A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation

TORBEN WENDT[1,2], STEVEN VAN DE PAR,[1] *AES Member*, AND STEPHAN D. EWERT[2]
(torben.wendt@uni-oldenburg.de) (steven.van.de.par@uni-oldenburg.de) (stephan.ewert@uni-oldenburg.de)

[1] *AG Akustik and Cluster of Excellence Hearing4all, Universität Oldenburg, Germany*
[2] *Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, Germany*

A fast and perceptually-plausible binaural room reverberation rendering method is presented, suited for application in interactive evaluation environments, e.g., for hearing aid development, and for psychophysical studies, room simulation, and computer games. A hybrid approach was used to achieve a high computational efficiency and perceptual plausibility: early reflections up to a restricted low order were calculated by the image source model (ISM; Allen and Berkley [J. Acoust. Soc. Am. Vol. 66(4), 1979]). The simulation of the reverberant tail was based on a feedback delay network (FDN; Jot and Chaigne [Proc. 90th AES Conv., 1991]), which is computationally very efficient and allows explicit control of the frequency-dependent decay characteristics. The FDN approach was modified to be adaptable to room dimensions and to different wall absorption coefficients. Furthermore, it was extended to create spatially distributed reverberation for arbitrary source positions as well as arbitrary receiver orientation and position, using head related impulse responses. To evaluate the simulation method, binaural room impulse responses were measured and synthesized for various rooms. Subjective ratings of perceived room acoustical attributes, and assessment of various common room acoustical parameters, generally showed a good correspondence between simulated and real rooms.

0 INTRODUCTION

The simulation of the reverberant acoustics of rooms has numerous applications ranging from the creation of acoustic scenarios for development and evaluation of signal-processing algorithms, to the presentation of sound sources in a virtual room to a listener, for example, to assess speech intelligibility impairment by reverberation. Room acoustical simulations are also of interest for audio-visual simulation environments (e.g., for training and rehabilitation [1]) and in entertainment, e.g., in computer games. Here virtual visual environments are typically accompanied with sound field simulations (auralization) including virtual room acoustics. As a consequence of the interactive nature of such applications, the virtual environment should be adaptive in real-time. Thus computational efficiency is required to achieve real-time updates of the room acoustical simulation, depending on the movement and head orientation of the user (or listener) or the sound sources, each of them preferably with six degrees of freedom (6-DOF).

The traditional way to emulate room acoustics is the measurement of binaural room impulse responses (BRIRs) and convolution of dry (i.e., reverberation-free) source signals with the BRIRs, yielding a virtual sound as perceived in a specific room. However, the measurement of room impulse responses is time consuming and lacks flexibility, i.e., it is restricted to static scenarios with a fixed listener position and orientation. Furthermore, this method is restricted to actually existing rooms. Alternatively, room acoustics can be simulated by computer programs enabling a variable degree of realism, ranging from simple artificial reverb to complex room acoustical simulation (image source reverberation [2], CATT-Acoustic [3], ODEON [4]). The use of such simulation methods can also overcome the restriction of static listener position by sampling virtual (acoustic) environments on a grid in space and interpolating between pre-computed BRIRs for different listener positions (e.g., CATT Walker module [5]). More computationally complex and acoustically faithful methods have been suggested in Schröder et al. (2007) [6].

For some applications, a physically correct rendering of a sound field is required, for example to correctly simulate the directivity of a beamformer. For applications mainly involving perception of human listeners, however, an auralization only has to be perceptually convincing, which means it has to be plausible and authentic.

Virtual acoustics is plausible if it sounds natural and if the listener has the impression that the rendering represents a real room. Authenticity implies that the simulated room sounds exactly like a specific real room. Authenticity will require the accordance of objective room acoustical parameters, such as reverberation time T_{60}, definition D, and others. When room simulations are used in psychoacoustic research, it is important that authenticity is provided in order to obtain measurement results that can be considered to be representative for real acoustic environments. Assuming that meeting perceptual criteria such as authenticity and plausibility are less restrictive than achieving physically correct simulations, high computational efficiency might be achieved by using profound physical simplifications in the room acoustic simulation. This becomes particularly important when real-time rendering of dynamic acoustic scenes is required where the positions of sources and receivers can be changed interactively.

Several approaches have been reported in literature for binaural room simulation. If the wavelength of a sound is small in comparison to the characteristic dimensions of sound reflecting objects, concepts of geometric acoustics can be applied to describe the sound field at a certain position. In that case, a standard approach is the image source model (ISM) [2], where the basic idea is to consider the reflection of a sound as the direct field of an image source. The sound of an image source differs from that of the original source by its delay and attenuation caused by the distance to the receiving point, and its frequency-dependent attenuation related to the reflective properties of the respective wall. Since all image sources act as sound sources as well, their sound can be reflected again at other walls, creating higher-order image sources. In this way, arbitrarily complex reflection paths can be modeled. The computation of all image source positions up to a certain order becomes quite simple if an empty shoebox-shaped room is assumed. Due to symmetry, a regular pattern of image sources emerges. Furthermore, all image sources are "visible" to the receiver, which cannot be assumed for arbitrary room geometries. Thus in the case of arbitrary geometry visibility- and validity tests would be necessary for each image source, increasing the computational cost considerably [7]. While the geometric solution for shoebox-shaped rooms is straightforward, the number of image sources, and thus computational complexity, increases fast with the third power of the image source order and accordingly with the third power of the duration of the impulse response. As a consequence, the image source model is not suited for a fast synthesis of complete BRIRs. However, sound reflection rendering based on the ISM is well investigated and has been applied for room acoustical simulations, mostly in combination with other algorithms in so-called hybrid approaches (e.g., [4, 8–16]).

Another geometric approach is the ray-tracing method [17]. The idea is to use a large number of sound particles that are emitted from a source into various directions. While traveling through the room, the particles' energies and paths are traced. When a particle hits a boundary, its propagation direction is changed according to a mirror reflection (e.g., [18]), and its energy is changed according to the absorption coefficient of the reflecting surface. Instead of specular reflections (as inherently assumed in the ISM), diffuse reflections can be simulated as well. Usually, this is achieved by partitioning the incident ray energy into a certain part that is carried by a specular reflection and the remaining part carried by further rays emitted from the point of incidence into all directions. Finally, sound particles that pass a certain area around the receiver point are collected to estimate the sound field. Although this technique is applicable to arbitrary room geometries and allows for diffuse reflections, it has the disadvantage that large numbers of particles are necessary to simulate room acoustics adequately making the method computationally highly demanding. Together with other techniques (e.g., the image source model), the ray tracing method has been further developed and applied in various room acoustics simulation algorithms (e.g., [4, 16, 19]).

To achieve reduced computational complexity for generating BRIRs, several approaches have been suggested involving physical simplifications. Kendall et al. (1986) [10] proposed a reverberation renderer that is based on the ISM but computationally much more efficient. Whereas the first and (most of the) second order reflections are rendered purely ISM based, the subsequent late reflections are simulated by an "inner reverberation network" containing different types of recirculating delay units. The delays are chosen based on delay patterns occurring in the ISM. By creating multiple interconnections between the higher-order reflections it is possible to capture the general spatio-temporal pattern of sound reflections in the shoebox-shaped room. In order to create spatial reverberation, the network produces 18 "reverberation streams" that are mapped to 18 carefully chosen incidence directions in three-dimensional space.

In order to synthesize real-time artificial reverberation with predefined reverberation time per frequency-band, but neglecting room geometry and realistic simulation of early reflections as provided by the above-mentioned approaches, considerably more simplified artificial reverberation algorithms can be used. A common and useful approach is based on feedback delay networks (FDNs). This class of algorithms offers low computational cost and explicit control over the decay characteristics. A very basic FDN consists of a set of parallel delay lines whose outputs are fed back to their inputs, redistributed according to a feedback matrix. As discussed in detail in [20], the generalized approach is based on Schroeder's pioneering work on parallel delay lines with feedback [21] and was further developed (among others) by Stautner and Puckette (1982) [22]. Stautner and Puckette introduced the generalization of a feedback matrix and a multichannel input, resulting in a faster increase of the number of acoustic modes, i.e., a higher pulse density, that prevents flutter-echo effects. The system was found to be stable if the feedback matrix is orthogonal. Further improvements were achieved by Jot and Chaigne (1991) [23], who derived a relation between the obtained reverberation time and the gain factors, which were then also generalized to "absorptive filters" in order to control the reverberation time in a frequency-dependent way. Jot (1997)

[13] provided further extensions (see below) and the role of feedback matrices was investigated by Rocchesso and Smith (1997) [24] and by Menzer and Faller (2010) [25]. As a similar but more general approach, digital waveguide networks were proposed by Smith (1985) [26]. While the FDN approach is computationally efficient, it does not directly provide a binaural output or any form of spatially distributed reverb. Menzer and Faller (2009) [27] have proposed to overcome this problem by including an interaural coherence filter at the output of the FDN. This allows to integrate the perceptually important frequency dependence of the interaural coherence in the synthesized BRIRs by matching them to that of measured BRIRs. An approach was also suggested that does not require a measured reference BRIR but no evaluation was reported. A limitation of the approach by Menzer and Faller (2009) [27] is that it does not model the spatial and temporal pattern of the early reflections resulting from the specific room geometry. Furthermore the position and orientation of the receiver cannot be directly taken into account.

A possible solution to achieve real-time performance while maintaining the advantages of the more "accurate" geometric BRIR synthesis and reverberation algorithms is a hybrid approach (e.g., [8–10, 12–14, 20]). Here, the initial part of the impulse responses, mainly composed of distinct reflections, is computed based on geometric methods and is rendered in real time using an FIR filter, whereas the later reverberation part is generated by a computationally efficient artificial reverberation algorithm. From a perceptual view, the motivation for such a hybrid approach is that the early sound reflections, typically considered as being those received within the first 50 ms after the direct sound, determine the perceived source position and width. Early reflections further influence the perceived timbre of the source [28] and support speech intelligibility [29]. The subsequent reverberant tail contains later reflections whose time-related density is very high. Typically the reverberant tail contains diffuse reflections, i.e., they are not contributing to the perceived source direction, but contribute to the perceived listener envelopment [30]. The frequency dependent energy and decay characteristics of the reverberant tail convey information about the wall absorption and room size. The source-to-receiver distance is conveyed by the energy ratio between direct sound and reverberation [31].

A number of hybrid approaches based on the FDN or related structures have been suggested: Moorer (1979) [8] did some advancements of Schroeder's reverberator networks and introduced the combination with an FIR filter for the rendering of early reflections. The parameters of that filter were chosen according to the image source model. Jot (1997) [13] proposed a highly modularized algorithm combining three (unitary) delay networks in cascade, where the first ones (without feedback) represent early reflections and the last one (with feedback) renders the reverberation part. A panning stage enables directional rendering for arbitrary playback setups. Realizations and simplifications of directional and binaural rendering are discussed in Jot et al. (1995) [12]. Savioja et al. (1999) [14] presented a simulation software called DIVA, which performs image-source-based early reflection rendering in a very elaborate way. Late reverberation is generated by a simplified FDN, fed by the (delayed) direct sound after application of an air absorption filter. For binaural reverberation, sums of different output channels of the simplified FDN are used in order to retrieve uncorrelated ear signals. As far as reported, no further adjustment of interaural cues, such as the interaural correlation, was performed. For all of the above hybrid approaches real-time performance was stated. As described, most of the approaches also feature binaural or directional rendering of the late reverberation part, which is important for the spatial impression.

Nevertheless, a remaining limitation of these approaches is the lack of direct relation of the reverberation part to room geometry and wall absorption, which will become increasingly relevant if only a few reflections are generated by a geometrically motivated early reflection renderer. The lack of spatial properties of the reverberant tail may specifically be a problem for receiver positions in proximity of a strongly absorbing wall. The suggested FDN-based hybrid approaches cannot render the spatial sound field for arbitrary receiver (head) rotations, thus not allowing 6-DOF movement, as would occur for natural movement and behavior in virtual environments.

