Werner, Stephan; Klein, Florian; Neidhardt, Annika; Sloma, Ulrike;
Schneiderwind, Christian; Brandenburg, Karlheinz:

Creation of auditory augmented reality using a position-dynamic binaural synthesis system - technical components, psychoacoustic needs, and perceptual evaluation

Original published in:	Applied Sciences. - Basel : MDPI. - 11 (2021), 3, art. 1150, 20 pp.
Original published:	2021-01-27
ISSN:	2076-3417
DOI:	10.3390/app11031150
[*Visited:*	2021-06-02]

applied sciences

MDPI

Article

Creation of Auditory Augmented Reality Using a Position-Dynamic Binaural Synthesis System—Technical Components, Psychoacoustic Needs, and Perceptual Evaluation

Stephan Werner [1,†,*], Florian Klein [1,†], Annika Neidhardt [1,†], Ulrike Sloma [1,†], Christian Schneiderwind [1,†] and Karlheinz Brandenburg [1,2,†,*]

[1] Electronic Media Technology Group, Technische Universität Ilmenau, 98693 Ilmenau, Germany; florian.klein@tu-ilmenau.de (F.K.); annika.neidhardt@tu-ilmenau.de (A.N.); ulrike.sloma@tu-ilmenau.de (U.S.); christian.schneiderwind@tu-ilmenau.de (C.S.)
[2] Brandenburg Labs GmbH, 98693 Ilmenau, Germany
* Correspondence: stephan.werner@tu-ilmenau.de (S.W.); khb@brandenburg-labs.com (K.B.)
† These authors contributed equally to this work.

Featured Application: In Auditory Augmented Reality (AAR), the real room is enriched by virtual audio objects. Position-dynamic binaural synthesis is used to auralize the audio objects for moving listeners and to create a plausible experience of the mixed reality scenario.

Abstract: For a spatial audio reproduction in the context of augmented reality, a position-dynamic binaural synthesis system can be used to synthesize the ear signals for a moving listener. The goal is the fusion of the auditory perception of the virtual audio objects with the real listening environment. Such a system has several components, each of which help to enable a plausible auditory simulation. For each possible position of the listener in the room, a set of binaural room impulse responses (BRIRs) congruent with the expected auditory environment is required to avoid room divergence effects. Adequate and efficient approaches are methods to synthesize new BRIRs using very few measurements of the listening room. The required spatial resolution of the BRIR positions can be estimated by spatial auditory perception thresholds. Retrieving and processing the tracking data of the listener's head-pose and position as well as convolving BRIRs with an audio signal needs to be done in real-time. This contribution presents work done by the authors including several technical components of such a system in detail. It shows how the single components are affected by psychoacoustics. Furthermore, the paper also discusses the perceptive effect by means of listening tests demonstrating the appropriateness of the approaches.

Keywords: auditory augmented reality; position-dynamic binaural synthesis; quality evaluation; system development; spatial listening

Citation: Werner, S.; Klein, F.; Neidhardt, A.; Sloma, U.; Schneiderwind, C.; Brandenburg, K. Creation of Auditory Augmented Reality Using a Position-Dynamic Binaural Synthesis System—Technical Components, Psychoacoustic Needs, and Perceptual Evaluation. *Appl. Sci.* **2021**, *11*, 1150. https://doi.org/10.3390/app11031150

Academic Editor: Hyunkook Lee
Received: 21 December 2020
Accepted: 25 January 2021
Published: 27 January 2021

1. Introduction

Fully immersive reproduction of spatial audio in a way that both artificial and real audio objects are perceived as plausible audible events in a virtual and/or augmented environment is something researchers have tried for many years. Recently, this has been demonstrated under certain conditions [1]. One method which has been proposed to achieve an auditory illusion of a spatial acoustic environment is via the help of an existing position-dynamic binaural synthesis system [2]. Even then, the occurrence of a plausible auditory illusion depends on many parameters. Beyond an adequate technical realization there are several context dependent quality parameters like congruence between synthesized scene and the listening environment or individualization of the technical system.

The goal of our research and this paper is to find efficient solutions to merge real and virtual acoustics in a plausible manner. Our approach is to start with measurements in

real rooms and to stepwise simplify these measurements to identify the relevant cues and information which are mandatory to retain plausibility. The advantage of this approach is that we always can compare to the upper reference: The actual measurement in the real scenario. After identifying these cues, measurement efforts can be minimized and the computational efficiency of algorithms to create an auditory augmented reality (AAR) system can be improved. The primary question is to what extent simplifications in the technical realization of an audio-AR system are permissible without leading to an intolerable impairment of spatial auditory perception.

In the following, we describe a basic scheme of an AAR system built from several different functional blocks which realize a position-dynamic binaural synthesis. This includes:

1. provision of binaural room filters from measurements or simulations,
2. representation of the scene,
3. creation and/or adaptation of binaural room filters related to the pose and position of the listener, and
4. real-time rendering to make a position-dynamic auralization possible.

Figure 1 shows these blocks in a basic scheme. In each block, different techniques and approaches are listed. Bold entries are techniques which will be described in greater detail.

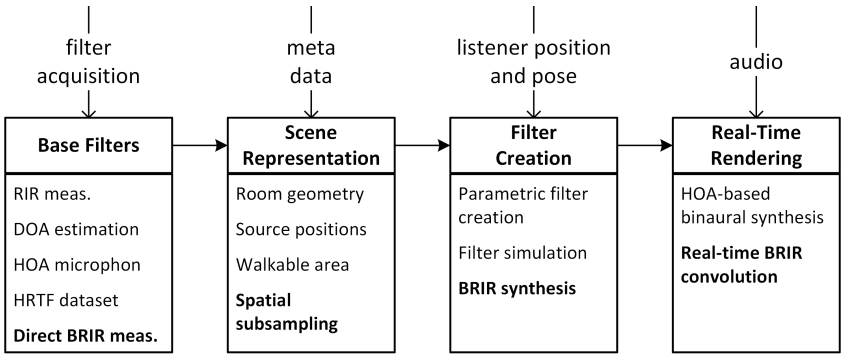

Figure 1. Basic block diagram of a position-dynamic binaural synthesis system showing different functional blocks with examples of realization. Bold marked realizations are related to the approaches presented.

The reproduction of an audio object in a reverberant room can be realized using BRIRs. The audio signal of the source is convolved with binaural room impulse responses (BRIRs) of the current source–receiver position and head orientation of the listener. For the acquisition of binaural room filters, several popular approaches exist. BRIRs can be simulated [3,4] or directly measured, for example, with a dummy head. To create a position and head-pose-dependent binaural synthesis, measurements for each dummy head orientation and position must be conducted. The separation of the head-related transfer functions (HRTFs) and room measurements reduces the effort significantly. One way to do this is to just measure the room impulse responses (RIRs) and to estimate or measure the direction of arrival (DOA) in a separate calculation or measurement. An estimation can be done by assuming a certain room geometry, such as a shoe-box model, and by assigning a direction to the measured reflections [5]. Information about the room can either be predefined or measured by other sensors (visual, radar, etc.). The DOA can also be measured with an appropriate microphone array. Once the DOA is available, a direction and an HRTF can be assigned to each reflection and a BRIR can be calculated. The Spatial Decomposition Method (SDM) is one approach in this domain [6,7]. An alternative way is to employ B-format or higher order ambisonic arrays to record transfer functions in the ambisonics signal representation. The spatial resolution directly depends on the number of microphones as well as on the array design. To create binaural signals the combination with a spherical HRTF data set is necessary as well [8].

The next step is the creation of a scene. Depending on the synthesis approach, the room geometry, source position, and the walkable area have to be defined. Most of the existing

systems and evaluations thereof are limited to simple room geometries like shoe-box rooms (for example, in [5,9]). Furthermore, the walkable areas are often placed centrally in the room, presumably to avoid special acoustic effects when being close to walls or corners. However, these restrictions are not necessarily system-related. As this research field is still emerging, more complex room acoustics simply have not been evaluated yet. Depending on the real-time capability of the succeeding filter creation step, a fine or coarse sampling of the walking area has to be considered. Sub-sampling the area based on psychoacoustic assumptions, for example, by considering just noticeable differences (JND) for direction and distance changes [10,11], can avoid the effort for a continuous adaptation of filter coefficients. In this paper, we will discuss the Maximum Allowed Error Method (MAEM) as one possibility to sample the walkable area [12].

