The room acoustical simulations run on Athlon 3000+ machines with 2 GB of RAM. This hardware configuration is also used as a test system for all performance measurements. As audio hardware, an RME Hammerfall system is used which allows sound output streaming with a scalable buffer size and a minimum latency of 1.5 milliseconds. In our case, an output buffer size is chosen to 256 taps (5.8 milliseconds). The network interconnection between all PCs was a standard Gigabit Ethernet.

5.1. Real-time requirements

Central aspects of coupled real-time systems are latency and the update rate for the communication. In order to get an objective criterion for the required update rates, it is mandatory to inspect typical behavior inside CAVE-like environments with special respect to head movement types and magnitude of position or velocity changes.

In general, user movements in CAVE-like environments can be classified in three categories [61]. One category is identified by the movement behavior of the user inspecting a fixed object by moving up and down and from one side to the other in order to accumulate information about its structural properties. A second category can be seen in the movements when the user is standing at one spot and uses head or body rotations to view different display surfaces of the CAVE. The third category for head movements can be observed when the user is doing both, walking and looking around in the CAVE-like environment. Mainly, the typical applications we employ can be classified as instances of the last two categories, although the exact user movement profiles can be individually different. Theoretical and empirical discussions about typical head movement in virtual environments are still a subject of research, for example, see [61–63] or [64].

As a field study, we recorded tracking data of users' head movements while interacting in our virtual environment. From these data, we calculated the magnitude of the velocity of head rotation and translation in order to determine the requirements for the room acoustics simulation. Figure 11(a) shows a histogram of the evaluated data for the translational velocity. Following from the deviation of the data, the mean translational velocity is at 15.4 cm/s, with a standard deviation of 15.8 cm/s and the data median at 10.2 cm/s, compare Figure 11(c). This indicates that the update rate of the room acoustical simulation can be rather low for translational movement as the overall sound impression does not change much in the immediate vicinity (see [65] for further information). As an example, imagine a room acoustical simulation of a concert hall where the threshold for triggering a recalculation of a raw room impulse response is 25 cm (which is typically half a seat row's distance). With respect to the translational movement profile of a user, a recalculation has to be done approximately every 750 milliseconds to catch about 70% of the movements. If the system aims at calculating correct image sources for about 90% of the movements, this will have to be done every 550 milliseconds. A raw impulse response contains the raw data of the images, their amplitude and delay, but not their direction in listener's coordinates. The slowly updated dataset represents, thus, the room-related cloud of image sources. The transformation into 3D listener's coordinates and the convolution will be updated much faster, certainly, in order to allow a direct and smooth responsiveness.

CAVE-like environments allow the user to directly move in the scene, for example, by walking inside of the boundaries of the display surfaces and tracking area. Additionally, indirect navigation enables the user to move in the scenery virtually without moving his body but by pointing metaphors when using hand sensors or joysticks. Indirect navigation is mandatory, for example, for architectural walkthroughs as the virtual scenery is usually much larger than the space covered by the CAVE-like device itself. The maximum velocity for indirect navigations has to be limited in order to avoid artifacts or distortions in the acoustical rendering and perception. However, during the indirect movement, users do

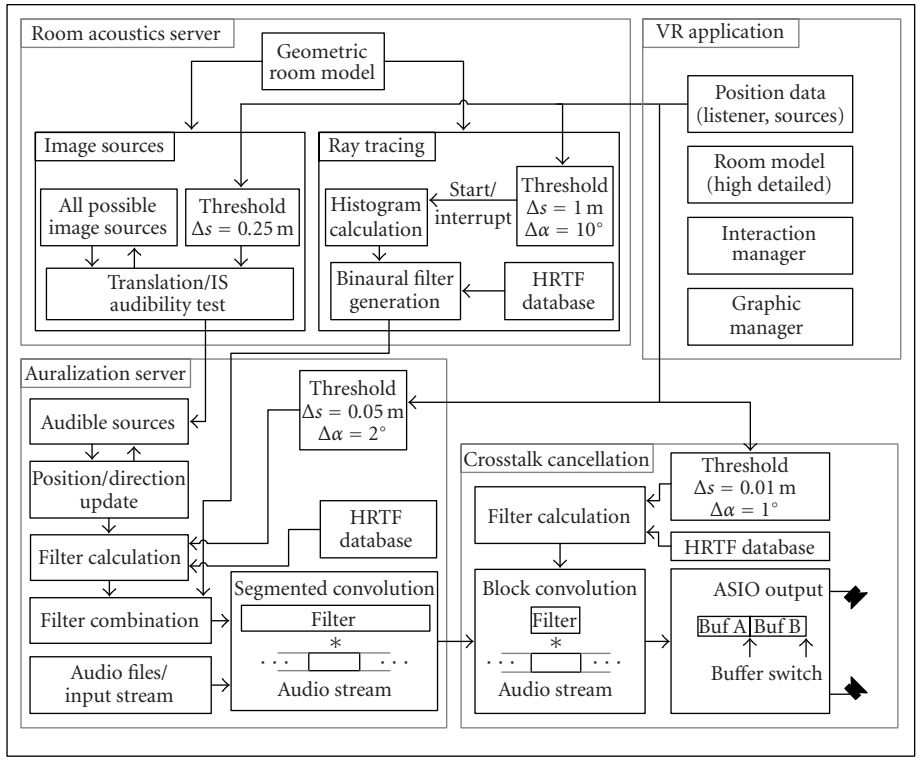
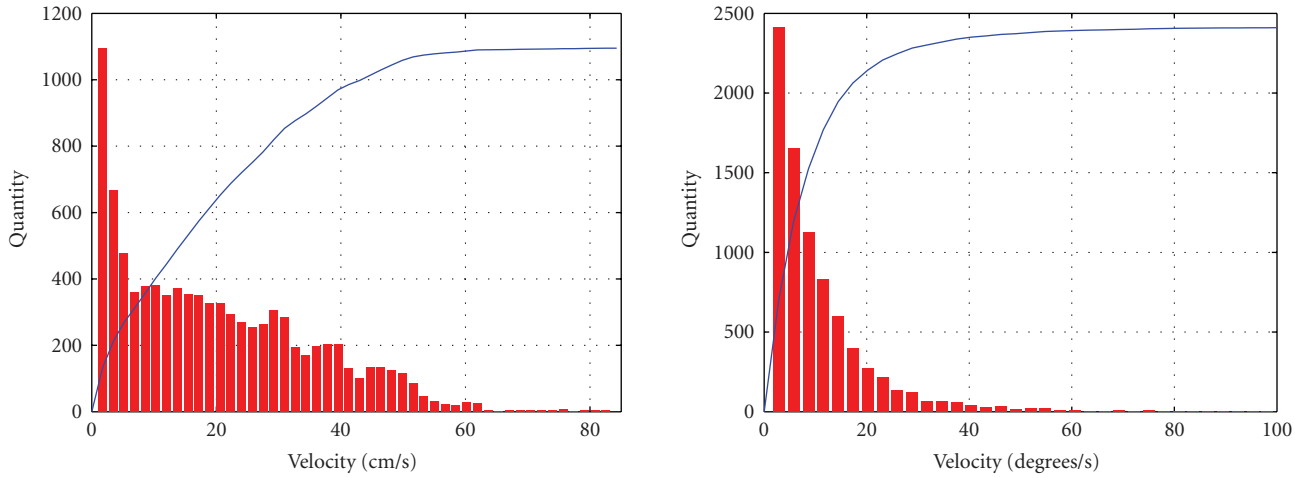