Moreover, none of the above mentioned approaches to binaural reverb rendering were evaluated in comparison to real rooms with regard to classical room acoustical parameters or perceptual attributes in subjective psychoacoustic listening tests.

In the current study, a hybrid approach is proposed based on an ISM for the early reflections combined with the FDN for the reverberant tail. An improved version of the method of Kendall et al. (1986) [10] was used to spatially render the FDN output. The aim is to render perceptually accurate binaural reverberation depending on room geometry, source position, and 6-DOF receiver position at low computational complexity. The ISM for shoebox geometry was used up to a certain (low) order. The connection between the ISM and FDN component was designed such that the maximum reflection order of the ISM is freely adjustable. Moreover, the use of shoebox image sources could be in principle exchanged by any other early reflection renderer. Instead of capturing the spatio-temporal pattern of reflections within the late reverberation in detail, as realized in Kendall et al., the FDN was used with delays determined by the room geometry and with explicit control of the frequency-dependent reverberation time to capture the temporal pattern. Consequently, static delays can be used for any source or receiver position enabling lower computational complexity. The spatial pattern of reflections was achieved in a similar way as in Kendall et al., by assigning different FDN outputs to spatial positions that can be rendered with the corresponding head-related-transfer-function (HRTF). In contrast to Kendall et al. a more uniform spatial distribution of sources was achieved independent of room dimensions such that a more realistic rendering of the diffuse sound field is obtained. In objective and subjective evaluations, the ISM reflection

order was varied in order to investigate its influence on both room acoustical parameters and perceptual attributes.

The suggested binaural reverberation rendering algorithm runs on an ordinary desktop computer and without need for a dedicated signal processor (thus enabling easy portability). Although the suggested algorithm is focused on generating a binaural rendering of the simulated room acoustics, it is easily portable to different auralization systems, such as loudspeaker-based higher-order ambisonics.

This article is organized as follows: in Sec. 1 the components of the proposed simulation algorithm (1.1, 1.2) and their combination (1.3) are described in detail. Section 2 contains the description of the measurements and simulations (2.1) used for the objective (2.2) and subjective (2.3) evaluations of the proposed simulation method. Finally, a summary and conclusions are given in Sec. 3.

1 SIMULATION METHOD

A hybrid approach (see also [8, 10, 12–14]) for synthesizing binaural room impulse responses (BRIR) is proposed. The early reflections are generated by an FIR filtering module controlled by the ISM that computes geometrically derived sound reflections up to a certain reflection order. To increase computational efficiency, and since not all details in the late reflections are perceptually relevant, the late-reverberation part is generated by an FDN.

The ISM implementation is restricted to empty shoebox-shaped rooms allowing for a considerable reduction of computational cost for calculating the positions of image sources. With these shoebox-shaped rooms an important class of room geometries is represented. The six wall surfaces are characterized by frequency-dependent absorption coefficients, where each wall is assumed to absorb homogenously (same absorption coefficient over its whole surface).

The auralization is described explicitly for the case of headphone presentation, which was realized by the application of head-related-impulse-responses (HRIRs) or, equivalently in the frequency domain, head-related-transfer-functions (HRTFs). A pair of HRTFs describes the sound transmission from one point in the free field to a point in the ear canal of the left and right ear, respectively [32]. Thus, such a pair of HRTFs contains the cues for sound localization, which are the interaural time difference (ITD) due to sound propagation speed, the interaural level difference (ILD) due to head shadowing, and coloration due to the filter effect of pinnae, head and torso [33].

1.1 Image Source Model (ISM)

The basic idea of the image source model is illustrated in Fig. 1 for a shoebox-shaped room (plan view). Here, S_0 represents the sound source and P is the receiver's position. Sound propagation paths are illustrated as rays. For clarity, only the direct sound and first order reflections at the side walls are shown. They are modeled by first order image sources, denoted by $S_{(n_x,n_y,0)}, n_x, n_y \in \{-1, 0, 1\}$, which are obtained by mirroring the original source at the respective walls. For each image source, its distance to the receiver

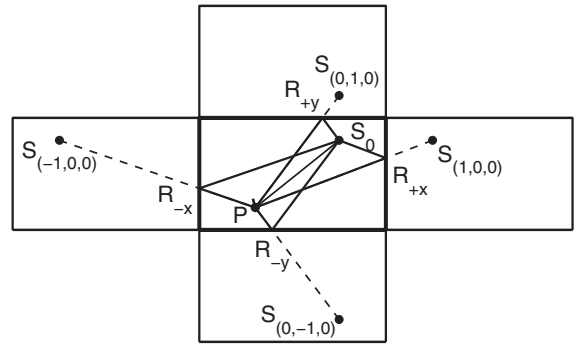

Fig. 1. Top view on a rectangular room (center position, bounded by bold lines) containing a sound source S_0 and a receiving point P (short line indicates orientation). Direct sound and first-order reflection paths are plotted, where the latter are modeled as direct sound paths of first-order image sources $S_{(n_x,n_y,0)}$, n_x, $n_y \in \pm\{0, 1\}$, located inside mirrored versions of the original room. With each wall a frequency-dependent reflection coefficient R_ℓ, $\ell \in \pm\{x, y\}$ is associated.

and the intervening wall reflection coefficient must be taken into account.

As only empty shoebox-shaped rooms are considered, the positions of image sources up to an arbitrary order are obtained as a symmetric pattern, with their overall number given by Eq. (4) in Sec. 1.4. The rendering parameters for a particular image source are determined by its distance to the receiver and the number of reflections in all specific walls in the reflection path.

The actual implementation of the early reflection renderer is shown as the left part of the block diagram in Fig. 2 showing the whole algorithm. Each branch renders a particular image source, indexed by i. This also includes the direct sound source (first branch, $i = 0$), which is assigned the reflection order $N = 0$. In the following, the processing steps for one particular image source are described (see also [11, 14]).

For each image source, the distance r_i to the receiver results in an attenuation factor $1/r_i$ and a delay $\Delta\tilde{\tau}_i$. The delay $\Delta\tilde{\tau}_i$ is the distance of the image source to the receiver divided by the speed of sound since this distance equals the length of the entire reflection path.

Subsequently, a reflection filter $\tilde{R}_i(f)$ is applied. Given that each wall surface has its own frequency dependent reflection coefficient, the product of all coefficients that contribute to the actual reflection path results in an "effective" reflection filter for this image source. The reflection coefficients for this filter are specified in octave bands between 250 Hz and 4 kHz. To emulate the frequency response for one reflection coefficient, a 13th-order filter is used, being a composition of single shelving peak filters, designed after [34]. A second-order Butterworth bandpass filter with heuristically chosen edge frequencies of 20 Hz and 14 kHz is applied additionally. The passband of the filter covers the frequency range where reflection coefficients are specified. The high frequency cutoff can be regarded as a rough approximation of the lowpass characteristics of air absorption in combination with typically increasing absorption

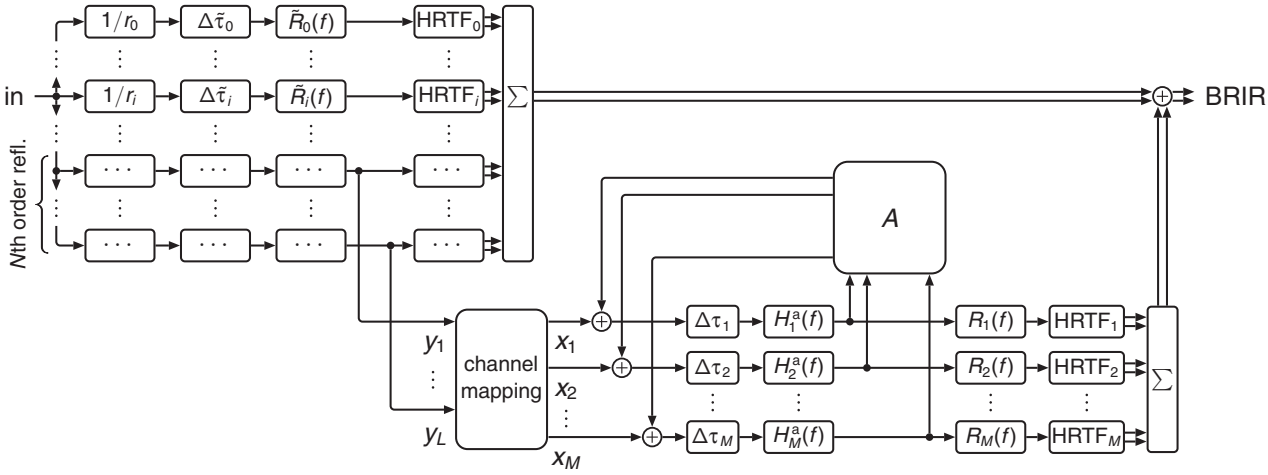

Fig. 2 . Block diagram of the hybrid algorithm containing the early-reflection renderer on the left and the extended feedback delay network on the right. In the early reflection renderer, each branch renders one single image source and includes the following processing steps: distance-related attenuation, distance-related delay, effective reflection coefficient, binauralization via HRTF (see text for explanation, Sec. 1.1). The direct sound source (first branch, representing the special case of an image source with order $N = 0$) is indexed with "0." The contributions from all image sources of order N are collected separately as signal vector y. For the feedback delay network see the text in Sec. 1.2 for explanation of all processing steps.

coefficients of the wall materials for high frequencies. Since "leaky" rooms are assumed, no static (i.e., DC) pressure persists, which is accounted for by the low-frequency cut off.

The spatial position of each image source is finally introduced by convolution with an HRIR, according to the direction of the image source with respect to the listener's orientation in space.

Finally, the contributions of all image sources are added up to form the (two channel) output. The multichannel signal (y_i), $i \in \mathbb{N}_{\leq L}$ collects the contributions of all image sources of order N and is used as input for the FDN module as described in Sec. 1.3.

1.2 Extended Feedback Delay Network (FDN)

The extended FDN used in the proposed hybrid room simulation algorithm is shown as the right part of the block diagram in Fig. 2. It mainly follows the structure of the multichannel network proposed by Jot and Chaigne (1991) [23]. Two additional processing blocks are inserted in each output branch: $H_j^{\mathrm{r}}(f)$ and HRTF_j, $j \in \mathbb{N}_{\leq M}$, which will together be referred to as "binauralization steps" and will be introduced later in this section.

The number of channels, M, is set to 12. This choice is in line with Jot (1997) [13], who proposed a number of 8 to 16 as sufficient, if the feedback matrix A is chosen conveniently. In addition, this value was selected because it allows the assignment of each channel to one specific wall of the room. Thus, because a shoebox-shaped room has six surfaces, there are two channels per specific wall (four channels per room dimension), which will be reflected in some of the following parameter choices.

The delay lengths $\Delta\tau_j$ are scaled according to the room dimensions (d_x, d_y, d_z) via sound propagation speed c, so that the echo density in the synthesized reverberation will be smaller if the room is larger (and vice versa). If the j-th channel is associated to the i-th room dimension, $i \in \{x, y, z\}$, its delay length is set to

$$\tau_j = \frac{1}{c}(d_i + \bar{d}\varepsilon_j), \tag{1}$$

where the second term is a channel-dependent jitter, which prevents multiple occurrences of the same delay values in different channels. It is created by a random number ε_j from a uniform distribution in the open interval $]-0.1, 0.1[$, multiplied with the arithmetic mean $\bar{d}$ of the room dimensions. Informal listening tests showed a certain robustness of the sound quality against the actual choice of the delay values: no notable audible differences occurred when, for instance, linearly increasing or random-jittered delays between certain minimum and maximum values, derived from the room dimensions, were chosen. Some approaches and guidelines to choose the FDN delays are reported in literature (e.g., [13, 24]). Jot (1997) [13] suggested that the total delay length, i.e., the sum of all delay values, should be "at least equal to one fourth of the decay time" in order to achieve a sufficient modal density. Rocchesso and Smith (1997) [24] proposed to choose the delays based on the lowest normal modes in a rectangular room. The approach used here is somewhat similar to that of [24]. However, the Jot criterion may have been infringed if a very small room with very low wall absorption, i.e., a high reverberation time, is assumed. In contrast to [24], the current approach may be problematic for extreme room dimensions such as long and narrow corridors.