When the listener changes his/her pose or position new filters need to be computed (or loaded). The most flexible solution would be a parametric filter creation approach. These are usually based on a measured or assumed model of a room or scene [5,13,14]. The room impulse response is decomposed into modifiable parameters which can be changed depending on the listener movement. This allows an efficient adaptation of filter coefficients, but the success relies on the quality of the model. Filter shaping approaches rely on a BRIR measurement on one position, and only certain properties of filters are adapted when the listener moves such as energy decay curve (EDC), level of direct sound, or initial time delay gap (ITDG) [15–17]. Often these changes are empirically determined, but they also can be estimated by simple models (such as inverse square law). These algorithms will be discussed in Section 2.2. The idea of these approaches are quite similar to ambisonics-based approaches. Auralization for different listener positions are realized by transformations based on simple models, e.g., distance-dependent attenuation or angular attenuation to mimic the directivity of sound sources [9,18].

The last step is the real-time rendering. A common solution for ambisonics-based approaches is the rendering of a virtual spherical loudspeaker setup [8,18]. The ambisonics audio signals are then converted to binaural signals. This is done by assigning an HRTF to each virtual loudspeaker. During the real-time rendering, only the gains of the virtual speakers are adjusted and HRTFs have to be applied. Ambisonics-based approaches are especially suited for realizing 6-DoF experience on the basis of recorded sound fields. Other approaches (especially the ones mentioned here) usually deliver binaural room impulse responses. These need to be convolved with the desired audio signals. The filters have to be changed each time the position or the pose of the listener changes.

As it has become clear, the plausible creation of virtual audio objects fused into a real room is the main challenge for AAR. To recognize audio objects, the human brain acts as a great pattern recognizer comparing learned and thus expected audio cues with the ones from the real surrounding. Additionally, presented audio objects have to fit these expectations [19,20]. A too large acoustic deviation between the acoustic properties of the real space and the virtually reproduced environment leads to a cognitive mismatch and thus to a collapse in the plausibility of the overall auditory scene. In the context of binaural synthesis, we call this the room divergence effect. The most prominent listening impression in this case is a collapse of the externalization of auditory events [21,22]. This is true for all auralization efforts using binaural technologies, but the effect is most prominent in an AAR environment. A comparison of the perceived sound events in real space with the virtual sound events is always possible in AR scenarios. The cognitive model of the environment created by experience is continuously updated and, in our brain, compared with the virtual auditory events [23]. As a corollary to this model of spatial hearing, judging such systems is conventionally done via listening tests. As, in most cases, no reference is available, listening test paradigms as known from audio coding to evaluate audio quality [24,25], etc. cannot be used.

This paper gives a summary on a specific AAR system which is used at the Technische Universität Ilmenau as a research demonstrator. The system is one possible technical realization to address some of the challenges in spatial listening. The main components of

the system are a psychoacoustically motivated spatial provision of the needed BRIRs in the walkable area, a synthesis of BRIRs using spatially sparse measurements of the room, and a real-time processing of the tracking data and filter convolution. The structure of the paper is as follows. Section 2 describes the system components used in our system in detail. Section 3 shows the perceptual evaluation of these components. Therefore, the names of the subsections are identical. Section 4 gives a summary and discussion of open questions.

2. Proposal of a Position-Dynamic Binaural Synthesis System

A basic feature of the presented system is the provision of synthesized BRIR data sets for discrete positions in the room. These discrete areas can either have a uniform distribution (e.g., grids with rectangular or triangular grid areas/cells) [26] or a nonuniform distribution of the single grid cells [12]. The sizes and shapes of the grids are motivated by psychoacoustic features like localization and distance blur.

A block diagram of the proposed audio system is shown in Figure 2. For each block, a number is given which is referenced in the text. On the right part of the figure, several blocks, which include input data to the blocks on the left side, are shown. The numbers in curly brackets show the connections.

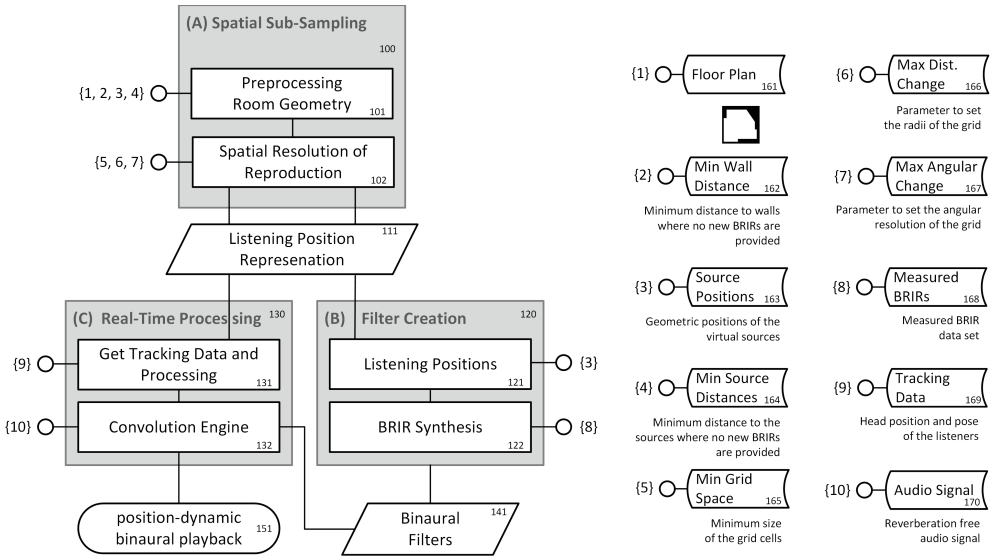

Figure 2. Block diagram of a specific position-dynamic binaural synthesis system including the provision of BRIRs in the room (100), filter creation (120), and real-time processing (130). The separated blocks on the right indicate input data to blocks on the left side. The descriptions in the text refer to the several blocks indicated by the number of each block.

2.1. Spatial Sub-Sampling

In the exemplary audio system, the different BRIR data sets, which are required for the listener's movement in the room, are created for predefined discrete areas in the room. These areas are called cells. The totality of all cells is called grid. Depending on the chosen spatial resolution, a BRIR data set is valid for one of these cells. A BRIR data set consists of BRIRs for the left and right ear for all available head orientations and for one audio object. Only in the center of this area, the BRIR data set corresponds to a correct mapping of distance and direction to the sound source. Towards the edges, the deviation of direction and distance cues and thus the reproduction error increases. If the spatial resolution is chosen to be low, perceptible errors will increase. The shapes of these cells can be uniform in the sense of a position-independent shape, for example, quadratic. However, the shapes of the cells can also be nonuniform in the sense of a position-dependent shape.

The use of a uniform grid is not very good at taking perceptive inaccuracies in localization [27] and distance perception [28,29] into account. Furthermore, it was shown that for a range of signal types a uniform positional BRIR grid requires a 5 cm resolution

or higher to provide a smooth transition without noticeable discontinuities [26]. If the listener is very close to the virtual audio object, a high spatial resolution must be selected to minimize perceptional errors. However, this high resolution is unnecessary at greater distances from the source which allows a reduction of the required number of BRIRs. The approach which is discussed here is called Maximum Allowed Error Method (MAEM) and was developed by Georg Götz and Samaneh Kamandi [30].

The MAEM (see block 100 in Figure 2) describes a method to parameterize the size and shape of an area in the listening room which is represented by one BRIR data set. The parameterization is motivated by perceptual thresholds in spatial hearing. Figure 3 shows the principles of localization blur and distance blur in the horizontal plane. If a sound source is perceived from a certain direction and distance, the acoustic position of this sound source can change within certain limits without changing the perception of direction and distance. The displacement is inaudible. In a similar way, other sound sources located in this range are perceived from the same direction and distance. The size of this range is determined by the just noticeable differences (JNDs) for direction and distance perception of the direct sound of the source. If a fixed localization blur or minimum audible angle (e.g., 5°) is assumed, the density of the cells (in the sense of the width of the cells) should increase at small distances to the source and decrease at larger distances. The situation is similar regarding distance blur. For small distances, the density (in the sense of the length of the cells) should be higher than for large distances. This approach leads to a grid of nonuniform cells which can effectively reduce the number of BRIRs required without causing increased errors in direction and distance perception. An extension of this approach could also consider the reflections to determine the JND and not only the direct sound.