Figure 10: The complete binaural auralization system.



(a) (b)

	v_t (cm/s)	v_r (deg/s)
$\bar{x}$	15.486	8.686
σ	15.843	11.174
$\tilde{x}$	10.236	5.239
$x_{\max}$	84.271	141.458

(c)

Figure 11: Histogram of translational (v_t) and rotational (v_r) velocities of movements of a user acting in a CAVE-like environment. The blue line depicts the cumulative percentage of the measurements. In (b), we limited the upper bound to 100 degrees/s for better readability, (c) shows the descriptive statistics about the measurements.

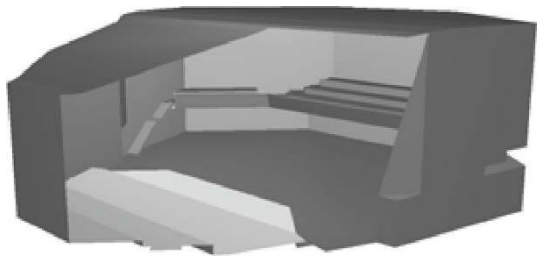


Figure 12: Sliced polygon model of the concert hall of Aachen's Eurogress convention center.

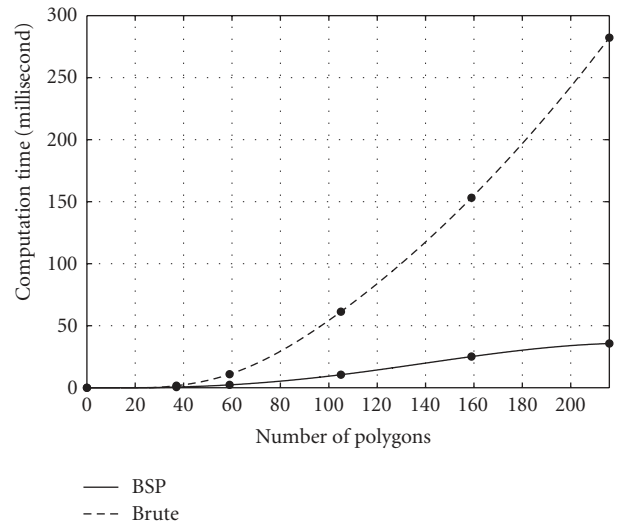


Figure 13: Comparison of required computation time for the ISs audibility test up to second-order ISs for different Eurogress models which differ in their level of detail (see [38] for details). With the growing number of polygons for the model's different levels of detail, the number of ISs grows exponentially, which leads to an exponential growth of the computation time for the brute-force approach. The computation time demands of the BSP-based method grows only linear due to the drop of search complexity up to $\mathcal{O}(\log N)$, N number of polygons.

not tend to move their head and the overall sensation reduces the capability to evaluate the correctness of the simulation. Once the users stop, it takes about 750 milliseconds as depicted above to calculate the right filters for the current user position. We made the experience that a limitation of the maximum velocity for indirect navigation to 100 cm/s shows good results and user acceptance.

In addition to the translational behavior, Figure 11(b) shows the rotational profile for head movements of a user. Peak angular velocities can be up to 140 degrees per second although these are very seldom. The mean for rotational movement is at 8.6 degrees/s with a standard deviation of 11.1 degrees/s and a data median at 5.2 degrees/s, compare Figure 11(c). Data sets provided as standard material for research on system latency, for example, by [66] or [61], show comparable results.