The next processing steps are the absorption filters with transfer functions H_j^{a}, which simulate the frequency-dependent sound attenuation caused by wall reflections and air absorption. The absorption filters allow for an explicit control over the resulting reverberation time $T_{60}(f)$. Under the assumption that all other processing steps within the

FDN do not attenuate the signal energy, the following relationship holds [23]:

$$20 \log_{10} \left| H_j^{\mathrm{a}}(f) \right| = -\frac{60\tau_j}{T_{60}(f)} . \tag{2}$$

Although this equation contains the delay element value of the respective channel, it describes a global attenuation and is not motivated by the association of each delay line to a room dimension. In the current simulation method, the reverberation time is predicted from the wall absorption coefficients $\alpha_i(f), i \in \mathbb{N}_{\leq 6}$ using Sabine's formula

$$T_{60}(f) = \frac{55.3 \cdot V}{c \cdot \sum_i \alpha_i(f) S_i} , \tag{3}$$

with V being the room volume, c the sound propagation speed, and S_i the surface area of the i-th wall. As for the reflection filters in the early reflections rendering module, peak-shelving filter compositions are used in order to approximate the specified frequency responses of the absorption filters. Again all frequency responses are finally bandlimited by the same bandpass filter as described above. The resulting frequency-dependent reverberation time is not affected within the range where absorption is defined.

Via the feedback matrix A the output of each delay line is redistributed to the FDN input channels. In order to obtain an energy preserving feedback, an orthogonal matrix is chosen. The general necessary condition to A is to be "lossless" as described, e.g., in [13, 20, 24]. Thus the absorption filters are the only processing steps affecting the decay characteristics of the impulse response. Additionally, in order to achieve a faster increase of pulse density at the output, it is advantageous if A has no zero entries [13]. There are several types of matrices that fulfill these requirements, such as Householder matrices, Hadamard matrices, or circulant matrices [13]. In contrast, Menzer and Faller (2010) [25] state that it is rather disadvantageous if the unitary feedback matrix has only nonzero elements, because this can practically lead to cancellation of certain reflections after multiple applications. Rocchesso and Smith (1997) [24] did extensive theoretical investigations on FDNs and digital waveguide networks, proposing circular feedback matrices as being efficient and versatile. Issues of computational complexity are discussed, too, in [13, 25].

In the current simulation method, it is assumed that the efficiency of A is not critical within the whole hybrid algorithm, such that a randomly created unitary matrix is chosen. For the 12×12 matrix, vectors of respective dimensionality were chosen from a normal distribution and then orthonormalized by the Gram-Schmidt process. In this way, a higher variability of pulse amplitudes is obtained and the output signal is expected to simulate a higher degree of diffuseness. Furthermore, informal listening tests yielded a consistently more natural sounding reverberation in comparison to other matrix types.

Finally, the binauralization steps, consisting of reflection filters and HRTFs, introduce a spatial rendering of the reverberation, ensuring that binaural aspects like a frequency-dependent interaural correlation are accounted for. Here the association of each FDN channel to a certain wall of the room is used: to achieve uniform spatial distribution of sound incidences around the head the incidence directions are mapped to points on a cube surface. The imaginary cube is always centered on the listener and oriented such that its surfaces are parallel to the room walls. Two points per side face are positioned at 1/3 and 2/3 of the length of the diagonal, and diagonals of opposing side faces are perpendicular to each other (see Fig. 3 for illustration). The respective HRTFs, according to the angle of incidence, are applied as indicated in Fig. 2. In comparison to the similar way of spatially mapping diffuse reverberation in Kendall et al. (1986) [10] the method is improved in two ways here: (i) less HRTFs are used here, implying less computational effort; (ii) incidence angles are more equally distributed especially if the room dimensions considerably differ across walls. Such an equal distribution is also apparent in real rooms because for any given point in time, all reflections can be considered to originate from a sphere surface centered on the receiver. The later the reverb, the larger the sphere radius becomes in comparison to the room dimensions and the spatial distribution of reflections becomes independent of the room dimensions. Particularly in cases of missing walls (no reflection) or strongly absorbing walls, and in the case of room dimensions considerably deviating from a cube, the current method is more realistic for arbitrary receiver positions in the room.

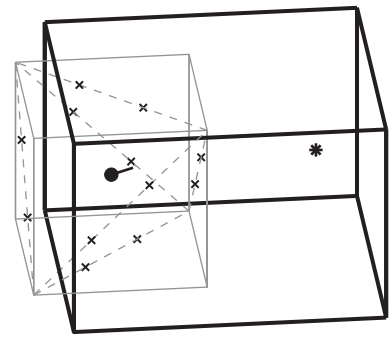

Fig. 3. Illustration of the incidence directions of the virtual reverberation sources for the binauralization steps. Shoebox-shaped room containing a sound source (asterisk) and a receiver (dot with "nose" indicating orientation). The virtual reverberation sources (crosses) are positioned on diagonals on the surfaces of an imaginary cube (axis-aligned with the room) around the receiver.

By the channel-wise application of reflection filters H_j^{r}, the signal of each incidence direction is weighted with the reflection coefficient of the room wall corresponding to the respective parallel cube side face. By this, together with the HRTFs, the spatial distribution of reverberation energy is simulated caused by different acoustical wall properties. For example, in a room with one strongly absorbing wall (or an open side) there will be no energy in the spatial reverberant field from that direction. In comparison to Menzer and Faller (2009), who propose an FDN with frequency dependent coherence matching (see introduction) the approach described here only requires knowledge about the room properties and will automatically lead to frequency-dependent realistic interaural coherence, without the need

for a reference BRIR. Here interaural differences are a consequence of the spatial distribution of the virtual sound sources. Additionally the current approach is easily adaptable to multichannel loudspeaker systems because it does not assume a two-channel output signal.

1.3 Combination of ISM and FDN

The combination of the early-reflections renderer (ISM) and late-reverberation renderer (FDN) was one of the main challenges in this hybrid approach. The resulting transition between both can be characterized by two properties: (i) the time of transition and (ii) the continuity of the reverberation energy decay.

The time of transition is indirectly controlled by the maximum image source order N. Since both accuracy and computational cost increase with increasing N, a trade-off has to be found. The smallest acceptable value will be derived from the objective and subjective evaluation (presented in Secs. 2.2 and 2.3, respectively). When considering the ability to localize a sound source, even the value $N = 0$ should be sufficient because of the precedence effect [35, 36]. As a standard value, N is set to 3 in the following, leading to a number of 63 sound sources to be simulated (see Eq. (5) in the following subsection) covering the most relevant early reflections.

For a given FDN, the shape of the transition is strongly influenced by the energy and the initial delay of the FDN input signal. They have to be chosen such that the resulting energy decay curve (EDC) [37] shows, in line with theory, a linear decay on a dB scale. To derive a suitable FDN input signal, it should be kept in mind that in the ISM, with each reflection order, a certain amount of acoustic energy is removed from the system (room). Since this behavior is represented by the absorptive filters in the FDN, using properties of the actual room, the energy of the FDN input must equal the total energy of the Nth-order reflections. This is achieved easiest if the FDN is directly fed with these reflections. The appropriate ISM output is the separate multichannel signal $(y_i), i \in \mathbb{N}_{\leq L}$, shown in Fig. 2 prior to application of the HRIR filters (HRIR auralization is the final step of the FDN itself). As already stated by Jot (1997) [13], it is important to use these pulses as a multichannel signal in order to avoid comb-filter coloration effects, which may occur if multiple delayed reflections are combined into one of the FDN input channels. Because the number L of image sources of order N does, in general, not equal the number M of FDN channels, the ith reflection contribution is fed into the FDN channel $((i-1) \bmod M) + 1$, which is indicated by the channel-mapping step in Fig. 2. This results in a certain set of prior summed early reflections as input to the same channel, but their number is bounded by $\mathrm{ceil}[(n_3(N) - n_3(N-1))/M]$ (see Eq. (4) in Sec. 1.4) per channel (i.e., 4, if $N = 3$). Informal listening tests showed them to be negligible in terms of comb-filter coloration, even for values of N larger than 3.

Additionally, feeding the FDN with Nth-order reflection signals has the convenience of creating a suitable initial delay of the FDN output for arbitrary choices of the maximum image source order. Given that the delay elements in the FDN are related to the room dimensions, the output pulses of the FDN seamlessly blend with the ISM output. Fig. 4 shows example echograms for illustration.

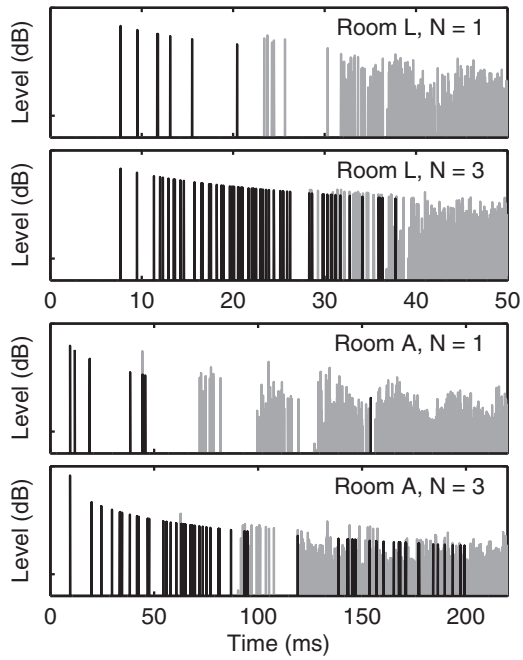

Fig. 4. Echogram plots showing time structures of simulated reflections for two rooms, L and A, and for two choices of N. The pulses generated by the early reflection (ISM) renderer are coded in black, those of the FDN in gray.

1.4 Computational Complexity

On the basis of the simulation method description, its computational complexity and especially the benefit from using the FDN can be theoretically estimated. Here, the computational complexity is regarded as a function of the room impulse response (RIR) duration.

For the ISM and in the case of a shoebox-shaped room, the number n_3 of image sources (including the direct sound treated as a 0th order image source, cf., Sec. 1.1) up to reflection order N can be calculated geometrically as a function of N:

$$n_3(N) = \frac{4}{3}N^3 + 2N^2 + \frac{8}{3}N + 1\,. \tag{4}$$

For a sufficiently large RIR duration, the average number $\mathrm{d}n$ of sound reflections arriving at the receiver point per time interval $\mathrm{d}t$ is

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{4\pi c^3 t^2}{V} \tag{5}$$

(see, e.g., [38]), c being the speed of sound and V the room volume. Thus, the absolute number of reflections up to time t is:

$$\int_0^t \frac{\mathrm{d}n}{\mathrm{d}t'}\,\mathrm{d}t' \sim t^3\,. \tag{6}$$

According to Eq. (6), the overall complexity of the early reflection renderer increases with the third power of the

RIR duration (t^3). The reflection filters for the higher order reflections can either be obtained by creating them on the basis of the prior multiplied respective reflection coefficients or by cascading the filters of the involved first order reflection filters. For the latter approach the filter order grows with reflection order. Although this is computationally more demanding, it was chosen because of its ability to account for sharp edges in the frequency response, which may occur for higher order reflections.

In the FDN (right part of Fig. 2, the number of parallel delay lines is fixed and independent of the resulting RIR duration, thus increasing complexity by a fixed amount. As an example, on a current desktop computer (Intel Core 2 CPU, each with 2.13 GHz clock rate) using MATLAB, a BRIR of a length of 0.73 and 14.0 s is synthesized in 0.71 ± 0.01 and (6.80 ± 0.03) s, respectively for an image source order of $N = 3$. All FDN processing steps but the binauralization steps are implemented in time domain as a MATLAB executable C function.

