3D REPRODUCTION OF ROOM AURALIZATIONS BY COMBINING INTENSITY PANNING, CROSSTALK CANCELLATION AND AMBISONICS

Sönke Pelzer, Bruno Masiero, Michael Vorländer

Institute of Technical Acoustics,
RWTH Aachen University,
Aachen, Germany
{spe,mvo}@akustik.rwth-aachen.de

ABSTRACT

Popular room acoustic simulations use hybrid models for precise calculation of the early specular reflections and stochastic algorithms for the late diffuse decay. Splitting the impulse response into early and late parts is also psychoacoustically reasonable. The early part is responsible for the localization and the spatial and spectral perception of sources, which makes the correct reproduction of its time-frequency structure important. In contrast the later part is responsible for the sense of spaciousness and envelopment, properties related to the room and its diffuse decay.

Nevertheless, in auralization systems the reproduction of the whole impulse response is done through the same reproduction system and method, even though there are systems better suited to coherent reproduction (important for the early arrivals of an impulse response) and others better suited for the reproduction of incoherent fields (the reverberant tail of an impulse response).

A hybrid approach is presented which uses one common loudspeaker system for the simultaneous rendering of different reproduction methods. A method with strong localization cues such as binaural via crosstalk cancellation or VBAP is used for the direct sound and early reflections, while a method with higher immersion and envelopment such as Ambisonics is used for the diffuse decay.

1. INTRODUCTION

The challenge of generating high quality artificial reverberation has been dealt with since the late 1950's with many studies and publications. Important insights about properties of a room impulse response were already derived, e.g. by Schroeder in 1954 [1]. Whilst this was mostly constrained to theoretical thoughts, early work in the field of artificial reverberation based on simple analogue feedback loops. The main goal in these times aimed at producing natural sounding reverberation [2]. Only after the introduction of the first computational algorithms for the estimation of real reflections, already known algorithms such as ray tracing (RT) and the image source method (ISM) were applied in acoustics by Krokstad in 1968 [3] and Allen and Berkley in 1979 [4]. From then on the focus shifted towards the replication of real and complex shaped rooms. Nevertheless, due to the high computational demand, it was not until 1984 that Borish [5] extended the popular image source model to arbitrary polyhedra. Until today, the combination of these two models in hybrid algorithms mark the state-of-the-art in room acoustics simulation and auralization techniques [6, 7, 8, 9], although more accurate approaches for the estimation of sound propagation in rooms are known. They base on Finite-Element-Methods (FEM), Boundary-Element-Methods (BEM) or Finite-Time-Differences (FDTD), but they suffer from high numerical demands on computation power and are thus hardly applicable for normal to larger rooms or broadband simulations including higher frequencies. Recent approaches used a combined wave and ray based simulation method, which calculates the lower end (e.g. below the Schroeder frequency) using the FEM [10]. Geometrically based simulations, such as the described RT or IS methods have, on the other hand, highly developed representatives that already realize real-time capabilities [11].

2. ROOMS ACOUSTICS, EARLY/LATE REFLECTIONS AND MIXING TIME

The room impulse response can be divided into an early part which is dominated by distinct strong early reflections and a late part that mainly consists of reflections which have been reflected and scattered several times, so that they thoroughly overlap due to increased reflection density over time and the broadening of the impulses with higher reflection orders. Many attempts have been made to define the transition time between these two parts on a physical basis, but recent conclusions show that physical mixing does not explain diffusion and does not define the moment when a sound field turns diffuse [12]. It is in question if a perfectly diffuse reverberation exists at all in a real room. As the motivation for the separation of the impulse response is based on a psychoacoustic effect, it can be concluded that the human auditory system is not able to distinguish single reflections anymore as from a certain reflection density, a consensus in literature [13, 14]. Thus the transition time can still be determined in perceptual investigations, of which many have been conducted in the recent years. Unfortunately, most of them were restricted to only one room [13, 14], so that generalized conclusions cannot be drawn.

A detailed comprehensive overview of physical predictors for the estimation of the transition time as well as their evaluation on a perceptual basis can be found in a recent publication by Lindau [15]. The investigated predictors comprised model based ones (deriving the transition time from room parameters such as volume and mean free path length) as well as impulse response based ones (analyzing the time domain impulse response).

Shoebox shaped rooms usually have longer mixing times, due to their long unobstructed path length and regular shape. For these enclosures, Lindau found a transition time t_m, proportional to the mean free path length, with

$$t_m = 20V/S + 12 \quad [ms], \tag{1}$$

V being the room volume and S the room's surface area. Absorption and reverberation time were not found to have significant influence.