The orientation of the user's head in the sound field is very critical as reflections have to be calculated for the head-related impulse response in listener's coordinates. The changing ITD of the HRTFs during head rotation may cause a significant phase mismatch of two filters. In cross-fading from one room impulse response to the next, these differences should not be too big as this might result in audible comb-filter effects. To reduce these differences, a filter change every 1-2 degrees is necessary here. In order to be precise for almost all possible rotational velocities, we consider a timing interval for a recalculation every 10–20 milliseconds as mandatory. As a consequence, the block size configured in the audio processing hardware should not be bigger than 512 samples as this limits the minimal possible update time to 11.6 milliseconds at a 44.1 kHz sampling rate.

5.2. Performance of the room acoustical simulation

To evaluate the implementation and to determine its real-time capabilities, several experiments were carried out on the test system. For a realistic evaluation, a model of the concert hall of Aachen's Eurogress (volume about 15 000 m^3) convention center was constructed, which is shown in Figure 12. All results presented in this contribution are based on this model.

The model is constructed of 105 polygons and 74 planes, respectively. Although it is kept quite simple, the model contains all indoor elements of the room which are acoustically of interest [67], for example, the stage, the skew wall elements, and the balustrade. Details of small elements are neglected and represented by equivalent scattering [68]. Surface properties, that is, absorption and scattering coefficients are defined through standardized material data [69, 70].

5.2.1. Image source method performance

The computation time for the translational movement of primary sound sources and their respective image sources depends solely on the number of image sources. An average computation time of about 1 millisecond per 1000 image sources was measured. The main part of the computation time is needed for the audibility test.

To give a better idea of the achieved speed up by the use of BSP trees, a brute-force IS audibility test has been implemented for comparison purpose. This algorithm tests every scene's polygon on intersection instead of testing only a few room's subpartitions by means of a BSP-tree structure. Figure 13 shows a comparison of measured computation times for the IS-audibility test up to second IS order of both approaches. As expected, the computation time of the brute-force method rises exponentially with the exponentially growing number of ISs, whereas the BSP-based approach has only a quite linearly growing computation time demand due to the drop of search complexity up to $\mathcal{O}(\log N)$, N number of polygons.

Table 2: Comparison of the measurement results of the IS audibility test.

IS order	Number of IS		IS audibility test	
	All	Audible	BSP [ms]	Brute [ms]
1	75	9	0.153	0.959
2	4,827	32	10.46	61.27
3	309 445	111	710.07	3924

Table 3: Calculation time of several parts of the filter.

Processing step	Time
Direct sound (512 taps)	300 μs
Single reflection (aver.)	50 μs
Preparation for segmented convolution (6000 samples)	1.1 ms

With the assigned time slot (see Section 5.1) of 750 milliseconds for the simulation process, real-time capability for a room acoustical simulation with all degrees of freedom such as movable sound sources, movable receiver, changing sources' directivities, and interaction with the scenery is reached for about 320 000 ISs to be tested during runtime. Applying these constraints to the measurement results of the IS audibility test (see Table 2) makes the simulation of the Eurogress model real-time capable up to order 3.

Besides the performance of the room acoustical simulation, the processing time of the filter is very important. All time measurements of the calculation routines presented in this section are performed on our test system.

Calculating the image sources of the Eurogress model up to the third order, 111 audible image sources can be found in the first part of the impulse response of 6000 samples length corresponding to 136 milliseconds. In this case, one source is placed on the stage, and the listener is located in the middle of the room. The complete filter processing (excluding the audibility test) is done in 6.95 milliseconds. Note, that the filter processing has different entry points. The rotation of the listener or a source does not cause a recalculation of the audible sources, only the filter has to be processed.

5.2.2. Ray-tracing performance

For measuring the performance of the ray-tracing algorithm, all materials of the Eurogress model were replaced by a single one in order to avoid influences of different scattering and absorption coefficients on the results.

As in the previous section, a brute-force ray tracing algorithm has been implemented to compare the results to the BSP-based method we use in our framework. While the brute-force approach has a linearly growing computation time, that is, a complexity of $\mathcal{O}(N)$, N number of polygons, the BSP-based algorithm grows only logarithmically with increasing time due to the drop of search complexity to $\mathcal{O}(\log N)$ (see Figure 14, $t < 0.8$ second). A ray gets terminated if a minimum energy threshold is reached. Thus, both approaches get faster with increasing time due to the growing number of reflections, that is, the growing rays' loss of energy and ray termination, respectively. As an example, the algorithm needs an average of about 2.6 second per 80 000 rays (10 000 rays per frequency band, the first two octave bands are skipped) for determination of an impulse response with the length of 1 secone. As the processing time of the ray-tracing algorithm increases linearly with the number of rays used, a comparison of these results is redundant. It is obvious that the algorithm is able to cope with the real-time requirements, especially when using small numbers of rays at first to get a low-resolution histogram. If the listener stays at one place for a longer period of time, the ray tracer can update the histogram with more rays to get a higher resolution and determine a longer impulse response, respectively.

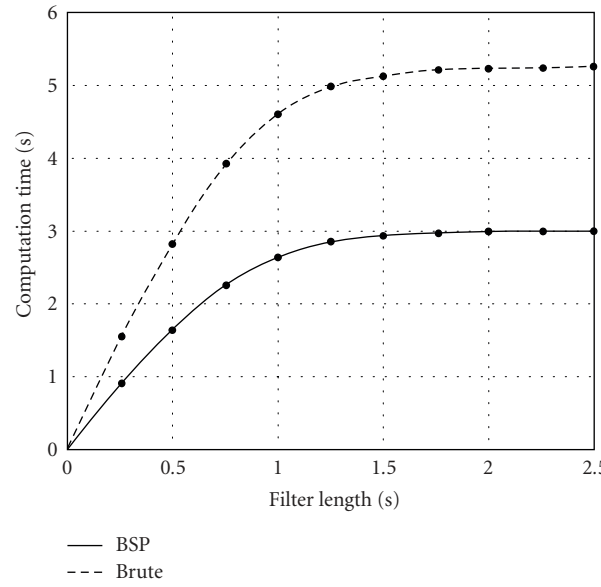


Figure 14: Comparison of required computation times for the determination of impulse responses with increasing length using 80 000 rays for the simulation.