2 EVALUATION

The suggested simulation method was evaluated in the following way: for a set of real existing rooms, measured and correspondingly synthesized BRIRs were compared with respect to objective measures and subjectively rated sound properties. For this purpose, a test-database of measured and synthesized room impulse responses was created. Where possible, BRIR synthesis was performed using the same room configuration as for the real rooms, including the source- and receiver configuration, for varying synthesis parameters. The main synthesis parameter to be evaluated was the maximum image source order as the trade-off measure between BRIR accuracy and computational efficiency. This and other synthesis parameters are introduced in Sec. 2.1.3.

2.1 Evaluation Data Base

2.1.1 Rooms

Two real rooms available at the University of Oldenburg ("UOL") were selected for BRIR recordings. In addition, two rooms whose measured BRIRs are provided within the AIR-database [39] ("AIR") were selected to cover a wide range of room types. Using short labels for later reference (the origin of the BRIRs is given in parentheses), the rooms were:

- Room H (UOL): an almost shoebox-shaped, empty room previously used as reverberation chamber, with uniform and very even finish coat (plaster) at all walls including ceiling and floor. All surfaces form a cuboid, except for the ceiling which is at a slope, resulting in a room height varying between 2.65 m and 2.98 m. Although the room is very small ($1.88 \times 2.74\,\mathrm{m}^2$; $14.5\,\mathrm{m}^3$ volume) it features a high reverberation time T_{60} of about 2.5 s. Being rectangular in shape and having six plain and even walls, H comes closest to the shoebox-shape approximation in the ISM method.
- Room L (UOL): a rectangular room ($4.97 \times 4.10\,\mathrm{m}^2$, 3.00 m height, $61.1\,\mathrm{m}^3$ volume), which is used as a laboratory. It has a carpet (felt, about 4 mm) on the floor and a suspended ceiling consisting of acoustic tiles. The side walls are made of concrete. The room is occupied with some furniture and has a large window at one side. Thus, by the presence of many different acoustic properties within one wall (including diffusion) and deviations from the assumed shoebox geometry, the assumptions of the synthesis method are partly infringed. The reverberation time T_{60} is about 0.3 s.
- Room A (AIR): a large aula (Aula Carolina, Aachen), roughly shoebox-shaped, but with some pillars inside and arches at the ceiling. The walls are mostly rigid (bricks, windows), except for some draperies. Compared to the room dimensions (approx. $19 \times 30\,\mathrm{m}^2$, 10 m height, $5700\,\mathrm{m}^3$ volume), the distance between source and receiver is quite small (approx. 2.7 m), which should result in a relatively high ratio of direct to reverberant sound. The reverberation time T_{60} is about 4.7 s.
- Room S (AIR): a shoebox-shaped seminar room ($10.8 \times 10.9\,\mathrm{m}^2$, 3.15 m height, $370.8\,\mathrm{m}^3$ volume) with concrete walls, windows at three sides and parquet floor. Interior objects are tables and chairs. The reverberation time T_{60} is about 0.8 s.

For the UOL rooms, BRIRs were measured for several source- and receiver configurations. However, here only one configuration per room is presented given that no distinct dependencies of the objective measures on positions were observed.

2.1.2 BRIR Measurements

Room impulse responses were measured using an omnidirectional loudspeaker based on a ring-radiator [40] to excite the rooms equally in all directions over a wide spectral range. For recording, the artificial head MK2 by Cortex with the corresponding measurement amplifier Manikin MK1 was used.

The excitation signal was a logarithmic sweep [41] ranging from 50 Hz to 18 kHz. The log-sweep method offers the advantage of discarding possible nonlinear harmonic distortions of the loudspeaker from the recorded and inverse filtered signal [42]. Moreover, a high signal-to-noise ratio is obtained, which was further increased by deriving each BRIR as a mean of 10 single BRIR measurements. It was checked that averaging did not lead to systematic alteration of the BRIRs. Finally, for equalization the complex spectra of the obtained BRIRs were divided by the complex spectrum of the anechoic loudspeaker impulse response. The anechoic response was measured with an omnidirectional microphone (B&K 4133) in 3 m distance on the horizontal plane in the anechoic chamber of the University of Oldenburg using the same sweep method. The resulting magnitude spectrum of the loudspeaker is shown in Fig. 5.

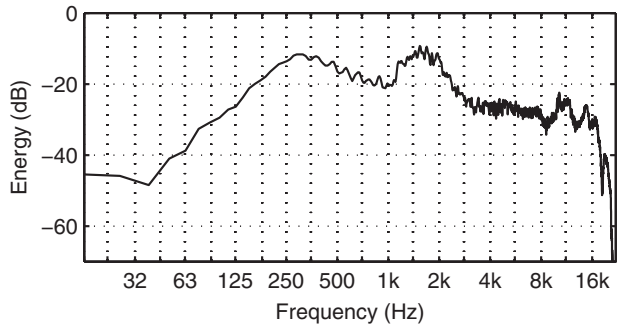

Fig. 5. Frequency response of the omnidirectional loudspeaker, measured in 3 m distance in an anechoic room (same sweep measurement technique as for the BRIRs; see description in the text).

2.1.3 BRIR Synthesis

For the synthesis of BRIRs with the proposed method, the room dimensions, the source and receiver positions, and the frequency-dependent absorption coefficients of the wall surfaces are required as input parameters. The room dimensions and source-receiver configurations were taken from the room geometry descriptions above (Sec. 2.1.1). Since the synthesis method assumes empty shoebox-shaped rooms, interior objects such as furniture were neglected. The absorption coefficients required in the synthesis method can, in general, be looked up in datasheets. However, these values will in most cases differ considerably from those of the actual walls, as only typical values are represented. To maximize the comparability of the resulting synthesized BRIRs to the measured ones with regard to reverberation time and overall frequency response, the measured reverberation times in frequency bands were taken into account: to obtain a reverberation time according to the respective measured BRIR, one absorption coefficient as a mean value over all walls was calculated from the inverse form of Sabine's formula Eq. (3). It should be noted that with a single average absorption coefficient for all walls, sound absorption is direction independent. Alternatively, Jot et al. (1997) [43] suggested to take T_{60} directly as an input parameter. However, the idea here is that for later applications all parameters for the simulation should be directly taken from the geometrical arrangement and the wall absorption coefficients, such that no BRIR measurement is required. Beside these "fixed" input parameters based on the measured rooms, the maximum image source order N was a variable parameter of interest. For the evaluation, N was varied in the range between 0 and 3, defining the "standard synthesis conditions." Additionally, "$N = \text{Inf}$" denotes that BRIRs were synthesized by only the ISM. In this case an actual value $N = 20$ was used that appeared sufficient for rooms with small reverberation times. (Therefore, it was only used to create a control sample BRIR for room L.) Thus in the cases "$N = 0$" and "$N = \text{Inf}$" the outputs of both components, the ISM and the FDN respectively, can be investigated separately.

Furthermore, two additional synthesis conditions were introduced that simplify spatialization in two ways. Instead of HRTFs only broad-band interaural level differences (ILD) were used, realized by a scaling factor of the left and right channel of each image source signal and FDN delay line, respectively. In the first additional synthesis condition, referred to as "P1," all ILDs were created by vector base amplitude panning: with P and S denoting the position vectors of the receiver and sound source, respectively, and e being the normalized direction vector of the interaural axis, the left and right channel were multiplied with the scaling factors:

$$s_{\mathrm{L,R}} = 1 \pm e \cdot \frac{P - S}{\|P - S\|} . \tag{7}$$

In the second additional synthesis condition, referred to as "P2," the same panning was used for the image sources, but for the FDN the left and right scaling factors were rounded to the values $(s_{\mathrm{L}}, s_{\mathrm{R}}) \in \{(2, 0), (1, 1), (0, 2)\}$. In the result, each FDN channel was mapped to the left, center, or right position. The P1 and P2 conditions were used to add some unnatural sounding examples in the evaluation in order to ensure a use of the complete rating scale for naturalness (see Sec. 2.3.3). P1 and P2 sound less natural because no sound externalization is achieved for the reverberation. Furthermore, the frequency dependency of ILD, which is observed in reality due to acoustic head shadowing, is not represented and large low frequency ILDs are introduced that do not occur in nature. Additionally, no interaural time differences are present, as well as coloration due to the sound source elevation.

Fig. 6 shows examples of measured and synthesized BRIRs for rooms L and A. For the synthesized BRIRs, $N = 0$ (mid panels) and $N = 3$ (lower panels) were used. It is observed that the waveforms for the early reflections are similar to some degree, though some small amplitude reflections are not recreated. These may be diffusely reflected parts and/or reflections from the interior objects. However, it should be pointed out that it is not the aim to recreate the particular BRIR waveform but rather the accordance of room acoustical parameters and perceptual attributes.

2.2 Objective Evaluation

In the following, the measured and synthesized BRIRs are analyzed in terms of objective parameters as defined in ISO 3382-1 [44]. For parameters that require a single channel (monaural) RIR, the BRIRs were averaged over the left and right channel before parameter calculation.

2.2.1 Reverberation Time

The reverberation time T_{60} is the most common decay rate measure and is usually obtained via the Schroeder integration method [37, 44]. In order to omit the noise floor, in accordance with ISO 3382-1 [44], the Lundeby method [45] was chosen to find a suitable truncation point. The T_{60} calculation was carried out in octave bands between 250 Hz and 8 kHz. Additionally, synthesized BRIRs were analyzed in the frequency ranges below and above these edges (that was, due to insufficient signal-to-noise ratios in those frequency regions, not possible for the measured BRIRs).

Fig. 7 shows the T_{60} results for the four rooms, calculated from measured BRIRs (black solid squares) and BRIRs of

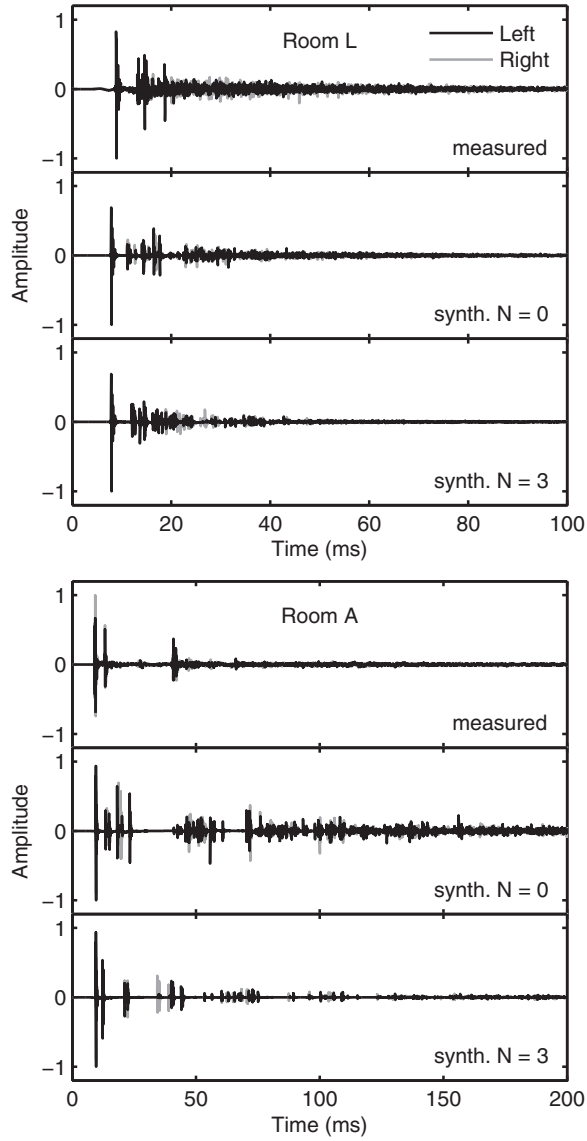

Fig. 6. Comparison of measured and synthesized BRIRs (normalized amplitudes) for room L and A.