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

$$O_{IS} \cdot \bar{t} = t_m \tag{2}$$

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

$$O_{IS,min} = \frac{t_{m-12}}{\bar{t}} + 1 \approx 2.7 \tag{3}$$

It can be concluded that for rooms, as selected by Lindau, which had shoebox shape and volumes in a wide range from $182m^3$ up to $8500m^3$, each with varied mean absorption, a general minimum IS order can be defined that results to three. After this third reflection, the sound field can be expected to be mostly mixing, uniform and isotropic, yielding a diffuse late reverberation. Similar observations were found by Kuttruff [16] when he analyzed the contributions of specular und diffuse energy in a room impulse response (RIR), as shown in Figure 1.

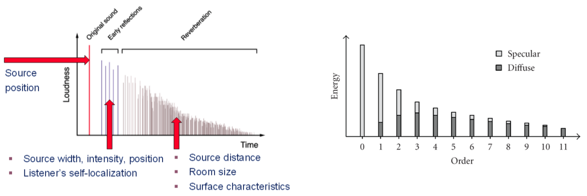

Figure 1: Left: *Perceptual and physical division of the room impulse response.* Right: *Relation of specularly and diffusely reflected sound in a typical room[16].*

Scattered reflections in the early part und all reflections after the image sources cut-off time (which should at least cover the mixing time) are then calculated using the ray tracing technique which calculates the temporal energy envelopes for each frequency band. The reflection modeling of image sources (IS) and RT are illustrated in Figure 2.

3. ROOM AURALIZATIONS USING SPATIAL 3D SOUND REPRODUCTION

To provide immersive auralizations the simulation results are processed so that they can be reproduced over headphones or loudspeakers. On the reproduction side it is important to remember the psychoacoustic motivation of the separation of components in the room impulse response. Early reflections and especially the direct sound have to be reproduced with highest precision in terms of time and direction of arrival and frequency spectrum. Due to the precedence effect, the direct sound has a major influence on the localization of a source and the early reflections will affect the perceived source width. The reproduction system has to make sure that localization is as natural as possible, including exact compliance with frequency-dependent interaural level and time differences [17].

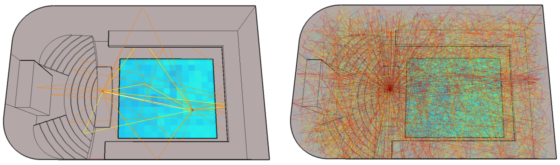

Figure 2: *Visualization of modeled reflections in a concert hall. Early reflections are constructed using the Image Source method (left) while late reverberation is modeled using ray tracing (right).*

The point of full three-dimensionality is missed in many spatial reproduction techniques. A 3D reproduction should include not only horizontally distributed sources, but also the incidence from elevated angles and near field effects for sources that are close to the head of the listener [18]. Even large and expensive wave-field synthesis (WFS) systems mostly do not provide height information. More commonly used and more affordable systems such as vector-base amplitude panning (VBAP) and Ambisonics can theoretically reproduce elevated sources, but there are only few implementations that support realistic distance perception. VBAP has no support for close-by sources and Ambisonics only in near-field compensated higher-order ambisonics (NFC-HOA) setups [19].

Regarding this, binaural technology has a lot of advantages in 3D rendering, being very close to the way how the human ear perceives sound in nature. But as a major disadvantage it is difficult to reproduce binaural cues using loudspeakers. Using headphones on the other hand is not only problematic in terms of comfort and externalization, but also usually not able to impart the feeling of envelopment in diffuse sound fields. Additional problems such as the necessity to compensate for individual headphone transfer functions accrue.

A popular method to reproduce binaural signals is the crosstalk cancellation (CTC) [20], also called transaural in some publications. It uses a regular loudspeaker system, with only two speakers required, and takes advantage of wave interference to achieve a sufficient channel separation between the left and right ear of the listener. The main drawback of this system is the requirement to accurately know the current position of the user, which is typically solved using a tracking system and continuous adaption of the CTC filters [21]. Thus, this technique is often found in virtual reality systems, when the user is already tracked for interaction or 3D visualization [22].

Guastavino et al. [23] compared different reproduction techniques (CTC, Ambisonics, Panning) and came to similar results as described above and summarized in Table 1 (with additional comments by this author). It can be concluded that the reproduction method must also account for the psychoacoustics that define our hearing in rooms.