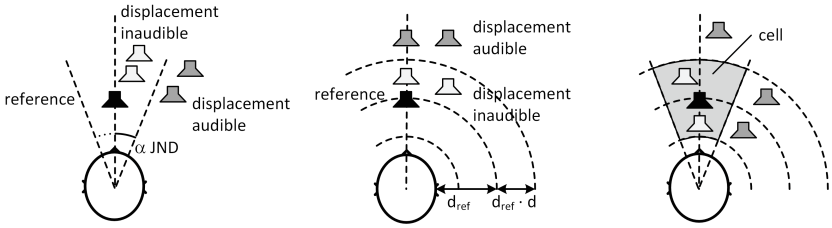

Figure 3. Localization blur (**left**), distance blur (**middle**), and assembled cell (**right**) in the horizontal plane. A displacement of the reference (black speaker symbol) by an angle or distance below the threshold is inaudible (light gray symbols), whereas a bigger displacement is audible (dark gray symbols). An area or cell is formed where no displacement is audible (after [30]).

The maps generated by the MAEM approach are based on a Voronoi diagram (Figure 4). First, a floor plan of the room can be loaded as 1 bit graphic (blocks 101 and 161 in Figure 2). In the immediate proximity of the walls, the calculation of the cells can be influenced by specifying a minimum wall distance (162). Acoustic influences of a very close wall cannot be represented, as in the process of filter creation (120) all BRIRs are extrapolated from only one measurement recorded in the middle of the room. The minimum wall distance therefore represents a distance from which no new cells are generated and thus no new BRIRs are made available. A minimum grid size (165) can be specified to avoid unnecessarily small cells, especially near the sources. The MAEM approach calculates the possible distances of the individual cells starting from the source position (163), using a minimum source distance (164), the maximum allowed distance change (166), and the maximum allowed angular change (167). Specifying a minimum source distance prevents too small and too many cells in the immediate proximity of the source. The maximum allowed distance change is the main parameter for the calculation of the map and it represents the distance blur described above and shown in Figure 3. The second main parameter is the maximum allowed angular change (167) in degrees which corresponds to the minimum audible angle (see Figure 3).

Figure 4 shows examples of nonuniform grids based on a Voronoi diagram. The main parameters for generating the grids are the maximum allowed angle error α and the

maximum allowed relative distance change d: Grid 1 $\alpha = 5°$, $d = 0.25$; Grid 2 $\alpha = 10°$, $d = 0.5$; Grid 3 $\alpha = 15°$, $d = 0.75$; and Grid 4 $\alpha = 20°$, $d = 1.0$. The distance parameter for Grid 1 is based on results from Spagnol et al. [10] where distance blurs of $d = 0.25$ were found. The angle parameters are estimates which have been collected from the literature [11,27] . It is assumed that the localization blur ranges from 1° to 10°.

The output of MAEM (100) is the provision of a list of all possible listening positions for each audio object (111). For the MAEM approach a listening position map (grid) for each object and the spatial positions of the individual cells are provided. The map is used to assign the listeners position in the room to the correct BRIR filter selection in the real-time processing block (130). The cell positions are needed for the filter creation (120).

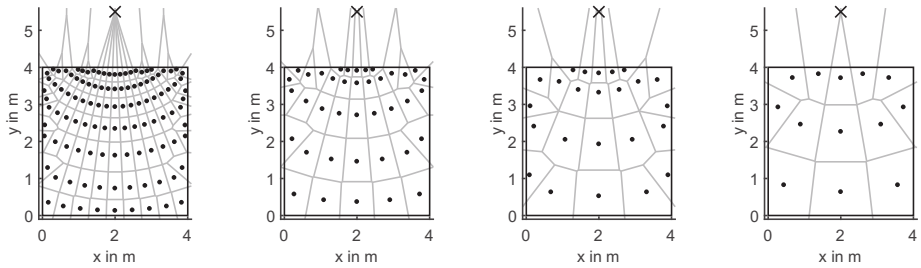

Figure 4. Examples of nonuniform grids for a quadratic room and frontal sound source (x); f.l.t.r.: Grid 1 $\alpha = 5°$, $d = 0.25$; Grid 2 $\alpha = 10°$, $d = 0.5$; Grid 3 $\alpha = 15°$; $d = 0.75$; and Grid 4 $\alpha = 20°$, $d = 1.0$.

2.2. BRIR Synthesis

A set of BRIRs must be provided for each possible listening position. If a dataset of BRIRs is measured, e.g., with a head-and-torso-simulator at one or few selected positions in a room, then BRIRs for further positions can be generated by interpolation and extrapolation. In our group we pursued two different approaches for that which are presented below. The first approach is based on a quite strong simplification and thus allows for an efficient implementation. The second approach manipulates more details with the goal to provide a better quality. Section 2.2.3 adds further functionality to both approaches in order to add sound source directivity.

2.2.1. Constant Reverberation

This first approach is based on the simple idea to keep the reverberation constant throughout the different positions. There have been several earlier studies, like that in [31], indicating that around 50 ms after the direct sound, at least in small rooms the reverberation can be kept constant for direction-dependent reproduction. It was shown that at least for an approaching motion towards a virtual sound source a simple adjustment of energy of the direct sound according the distance to the sound source was sufficient to achieve a plausible reproduction. Moreover, it was perceived as plausible as the original set of BRIRs measured along the walking line [17,32].

The basic idea to keep the reverberant part constant is not new, but due to its simplicity a very interesting one. However, it is likely that it has its limitations and will deliver perceptually satisfying results only under certain conditions. For this reason, we started to investigate this approach for the given application scenario of synthesizing BRIRs for an AAR-system with 6-DoF. So far, two studies have been conducted for the case of walking towards and away from a virtual loudspeaker in two different rooms with a similar size but quite different reverberation times (RT60 of 0.27 s and 1.0 s). In both cases, using the reverberation measured at one position of a considered walking line, keeping the reverberant part of the BRIRs constant over the different positions did not significantly reduce the plausibility. Two different cases of direct sound were taken into account. In one version, the direct sound originally measured at all the different positions was used. The other version was built on the measurements from one specific position in the room. BRIRs

for other positions were created by simply adjusting the level of the direct sound. Both realizations were perceived as plausible as the original fully measured set of BRIRs. More details about the experiments are provided in Section 3.2.1.

In both experiments, the translation line with a length of 2 m was located in front of the loudspeaker. Therefore, it remains an open question whether this very simple approach still creates convincing results for the cases of walking past and behind a virtual sound source. For these cases, the source directivity has to be considered when modeling the direct sound. Moreover, there is still a lack of knowledge about the perception of room acoustical details and their relative changes for the cases of low direct sound energy.

2.2.2. Acoustical Shaping

Acoustic shaping describes an approach to adapt a BRIR by changing single distant dependent acoustic parameters. The aim is to reach perceptual suitability of the synthesized BRIRs regarding spatial auditory perception. The basic idea is to change the acoustic structure of the early reflections of measured BRIRs based in a shift of the initial time delay gap (ITDG). The ITDG is the time between the incoming direct sound and the first reflection. The ITDG is a distance-dependent acoustic parameter. The ITDG is small for distant sources and bigger for closer ones. Figure 5 shows this distance dependency of the ITDG if the first reflection is a ground reflection.

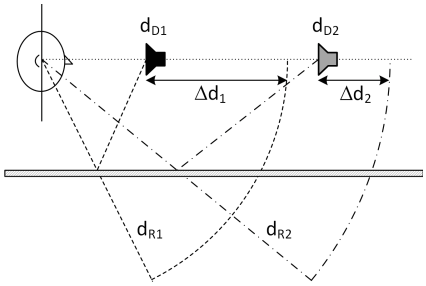

Figure 5. Principle of the distance dependency of the initial time delay gap (ITDG), with d_{Dx} ... distance of direct sound, d_{Rx} ... distance of reflection, Δd_x ... distance of the ITDG; (after [15]).

The principle of filter shaping (122) based on a manipulation of the ITDG is shown in Figure 6 as a block diagram. The presented approach is based on the work from Füg et al. [15], and it uses one measured data set of BRIRs at one position in the room. Depending on the new listening position and head orientation to be calculated, the BRIRs corresponding to the new yaw orientation are selected from the recorded data set. The BRIRs are split into a direct part, the early reflections, and the late reverberation. The transition point between early reflections and late reverberation is defined by the perceptual mixing time [31] of the auralized room. The mixing time defines a point in time of a BRIR after which its content is perceptually independent from the head pose or the position in the room. The beginning of the early reflections is defined by the choice of the time index which is $(0.1 \cdot ITDG)$ before the first reflection. The samples within the defined area of early reflections are now rearranged in time regarding the distance dependency of the ITDG. Therefore, the time of the first reflection is changed depending on the distance between source and listening position. In the presented case, the ITDG change originates from a measurement in a TV studio at our university (TU Ilmenau), but it can be adapted for a specific room or geometric arrangement. If a BRIR for a closer distance than the recording position is needed, the samples before the first reflections are stretched in time while the samples up to the mixing time are compressed in a linear manner. If a greater distance is desired, it behaves in the opposite sense. The last adaptation is a change of the energy of the shaped BRIR depending on the new distance following the inverse-square law for energy distributions.