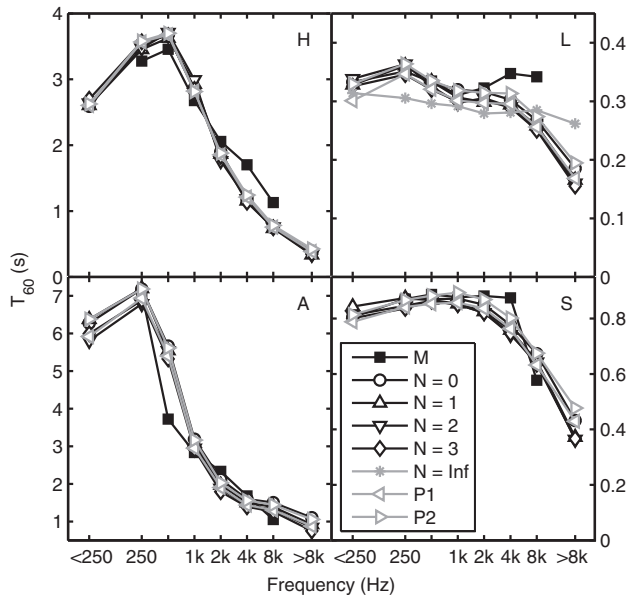

Fig. 7. Reverberation times as functions of octave band center frequency (and spectral edges <250 Hz, >8 kHz) for different rooms (panels), obtained from measured RIRs (black solid squares) and synthesized RIRs (different synthesis conditions, see legend).

all synthesis conditions (open symbols). It can be observed that the reverberation time curves for synthesized BRIRs match the measured BRIRs very well in most cases. In room S the strong decay beyond 4 kHz is not that well reproduced, and in L, with increasing frequency, all synthesized curves drop faster than the measurement.

With the exception of the condition $N = \text{Inf}$, which is separately discussed later, no considerable differences are observed between the synthesis methods.

The good match between the T_{60} values of the synthesized and measured BRIRs could be expected, since the FDN was designed to yield reverberation with a specified frequency dependent T_{60} and a smooth transition between the ISM and FDN was achieved. Remaining deviations between synthesis and measurement can be attributed to the choice of the absorption filters in the FDN. Because the number of applications of one certain absorption filter grows exponentially with time, even a small deviation in its actual frequency response from the desired frequency response causes a larger deviation in the obtained T_{60} value. Especially the high frequency drop in room L is caused by the additional global bandpass within each absorption filter. As a consequence, future versions could improve the accuracy of the absorption filter frequency responses (see Secs. 1.1, 1.2), as well as an improvement of the global bandpass.

The T_{60} values of the $N = \text{Inf}$ condition (only applied for room L, upper right panel) show deviations to the measured T_{60} values at all frequencies but are nevertheless in the same range. The slightly larger deviations are not surprising in this case, given that T_{60} is not controlled explicitly in the ISM.

2.2.2 Early Decay Time

The early decay time (EDT), proposed by Jordan (1970) [46], is another decay-rate measure, which is calculated similarly as T_{60}, but taking into account only the first time interval, where the normalized BRIR energy (obtained by the Schroeder integration method) drops from 0 dB to −10 dB. EDT and T_{60} differ if the EDC deviates from a linear decay. Furthermore, it is considered to be a better measure for the subjective reverberance than T_{60} [44]. As the early reflections are more strongly weighted, the EDT is more sensitive to room geometry and source- and receiver-configuration.

Fig. 8 shows the EDT results for the four rooms in the same manner as in Fig. 7. Here, more deviations between synthesis and measurements are observed. However, all together, the synthesis still reasonably follows the measured EDTs (taking the different ordinate ranges into account). The weaker accordance of the curves is not much surprising,

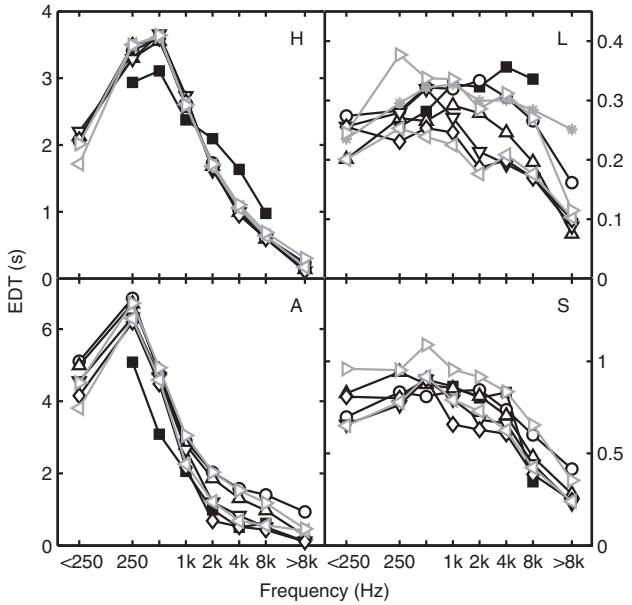

Fig. 8. Early decay time as functions of octave band center frequency (and spectral edges <250 Hz, >8 kHz) for different rooms (panels), obtained from measured RIRs (black solid squares) and synthesized RIRs (different synthesis conditions, see legend in Fig. 7).

since the simulation method is not explicitly tuned for correct EDTs as it was for T_{60}.

Except for room H, considerably larger variances in EDT depending on the synthesis condition are visible. These more profound variations can be explained by the higher weighting of early reflections, which vary strongly with the choice of the maximum image source order N.

In order to compare the different synthesis conditions and their deviations from the measurement condition, the results were analyzed in terms of the just noticeable difference (JND) of EDT as defined in ISO 3382-1 [44]. In the standard, the JND is given as a relative value of 5% of the arithmetic mean of EDTs for the 500-Hz and 1-kHz octave bands. Calculating this mean for the results yields deviations from measurement condition ranging from about 2 JNDs (rooms S, L), with almost no dependency from synthesis condition, to a maximum of 10 JNDs (room A, $N = 0$) and 6 JNDs (room A, $N = 3$). It should be noted that besides ISO 3382-1 actual JND values should vary with the source signal the BRIR is convolved with. Whereas a small JND in the region of 5% generally occur for transient signals, it is expected that JNDs are considerably larger for signals with a more stationary character. For instance, Meng et al. (2006) [47] found JNDs for T_{60} up to about 30% for musical motifs as source signal. Similar results might be expected for EDT.

2.2.3 Definition and Clarity Index

Sound reflections in closed rooms can influence speech intelligibility in two ways: reflections with delays and/or levels below certain limits compared to the direct sound are not perceived as separate echoes but rather result in higher intensity of the early sound components. In this case early reflections support speech intelligibility and might cause a larger apparent source width. In contrast, reflections with larger delays and/or higher levels rather affect speech intelligibility by masking certain syllables (see, e.g., [38], Ch. 7.3).

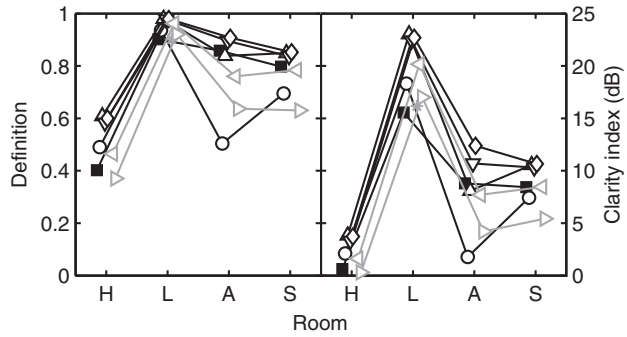

Fig. 9. Definition (left panel) and clarity index (right panel) (broadband each), calculated from measured (black solid squares) and synthesized (see legend in Fig. 7) RIRs of the four rooms.

As two objective measures for speech intelligibility and transparency of music, ISO 3382-1 [44] defines the definition D and the clarity index C, respectively. For the calculation of D and C, the measured BRIRs were truncated according to the Lundeby method [45] in order to discard the noise floor. For the corresponding synthesized BRIRs, the same truncation points were chosen for comparability.

The results for the four rooms are shown in Fig. 9. For most synthesis conditions and rooms a fair accordance with the measurement is observed. For definition, except for the two conditions "$N = 0$" in room A and "P2" in rooms A and S, deviations lie within a range of 0.2. For clarity index, deviations are mostly in the range of 3 dB, only for a few cases up to 7 dB.

A qualitative comparison between the rooms shows that D and C are higher in those rooms with a small T_{60}, as expected. A distinct example is room H with nearly $C = 0$ dB (measured) and maximum 4 dB (synthesized), in which case the reverberant energy equals almost the early-reflection energy. However, a clear exception for this relation is the aula (room A), which shows the highest T_{60} (see Fig. 7) but has almost the same definition and clarity index as room S. This can be explained by a very small distance between source and receiver, compared to the room dimensions (see Sec. 2.1.1), producing a high ratio of direct to reverberant sound. Here, the largest deviation in D and C is obtained for "$N = 0$," which clearly underlines the advantage of geometric early reflection simulation over a pure reverberation algorithm.

For further analysis, JNDs for D and C were assessed as defined in ISO 3382-1 [44]. In the standard, the JNDs are given as $D = 0.05$ and $C = 1$ dB of the arithmetic means of these measures for the 500-Hz and 1-kHz octave bands. For all rooms and for the standard conditions with $N > 1$, deviations from measurement condition were in a range of

about 2 JNDs (D) and 4 JNDs (C). Largest deviations occur for room A, $N = 0$: -10 JNDs (D) and -8.6 JNDs (C).

2.2.4 Interaural Cross-Correlation Coefficient

The spatial perception of sound depends on binaural aspects, which were not covered by the objective measures so far. As a subjective measure for the perceived spaciousness, the apparent source width (ASW) is often discussed. It was proposed by Hidaka et al. (1995) [28] to be "the apparent auditory width of the sound field created by a performing entity as perceived by a listener in the audience area of a concert hall." As a model for ASW, the same authors have proposed an approach based on the interaural cross-correlation function, which is in general defined as

$$\mathrm{IACF}_{[t_1,t_2]}(t) = \frac{\int_{t_1}^{t_2} h_\mathrm{L}(\tau)\cdot h_\mathrm{R}(\tau+t)\,\mathrm{d}\tau}{\sqrt{\int_{t_1}^{t_2} h_\mathrm{L}^2(\tau)\,\mathrm{d}\tau\cdot\int_{t_1}^{t_2} h_\mathrm{R}^2(\tau)\,\mathrm{d}\tau}}, \qquad (8)$$

with $h_{\mathrm{L,\,R}}$ being the left and right channel of the BRIR, respectively. The integration interval $[t_1,\ t_2]$ is typically chosen to select the early, late or whole part of a room impulse response. As a single-number value, the interaural cross-correlation coefficient,

$$\mathrm{IACC}_{[t_1,t_2]} = \max\left\{\mathrm{IACF}_{[t_1,t_2]}(t);\ -1\ \mathrm{ms} \le t \le 1\ \mathrm{ms}\right\}, \quad (9)$$

can be used, and Hidaka et al. [28] proposed the number $(1 - \mathrm{IACC}_{[0,\,80\,\mathrm{ms}]})$ as a measure of ASW. In the literature, the notation $\mathrm{IACC_E} := \mathrm{IACC}_{[0,\,80\,\mathrm{ms}]}$ is often found for the IACC of the early part and will be used in the following, too. However, the applicability of $(1 - \mathrm{IACC_E})$ as a measure of ASW is not unambiguous. For example, measurements and simulations by de Vries et al. (2001) [48] showed very strong fluctuations in $\mathrm{IACC_E}$ with source or listener placement, whereas such fluctuations in ASW were not found.