Table 1: *Comparison of different reproduction techniques, as published by Guastavino [23] with additional comments.*

Method	Advantages	Drawbacks
Binaural CTC	Precise localization, good readability, near field sources	Poor realism, lack of immersion/envelopment, needs individual HRTF
Ambisonics	Strong immersion and envelopment	Poor localization/readability
Stereo Panning	Precise localization	Lack of immersion/envelopment

4. HYBRID REPRODUCTION SYSTEMS

The idea to combine different systems for a separated reproduction of direct sound and reverberation was first mentioned in the Ambiophonics group in the early 1980s, mainly supported by Glasgal, Farina and Miller [24]. Their approach proposed a crosstalk canceled stereo-dipole playback for a wider stereo image and optional additional ambience speakers fed by the original signal convolved with an IR of a hall or similar reverberant space. The idea mainly aimed at an advanced reproduction of commercially available stereo recordings that were performed with certain popular microphone arrangements, such as ORTF or M/S, but the group also proposed their own microphone methodology and called it Ambiophone: two head-spaced omnidirectional microphones with a baffle behind them to muffle room reflections from non-frontal directions. Farina combined the stereo-dipole technique then with Ambisonics and had the chance to convolve his recordings with Ambisonics impulse responses of the hall where the recordings were actually made. The application of Ambiophonics can mainly be seen in the enhancement of stereo or 5.1 recordings, but the optional ambience channels have to be seen more as an artificial effect due to the fact that general recordings do not come with spatial impulse responses of the recording venue.

It was not before 2010 that Favrot proposed to apply the idea of hybrid reproduction that is matched to the events in a RIR to room acoustics prediction models which can generate spatial IRs for existing or virtual halls [25]. He used a variable Ambisonics order for the early and late part of the RIR to benefit from reduced computation load for late reverberation and better localization of the direct sound.

In this present contribution a combined hybrid system is introduced that uses one common loudspeaker system to play a CTC and an Ambisonics signal at the same time. The binaural signal will ensure high detail of temporal and spectral features of the direct sound and early reflections, while the Ambisonics signal is used to produce a spacious and enveloping diffuse sound field. The poor localization abilities of Ambisonics are published in a variety of studies [23, 26], and the poor immersion of binaural or transaural reproduction is documented as well [23]. Both observations clearly motivate the hybrid approach where binaural signals are used for the direct sound and early reflections and Ambisonics for the late decay.

Table 1 shows how the Pros and Cons of these two technologies are close to being perfectly complementary. The benefit and effort of binaural signals that use individual head-related transfer function (HRTF) are recently discussed. An investigation by Majdak found that a mismatch and lack of individualization substantially degraded the localization performance of targets placed outside of the loudspeaker span and behind the listeners, showing the relevance of individualized CTC systems for those targets [27].

The earlier introduced transition time is perfectly qualified to define the crossover between the two reproduction systems, with the same motivation as for the simulation. Therefore the CTC is used to reproduce the direct sound and specular reflections up the order of 3. Further reflection paths and all scattered reflections are fed into the Ambisonics engine. The presented idea is not meant to replace any cinema or public address system, due to the fact that the CTC is a single-user experience. It is more aimed at sophisticated room acoustics simulation and reproduction in virtual acoustics applications, such as virtual concert hall prototyping or fully immersive virtual environments [22].

4.1. Binaural Synthesis

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

Virtual sound sources closer than 2 m need appropriate HRTF data that is measured in the near field (0.2, 0.3, 0.4, 0.5, 0.75, 1.0, 2.0 meters) [21]. If no such near field data is available, a range extrapolation should be applied, as proposed by Pollow [28]. If a NFC-HOA decoder is available, the range extrapolation can be implemented by interpreting a set of fixed-range HRTF as a virtual loudspeaker array [29].

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

4.2. Crosstalk Cancellation

A loudspeaker-based binaural reproduction chain starts with a binaural signal that can be either recorded using an artificial head or, in case of the presented method, simulated and synthesized by convolution of a HRTF or binaural RIR with an anechoic signal. A crosstalk cancellation filter network makes sure that the original binaural signal arrives at the listeners eardrums. Ideally, the CTC filters have to be constantly adjusted to the listener's head position and rotation. Combined with a dynamic binaural synthesis, a dynamic CTC allows a realistic spatial reproduction with only few loudspeakers. Dynamically adjusting CTC filters and binaural IRs in real-time requires considerable system complexity and low-latency convolution, both available since a few years [21]. The tracking devices often base on electromagnetic or optic input, but current developments aim at contact-free 6-degrees-of-freedom tracking by using infrared depth maps or video-based face detection.