5.3. Network

With respect to the timing, the optical tracking system is capable of delivering spatial updates of the position and orientation of the user's head and an additional interaction device to the VR application in 18.91 milliseconds. This figure is a direct result from the sum of the time needed for the visual recognition of two tracking targets as well as the transmission time for the measured data over a network link. For applications that must have a minimum latency time and do not need wireless tracking, the usage of an electromagnetic tracking system can reduce the latency to ≈ 5 milliseconds.

However, the VirKopf system distinguishes between two types of update messages. One type deals with low-frequency state changes such as commands to play or stop a specific sound. The second type updates the spatial attributes of the sound source and the listener at a high frequency. For the first type, a reliable transport protocol is used (TCP), while the

latter is transmitted at a high frequency over a low overhead but possibly unreliable protocol (UDP).

In order to get an estimate of the costs of network transport, the largest possible TCP and UDP messages produced by the VirKopf system were transmitted from the VR application to the VirKopf server many times and then sent back. The transmission time for this round trip was taken and halved for a single-trip measurement. The worst case times of the single trips are taken as a basis for the estimation of the overall cost introduced by the network communication. The mean time for transmitting a TCP command was 0.15 millisecond ± 0.02 millisecond. The worst case transmission time on the TCP channel was close to 1.2 millisecond. UDP communication was measured for 20 000 spatial update tables for 25 sound sources, resulting in a transmission time for the table of 0.26 millisecond ± 0.01 millisecond. It seems surprising that UDP communication is more expensive than TCP, but this is a result from larger packet sizes of an spatial update (≈ 1 kB) in comparison to small TCP command sizes (≈ 150 bytes).

5.4. Overall performance

Several aspects have to be taken into account to give an overview of the performance of the complete system, the performance of several subsystems, the organization of parallel processing, the network transport, but also of the scenery, namely, the simulated room (dimension and complexity of the geometry), the velocity of sources, and finally the user. Updating the room acoustical simulation is the most time-consuming part of the system and requires a strategy of achieving the best perceptual performance. Image sources and ray tracing are processed independently on different CPUs. The binaural filter of the ray tracing process will be calculated directly on the ray-tracing server. The auralization server has to calculate the image source filter and combine all filter segments of the ray-tracing process. Figure 15 describes one possible segmentation of the ray tracing and combination of the image source filter. It should be mentioned that the length of the specular part is room dependent. The ray-tracing interrupt point will be adjusted based on the movement velocity of the listener and the sources. This means that the audio signal is filtered with the updated first part of the room impulse response while the generation of the late part by ray tracing is still in progress. The filter segment to be updated will be cut off from the complete filter with a short ramp of 32 samples ≈ 0.72 millisecond, and the new segment will be placed in with the same ramp to avoid audible artifacts.

Due to the dependency of all these factors, update times cannot be estimated in general. For this reason, we will give some detailed examples with respect to the performance measurements (see Tables 4 and 5) made in several sections above. It should be noticed that the image source filter will be updated at any time the source or the head moved more than 2 cm or turned more than 1 degree, respectively. The image source filter will be calculated on the current list of audible sources (positions updated). The resulting filter only contains a few wrong reflections which will be removed after the audibility test. Thus, the specular reflections at the first part of the impulse response become audible with the correct spatial representation already after 35 milliseconds (tracking + UDP transport + CTC filter generation IS filter generation + audio buffer swap). This is also the time needed to react to a listener's head rotation (see Table 5).

Table 4: Overview of performance measurements of the several subsystems.

Action	Time
Tracking	18.90 ms
UDP transport	0.26 ms
CTC filter generation	3.50 ms
Audio buffer swap	5.80 ms
IS audibility test	710.00 ms
IS filter (2 × 6.95 ms)	13.90 ms
Ray tracing	
500 ms impulse response length	1600.00 ms
1 s impulse response length	2600.00 ms
2 s impulse response length	3000.00 ms

6. SUMMARY

In this contribution, we introduced a quite complex system for simulation and auralization of room acoustics in real time. The system is capable of simulating room acoustical sound fields in any kind of enclosures without the prerequisite of any diffuse-field conditions. The room shape can hence be extremely long, flat, coupled, or of any other special property. The surface properties, too, can be freely chosen by using the amount of wave scattering according to standardized material data. Furthermore, the system includes a sound field reproduction for a single user based on dynamic crosstalk cancellation (virtual headphone). The software is implemented on standard PC hardware and requires no special processors. The performance (simulation processing time, filter update rates, tracker, and sound hardware latency) was evaluated and considered sufficiently in the case of a concert hall of medium size.

Particular features of the system are the following.

(i) It is not based on any assumption of an ideal diffuse sound field but on a full room acoustic simulation in two parts. Specular and scattered components of the impulse response are treated separately. Any kind of room shape and volume can be processed except small rooms at low frequencies.

(ii) The decision with regard to the amount of specular and diffuse reflections is just room dependent and purely based on physical sound field aspects.