Nevertheless, for the evaluation in the current study, $\mathrm{IACC_E}$ was calculated in octave bands as a pure objective measure for all BRIRs. In order to obtain a single-number measure, the mean of $\mathrm{IACC_E}$ values for the 500 Hz, 1 kHz, and 2 kHz bands is commonly regarded and denoted by the symbol $\mathrm{IACC_{E3}}$. The results for $\mathrm{IACC_{E3}}$ are shown in Fig. 10 for the four rooms. The symbols are the same as in Fig. 7. For the standard synthesis conditions with $N > 0$ a good accordance between measurement and synthesis is observed. For some rooms (H, A) a slightly better match with the measurements can be observed toward higher maximum image source orders. The $N = 0$ synthesis almost always results in very low $\mathrm{IACC_{E3}}$ values, which means that the binaural FDN itself creates a very high degree of interaural diffuseness. Because the early parts of the BRIRs were analyzed, these results show the importance of simulating early reflections more accurately. The P1 and P2 conditions produce overall high values, which is plausible as only ILDs were applied. For IACC, ISO 3382-1 [44] defines a constant JND of 0.075. For the broadband IACC values, differences between measured and synthesized BRIRs were calculated and averaged over all rooms. For the standard synthesis conditions with $N > 0$ absolute differences were smaller than 1 JND. For $N = 0$ an average difference of -1.5 JNDs was obtained. However, it should be noted that in contrast to ISO 3382-1 the actual JND depends on the reference IACC and increases with decreasing IACC, as shown by several authors [49, 50].

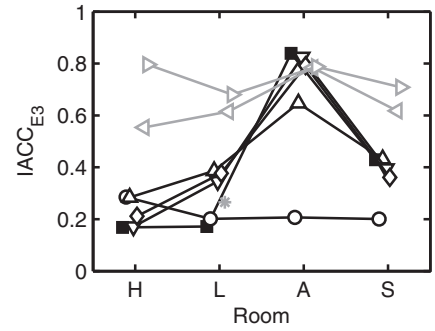

Fig. 10. Interaural cross correlation coefficient calculated from the early parts (0–80 ms) of measured (black solid squares) and synthesized (see legend in Fig. 7) BRIRs of the four rooms. Each data point is the average of IACCs for the 500-Hz, 1-kHz, and 2-kHz octave bands, indicated by the symbol $\mathrm{IACC_{E3}}$.

2.3 Subjective Evaluation

The simulation method was evaluated subjectively by rating dry test signals convolved with both measured and synthesized BRIRs. The subject's task was to rate several perceptual attributes that represent both specific perceived room properties and more qualitative properties. For simplicity, in the following the convolved test signals will also be referred to as the BRIRs.

2.3.1 Subjects and Apparatus

Fifteen subjects (7 female, 8 male) aged 24 to 32 years participated in the experiment. All reported to be normal-hearing, except for one subject with a very slight impairment. Twelve subjects had prior experience with psychoacoustic experiments. Subjects were seated in a sound attenuating listening booth. Playback and rating of the test signals were controlled via a graphical user interface. Sounds were presented using Sennheiser HD 650 headphones driven by an RME Fireface UC at a sampling rate of 48 kHz.

2.3.2 Synthesis Conditions and Source Signals

BRIRs were chosen from the four rooms as described in Sec. 2.1.1. In addition to the measured BRIRs, synthesized BRIRs were chosen for all standard synthesis conditions $N \in \{0, 1, 2, 3\}$ as described in Sec. 2.1.3. Additionally, the P1- and P2-conditions were used. The condition $N = \mathrm{Inf}$ was added for room L only because of time constraints for the rating procedure.

To obtain the actual test signals, two dry source signals, a speech and a music signal, were convolved with the BRIRs. The speech signal contained an approximately 4-s long female spoken sentence and the music signal an approximately 10-s long sample of a guitar piece. The latter was a natural sounding MIDI-controlled software instrument, simulating an acoustic guitar with steel strings. By these

choices, ordinary signal types of everyday life were represented and—especially for the music signal—the rooms were excited in a wide frequency range. Furthermore, the music signal contained many transients as well as stationary parts. The RMS value of measured BRIRs was equalized to the respective synthesized one with $N = 3$. Among all synthesized BRIRs per room these ($N = 3$) were chosen as reference. Given that the spectra and temporal decay of the synthesized and measured BRIRs are similar, it is assumed that RMS equalization also equalizes loudness to a good approximation, as also supported by informal listening. Thus no further attempt to normalize loudness was considered necessary. Likewise the RMS values of the source signals were equalized. Only signal parts with energy levels above −60 dB under maximum were considered for computing the RMS to avoid a strong impact of signal pauses. Overall presentation sound pressure levels of 60 to 65 dB SPL were obtained.

2.3.3 Rated Properties and Procedure

The following sound properties were rated by the subjects on a discrete rating scale with steps from 1 to 7: naturalness (high—low), reverberance (low—high), room size (small—large), coloration (dark—bright), metallic character (none—strong), source width (small—large).[1] These attributes are a subset of perceptual qualities proposed in [51] in order to characterize virtual acoustic environments. This specific subset was chosen to represent the main aspects naturalness ("naturalness"), room acoustical properties ("reverberance," "coloration," "source width"), and potential artifacts of the simulated BRIR with regard to the measured BRIR ("metallic character," "coloration"). Especially the metallic character is a well-known artifact in virtual acoustics and reverberation algorithms (e.g., [8]). The perceived room size, regarded as a room acoustical property, but not directly proposed in [51] was added, in order to investigate whether it is perceived separately from reverberance.

The experiment was performed by each subject in two sessions on different days. The first session contained a standardized verbal introduction, the presentation of several sound examples (see below), and the rating task itself. The second session was a retest session and therefore identical to the first one, except for a shorter introduction and a different random order of the presented sounds. One session had an average duration of 90 minutes, including breaks. Before starting the first session, the subjects were told, among others, that speech and music signals are presented that were rendered in different rooms. The source signals were always the same, thus sound differences are only due to different room properties. Subjects were told to imagine being inside those rooms with sound sources outside their heads. It was noted that the rooms can contain interior objects or be empty. Subsequently selected sound examples were presented to give an impression of the total range in which the perceptual attributes could vary. BRIRs were chosen from the set of those, which were to be rated in the main task, but a different source signal from the main task was used (short spoken sentence, male speaker). Thus the test signals could not be recognized directly in the later main task. The subjects could listen to the examples as often as desired for the introducing examples and in the main task. In the main task, sounds were presented for each property in a random order. After all samples were rated concerning one property, the rating of the same samples (in a different random order) started concerning the next property.

2.3.4 Results and Discussion

The subjective evaluation results, averaged over subjects and source signals, are shown in Fig. 11. For each property, shown in the separate panels, mean values are plotted against rooms with error bars indicating the inter-subject standard errors (standard deviation across all subjects and source signals, divided by the square root of the number of subjects).

Comparing the pattern of results as a function of room for each of the attributes, overall a good agreement between the measured and synthesized BRIR conditions is achieved. One exception is the attribute coloration. Here, the results for the different synthesis conditions are considerably spread over the rating scale. Another exception is the P2 panning condition (gray right-pointing triangle; ILDs instead of HRTFs, see Sec. 2.1.3), which shows deviation from all other BRIR conditions for all attributes and the attribute metallic character for which all BRIR synthesis conditions show some deviation from the data. A similar pattern of results is observed for naturalness and metallic character. Similarly, reverberance and room size were rated very similar to each other across rooms. For coloration and source width the pattern of results is different, showing considerably more spread for coloration. In addition, subjects did not use the full range of the rating scale for coloration and source width, while they did for all other attributes.

To assess the effect of BRIR condition in more detail, a repeated-measures ANOVA was applied to each combination of room and attribute. A significant main effect (applying Greenhouse-Geisser sphericity correction) of BRIR condition (measured and synthesized) on ratings for most perceptual attributes and rooms (see Table 2 in the Appendix) was found. No significant main effect was found only for reverberance in room H ($p = 0.057$), room size in room L ($p = 0.077$), and source width in rooms A, H, S ($p = 0.108$, $p = 0.081$, $p = 0.204$). Post-hoc comparisons (Bonferroni correction, $p < 0.05$) were performed to assess the differences between BRIR conditions. The results are summarized in Table 1 in a matrix form with the upper and lower triangle showing pairwise comparisons for two different perceptual attributes (indicated at the top and bottom of each sub table). The rooms for which the pairwise comparison showed a significant difference are indicated by the letters.

[1] The original german attributes were Natürlichkeit (hoch—niedrig), Raumklang (trocken—hallig), Größe des Raumes (klein—groß), Klangfärbung (dunkel—hell), Metallizität des Klanges (niedrig—hoch), Breite der Schallquelle (klein—groß).

Table 1. Post-hoc comparisons (Bonferroni correction applied) of the subjective evaluation results. The top and bottom triangles of each matrix are for different attributes as indicated at the top and bottom. Each cell represents a comparison between two conditions and contains all names of rooms for which these conditions were rated. If a significant difference was found ($p < 0.05$), the room name is printed black (otherwise gray).

	M	0	1	2	3	Inf	P1	P2	
							↓ Naturalness		
M		**AH**SL	AHSL	AHSL	AHSL	L	ASL	**ASL**	M
0	AHSL		**AS**L	**A**SL	**AHS**L	L	**A**SL	A**L**	0
1	AHSL	AHSL		AHSL	AHSL	L	ASL	**ASL**	1
2	AHSL	**A**HSL	**A**HSL		AHSL	L	ASL	**ASL**	2
3	AHSL	**A**HSL	**A**HSL	AHSL		L	ASL	**ASL**	3
Inf	L	L	L	L	L		L	L	Inf
P1	ASL	**A**SL	**A**SL	ASL	ASL			**ASL**	P1
P2	ASL	**AS**L	ASL	ASL	ASL	L	ASL		P2
	Reverberance ↑						↓ Room size		
M		A**H**SL	AHSL	AHSL	AHSL	L	ASL	A**S**L	M
0	AH**S**L		AHL	AHSL	AHSL	L	**A**SL	**A**SL	0
1	AH**S**L	AHSL		AHSL	AHSL	L	ASL	A**S**L	1
2	**AHS**L	A**H**SL	A**H**SL		AHSL	L	ASL	ASL	2
3	**AHS**L	AHSL	**A**HS**L**	AHSL		L	ASL	ASL	3
Inf	L	L	L	L	L			L	Inf
P1	AS**L**	ASL	A**S**L	A**SL**	**ASL**	**L**		ASL	P1
P2	ASL	ASL	A**S**L	**ASL**	**A**SL	L	ASL		P2
	Coloration ↑						↓ Metallic character		
M		**AHS**L	**AHS**L	**A**HSL	**A**HSL	L	A**S**L	**AS**L	M
0	AHS**L**		**A**HSL	**AHS**L	**AH**SL	L	**A**SL	ASL	0
1	AHSL	AHS**L**		A**H**SL	**AH**SL	L	**A**SL	ASL	1
2	AHSL	AHS**L**	AHSL		AHSL	L	ASL	A**S**L	2
3	AHSL	AHS**L**	AHSL	AHSL		L	ASL	A**S**L	3
Inf	L	**L**	L	L	L		L	L	Inf
P1	ASL	AS**L**	ASL	ASL	ASL	L		**AS**L	P1
P2	ASL	ASL	ASL	ASL	ASL	L	ASL		P2
	Source width ↑								
	M	0	1	2	3	Inf	P1	P2	

Naturalness was rated quite consistently across subjects and ratings show a very good agreement between synthesized and measured conditions. Even higher naturalness ratings for the synthesized conditions than for the measured conditions occurred. The post-hoc comparisons (Table 1) show that the $N = 0$ condition shows ratings significantly different from those for the measured BRIR and the other synthesized ones for some rooms. The P2 condition, which was intended to sound spatially unnatural, shows significant deviations to the measured BRIR for A, S, L. For the P1 almost no significant differences for naturalness were found. Though subjects were also told to pay attention to the spatiality of the sound, they obviously did not detect the spatial errors that were introduced by the usage of broadband ILDs instead of HRTFs in condition P1 (see Sec. 2.1.3). A reason might be a certain familiarity with headphone listening to music recordings where stereo panning is a usual technique that thus might have been used as a

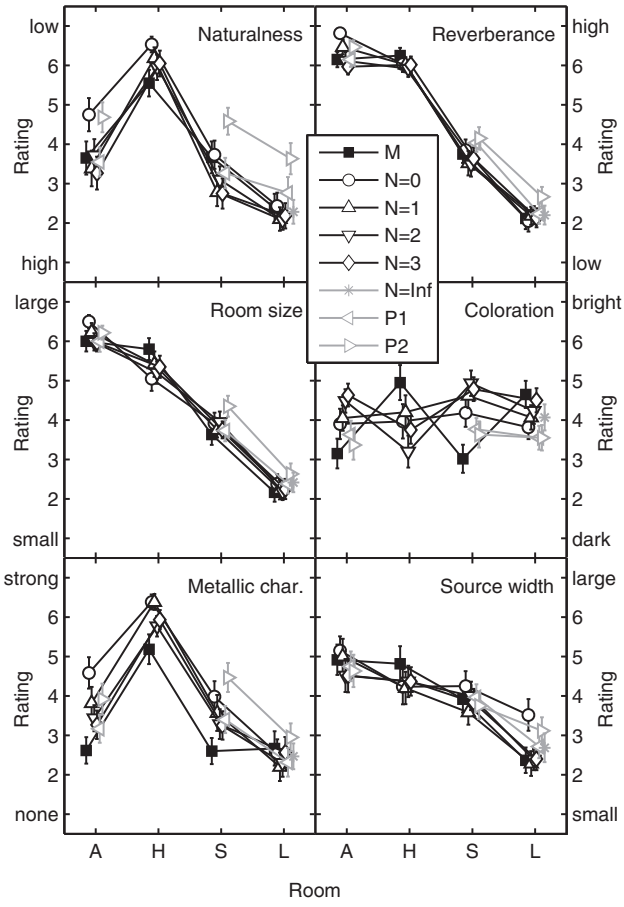

Fig. 11. Subjective sound property rating (panels) of measured (black solid squares) and synthesized (see legend) BRIRs of the four rooms, averaged over all subjects. Error bars indicate standard errors over all subjects.

further internal reference for naturalness with respect to spatiality.