When combined with an Ambisonics or other reproduction setup that already offers a multiple loudspeaker installation, it is possible to select two speakers of this setup which will serve the best possible channel separation. In a dynamically tracked system, this loudspeaker pair can be continuously exchanged dependent on the user's head position, without noticeable switching artifacts, as shown by Lentz [21]. Equation (4) describes the CTC as a closed-form solution, with $Z_{L/R}$ denoting the perceived signal:

$$\begin{bmatrix} Z_L \\ Z_R \end{bmatrix} = \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix} \begin{bmatrix} Y_L \\ Y_R \end{bmatrix} = \mathbf{Hy} = \mathbf{z} \tag{4}$$

The filters for the CTC are placed prior to the loudspeakers, so that $y = C \cdot x$. The transfer function of the complete system is given in matrix form as

$$\mathbf{z} = \mathbf{H} \cdot \mathbf{C} \cdot \mathbf{x} \tag{5}$$

For a binaural reproduction the output z should be equal to the input x apart from a time delay. Thus, the following equation has to be valid:

$$\mathbf{H} \cdot \mathbf{C} = \mathbf{e}^{-\mathbf{j}\boldsymbol{\Delta}} \cdot \mathbf{I} \tag{6}$$

with I being the identity matrix. The transfer matrix with the crosstalk cancellation filters C can easily be obtained by means of a pseudo-inverse of the transfer matrix H, resulting in:

$$\mathbf{C} = \mathbf{e}^{-\mathbf{j}\boldsymbol{\Delta}} \cdot \mathbf{H}^{+} \tag{7}$$

The closed-form solution according to equation (7) is the exact solution for the entire crosstalk cancellation. It requires, however, infinitely long filters that are also prone to have stability problems. The later issue can be dealt with through a regularized matrix inversion approach.

The regularization applies a constraint at the maximum gain allowed to the filters and can be expressed as follows:

$$\mathbf{C} = \mathbf{e}^{-j\Delta} \cdot (\mathbf{H}^H\mathbf{H} + \beta(f)\mathbf{I})^{-1} \cdot \mathbf{H}^H \tag{8}$$

This approach has no requirement on the matrix $\mathbf{H}$ to be square, meaning that more than two loudspeakers could be simultaneously used to achieve improved channel separation [30].

4.3. Higher-Order Ambisonics

The Ambisonics technique was initially designed in the early 1970s to perform spatial recordings and multi-channel broadcasting [31]. The known recording method of intensity stereophony with two perpendicularly superposed cardoid microphones (XY-arrangement) was upgraded by the use of an extra figure-of-eight microphone perpendicular to the other two microphones – note that the XY-arrangement with two cardioids can be substituted by two figure-of-eight and an omnidirectional microphone. This configuration corresponds already to the 0th and 1st spherical harmonics (SH) orders. Moreover, the original formulation of 1st order Ambisonics can be expanded to higher spherical harmonics orders, the so called higher-order ambisonics (HOA). This improves the usually imprecise localization of only 1st order reproduction at the cost of a more complex recording and reproduction system.

An Ambisonics microphone with three figure-of-eight microphones plus one omnidirectional microphone at the same position is unpractical. However, a set-up with four omnidirectional microphones on the faces of a tetrahedron can be used instead by later transforming the signals into the desired omnidirectional and figure-of-eight patterns using spherical harmonics transformation. The microphone signals are called A-format while the transformed signals are called B-format. The B-format signals can be independently stored or broadcasted and for playback they are adequately decoded into the G-format which is directly fed into the speaker set-up available for reproduction. The B-format guarantees storage and transmission of spatial audio data, independent of the decoding stage. The decoding step is then only dependent on the available loudspeaker setup, which has to be dimensioned to fulfill the requirements of the Ambisonics order N, i.e. the number of loudspeakers L has to be at least:

$$L \approx (N+1)^2 \tag{9}$$

In the proposed hybrid system Ambisonics is only used for late reflections, therefore the usual implementation of plane wave sources would be sufficient and near-field compensation (NFC) [19] is not essential. However, to include comparisons of direct sound rendering using the different methods, also a NFC decoder has been implemented.

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

4.3.1. Generation of Ambisonics B-format Impulse Responses

The late reverberation is predicted using a ray tracer. Thus, the simulation result is a data structure that contains the amount of energy that is arriving from a certain direction in a certain time interval in a certain frequency band, as shown in Figure 3. The temporal, spectral and spatial domains are discretized, usually in accordance with the number of rays for the desired resolutions.