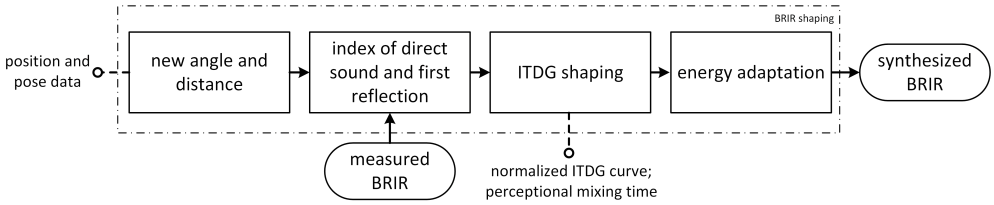

Figure 6. Block diagram of the synthesis approach adapting the initial time delay gap (ITDG) and using BRIRs of one measured position.

Figure 7 shows the ITDG-approach exemplarily on a BRIR shaping for a measured BRIR at approximately 4.2 m and a new BRIR at approximately 1.6 m. For clarity of the figure and approach, the traveling time between source and listening position is shown for the new BRIR and the measured BRIR is shifted to this time in the figure. In a real scenario, BRIRs would also be shifted along the time axis. In the case shown, the measured BRIR would have a longer traveling time than the closer, new BRIR. Furthermore, no distant-dependent energy adaptation is shown. The area where the BRIR is manipulated is indicated as a box in the figure. It covers the first reflections up to the mixing time (endpoint). Within this range the first reflection (usually the ground reflection) is detected and the ITDG is determined. According to the new distance all samples are mapped to their new times using a linear compression and expansion characteristic curve within the range of the early reflection. The transition point of the two curves is the time of the first reflection. The first curve shifts all samples before the first reflection, the second shifts all samples after the first reflection until the mixing time. In terms of Figure 7, the samples are compressed before the first reflection and expanded afterwards.

The yaw orientation is realized by selecting the corresponding direct sound of the measured or otherwise calculated BRIRs. An improvement can be realized by including an interpolation of the direct sound to synthesize a finer yaw resolution for example. The intensity of the direct sound and reverberation sound is adapted according to the new synthesized grid position.

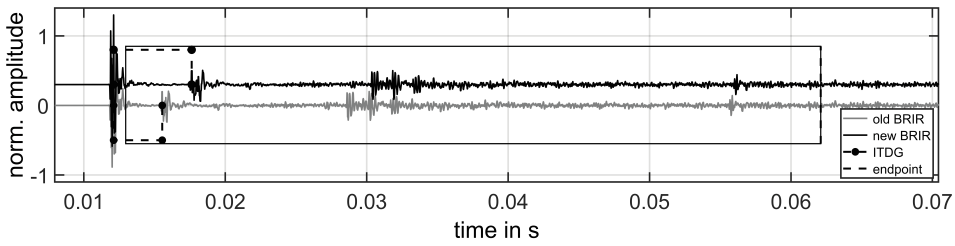

Figure 7. Shaping of a BRIR based on the initial time delay gap (ITDG) as an exemplary illustration. The box indicates the audio samples of the early reflections which are manipulated. Dashed lines indicate direct sound, first reflection, and endpoint of manipulation. A distant-dependent energy adaptation is not shown (figure after Füg et al. [15]).

The assumption of this ITDG-based approach is that the detected first reflection is suitable for a distance adjustment. This is the case if it corresponds to a ground reflection, for example. It is not the case if the detected reflection is a wall reflection directly behind the measured sound source on a line from source to measurement position. In this case, the determined ITDG is not distance-dependent. The challenge lies in correctly determining the first valid reflection. Furthermore, the adjustment of the BRIR causes a more or less strong change of the total time range of the early reflections and thus a change of the acoustic mapping of the room. In detail this means that, e.g., with the synthesis of a closer position the ITDG is correctly enlarged but also that the later early reflections are also shifted in time. In general, this is justified by the change of location in space. However, it does not correspond to the actual change of the reflection pattern. However, it is conjectured that the validity of this approach is due to the ability of preserving the overall relative structure

of the occurrence of the reflections. Timbre characteristics of the single early reflections are conserved which keeps the synthesis plausible (see Section 3.2.2).

2.2.3. Sound Source Directivity

For both presented approaches of creating BRIRs for additional listening positions, an adequate modeling of the direct sound with its directivity is essential, as it is also relevant for the progress of the DRR within a given listening area. The shaping of the ITDG described in Section 3.2.3 does not include a correct representation of the sound source directivity (SSD) pattern. The BRIRs of the measured position contain the directional characteristics of the sound source at that position. If these BRIRs are used to synthesize another position in the room, the directional characteristics remain unchanged. A correction is therefore desirable, to minimize the physical and the perceived differences between the measured and synthesized BRIRs. This is expected to result in a more plausible listening experience.

The shaping algorithm is therefore enhanced in a further implementation and study. An additional processing step to consider the SSD is shown in Figure 8. A predefined source-directivity pattern of the sound source used for the measurements is taken into account to vary the direct sound part of the BRIRs. In this setup, a Geithain MO-2 loudspeaker is used for the measurements. The frequency and angular-dependent changes in the amplitude for different orientations are adopted for the calculations of the extrapolated BRIRs. For that, the algorithm determines the angular relation between the position of the loudspeaker to the measurement position and to the synthesis position. The amplitude of the direct sound part of the BRIR at the position to be synthesized is then boosted or attenuated according to the change in amplitude in the pattern between these angles. Additionally, the inverse distance law is applied to take care of the different distances to the position of the loudspeaker. Further details are described in [33].

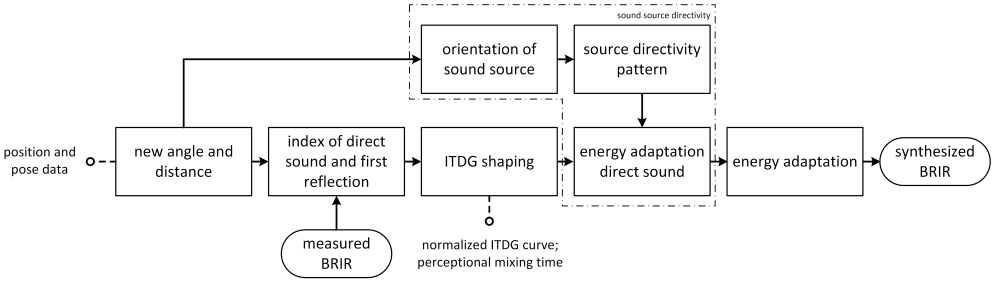

Figure 8. Block diagram of the synthesis chain to include directivity patterns in the ITDG-shaping approach.

The evaluation of physical parameters shows an improvement in the deviation of the direct to reverberation energy ratio (DRR) between a measured reference BRIR and the synthesized BRIR considering the SSD in comparison with BRIRs not considering the SSD. Figure 9 shows a heat map to display the deviations. In an investigation using the position "D1" for measurement for the most synthesized positions deviations smaller than 2 dB are reached. Especially for the positions off-axis to the sound source, an improvement is seen. However, the reduction of the DRR deviation is in a range from 0.5 dB to around 4 dB dependent on the difference between the synthesized and measured position. The deviations are mostly within the JND for the examined DRR range and therefore are not perceptible.

Even though good results for the DRR deviations are reached, there is still room for improvement. The perceived differences, shown in Section 3.2.3, are probably based on other physical parameters. Further research should, for example, investigate whether changes of the first reflections and the reverberant parts of the BRIRs due to SSD need to be considered for the calculations as well to address spatial perception and coloration effects in more detail.

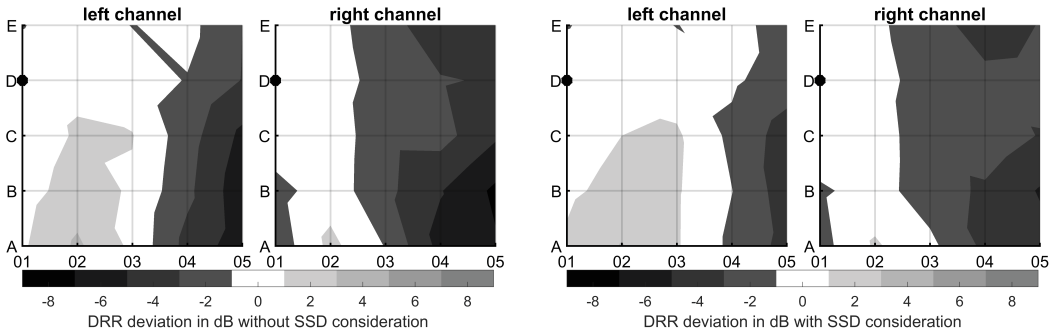

Figure 9. Deviation of broadband DRR between synthesized and measured BRIRs for incidence angle of 30°. Left the results for BRIRs without SSD consideration, and right with SSD consideration are shown. The Black dot represents the measured position for the synthesis "D1", active source "S1" is placed 1 m below of position "A2".