While the ratings for metallic character show a similar pattern of results as for naturalness in Fig. 11, synthetic BRIRs were mostly rated to be more metallic (rooms A, H, S). This is supported by the post-hoc test (Table 1), with significant differences between the measured and the synthetic BRIRs for A, H, and S, mostly for lower values of N. This might identify a weak point of the FDN implementation. For room L, which has the overall lowest ratings for metallic character, no significant differences between conditions were found. The similar patterns of results for the metallic character and naturalness in Fig. 11 might indicate that the metallic character could be one main factor of unnaturalness. Later investigations identified the delay elements in the FDN to influence the metallic character. It turned out that the delay values for room H are too small to fulfill the Jot criterion [13] (cf., Sec. 1.2), whereas those values for all other rooms do fulfill the criterion. Informally, it was shown that a metallic sound can be avoided by choosing larger delays. For improvement, paradigms for choosing delays proposed by other authors (e.g., [13, 52, 53]) should be incorporated. An open question is still whether a real room of this size would sound that metallic if walls were perfectly plain and rectangular to each other. Nevertheless, because this is usually not observed in real cases, such metallic character should be avoided.

Reverberance and room size are discussed together because their results are again similar. In a qualitative comparison, the order of rooms with respect to rated reverberances match the order with respect to T_{60} and EDT (see Secs. 2.2.1, 2.2.2). The same holds for room size and actual room volumes with the exception of room H, which is the smallest room (see Sec. 2.1.1) but nevertheless very reverberant. The ratings for most synthesis conditions are very similar supported by the post-hoc comparisons (Table 1) where significant differences are rather sparse. The $N = 0$ condition is different to the measured BRIR for a single room (A for reverberance, H for room size). Otherwise a few differences occurred mainly for room A. Thus, the simulation method is able to represent these two acoustic properties quite consistently. For reverberance, this result is not surprising since T_{60} and EDT of synthesis conditions match in the same manner the values for respective measurements (see Secs. 2.2.1, 2.2.2). For room size it is not surprising, too, since this measure is expected to be subjectively strongly related to reverberance.

In contrast to the other attributes, coloration was rated to be in the medium range of the scale. For the measured BRIRs, a small dependency of coloration from room in the range of about two rating points is obtained. For the synthesized BRIRs, such a dependency is either not obtained (e.g., $N = 0$, P1, P2) or occurs partly even in an opposing way (e.g., $N = 2$). In fact, for the synthesis conditions, deviations toward a brighter sound (rooms A and S) were expected from informal listening tests, and in future versions, a correcting filter, probably instead of the band-pass within the absorption filters (see Sec. 1.2), shall be applied. Changes in coloration with synthesis condition do not seem to be systematic, for instance in such a way that coloration is changed in a certain direction with increasing N. The post-hoc comparisons (Table 1) show significant differences for many conditions scattered across all rooms. The not systematic effects might be a hint that the subjects had some difficulties to rate coloration, possibly because differences were very small or because subjects had difficulty with interpreting what coloration means. However, since the inter-subject standard errors are not extremely high it appears that subjects had a similar judgment of coloration.

For the source width, no large differences between measured and synthesized conditions were observed. The largest deviation from the measured condition occurs for room L, $N = 0$, and is about 1.5 rating points toward a larger source width. The post-hoc comparison showed significant rating differences for room L between the $N = 0$ condition and all other BRIRs and between $N = 0$ and the measured BRIR. Otherwise no significant differences were found. Similarly as for coloration, ratings lie within the medium range of the scale. Thus, source width is not perceived as being that much different between rooms (like, e.g., for reverberance), though a certain dependency is visible. It seems to be related to reverberance/room size

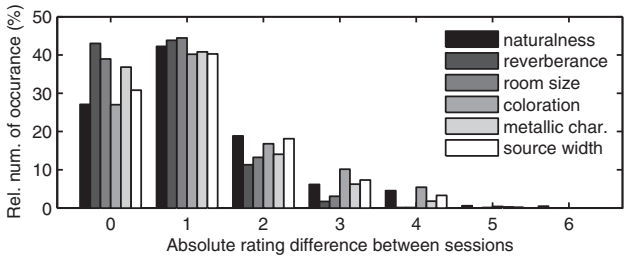

Fig. 12. Relative occurrences of absolute rating differences between test and retest session for all perceptual attributes.

in such a manner that perceived source width increases with increasing perceived reverberance/room size.

To assess the reliability of the data, it was investigated whether subjects gave same or similar ratings for same conditions in both sessions of the experiment. Therefore, absolute rating differences between both sessions per subject and condition were counted and normalized to the total number of ratings per condition. The obtained values are shown in Fig. 12. The majority of rating differences was 1 and 0 for all perceptive attributes, whereas ratings that differed in more than 4 points between the sessions did almost not occur. Hence, an overall good stability in the ratings was observed.

3 SUMMARY AND CONCLUSIONS

A hybrid approach for binaural room reverberation rendering was presented to binaurally auralize a room based on absorption coefficients or T_{60} times and HRIRs. The proposed synthesis method renders geometrically derived early reflections using the ISM under the assumption of shoebox-shaped rooms. To limit the computational complexity the ISM is only used up to a low reflection order, whereas the reverberant tail of the BRIR is reproduced by an FDN. The FDN is computationally very efficient and allows explicit control of the spectral decay characteristics.

For binaural auralization, the application of HRIRs is straightforward in the ISM-based early-reflection renderer, whereas the FDN was extended by a binauralization stage enabling a spatial distribution of the reverberation. Here, HRIRs from twelve directions spatially distributed on a cube axis-aligned with the shoebox room were chosen to render the output of the twelve delay lines in the FDN with delays according to the room dimensions matching the cube surfaces. In comparison to Menzer and Faller (2009) [27], who generated binaural FDN outputs by decorrelation, the suggested binauralization is more physically motivated taking into account the room geometry. The current approach improves the conceptually similar spatial distribution of reverb streams in Kendall et al. (1986) by creating uniform spatial reverb for rooms with unequal room dimensions and for arbitrary receiver positions. In contrast to Menzer and Faller (2009) the current approach is suited for six degrees-of-freedom movement of the receiver and takes the receiver and source position into account, which could be even further improved taking aspects of [27] into account.

In order to investigate to what extent the simulation method is able to represent room acoustical measures correctly on the one hand, and to be plausible on the other hand, both objective and subjective evaluations were performed. For that purpose, measured and respectively synthesized BRIRs were compared with respect to objective and subjective properties. Here, one main varying synthesis parameter was the maximum order N of reflections to be computed by the ISM.

In the objective evaluation, good agreement between measured and synthesized BRIRs was verified for the reverberation time, independently from synthesis condition. Since the FDN is designed to create a defined frequency dependent decay rate, this result was not surprising. For early decay time (EDT), definition, and clarity index mostly good or fair agreements between measured and synthesized BRIRs were found. However, in comparison to constant standard JNDs after ISO 3382-1 [44], depending on the synthesis condition, deviations between 2 and 10 JNDs occurred. The interaural cross correlation coefficient $IACC_{E3}$ was reproduced well for many conditions. Deviations from measured BRIRs lie mostly in the range of the broadband IACC JND after ISO 3382-1 [44].

In the subjective evaluation sound quality properties were rated for measured and corresponding synthesized BRIRs, convolved with a dry speech- and music signal, respectively. Here, plausibility was evaluated in terms of naturalness and metallic character. Ratings for synthesized BRIRs did not differ significantly from those for measured BRIRs if at least the first reflection order was simulated by the ISM ($N > 0$). Synthesized BRIRs were even rated to be slightly more natural for some conditions. Metallic character showed ratings mostly inverse to naturalness, indicating to be one aspect of unnaturalness.

Authenticity was evaluated indirectly by the comparison of ratings between measured and synthesized BRIRs in terms of perceived room size, reverberance, spectral coloration, metallic character, and source width. The results showed mostly very similar ratings for measured and synthesized BRIRs, independently from synthesis condition. Pair-wise comparisons yielded only a few conditions with significant differences, especially for the panning conditions and small values of N. Two weak points were the perceived metallic character that was found to be a bit more distinct for synthesized BRIRs compared to measured ones and also coloration that deviated between different synthesis conditions.

By experimental design, the subjective evaluation did not investigate whether sound sources are localized to equal directions and distances for measured and synthesized BRIRs, which was beyond the scope of this paper. Second, authenticity was investigated only indirectly and not directly by room mapping tasks (which could be done by pictures of the rooms or actual inspections by subjects). Nevertheless, for the investigated sound properties, it was found that the simulation method is able to reproduce their perception in most cases. It should be emphasized that this

was achieved with some drastic physical simplifications, even if low maximum reflection orders are simulated accurately.

In conclusion, the proposed method can successfully synthesize BRIRs for the tested room types resulting in similar acoustical properties as assessable by objective measures and mostly good agreement of subjective ratings with the respective real rooms. The proposed method is already used in a psychoacoustic task with adaptively changing BRIR and is currently implemented in a real-time system. Extensions for more complex scenarios such as coupled rooms are currently under investigation. Different wall absorption coefficients on the walls can potentially be used in the binauralization stage of the FDN to introduce a more realistic spatial energy distribution in the reverberant field.

4 ACKNOWLEDGMENTS

This work was supported by the DFG FOR 1732 and Cluster of Excellence EXC 1077/1 Hearing4all.

REFERENCES

[1] P. N. Wilson, N. Foreman, and D. Stanton, "Virtual Reality, Disability and Rehabilitation," *Disability and Rehabilitation*, vol. 19, pp. 213–220 (1997).

[2] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 66, no. 4, pp. 943–950 (1979).

[3] B.-I. Dalenbäck, "Engineering Principles and Techniques in Room Acoustics Prediction," in *BNAM*, Bergen, Norway (May 2010).

[4] G. M. Naylor, "Treatment of Early and Late Reflections in a Hybrid Computer Model for Room Acoustics," in *124th ASA meeting New Orleans* (November 1992).

[5] B.-I. Dalenbäck and M. Strömberg, "Real Time Walkthrough Auralization—The First Year," Technical report, CATT (Dalenbäck), Valeo Graphics (Strömberg) (2010).