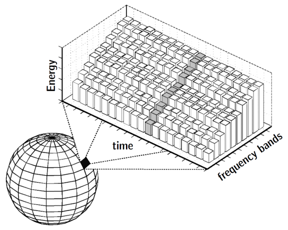

Figure 3: *Acoustic ray tracing results in a spatial data structure with time-frequency information of the energy of incident rays for each detection sphere.*

As the number of rays is meant to be kept a variable parameter and is usually chosen to a much lower number compared to the amount of real reflections in a room, the late reverberation is modeled by a synthetic sequence of Dirac pulses. If these pulses are arranged in accordance with the exponentially growing actual reflection densities over time, as derived by Kuttruff [16], then the whole sequence describes a Poisson process. The spectrum of this Poisson sequence is flat, so that no coloration occurs. In order to apply the temporal envelope of the ray tracing result for a certain frequency band, the noise sequence is filtered through an octave filter bank. Before the distinct pulses are smeared and overlapped by the filter bank, they are at first weighted by spherical harmonics. Therefore each single pulse is inserted in an own channel for each spherical harmonics order/degree with an amplitude according to the direction of arrival which is known by the ray tracing results. The band filtered Poisson sequences, which contain already temporally and spatially weighted pulses, are then superposed and result in the channels of a broadband B-format impulse response.

To enable comprehensive listening tests, it is also possible to render all image sources additionally into the B-format, so that the whole room impulse response (including direct sound and early reflections) can be played back through the Ambisonics system.

As for the IR synthesis algorithm, there is no limitation to the maximum SH order. In practice, 1st order reproduction can be sufficient in a hybrid system, while the cues that are important for localization are covered by the CTC. Anyway, if the late decay is not spatially homogeneous, e.g. in case of late echos from certain directions, a higher order encoding will improve the localization of late reflections.

In preliminary informal listening tests, the 2nd order signals were judged better than 1st order signals, even when only applied for late reflections during combined hybrid CTC/Ambisonics playback. Especially for cases when unusual room shapes should be simulated, such as long L-shaped corridors, it will become important to have spatially coded late reverberation. Problems as reported by other authors [32, 33], who encountered phenomena such as coloration or inside-head localization due to the correlation of the loudspeaker feeds, were not noticed in the test trials with the presented method. However, the decision if late reverberation should be spatially coded and thus have correlated loudspeaker feeds or if uncorrelated noise should be used which assumes a perfectly diffuse sound field and misses out on any spatial cues is subject to upcoming listening tests in the near future.

4.4. Calibration

To be able to seamlessly mix the early and late part of the IR with different reproduction techniques it must be ensured that their levels are accurately adjusted. In an ideal case under free field conditions it is possible to calculate the resulting sound pressure levels at a single position for any of the presented techniques. Under real world conditions the perfect sweet spot does not exist because the signals are presented to a human listener with two ears. In case of Ambisonics and CTC the assumption of ideal interference of the signals in a single point (Ambisonics) or two ear drums (CTC) usually does not hold true.

Dependent on the actual speaker layout the position of a virtual source has also an impact on the resulting level. This is especially the case if the layout is not regular or if the mounting conditions of each speaker are not exactly the same (which is nearly impossible to achieve in a normal room).

Therefore it is hard to calculate the accurate binaural sound pressure levels. To equalize the levels as best as possible without any knowledge about the virtual scene or real listening room, an equal distribution of virtual sources on a sphere (>900 sources) was used. The listening room with installed loudspeaker system is then measured or simulated and the loudspeaker impulse responses are used for CTC, HOA and VBAP decoding. To prepare signals for an unknown listening room, the loudspeaker IRs can be simulated for free field conditions. However, the impact of the listening room on the final levels and the calibration between the different formats has not been analyzed yet. The author assumes that the impact is different for CTC compared to Ambisonics or VBAP.

A general investigation on the impact of the listening room on loudspeaker reproduction of auralizations that include reverberation was performed by the author and will be published [34].

5. LOCALIZATION PERFORMANCE LISTENING TESTS

The accuracy of localization of virtual sources was measured for different reproduction methods. A listening test was conducted in a fully anechoic room with a 24-channel loudspeaker array, as shown in Figure 4). The loudspeakers were arranged in three layers with elevations of 0° and ±30° and an azimuth angle of 45° between each loudspeaker starting with a frontal direction of 0° azimuth.