2.3. Real-Time Processing

Real-time convolution for binaural synthesis requires low delay in order to keep the overall system latency (SL) below perceptual thresholds. When the latency is above a certain threshold, static sound sources will no longer be stable when the listener moves his head. Therefore, high SL can be a cue for listeners to detect whether a sound source is real or virtual. For dynamic synthesis considering head rotation, Lindau [34] mentions a threshold for SL around 100 ms. The measured thresholds depend on the signal and the test paradigm which is used for testing.

For pose and position dynamic systems in AR, additional parameters become evident: Typically, visual cues are available as reference for the virtual sound sources which could lower the SL threshold depending on the positional and temporal precision of the visual cue in terms of a temporal and spatial Ventriloquism-effect [35]. However, the presence of a visual object may increase sensitivity to latency effects, as matching can always take place especially in an AR scenario. When the listener is able to change his position in addition to head movements, higher movement velocity and acceleration are expected in comparison to head movements only. No studies could be found on this subject.

In case of BRIR rendering, long filters have to be convolved with the source signal in real-time for several sound sources at the same time. The state-of-the-art solution for this use case is a blocked convolution (overlap-add or overlap-save) with uniformly partitioned filters. This solution is significantly faster than using non-partitioned filters, but the computational complexity increases linearly with increased filter size and decreased block size. In case of limited computational power, a compromise between filter length and block size (which directly relates to the delay induced by the convolution) has to be found. To overcome this limitation, filters can be partitioned nonuniformly (e.g., short segments for the direct sound and early reflections and long segments for the late reverb). The drawback of this solution is the implementation effort, because each sub-convolution needs to be scheduled correctly. This may require fine tuning to a specific hardware configuration [36].

Another way to reduce the computational load is to make use of the perceptual mixing time [31]. Depending on the room volume, this value can range between 30 to 100 ms [31]. It has to be noted that these values were only evaluated for a change of the head orientation but not position. This means that only parts of the filter have to be exchanged when the listener moves around or changes his head pose. Even though this reduces the amount of data which needs to exchanged, the load for the convolution itself remains the same. Meesawat and Hammershøi [37] conducted a small study considering different source and receiver positions in the room. For this specific room and these positions, they mention a time frame of 40 to 60 ms after which the BRIRs could be exchanged without perceptual consequences. However, the listening positions were always located in a close distance in front of the (virtual) loudspeaker. In 6-DOF scenarios, listeners can also walk to positions

with very low direct sound energy. In such cases, listeners may be more sensitive to small changes in the room acoustics.

When BRIRs are not precomputed, additional processing power is needed to synthesize or simulate them on-the-fly. While real-time capability depends on the specific algorithm in the first place, some techniques to save processing power can be applied to any algorithm. One of these techniques could be the aforementioned Maximum Allowed Error Method. Filters for a new position only need to be generated when a perceptual difference is expected. Another technique could be the prediction of listener movement. As we cannot change our position arbitrarily fast, some movements can be predicted and thus corresponding filters can be computed dynamically. This might increase the overall number of calculations, but it would help to balance the processing load. When the listener moves and the prediction was successful, a new filter only needs to be loaded instead of being computed rapidly. As a result, BRIR computation can run at a constant pace without strong load peaks.

For the systems and experiments described in this article the partitioned convolution and filter management was realized with the open source Python tool *pybinsim* [38]. It is based on uniformly partitioned convolution with the overlap-save approach. In the different setups, the block size was either 256 or 512 samples at a sampling rate of 48 kHz

3. Perceptual Evaluation

This section addresses the evaluation of the approaches and methods described in Section 2. The focus is on the quality assessment of individual quality features as well as of the overall impression of position-dynamic binaural synthesis.

3.1. Spatial Sub-Sampling

The main aim of the evaluation was to investigate the occurrence and perception of unanticipated events while walking. By this we assumed to find errors caused by different spatial resolutions of the provided grids. One sound source position outside the walkable area was synthesized playing a music or speech audio signal. The creation of the BRIRs for each grid cell was based on the ITDG-shaping approach described in Section 3.2.3. A BRIR data set from the mid position of the walkable area with a yaw resolution of 5 degrees was applied. A KEMAR dummy head was used for the measurement. The perceptual evaluation was performed for different quality features and for four different grids with different spatial resolutions as shown previously in Figure 4. In addition to single quality features, the evaluation also included the rating of the overall impression. The individual features included localization, externalization, and timbre perception. The ratings of 21 test persons with a mean age of 29.1 years (standard deviation 8.2) have been used for the evaluation. Fifteen of 22 participants (five women and 17 men) were experienced in listening tests and ten persons had special experience with binaural synthesis systems.

The task for the test persons was to walk twice along a given path through the accessible area. The mean walking speed was at 0.4 m/s, with a length of the whole path of 48 meters. After passing the path, the test persons had to evaluate the quality features externalization, ability to localize the auditory event (LA), the stability of the position of the reproduced audio object (LS), the coloration during movement (CO), and the overall impression (OI). Quasi-continuous rating scales have been used for OI, LA, and LS with the extremes “poor” as minimum and “very good” as maximum. For CO the endpoints are “strong coloration” and “no coloration” . A discrete three-point scale with “in-head”, “outside but close to the head” , and “outside the head” was used for externalization.

Figure 10 shows the ratings for the externalization as index. The index is the ratio of the number of ratings for “outside the head” and the overall number of ratings. When using the grid with the highest resolution (Grid 1), high externalization indices close to 1 are visible. The indices decrease continuously but slightly to a value of around 0.75 when less high-resolution grids are used. Figure 11 shows the ratings for the individual quality features and for the overall impression on the music signal. The best ratings can be found

for Grid 1. There is a tendency for lower valuations to be given for the less resolved grids. The interquartile distances tend to rise slightly for Grids 3 and 4. This indicates that the test persons are less in agreement if grids with lower resolutions are used. Only the results for the music signal are shown here, as this was evaluated as more critical in comparison to the speech signal. The ratings for the speech signal show a very similar trend, with slightly better ratings across all quality features.

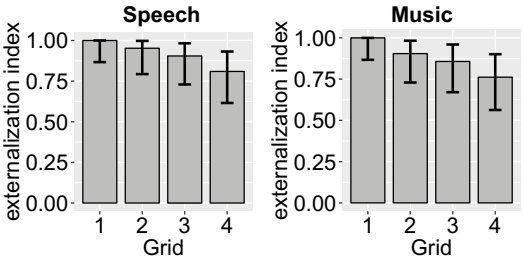

Figure 10. Externalization as index with 95% conf. interval; **left**: speech, **right**: music.

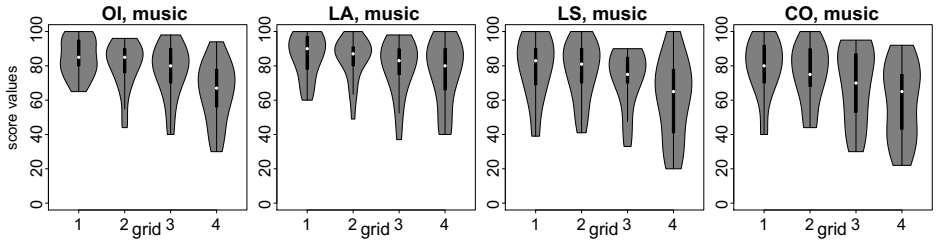

Figure 11. Ratings of quality features for the different grids and for the music signal as violin plots including box-plots; OI. . . Overall Impression with 0 = poor and 100 = very good, LA. . . Localization Ability with 0 = poor and 100 = very good, LS. . . Localization Stability with 0 = poor and 100 = very good, CO. . . Coloration with 0 = strong coloration and 100 = no coloration.

Overall, it must be noted that relatively high quality ratings have been given for both the overall impression and the individual quality features with respect to the reduced spatial resolution of the grids. However, it must be made clear that perceived quality also depends on the type of application and test paradigm. For example, a test scenario with a direct comparison of real and virtual sources with the intention of testing for authenticity would most likely lead to a much more critical evaluation [1].