[6] D. Schröder, F. Wefers, S. Pelzer, D. S. Rausch, M. Vorländer, and T. Kuhlen, "Virtual Reality System at RWTH Aachen University," in *Proceedings ICA 2010, 20th International Congress on Acoustics: 23–27 August 2010*, Sydney, New South Wales, Australia (2010).

[7] J. Borish, "Extension of the Image Model to Arbitrary Polyhedra," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836 (1984).

[8] J. A. Moorer, "About This Reverberation Business," *Computer Music J.*, vol. 3, no. 2, pp. 13–28 (1979).

[9] F. R. Moore, "A General Model for Spatial Processing of Sounds," *Computer Music J.*, vol. 7, no. 6, pp. 6–15 (1983).

[10] G. Kendall, W. Martens, D. Freed, D. Ludwig, and R. Karstens, "Image-Model Reverberation from Recirculating Delays," presented at the *81st Convention of the Audio Engineering Society* (1986 Nov.), convention paper 2408.

[11] A. Persterer, "A Very High Performance Digital Audio Processing System," in *Proc. 13th International Conf. on Acoustics*, Belgrad, Yugoslavia (1989).

[12] J.-M. Jot, V. Larcher, and O. Warusfel, "Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony," presented at the *98th Convention of the Audio Engineering Society* (1995 Feb.), preprint 3980.

[13] J.-M. Jot, "Efficient Models for Reverberation and Distance Rendering in Computer Music and Virtual Audio Reality," Technical report (1997).

[14] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen "Creating Interactive Virtual Acoustic Environments," *J. Audio Eng. Soc.*, vol. 47, pp. 675–705 (1999 Sep.).

[15] S. M. Schimmel, M. F. Müller, and N. Dillier, "A Fast and Accurate 'Shoebox' Room Acoustics Simulator," Technical report (2009).

[16] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual Reality System with Integrated Sound Field Simulation and Reproduction," *EURASIP J. Advances in Signal Processing* (2007).

[17] A. Krokstad, S. Strøm, and S. Sørsdal, "Calculating the Acoustical Room Impulse Response by the Use of a Ray Tracing Technique," *J. Sound Vib.*, vol. 8, no. 1, pp. 118–125 (1968).

[18] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 4th Ed. (John Wiley & Sons, Inc., 2000).

[19] T. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, J. E. West, G. Pingali, P. Min, and A. Ngan, "A Beam Tracing Method for Interactive Architectural Acoustics," *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 739–756 (2004).

[20] W. G. Gardner, *Reverberation Algorithms* (1998).

[21] M. R. Schroeder, "Natural Sounding Artificial Reverberation," *J. Audio Eng. Soc.*, vol. 10, pp. 219–223 (1962 July).

[22] J. Stautner and M. Puckette, "Designing Multi-Channel Reverberators," *Computer Music J.*, vol. 6, no. 1, pp. 52–65 (1982).

[23] J.-M. Jot and A. Chaigne "Digital Delay Networks for Designing Artificial Reverberators," presented at the *90th Convention* of the Audio Engineering Society (1991 Feb.), convention paper 3030.

[24] D. Rocchesso and J. O. Smith, "Circulant and Elliptic Feedback Delay Networks for Artificial Reverberation," *IEEE Trans. Speech & Audio*, vol. 5, no. 1, pp. 51–63 (1997).

[25] F. Menzer and C. Faller, "Unitary Matrix Design for Diffuse Jot Reverberators," presented at the *128th Convention of the Audio Engineering Society* (2010 May), convention paper 7984.

[26] J. O. Smith, "A New Approach to Digital Reverberation Using Closed Waveguide Networks," in *Proc. 1985 International Computer Music Conference*, Burnaby, BC, San Francisco (1985).

[27] F. Menzer and C. Faller, "Binaural Reverberation Using a Modified Jot Reverberator with Frequency-Dependent Interaural Coherence Matching," presented at the *126th Convention of the Audio Engineering Society* (2009 May), convention paper 7765.

[28] T. Hidaka, L. L. Beranek, and T. Okano, "Interaural Cross-Correlation, Lateral Fraction, and Low- and High-Frequency Sound Levels as Measures of Acoustical Quality in Concert Halls," *J. Acoust. Soc. Am.*, vol. 98, no. 2, pp. 988–1007 (1995).

[29] J. S. Bradley, H. Sato, and M. Picard, "On the Importance of Early Reflections for Speech in Rooms," *J. Acoust. Soc. Am.*, vol. 113, pp. 3233–3244 (2003).

[30] J. S. Bradley and G. A. Soulodre, "The Influence of Late Arriving Energy on Spatial Impression," *J. Acoust. Soc. Am.*, vol. 4, pp. 2263–2271 (1995).

[31] A. W. Bronkhorst and T. Houtgast "Auditory Distance Perception in Rooms," *Nature*, vol. 397, pp. 517–520 (1999).

[32] F. L. Wightman and D. Kistler, "Headphone Simulation of Free-Field Listening. I: Stimulus Synthesis," *J. Acoust. Soc. Am.*, vol. 85, pp. 858–867 (1989).

[33] J. Blauert, *Spatial Hearing* (MIT Press, Cambridge, 1983).

[34] M. Holters and U. Zölzer, "Parametric Higher-Order Shelving Filters," in *14th European Signal Processing Conference (EUSIPCO 2006)*, 2006.

[35] L. Cremer and H. A. Müller, *Die wissenschaftlichen Grundlagen der Raumakustik*, 1st Ed. (S. Hirzel Verlag Stuttgart, 1948).

[36] Wallach, "On Sound Localization," *J. Acoust. Soc. Am.*, vol. 10, pp. 270–274 (1949).

[37] M. R. Schroeder, "New Method of Measuring Reverberation Time," Technical report, Bell Telephone Laboratories (1965).

[38] H. Kuttruff, *Room Acoustics, 5th Ed.* (Elsevier Applied Science, 2009).

[39] M. Jeub, M. Schäfer, and P. Vary, "A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms," Technical report (2009).

[40] R. Kruse, A. Häußler, and S. van de Par, "An Omnidirectional Loudspeaker Based on a Ring-Radiator," *Applied Acoustics*, vol. 74, pp. 1374–1377 (2013).

[41] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," presented at the *108th Convention* of the Audio Engineering Society (2000 Feb.), convention paper 5093.

[42] S. Müller and P. Massarani, "Transfer-Function Measurements with Sweeps," *J. Audio Eng. Soc.*, vol. 49, pp. 443–471 (2001 June).

[43] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model," presented at the *103rd Convention of the Audio Engineering Society* (1997 Sep.), convention paper 4629.

[44] ISO 3382-1: Acoustics—Measurement of Room Acoustic Parameters—Part 1: Performance Spaces.

[45] A. Lundeby, T. E. Vigran, H. Bietz, and M. Vorländer "Uncertainties of Measurements in Room Acoustics," *Acta Acustica united with Acustica*, vol. 81, pp. 344–355 (1995).

[46] V. L. Jordan "Acoustical Criteria for Auditoriums and Their Relation to Model Techniques," *J. Acoust. Soc. Am.*, vol. 47, no 2, pp. 408–412 (1970).

[47] Z. Meng, F. Zhao, and M. He, "The Just Noticeable Difference of Noise Length and Reverberation Perception," in *International Symposium on Communications and Information Technologies, ISCIT* (2006).

[48] D. de Vries, E. M. Hulsebos, and J. Baan, "Spatial Fluctuations in Measures for Spaciousness," *J. Acoust. Soc. Am.*, vol. 110, pp. 947–954 (2001).

[49] C. Kim, R. Mason, and T. Brookes "Initial Investigation of Signal Capture Techniques for Objective Measurement of Spatial Impression Considering Head Movement," presented at the *124th Convention of the Audio Engineering Society (*2008 May), convention paper 7331.

[50] S. Klockgether and S. van de Par, "Just Noticable Differences of Spatial Perception in Directly Manipulated Binaural Room Impulse Responses," in *AIA/DAGA 2013*, Merano, Italy (2013).

[51] A. Lindau, V. Erbes, H.-J. Maempel, S. Lepa, F. Brinkmann, and S. Weinzierl, "A Spatial Audio Quality Inventory for Virtual Acoustic Environments (SAQI)," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 984–994 (2014) (special issue on Auralization and Ambisonics).

[52] D. Rocchesso, "The Ball within the Box: A Sound-Processing Metaphor," *Computer Music J.*, vol. 19, no. 4, pp. 47–57 (1995).

[53] F. Menzer, "Choosing Optimal Delays for Feedback Delay Networks," in *DAGA, Oldenburg* (2014).

APPENDIX

Table 2. Results of the repeated-measures ANOVAs conducted for each combination of room and perceptual attribute. The assumption of sphericity was violated in all cases (Mauchly's test) and a Greenhouse-Geisser correction (ε) was applied. The *F* statistic and *p* values are given and significant main effects of the synthesis condition (including measured BRIRs) on the ratings of the perceptual attributes are indicated by the number of stars in the last column.

Room	ε	*F*	*p*	ε	*F*	*p*
	Naturalness			Coloration		
A	0.677	16.145	0.000***	0.604	9.698	0.000***
H	0.499	5.341	0.011*	0.772	13.658	0.000***
S	0.817	14.065	0.000***	0.432	13.117	0.000***
L	0.618	13.273	0.000***	0.454	8.579	0.000***
	Reverberance			Metallic character		
A	0.700	16.595	0.000***	0.505	27.107	0.000***
H	0.805	2.637	0.057	0.527	15.748	0.000***
S	0.628	4.392	0.005**	0.588	12.983	0.000***
L	0.661	4.231	0.003**	0.631	3.676	0.008**
	Room size			Source width		
A	0.512	4.357	0.009**	0.433	2.241	0.108
H	0.584	5.576	0.006*	0.628	2.545	0.081
S	0.689	4.844	0.002*	0.649	1.541	0.204
L	0.601	2.202	0.077	0.456	9.812	0.000***

THE AUTHORS

Torben Wendt

Steven van de Par

Stephan D. Ewert

Torben Wendt was born in Germany in 1986. He studied physics at the Christian-Albrechts-Universität zu Kiel, Germany, and at the Universität Oldenburg, Germany, and received the Master's degree in 2013 from the Universität Oldenburg. During his Bachelor's thesis he did research in basic binaural psychoacoustics. In his master's thesis he started working in the field of room acoustical simulations, which he continues currently as a Ph.D. student.

●

Steven van de Par studied physics at the Eindhoven University of Technology, Eindhoven, The Netherlands, and received the Ph.D. degree in 1998 from the Eindhoven University of Technology, on a topic related to binaural hearing. As a Postdoctoral Researcher at the Eindhoven University of Technology, he studied auditory-visual interaction and was a Guest Researcher at the University of Connecticut Health Center. In early 2000, he joined Philips Research, Eindhoven, to do applied research in auditory and multisensory perception, low-bit-rate audio coding, and music information retrieval. Since April 2010 he holds a professor position in acoustics at the Universität Oldenburg, Germany, with a research focus on the fundamentals of auditory perception and its application to virtual acoustics, vehicle acoustics, and digital signal processing. He has published various papers on binaural auditory perception, auditory-visual synchrony perception, audio coding, and computational auditory scene analysis.

●

Stephan D. Ewert was born in Germany in 1972. He studied physics and received his Ph.D. degree in 2002 from the Universität Oldenburg, Oldenburg, Germany. During his Ph.D. project he spent a three-month stay as visiting scientist at the Research Lab of Electronics at the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. From 2003 to 2005 he was Assistant Professor at the Centre of Applied Hearing Research at the Technical University of Denmark (DTU), Lyngby, Denmark. He re-joined Medizinische Physik at the Universität Oldenburg in 2005 and there heads the Psychoacoustics and Auditory Modeling Group since 2008. Dr. Ewert started his interest in hearing and audio engineering by developing loudspeakers during his early undergraduate years. His field of expertise is psychoacoustics with a strong emphasis on perceptual models of hearing. Dr. Ewert has published various papers on amplitude modulation perception and processing, spectro-temporal processing, and binaural hearing. More recently, he also focused on perceptual consequences of hearing loss and hearing-aid algorithms.