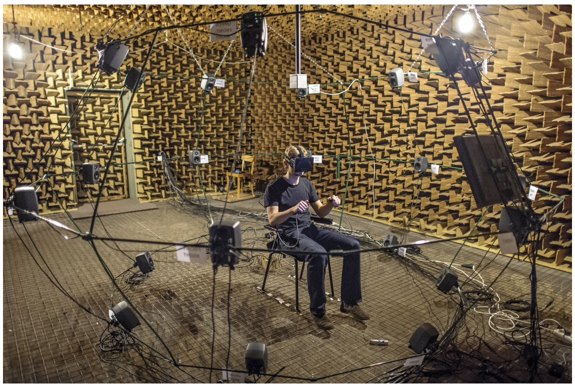

Figure 4: *Listening tests were conducted in an anechoic chamber equipped with 24 loudspeakers for spatial reproduction.*

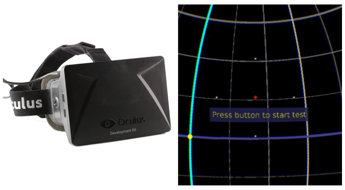

Figure 5: *Tracked head-mounted display and virtual environment for accurate pointing.*

A tracked head-mounted display (HMD, Oculus Rift) provided an accurate and bias-free pointing method [35]. A three-dimensional virtual sphere was rendered with a grid in 15° resolution and reference lines for horizontal and median planes, as shown in Figure 5. The listener's current view direction and head orientation/rotation were shown. In a training phase only real loudspeakers were driven with pink noise and the HMD displayed the actual source position. Averaged over all subjects an average pointing accuracy of 0.3° was measured.

Five reproduction methods were tested. Three pure implementations of CTC, VBAP and 4th-order Ambisonics and two hybrid variants using CTC or VBAP for the early party and 4th-order Ambisonics for the late reverberation. For the CTC the HRTFs of the artificial head *Fabian*[36] were used and 2 loudspeakers at an elevation of 0° and an azimuth of ±45°. These HRTFs were measured with a source distance of 1.7m and therefore matching the loudspeaker distance in the test chamber. Fourth order Ambisonics was used with plane wave max-$|r_E|$ Ambisonics decoding.

As virtual room a model of the Concergebouw in Amsterdam was used to provide a realistic environment. The receiver was placed over the first rows. Four sources were presented at a distance of 5.5m (critical distance in this room model) at positions shown in Figure 6. The positions were limited to the front direction and to an elevation of ±30° to ensure a valid reproduction for VBAP and avoid the subjects having to turn around. Each position was repeated three times while the order of all stimuli was randomized.

A sound file was only played as long as the subjects were looking directly in front. A deviation of more than 2° would pause the playback immediately. The samples could be repeated as often as desired.

In total 18 subjects participated in the test with an average localization accuracy of 16°. The deviation was calculated as the distance of the cone of confusion of the presented source and the chosen direction. Individual results of all subjects are shown in Figure 7 (left). Between the presented five systems no significant differences were found in a one-way ANOVA test, as shown on the right hand side of Figure 7. The ANOVA yields a significant main effect of source position (F(3,51)=6.695; MSE=0.222; $p < 0.001$; η_p^2 =0.283) between position 4 and all other positions. A two-way ANOVA revealed a significant main effect of system for positions 2 (F(4,64)=9.463; MSE=0.075; $p < 0.001$; η_p^2=0.372) and 4 (F(4,64)=11.538; MSE=0.019; $p < 0.001$; η_p^2=0.419), as shown in Figure 8. It can be concluded that VBAP and HOA have a higher dependency on the source position than the CTC.

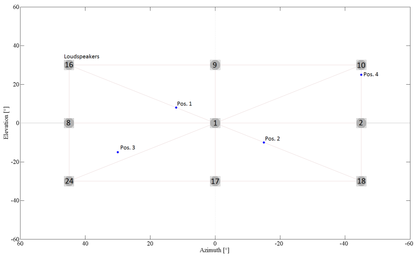

Figure 6: *Presented 4 source positions in the listening test. Frontal 9 loudspeakers are shown for reference.*

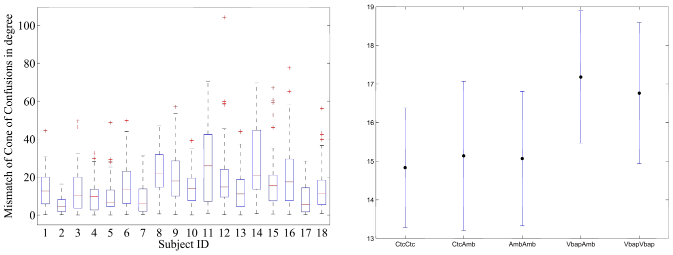

Figure 7: Left: *Results of the localization perfor-mance tests for the 18 individual subjects.* Right: *Results of the ANOVA for different reproduction systems (CTC, CTC+HOA, HOA, VBAP+HOA, VBAP). No significant differences were found in localization performance.*