(iii) The user will just be involved to create the room CAD model and the standard material data of absorption and scattering. Therefore, import functions of commercial non-real-time simulation software can be used. The fact that the auralization is performed in

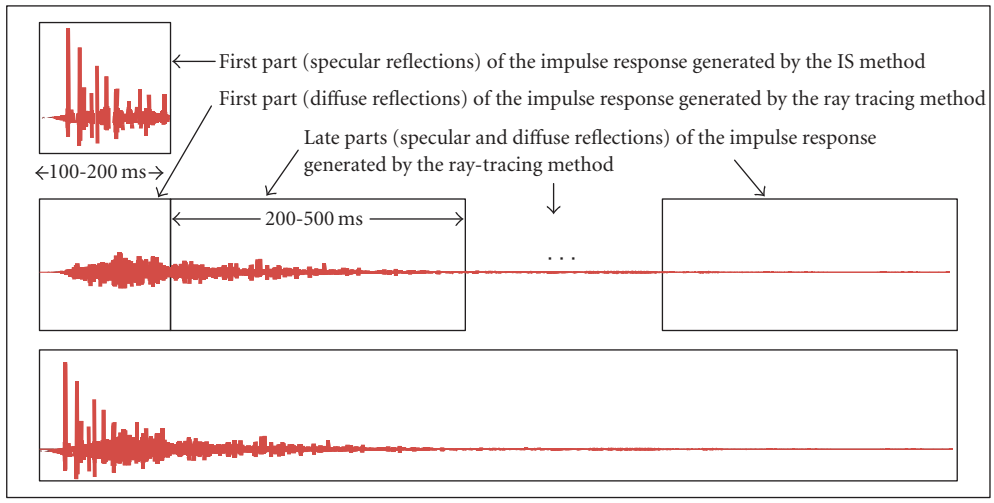


Figure 15: Combination of filter (or filter segments) for one ear generated by ray tracing and the first part of the impulse response generated by the image source model.

Table 5: Update intervals for different modes and conditions of head or source movements based on the measurements shown in Table 4.

Action	Update rate	Filter content to be updated
Head rotation	35 ms	Binaural processing in listeners coordinates
Translational head/source movement > 0.25 m	710 ms	Binaural processing in listeners coordinates Specular impulse response (3D image source cloud)
Translational head/source movement > 1.0 m (complete impulse response update)	3.0 s	Binaural processing in listeners coordinates Specular impulse response (3D image source cloud). Scattering impulse response (3D scattering matrix)
Fast translational head/source movement > 1.0 m (update of the first 500 ms)	1.6 s	Binaural processing in listeners coordinates Specular impulse response (3D image source cloud). Scattering impulse response (3D scattering matrix).

real time means that the user is not required to carry out any additional tasks. The system will adjust all relevant runtime parameters automatically and inherently, like division into specular and scattered parts and filter update rates.

(iv) The treatment of the components of the binaural impulse response is separated regarding the simulation itself, the update rate to the auralization server, and the convolution process.

(v) The decision regarding the update rate and depth of impulse response simulation is based on the interaction and speed of movement of the user in the VR system.

(vi) The precision of details in the impulse response, its exactness of delays, and its exactness of direction of sound incidence are just depending on the relative arrival time in the impulse response. This is in agreement with the ability of the human hearing system regarding localization and echo delays. Is should also be mentioned here that the system parameters of simulation depth and update rate are not controlled by the user but inherently treated in the system. This

way of processing will create full complexity and exact auralization in the very early part of the direct sound and the first reflections. Gradually, the sound energy will be transferred into the scattered component of the impulse. The precision and update rates are reduced, motivated by the limits due to psychoacoustic in masking effects. The system is open for further extension with respect to sound diffraction and sound insulation.

The real-time performance of the room acoustical simulation software was achieved by the introduction of a flexible framework for the interactive auralization of virtual environments. The concept of scene graphs for the efficient and flexible linkage of autonomously operating subscenes by means of so-called portals has been incorporated into the existing framework and combined with an underlying BSP-tree structure for processing geometry issues very fast. The use of this framework provides the possibility of a significant reduction of computation time for both applied algorithms (deterministic image sources and a stochastic ray tracer). Especially, the image source method is improved by the introduction of spatial data structures as portal states can be exploited so that the number of image sources to be processed can be reduced remarkably.

A fast low latency engine ensures that impulse responses regardless of their complete length will be considered by the filtering of the mono audio material after 5.8 milliseconds (block length 256 samples). Optimizations concerning modern processor extensions enable the rendering of, for example, 10 sources with filters of 3-second (132 000 taps) length or 15 sources with filters of 2-second length.

The reproduction of the binaural audio signal is provided by a dynamic crosstalk cancellation system with no restrictions to user movements. This system acts as a virtual headphone providing the channel separation without the need to wear physical headphones.

Gigabit Ethernet is used to connect the visual rendering system and the audio system. The visual VR system transmits the control commands as well as the spatial updates of the head and the sources. The control commands (e.g., start/stop) will be considered in the audio server after 0.15 millisecond so that the changes are served with the next sound output block for a tight audio video synchronism.

7. OUTLOOK

Despite the good performance of the whole system, there are many aspects that have to be investigated. To further enhance the quality of the room acoustical simulation, physical effects like sound insulation and diffraction are to be incorporated into the existing algorithms. In addition, the simulation of frequencies below the Schroeder frequency could be done by means of a fast and dynamic finite element method (FEM)-solver. The existing framework is already open to take these phenomena into account, the respective algorithms have only to be implemented. At present, the simulation software is implemented in a first version as a self-contained stable base. Thus, optimizing the algorithms is necessary to further increase their performance, especially with focus on the computing of processes in parallel. Position prediction could be a possibility of reducing the deviation of the position, the filter was calculated for, and the actual listener's position.