3.2. BRIR Synthesis

Different approaches to synthesize BRIRs have been presented in the previous sections. The methods based on constant reverberation, acoustical shaping, and the integration of the sound source directivity are discussed below.

3.2.1. Constant Reverberation

The idea to keep the whole reverberation part of the BRIRs constant when the listener changes the position is a very simple approach which is very efficient with regard to the required calculation power and memory. The physics actually occurring for listeners walking through a sound field are roughly approximated. This is likely to cause audible drawbacks and inaccuracies. In contrast, it is also known that listeners are not sensitive to all the small physical details in such scenarios. Moreover, a perceptually indistinguishable replication of the real sound field is not required in all applications. For this reason, it is necessary to define the desired perceptual quality before designing a reproduction system with an adequate, efficient synthesis method. In addition, a suitable test method for a perceptual evaluation of the proposed approach for simplified binaural rendering is required.

The minimum demand in terms of perceptual quality is the creation of a plausible auditory illusion. Plausibility refers to the agreement of the auditory impression with

an internal reference [39]. Lindau and Weinzierl [40] proposed a method to evaluate the plausibility of auditory illusion created with binaural synthesis by randomly providing the binaural version or its corresponding real version and asking the participants in a Yes/No test paradigm whether they are listening to the simulation or not. However, for practical reasons this method requires taking the occlusion effect of the headphones into account. This can affect the quality of the auditory image of the real sound source as well [41].

The occlusion effect of the headphones could also dilute the perceptual differences between different approaches to simplify a BRIR data set for a perceptually optimized reproduction. For this reason, a different method to evaluate plausibility was needed. One idea was to ask for it directly. Especially inexperienced subjects often indicate that they are not sure about their own criteria for plausibility. For this reason, further questions were added to evaluate the quality of the auditory illusion [17,32]. Besides plausibility the participants had to rate externalization, continuity, sound source stability, and impression of walking towards (and away from) a sound source in a multi-attribute absolute category rating. Direct comparison between the different test cases was not possible within this experiment.

In the experiment, Neidhardt et al. [17] realized an interactive approaching motion towards a virtual loudspeaker and away from it with a BRIR data set measured with a positional resolution of 25 cm. Additionally, systematic simplifications were applied to the originally measured BRIR data set. One of the approaches was to replace the reverberation parts of all BRIRs in the data set by the reverberation tail of BRIRs from one specific position on the line. Thus, the naturally occurring changes in the early reflection pattern between the different listening positions were removed in this reproduction. Another approach, further simplifying the synthesis, used the BRIRs from one position and kept their reverberation constant over the various listening positions. Only the energy of the direct sound was adjusted according to the energy progress of the direct sound in the measured data set. Consequently, the changes of the ITDG and of the early reflection patterns between the different listening positions were removed in this realization. Still, the listeners perceived this scene as plausible as the originally measured data set. As expected, there was generally a strong correlation between the ratings for the impression of walking towards a sound source and plausibility. Furthermore, all other attributes correlated more with plausibility than with any other attribute. The simplified scenes with constant reverberation and also with the very simple modeling of the level of the direct sound were perceived as plausible as the scene based on the originally measured BRIR data set. The ratings for the plausibility of the created auditory illusion were not affected by this simplification. The same was observed for the other attributes.

This first experiment was conducted in a quite dry listening laboratory with a reverberation time $RT60 = 0.27$ s. To verify this observation, a similar experiment was conducted in a more reverberant room, a seminar room with a reverberation time $RT60 = 1.0$ s. Again, the realization based on a constant reverberation and an adjustment of the direct sound level was rated with the same plausibility as the realization with the BRIR data set originally measured in this room [32]. The simple reproduction was realized in two different versions based on the BRIRs of two different positions, 1.25 m and 3.25 m, from the sound source. For both versions, the ratings for plausibility did not vary significantly from the original measured BRIR set.

So far, the constant-reverberation approach has only been evaluated for the case of walking towards and away from the sound source in a close distance in front of a loudspeaker. The case of walking past a sound source requires the consideration of the source directivity, when modeling the direct sound. This will not only affect the progress of its total energy, but also its spectral content because directivity is typically frequency-dependent. One approach could be the one proposed by Sloma et al. [33], which is discussed in Section 2.2.3. Additionally, it remains open, whether the sensitivity to changes or a lack of changes in the early reflection pattern increases if the listener walks past and behind a sound source, where the direct sound energy is low.

3.2.2. Acoustical Shaping

A listening test has been conducted to evaluate the ITDG-based BRIR synthesis approach described in Section 3.2.3. The ability of the technical system to create a plausible auditory augmented reality is indicated by evaluating single quality features. Figure 12 shows the rating scales and names of the single features.

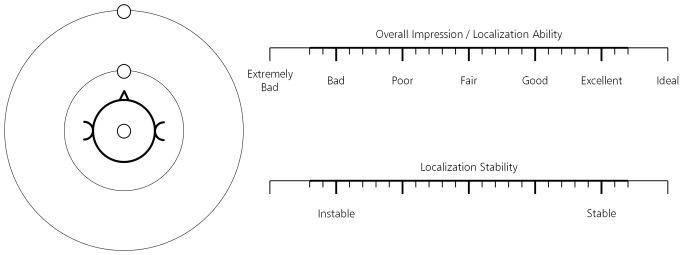

Figure 12. Rating scales for externalization, overall impression, localization ability, and stability. Scale design on the right from Jekosch-Bodden named in [42]. The designation of the upper scale with two quality features is only for this figure. In the evaluation, each feature was evaluated with its own scale.

The feature externalization describes the perception of an auditory event outside the head of a listener. This feature is assumed as critical if a spatial fusion of the virtual audio object and a real (visual) object in the room is aspired. Furthermore, this feature is strongly affected by the so-called room divergence effect which describes a decrease of externalization if the synthesized audio object does not fit to the expected auditory environment [22]. The localization ability is defined as the ability of the test participants to perceive a direction of the auditory event. The localization stability describes if the event is stable at a fixed position in the real room during the movement of the participant. An effect on stability is expected by the relatively low grid and yaw resolution. The individual overall impression can be rated as a further feature.

A discrete three-point scale is used for externalization with the naming "in-head", "outside but close to the head", and "outside the head". An extended continuous scale from Jekosch and Bodden is used for the other features (scale details in [42]). This scale type takes into account the fact that test participants tend to avoid using extremes on rating scales. The scale therefore provides extended scale categories at the lower and upper ends.

The test persons have to move on predefined paths through the auralized grid and rate each of the four paths (see Figure 13). The paths are selected to cover close/far distances to the virtual audio object, rapid/slow change of the relative angle between head and source position, front/back head orientations to the audio object, and most critical switching of the filter functions for the different grid positions.

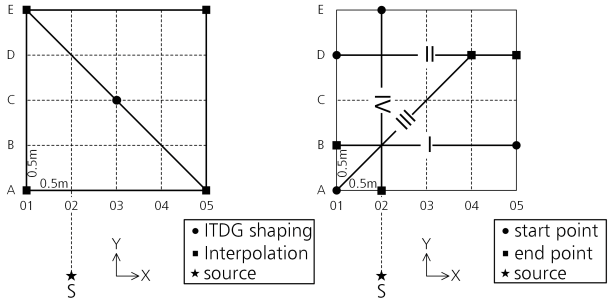

Figure 13. Auralized setup used for quality evaluation; **Left**: marked are the measurement points of BRIRs used for ITDG shaping and for the interpolation approach. **Right**: walkable paths (I to IV) used in the evaluation (data from the work in [43]).

An empty office room with dimensions 4.8 m × 4.7 m × 3 m and a reverberation time of ~0.5 s has been used for the recordings and as listening room for the test on system viability [43]. A HTC VIVE-based tracking using the HTC VIVE tracker [44] has been used for position and pose estimation. An extra-aural headphone has been used for playback [45]. The transfer characteristic of the positioned headphone was equalized using a KEMAR dummy head (same one as for the BRIRs measurement). Figure 13 shows the setup used in the quality evaluation of the BRIRs synthesis. The left part of the figure shows the measurement points of BRIRs used in the ITDG shaping. Furthermore, an additional approach was investigated, which creates new BRIRs from a weighted interpolation of three measuring points. In this approach, the BRIRs are shifted in their traveling time to the new position. The time range of the early reflections is separated from all three BRIRs and is interpolated in time domain. The weighting of the interpolation results from the distance of the new position to the measured positions. The direct sound component as well as the late reverberation for the new position is taken from the nearest measured BRIR. Finally, an energy adjustment of the direct sound component, the new interpolated early reflections, and the late reverberation takes place. This approach is mentioned here only briefly, and it is referred to in the literature [2,43]. In the experiment, BRIR data sets have been provided for a quadratic uniform grid with an edge length of 0.5 m. A reference system was established for which BRIR measurements at all intersections of the grid have been performed. The yaw resolution was 5° for all systems under test. The right part of the figure shows the walkable paths for evaluation.