6. CONCLUSIONS

A method was presented that combines different loudspeaker-based reproduction methods (such as CTC, Ambisonics or VBAP) to auralize a sound field. The sound field can consist of one or more sound sources and all reflections of these sources that bounce off walls in the virtual scene. The simulation of the sound field including all room acoustics reflections is done using the RAVEN framework. For the hybrid auralization, the room impulse response is divided into three parts (direct sound, early reflections and late reverberation), according to findings from psychoacoustic research. This division also suits the algorithms of geometrical acoustics that are commonly used in room acoustics simulations (image sources and ray tracing).

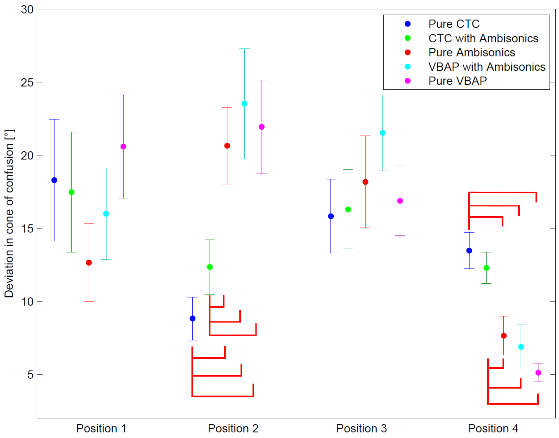

Figure 8: *Results of two-way ANOVA grouped by presented source position. The high impact of the presented source position is visible. Red brackets indicate significant differences.*

The hybrid approach can take advantages of the individual strengths of each reproduction method. Strong localization cues are necessary for direct sound rendering. The late diffuse sound field should be rendered using immersive reproduction methods. Both signals are calculated using RAVEN and are played back simultaneously through the same loudspeaker setup. To enable a seamless transition the average loudness of the different systems has to be calibrated accurately, which has to be done for each individual loudspeaker setup.

The moment of transition from the early to the late part of the impulse response is defined by the mixing time. It was shown that in typical cases after three reflection orders the sound-field can be expected to be mixing and diffuse. Then the renderer can switch from a method with strong localization to a method with high envelopment. The aim is to render a realistic and natural sounding high quality auralization of spatial sound with reverberation.

With the used loudspeaker setup (24-channel array) none of the tested systems provided an overall superior localization performance than the other systems. However, the binaural CTC provided a more homogeneous localization accuracy across different source positions. This behavior is typically preferred, especially for scenes with moving sources, making this technique suitable for the early part of the impulse response. A test for the immersiveness of different systems has to be designed and conducted in further research, to find an optimal method for the reproduction of late reverberation.

7. REFERENCES

[1] M.R Schroeder, "Die statistischen parameter der frequenzkurven von großen räumen," *Acustica 4*, vol. 4, pp. 594–

600, 1954.

[2] M.R Schroeder, "Natural sounding artificial reverberation," *13th AES Convention*, 1961.

[3] S. S. A. Krokstad and S. Sorsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *J. Sound and Vibration*, vol. 8, pp. 118–125, 1968.

[4] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. America*, vol. 65:943, 1979.

[5] J. Borish, "Extension of the image model to arbitrary polyhedra," *J. Acoust. Soc. America*, vol. 75, pp. 1827–1836, 1984.

[6] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, RWTHedition Series. Springer, 2011.

[7] G. M. Naylor, "Odeon - another hybrid room acoustical model," *Applied Acoustics*, vol. 38:131, 1993.

[8] CATT-Acoustic, *http://www.catt.se*.

[9] W. Ahnert and R. Feistel, "Ears auralization software," *J. Audio Eng. Soc. Vol.*, vol. 41 (11), pp. 897–904, 1993.

[10] S. Pelzer, M. Aretz, and M. Vorländer, "Quality assessment of room acoustic simulation tools by comparing binaural measurements and simulations in an optimized test scenario," *Acta acustica united with Acustica*, vol. 97, no. S1, pp. 102–103, 2011.

[11] D. Schröder, *Physically Based Real-time Auralization of Interactive Virtual Environments*, Ph.D. thesis, RWTH Aachen University, 2011.