Preliminary listening tests showed that the generated virtual sources could be localized at a low error-rate [59]. The room acoustical simulation was perceived as plausible and matching to the generated visual image. In the future, more tests will be accomplished to evaluate the limitation of the update rates and the number of sources. Perception based reduction such as stated in, for example, [71, 72] is also an interesting method of reducing the processing costs, and will be considered in the future.

ACKNOWLEDGMENTS

The authors would like to thank Frank Wefers, Hilmar Demuth, and Philipp Dross for their commitment during parts of the programming work, and also Torsten Kuhlen, Andreas Franck, and Mark-Oliver Güld for their support and discussion. Furthermore, thanks to the DFG for funding parts of the project (DFG-Project "The Virtual Headphone," 2004–2006). Finally, the authors would like to thank the anonymous reviewers for the extended work which helped a lot to improve this contribution.

REFERENCES

[1] D. R. Begault, "Challenges to the successful implementation of 3-D sound," *Journal of the Audio Engineering Society*, vol. 39, no. 11, pp. 864–870, 1991.

[2] M. Naef, O. Staadt, and M. Gross, "Spatialized audio rendering for immersive virtual environments," in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST '02)*, pp. 65–72, Hong Kong, November 2002.

[3] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart, "The CAVE: audio visual experience automatic virtual environment," *Communications of the ACM*, vol. 35, no. 6, pp. 65–72, 1992.

[4] D. A. Burgess and J. C. Verlinden, "An architecture for spatial audio servers," in *Proceedings of Virtual Reality Systems Conference (Fall '93)*, New York, NY, USA, November 1993.

[5] J. D. Mulder and E. H. Dooijes, "Spatial audio in graphical applications," in *Visualization in Scientific Computing*, M. Göbel, H. Müller, and B. Urban, Eds., pp. 215–229, Springer, Wien, Austria, 1994.

[6] Lake Huron, 2005, http://www.lake.com.au/.

[7] L. Savioja, *Modeling Techniques for Virtual Acoustics*, Ph.D. thesis, Helsinki University of Technology, Helsinki, Finland, December 1999.

[8] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *Journal of the Audio Engineering Society*, vol. 47, no. 9, pp. 675–705, 1999.

[9] T. Funkhouser, P. Min, and I. Carlbom, "Real-time acoustic modeling for distributed virtual environments," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*, pp. 365–374, Los Angeles, Calif, USA, August 1999.

[10] R. L. Storms, "Npsnet-3D Sound Server: An Effective Use of the Auditory Channel," 1995.

[11] H. Kuttruff, *Room Acoustics*, Elsevier Science Publisher, New York, NY, USA, 4th edition, 2000.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[13] J. Borish, "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–1836, 1984.

[14] B.-I. L. Dalenbäck, "Room acoustic prediction based on a unified treatment of diffuse and specular reflection," *The Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 899–909, 1996.

[15] P.-A. Forsberg, "Fully discrete ray tracing," *Applied Acoustics*, vol. 18, no. 6, pp. 393–397, 1985.

[16] T. Funkhouser, N. Tsingos, I. Carlbom, et al., "A beam tracing method for interactive architectural acoustics," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 739–756, 2004.

[17] G. M. Naylor, "ODEON—another hybrid room acoustical model," *Applied Acoustics*, vol. 38, no. 2–4, pp. 131–143, 1993.

[18] U. M. Stephenson, "Quantized pyramidal beam tracing—a new algorithm for room acoustics and noise immission prognosis," *Acta Acustica United with Acustica*, vol. 82, no. 3, pp. 517–525, 1996.

[19] D. van Maercke, "Simulation of sound fields in time and frequency domain using a geometrical model," in *Proceedings of the 12th International Congress on Acoustics (ICA '86)*, vol. 2, Toronto, Ontario, Canada, July 1986, paper E11-7.

[20] M. Vorländer, "Simulation of the transient and steady state sound propagation in rooms using a new combined sound particle—image source algorithm," *The Journal of the Acoustical Society of America*, vol. 86, pp. 172–178, 1989.

[21] I. Bork, "A comparison of room simulation software—the 2nd round Robin on room acoustical computer simulation," *Acta Acustica United with Acustica*, vol. 86, no. 6, pp. 943–956, 2000.

[22] M. Vorländer, "International round Robin on room acoustical computer simulations," in *Proceedings of the 15th International Congress on Acoustics (ICA '95)*, pp. 689–692, Trondheim, Norway, June 1995.

[23] H. Kuttruff, "A simple iteration scheme for the computation of decay constants in enclosures with diffusely reflecting boundaries," *The Journal of the Acoustical Society of America*, vol. 98, no. 1, pp. 288–293, 1995.

[24] C. L. Christensen and J. H. Rindel, "A new scattering method that combines roughness and diffraction effects," in *Forum Acousticum*, Budapest, Hungary, 2005.

[25] R. Heinz, "Binaural room simulation based on an image source model with addition of statistical methods to include the diffuse sound scattering of walls and to predict the reverberant tail," *Applied Acoustics*, vol. 38, no. 2–4, pp. 145–159, 1993.

[26] Y. W. Lam, "A comparison of three reflection modelling methods used in room acoustics computer models," *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2181–2192, 1996.

[27] M. Vorländer, "Ein Strahlverfolgungsverfahren zur Berechnung von Schallfeldern in Räumen," *Acustica*, vol. 65, no. 3, pp. 138–148, 1988.