Eighteen test persons with a mean age of 29 years (SD = 9) participated in the test. The ratings for externalization are shown as indices in Figure 14 for the different systems under test and investigated paths. The index is the ratio of the ratings for a perception of an auditory event outside the head (outer circle in Figure 12) and the overall number of ratings. No differences in terms of a significance level with $p < 0.05$ between the used BRIR synthesis systems or evaluation paths have been observed. However, there is a trend towards higher indices for path II and smaller indices for path IV. Path II includes the largest distance from listening positions to the source position as well as the presentation of the sound source from a lateral direction. Effects due to the relatively low spatial resolution of the grid are less significant due to the high distance. Furthermore, the perceived externalization of lateral sound sources is higher compared to frontal presentation. Path IV, on the other hand, includes close distances to the source and a predominantly direct frontal or rear source position.

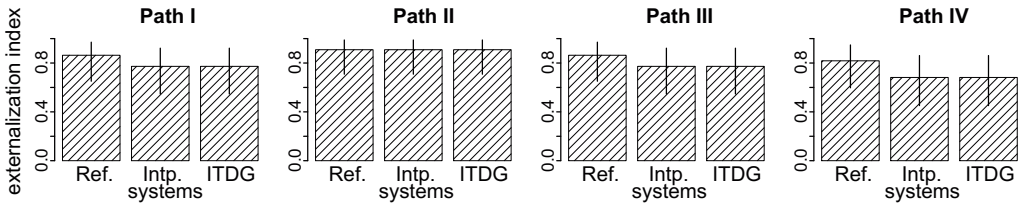

Figure 14. Externalization indices for the systems and paths with 95% binomial confidence intervals; Ref.: reference, ITDG: initial time delay gap, Intp.: interpolation (data from the work in [43]).

The overall impression is rated in the range of "good" (scale points > 4) for the systems and the paths (see Figure 15). Bigger differences are observed for paths I and III. These paths are most critical in terms of jumps in the perceived position of the auditory event caused by the low grid resolution at close distances to the audio object position. The synthesis approach using interpolation between BRIRs from three measurement positions was evaluated compared to the ITDG approach and the reference. The interpolation of the early reflections causes a change of the early reflection pattern compared to measured BRIRs. This is especially reflected in changes in the timbre of the early reflections. Other approaches which consider a maintenance of the reflections give reason to expect better quality ratings.

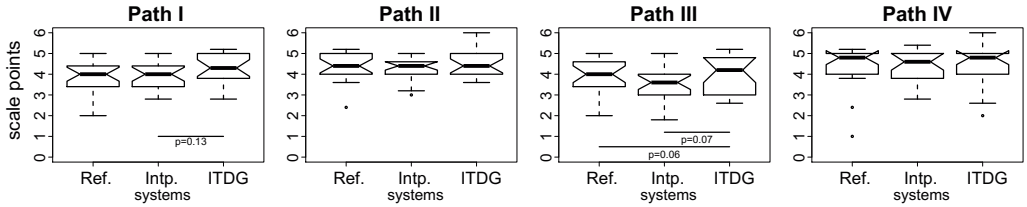

Figure 15. Boxplots for Overall Impression (OI) rated on the scale shown in Figure 12; Differences between distributions with $p < 0.2$ using a Wilcox-test are indicated; scale labels: 0 = extremely bad, 1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent, 6 = ideal (data from the work in [43]).

The stability of the perceived position of the auditory event is shown in Figure 16. Differences are visible between the paths and some systems. Paths I and III show lower grades compared to the other paths. These paths are assumed as most critical because of the close distance to the source and the low grid resolution. Especially at small distances and when moving diagonally through the grid, high errors in the perception of distance and angle changes are to be expected here. No statistical differences ($p < 0.05$) between the synthesis systems or evaluation paths are observed. Only slightly bigger interquartile distances for the interpolation approach compared to the ITDG approach are visible. The overall good localization ability at a scale value of "4" is the result of the usage of the unchanged (except of an intensity adaptation) direct sound of the recorded BRIRs.

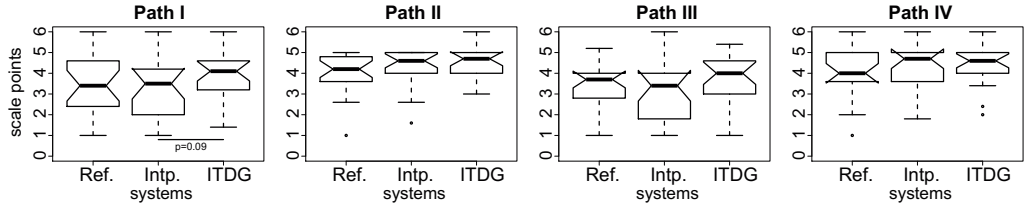

Figure 16. Box-plots for Localization Stability (LS) rated on the scale shown in Figure 12; Differences between distributions with $p < 0.2$ using a Wilcox-test are indicated; scale labels: 1 = instable, 5 = stable (data from the work in [43]).

3.2.3. Sound Source Directivity

The influence of adding a directional characteristic to the sound source was investigated in a listening test. Thereby, a data set of measured BRIRs was adjusted to correspond to a different position in the room. The BRIR synthesis from Section with the enhancement described in Section 2.2.3 was used for this purpose. The measured position corresponds to position "D1" in Figure 13. The loudspeaker used for the measurements was a Geithain MO-2. The directivity characteristic of the loudspeaker was simulated with the software VituixCAD [46] using data from the technical specifications of the loudspeaker. Of course, this is just a coarse approximation of the real speaker.

Eighteen listeners (14 male and four female) participated in a comparison test to rate the overall difference on a 6-point scale. The new synthesized positions "A2", "C2", "C5", and "D2" with and without directivity correction had to be compared with a reference, which was the real BRIR measurement at this position. The listening test was conducted without tracking of the head orientation, and therefore the participants had to stand on a specified position and were requested to listen without rotation of the head. A music and speech sample were used as audio content to auralize the measured and synthesized BRIRs for evaluation purposes. An analysis of the ratings using a Wilcoxon test showed no difference ($p < 0.05$) between the audio content, which were therefore evaluated together. Figure 17 displays the results for an incidence angle of 30° and the positions "A2", "C5", and "D2". A value of "0" indicates that there was no perceptual difference. The higher the value, the higher was the difference between the auralizations of the synthesized and the measured BRIRs. The ratings of the test participants show a clear decrease of the perceived overall difference when the SSD is included (D-1M) compared to the results without the

inclusion of the SSD (1M). The stimuli without directivity correction reach a median of 3.5 to 2 scale points compared to a median of 1 to 0 for the stimuli with directivity.

The evaluation showed that the integration of the SSD results in an auralization, which differs to a lesser degree from the original BRIR and therefore provides an enhanced listening experience. A more detailed description of the experiments and their results is discussed in [33].

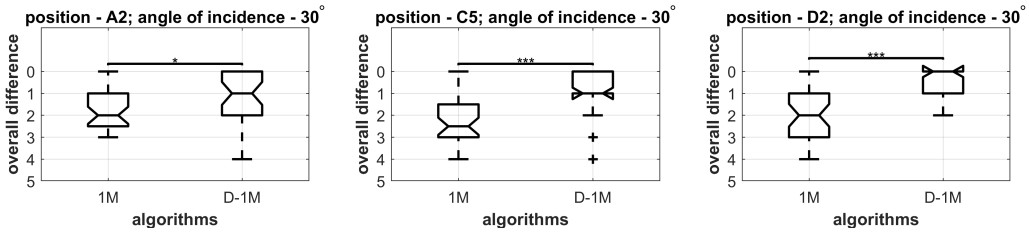

Figure 17. Perceived overall difference between the synthesized BRIRs and the measured BRIRs for two synthesis methods, without (1M) and with (D-1M) inclusion of the SSD, and for the synthesized positions "A2", "C5", and "D2" for an incidence angle of 30°. A value of "0" indicates no perceptual difference and "5" a very high difference.—*** $p \leq 0.001$, * $p \leq 0.05$ (Wilcoxon signed rank test).