[12] J.-D. Polack, "Is mixing the source of diffusion?," *J. Acoust. Soc. Am.*, vol. 129(4), pp. 2502–2502, 2011.

[13] A. Reilly, D. McGrath, and B.-I. Dalenbäck, "Using auralisation for creating animated 3-d sound fields across multiple speakers," *Proc. 99th AES Conv., New York*, vol. preprint no. 4127, 1995.

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

[15] A. Lindau, "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," *Proc. AES 128th Conv., London, UK*, 2010.

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

[17] A. Avni and B. Rafaely, "Sound localization in a sound field represented by spherical harmonics," *Proc. 2nd Internat. Symposium on Ambisonics and Spherical Acoustics, Paris, France*, 2010.

[18] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual reality system with integrated sound field simulation and reproduction," *EURASIP journal on advances in signal processing*, vol. 2007, pp. 70540, 19 S., 2007.

[19] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," *AES 23rd Internat. Conf., Copenhagen, Denmark*, 2003.

[20] B. B. Bauer, "Stereophonic earphones and binaural loudspeakers," *J. Audio Eng. Soc.*, vol. 9(2), pp. 148–151, 1961.

[21] T. Lentz, *Binaural technology for virtual reality*, Ph.D. thesis, RWTH Aachen University, 2011.

[22] D. Schröder, F. Wefers, S. Pelzer, D. Rausch, M. Vorländer, and T. Kuhlen, "Virtual reality system at rwth aachen university," in *Proc. ICA 2010, 20th Internat. Congress on Acoustics, Sydney, Australia*. 2010, Australian Acoustical Society, NSW Division, 1 CD-ROM.

[23] C. Guastavino, V. Larcher, G. Catusseau, and P. Boussard, "Spatial audio quality evaluation: comparing transaural, ambisonics and stereo," *Proc. 13th Internat. Conf. on Auditory Display, Montreal, Canada*, 2007.

[24] A. Farina, R. Glasgal, E. Armelloni, and A. Torger, "Ambiophonic principles for the recording and reproduction of surround sound for music," *Proc. AES 19th Internat. Conf.*, 2001.

[25] S. Favrot and J. M. Buchholz, "Lora: A loudspeaker-based room auralization system," *Acta Acustica united with Acustica*, vol. 96, pp. 364–375, 2010.

[26] V. Pulkki, "Evaluating spatial sound with binaural auditory model," *Proc. Internat. Computer Music Conference, Havana, Cuba*, pp. 73–76, 2001.

[27] P. Majdak, B. Masiero, and J. Fels, "Sound localization in individualized and non-individualized crosstalk cancellation systems," *J. Acoust. Soc. America*, vol. 133(4), pp. 2055–2068, 2013.

[28] M. Pollow, K.-V. Nguyen, O. Warusfel, T. Carpentier, M. Müller-Trapet, M. Vorländer, and M. Noisternig, "Calculation of head-related transfer functions forarbitraryfield points using spherical harmonics decomposition," *Acta acustica united with Acustica*, vol. 98(1), pp. 72–82, 2012.

[29] T. Musil M. Noisternig, A. Sontacchi and R. Höldrich, "A 3d ambisonic based binaural sound reproduction system," *AES 24th Internat. Conf. on Multichannel Audio, Banff, Canada*, 2003.

[30] B. Masiero, *Individualized binaural technology: measurement, equalization and perceptual evaluation*, Ph.D. thesis, RWTH Aachen University, 2012.

[31] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21(1), pp. 2–10, 1972.

[32] A. Solvang, "Spectral impairment for two-dimensional higher order ambisonics," *J. Audio Eng. Soc.*, vol. 56, pp. 267–279, 2008.

[33] J. Daniel, *Representation de champs acoustiques, application a la transmission et a la reproduction de scenes sonores complexes dans un contexte multimedia (in french)*, Ph.D. thesis, Universite Paris 6, 2000.

[34] S. Pelzer and M. Vorländer, "Auralization of virtual rooms in real rooms using multichannel loudspeaker reproduction," *J. Acoust. Soc. America*, vol. 134, pp. 3985–3985, 2013.

[35] P. Majdak, M. J. Goupell, and B. Laback, "3-d localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Attention, Perception, & Psychophysics*, vol. 72, no. 2, pp. 454–469, 2010.

[36] A. Lindau and S. Weinzierl, "Fabian-schnelle erfassung binauraler raumimpulsantworten in mehreren freiheitsgraden," *Fortschritte der Akustik 33.2: 633*, 2007.