[28] J. S. Suh and P. A. Nelson, "Measurement of transient response of rooms and comparison with geometrical acoustic models," *The Journal of the Acoustical Society of America*, vol. 105, no. 4, pp. 2304–2317, 1999.

[29] U. P. Svensson, R. I. Fred, and J. Vanderkooy, "An analytic secondary source model of edge diffraction impulse responses," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2331–2344, 1999.

[30] N. Tsingos, T. Funkhouser, A. Ngan, and I. Carlbom, "Modeling acoustics in virtual environments using the uniform theory of diffraction," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*, pp. 545–552, Los Angeles, Calif, USA, August 2001.

[31] U. M. Stephenson, *Beugungssimulation ohne Rechenzeitexplosion: die Methode der quantisierten Pyramidenstrahlen; ein neues Berechnungsverfahren für Raumakustik und Lärmimmissionsprognose; Vergleiche, Ansätze, Lösungen*, Ph.D. thesis, RWTH Aachen University, Aachen, Germany, 2004.

[32] M. Slater, A. Steed, and Y. Chrysanthou, *Computer Graphics and Virtual Environments: From Realism to Real-Time*, Addison Wesley, New York, NY, USA, 2001.

[33] L. Cremer and H. A. Müller, *Die wissenschaftlichen Grundlagen der Raumakustik—Band 1*, S. Hirzel, Stuttgart, Germany, 2nd edition, 1978.

[34] T. Akenine-Möller and E. Haines, *Real-Time Rendering*, A. K. Peters, Natick, Mass, USA, 2nd edition, 2002.

[35] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics, Principles and Practice*, Addison Wesley, Reading, Mass, USA, 2nd edition, 1996.

[36] R. Shumacker, R. Brand, M. Gilliland, and W. Sharp, "Study for applying computer-generated images to visual simulations," Report AFHRL-TR-69-14, U.S. Air Force Human Resources Laboratory, San Antonio, Tex, USA, 1969.

[37] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, Mass, USA, 2nd edition, 2001.

[38] D. Schröder and T. Lentz, "Real-time processing of image sources using binary space partitioning," *Journal of the Audio Engineering Society*, vol. 54, no. 7-8, pp. 604–619, 2006.

[39] R. Heinz, *Entwicklung und Beurteilung von computergestützten Methoden zur binauralen Raumsimulation*, Ph.D. thesis, RWTH Aachen University, Aachen, Germany, 1994.

[40] J. S. Bradley and G. A. Soulodre, "The influence of late arriving energy on spatial impression," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2263–2271, 1995.

[41] J. H. Rindel, "Evaluation of room acoustic qualities and defects by use of auralization," in *Proceedings of the 148th Meeting of the Acoustical Society of America*, San Diego, Calif, USA, November 2004.

[42] D. Schröder, P. Dross, and M. Vorländer, "A fast reverberation estimator for virtual environments," in *Proceedings of the AES 30th International Conference*, Saariselkä, Finland, March 2007.

[43] T. Brookes and C. Treble, "The effect of non-symmetrical left/right recording pinnae on the perceived externalisation of binaural recordings," in *Proceedings of the 118th Audio Engineering Society Convention*, Barcelona, Spain, May 2005.

[44] D. S. Brungart, W. M. Rabinowitz, and N. I. Durlach, "Auditory localization of a nearby point source," *The Journal of the Acoustical Society of America*, vol. 100, no. 4, p. 2593, 1996.

[45] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, no. 6713, pp. 747–749, 1998.

[46] H. Lehnert and M. Richter, "Auditory virtual environment: simplified treatment of reflections," in *Proceedings of the 15th International Congress on Acoustics (ICA '95)*, Trondheim, Norway, June 1995.

[47] G. Romanenko and M. Vorländer, "Employment of spherical wave reflection coefficient in room acoustics," in *IoA Symposium Surface Acoustics*, Salford, UK, 2003.

[48] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti, "Surround-screen projection-based virtual reality: the design and implementation of the CAVE," in *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '93)*, pp. 135–142, ACM Press, Anaheim, Calif, USA, August 1993.

[49] B. B. Bauer, "Stereophonic earphones and binaural loudspeakers," *Journal of the Audio Engineering Society*, vol. 9, no. 2, pp. 148–151, 1961.

[50] O. Kirkeby, P. A. Nelson, and H. Hamada, "Local sound field reproduction using two closely spaced loudspeakers," *The Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 1973–1981, 1998.

[51] H. Møller, "Reproduction of artificial-head recordings through loudspeakers," *Journal of the Audio Engineering Society*, vol. 37, no. 1-2, pp. 30–33, 1989.

[52] W. G. Gardner, *3-D audio using loudspeakers*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass, USA, 1997.

[53] T. Lentz and O. Schmitz, "Realisation of an adaptive cross-talk cancellation system for a moving listener," in *Proceedings of the 21st Audio Engineering Society Conference*, St. Petersburg, Russia, June 2002.

[54] T. Lentz and G. K. Behler, "Dynamic cross-talk cancellation for binaural synthesis in virtual reality environments," in *Proceedings of the 117th Audio Engineering Society Convention*, San Francisco, Calif, USA, October 2004.

[55] Steinberg, "ASIO 2.0 Audio Streaming Input Output Development Kit," 2004.

[56] T. Lentz, "Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments," *Journal of the Audio Engineering Society*, vol. 54, no. 4, pp. 283–294, 2006.

[57] T. Takeuchi, P. Nelson, O. Kirkeby, and H. Hamada, "The effects of reflections on the performance of virtual acoustic imaging systems," in *Proceedings of the International Symposium on Active Control of Sound and Vibration (ACTIVE '97)*, pp. 955–966, Budapest, Hungary, August 1997.

[58] D. B. Ward, "On the performance of acoustic crosstalk cancellation in a reverberant environment," *The Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1195–1198, 2001.

[59] T. Lentz, J. Sokoll, and I. Assenmacher, "Performance of spatial audio using dynamic cross-talk cancellation," in *Proceedings of the 119th Audio Engineering Society Convention*, New York, NY, USA, October 2005.

[60] W. G. Gardner, "Efficient convolution without input-output delay," *Journal of the Audio Engineering Society*, vol. 43, no. 3, pp. 127–136, 1995.

[61] J. J. La Viola Jr., "A testbed for studying and choosing predictive tracking algorithms in virtual environments," in *Proceedings of the 7th International Immersive Projection Technologies Workshop, 9th Eurographics Workshop on Virtual Environments*, pp. 189–198, Zurich, Switzerland, May 2003.

[62] R. Azuma and G. Bishop, "A frequency-domain analysis of head-motion prediction," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '95)*, pp. 401–408, ACM Press, Los Angeles, Calif, USA, August 1995.

[63] L. Chai, W. A. Hoff, and T. Vincent, "Three-dimensional motion and structure estimation using inertial sensors and computer vision for augmented reality," *Presence: Teleoperators and Virtual Environments*, vol. 11, no. 5, pp. 474–492, 2002.

[64] J.-R. Wu and M. Ouhyoung, "A 3D tracking experiment on latency and its compensation methods in virtual environments," in *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology (UIST '95)*, pp. 41–49, ACM Press, Pittsburgh, Pa, USA, November 1995.

[65] I. B. Witew, "Spatial variation of lateral measures in different concert halls," in *Proceedings of the 18th International Congress on Acoustics (ICA '04)*, vol. 4, p. 2949, Kyoto, Japan, April 2004.

[66] R. Azuma and G. Bishop, "Improving static and dynamic registration in an optical see-through HMD," in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '94)*, pp. 197–204, ACM Press, New York, NY, USA, July 1994.

[67] W. Pompetzki, *Psychoakustische Verifikation von Computermodellen zur binauralen Raumsimulation*, Ph.D. thesis, Ruhr-Universität Bochum, Bochum, Germany, 1993.

[68] M. Vorländer and E. Mommertz, "Definition and measurement of random-incidence scattering coefficients," *Applied Acoustics*, vol. 60, no. 2, pp. 187–199, 2000.

[69] ISO 354, "Acoustics, Measurement of sound absorption in a reverberant room," 2003.

[70] ISO/DIS 17497-1, "Acoustics Measurement of the sound scattering properties of surfaces—part 1: measurement of the randomincidence scattering coefficient in a reverberation room".

[71] N. Tsingos, "Scalable perceptual mixing and filtering of audio signals using an augmented spectral representation," in *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx '05)*, Madrid, Spain, September 2005.

[72] N. Tsingos, E. Gallo, and G. Drettakis, "Perceptual audio rendering of complex virtual environments," in *Proceedings of the 31st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '04)*, pp. 249–258, Los Angeles, Calif, USA, August 2004.

Tobias Lentz was born in Mönchengladbach, Germany, in 1971. He studied electrical engineering at RWTH Aachen, Germany, from where he received a Dipl.-Ing. (M.Sc.) degree in 2001. Since 2001 he has been working as a Research Assistant and is currently a Ph.D. candidate at the Institute of Technical Acoustics, RWTH Aachen University. His main focus is on three-dimensional audio technologies, architectural acoustics, crosstalk cancellation, binaural technology, and real-time applications for virtual reality. Currently, he is finishing his Ph.D. thesis on "Binaural Technology for Virtual Reality." He is a Member of the Audio Engineering Society (AES) and the German Acoustical Association (DEGA).

Dirk Schröder was born in Cologne, Germany, in 1974. He studied electrical engineering and information technology at RWTH Aachen University, Germany, and received a degree of Dipl.-Ing. (M.Sc.) in 2004. He has been working at the Institute of Technical Acoustics, RWTH Aachen University, as a Research Assistant since 2005 and is currently a Ph.D. candidate at RWTH Aachen University. His main research field is room acoustic simulation with special focus on interactive real-time applications such as virtual reality. He is a Member of the Audio Engineering Society (AES) and the German Acoustical Association (DEGA).

Michael Vorländer is a Professor at RWTH Aachen University, Germany. After university education in physics and doctor degree (Aachen, 1989 with a thesis in room acoustical computer simulation), he worked in various fields of acoustics at the PTB Braunschweig, the National Laboratory for Physics and Technology. In 1995 he finished the qualification as university lecturer (habilitation) with a thesis on reciprocity calibration of microphones. In 1996 he accepted an offer from RWTH Aachen University for a Chair and Director of the Institute of Technical Acoustics. He is President of the European Acoustics Association, EAA, in the term 2004–2007 and former Editor-in-Chief of the International Journal Acta Acustica united with Acustica (1998–2003). He is a Member of the German Acoustical Society, DEGA, of the German Physical Society, DPG, and a Fellow of the Acoustical Society of America, ASA.

Ingo Assenmacher was born in Düren, Germany, in 1974. He studied computer science at RWTH Aachen University, Aachen, and received a degree of Dipl.-Inform. (M.Sc.) degree in 2002. He is currently working at the Center for Computation and Communication, RWTH Aachen University, as a Research Assistant and is a Ph.D. candidate at RWTH Aachen University. His main research fields are interaction in immersive Virtual Environments, software methods for real-time environments and virtual-reality-based data visualization and exploration.