4. Discussion

This paper gives an insight into an audio system for the creation of an Auditory Augmented Reality (AAR) environment. Single system components were described, which are motivated by human spatial hearing. The objective is to minimize the technical effort while maintaining sufficiently good spatial auditory perception in an AR scenario. Specifically, methods for BRIR synthesis were presented that create new listening positions in a room based on very few measurements. The evaluations of these methods show that the spatial audio quality remains comparable to reference measurements and allows a high plausibility of the generated listening environment for a moving listener. Furthermore, a method was presented that optimizes the spatial deployment of BRIRs in space based on spatial auditory perception. This makes it possible to significantly reduce the number of BRIRs required without causing disruptive effects on auditory perception.

While the techniques described in this paper show significant progress towards the goal of a plausible synthesis of auditory events in an augmented reality, another important goal of such developments is not yet reached: "What is necessary to enable AAR systems, e.g., for consumer devices in a way that the illusion is perfect for every user?" While the authors are not aware of any technical systems which fulfill this goal, there are known limitations in our work, too.

The methods presented here interfere strongly with the structure of the BRIRs. Nevertheless, the evaluation shows that the perceived spatial audio quality is only moderately or partly not affected. It seems that essential characteristics and features are preserved even in the modified BRIRs. Further research in this area to determine relevant acoustic features and feature combinations which are used to build up a cognitive auditory model of our environment seems essential.

Our proposed algorithms can never extrapolate the first reflections correctly, but the negative impact seems to be weak in our listening tests. This observation questions the importance of first reflections in perception of the room. First reflections can have an influence on the perceived direction, timbre perception and coloration, apparent source width, and others [47]. A study by Brinkmann et al. claims that rendering the first six reflections is sufficient to minimize the overall difference compared to an auralization containing all reflections [48]. However, these study data are based on image source models and shoe-box-shaped rooms which may differ from the acoustics in real rooms. The perceptual metrics in our listening tests suggest that the effect of acoustically authentic reproduction of the first reflection on externalization is small. This could be an explanation why our approaches are good regarding this quality feature. However, it remains open to what extent these results can be generalized. However, it is a strong conjecture that

essential patterns of the auditory space are preserved in the adapted BRIRs to ensure a high match with the internal representation of the room in the auditory system (cognitive model of the room). The endeavor to answer these questions are of interest for our further research to cover also larger areas to be synthesized, rooms with complex geometries, and for all types of audio signals. Observations may also differ with the tasks set to the listener and the complexity of the listening scenarios.

The presented work still leaves other open questions:

- What is the influence of room modes regarding the measurement positions (avoided in the current measurements)?
- How do the two proposed methods compare to each other and to other approaches? We have substantial data on our approach, but a comparison to other approaches lacks standardized test methods. A measurement based approach like the one described in this paper has always the advantage of a basic match of the filter with the actual room characteristics, but is this really necessary?
- Can we propose listening tests which measure perceptual thresholds for this type of auralization? Could this be done using MAEM?

There is definitely much more research to be done in this field. Current results are promising enough to hope that at some point we will get a simple and highly plausible reproduction of sound in a room via headphones.

Author Contributions: Conceptualization by S.W. and F.K.; Methodology, investigations, and software in Section 2.1 S.W. and F.K.; Methodology, investigations, and software in Section 2.2.1 A.N. and C.S.; Methodology, investigations, and software in Section 2.2.3 U.S.; Methodology, investigations, and software in Section 2.3 F.K. and A.N.; Evaluation in Section 3.1 S.W. and F.K.; Evaluation in Section 3.2.1 A.N. and C.S.; Evaluation in Section 3.2.2 S.W. and F.K.; Evaluation in Section 3.2.3 U.S.; Formal analysis from all authors regarding the sections named before; Writing, review, and editing all authors; Supervision by K.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research has partially been funded by CYTEMEX project funded by the Free State of Thuringia, Germany (FKZ: 2018-FGI-0019); Project BR 1333/18-1 funded by German Research Foundation (DFG), CharakU-Project funded by the Free State of Thuringia, Germany (FKZ: 5575/10-16); Research Group founded by Free State of Thuringia, Germany (2015-FGR-0090); Project BR 1333/13 funded by German Research Foundation (DFG).

Institutional Review Board Statement: The studies presented here are scientific studies with human subjects that, according to common research ethics standards of the relevant scientific societies as well as survey research, does not pose any particular risks (so-called "no risk study"), i.e., participation in the study "does not produce harm or discomfort beyond everyday experience", as formulated in the ethics guidelines of the German Psychological Society (DGPs, 2004). According to the regulations for scientific work of the TU Ilmenau, there was no obligation to obtain a vote of an ethics committee at the time of the studies.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brinkmann, F.; Lindau, A.; Weinzierl, S. On the authenticity of individual dynamic binaural synthesis. *J. Acoust. Soc. Am.* **2017**, *142*, 1784–1795. [CrossRef] [PubMed]
2. Brandenburg, K.; Cano, E.; Klein, F.; Köllmer, T.; Lukashevich, H.; Neidhardt, A.; Sloma, U.; Werner, S. Plausible Augmentation of Auditory Scenes using Dynamic Binaural Synthesis for Personalized Auditory Realities. In Proceedings of the Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality, Redmond, WA, USA, 20–22 August 2018.
3. Savioja, L.; Huopaniemi, J.; Lokki, T.; Väänänen, R. Creating Interactive Virtual Acoustic Environment. *J. Audio Eng. Soc.* **1999**, *47*, 675–705.
4. Zimmermann, A.; Lorenz, A. LISTEN: A user-adaptive audio-augmented museum guide. *User Model. User-Adapt. Interact.* **2008**, *18*, 389–416. [CrossRef]

5. Pörschmann, C.; Stade, P.; Arend, J. Binaural auralization of proposed room modifications based on measured omnidirectional room impulse responses. In Proceedings of the Meetings on Acoustics, New Orleans, LO, USA, 4–8 December 2017; Volume 30, p. 015012. [CrossRef]
6. Tervo, S.; Pätynen, J.; Kuusinen, A.; Lokki, T. Spatial Decomposition Method for Room Impulse Responses. *J. Audio Eng. Soc.* **2013**, *61*, 16–27.
7. Garí, S.V.A.; Brimijoin, W.O.; Hassager, H.G.; Robinson, P.W. Flexible binaural resynthesis of room impulse responses for augmented reality research. In Proceedings of the EAA Spatial Audio Signal Processing Symposium, Paris, France, 6–7 September 2019; pp. 161–166. [CrossRef]
8. Zotter, F.; Frank, M. *Ambisonics—A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, 1st ed.; Springer: Heidelberg, Germany, 2019.
9. Plinge, A.; Schlecht, S.; Thiergart, O.; Robotham, T.; Rummukainen, O.; Habets, E. Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information. In Proceedings of the Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality, Redmond, WA, USA, 20–22 August 2018.
10. Spagnol, S.; Hoffmann, R.; Kristjansson, A.; Avanzini, F. Effects of stimulus order on auditory distance discrimination of virtual nearby sound sources. *J. Acoust. Soc. Am.* **2017**, *4*, 375–380. [CrossRef] [PubMed]
11. Blauert, J. *Spatial Hearing: The Psychophysics of Human Sound Localization*; MIT Press: Cambridge, MA, USA, 1997.
12. Werner, S.; Klein, F.; Götz, G. Investigation on spatial auditory perception using non-uniform spatial distribution of binaural room impulse responses. In Proceedings of the 5th International Conference on Spatial Audio (ICSA), Ilmenau, Germany, 26–28 September 2019.
13. Stade, P.; Arend, J.; Pörschmann, C. A parametric model for the synthesis of binaural room impulse responses. In Proceedings of the Meetings Meeting of Acoustical Society of America and 8th Forum Acusticum, Boston, MA, USA, 25–29 June 2017; Volume 30, p. 015006. [CrossRef]
14. Pörschmann, C.; Stade, P.; Arend, J. Binauralization of omnidirectional room impulse responses-algorithm and technical evaluation. In Proceedings of the 20th International Conference on Digital Audio Effects (DAFx), Edinburgh, UK, 5–9 September 2017; pp. 345–352.
15. Füg, S.; Werner, S.; Brandenburg, K. Controlled Auditory Distance Perception using Binaural Headphone Reproduction—Algorithms and Evaluation. In Proceedings of the VDT International Convention 27th Tonmeistertagung, At Cologne, Germany, 22–25 November 2012.
16. Mittag, C.; Werner, S.; Klein, F. Development and Evaluation of Methods for the Synthesis of Binaural Room Impulse Responses based on Spatially Sparse Measurements in Real Rooms. In Proceedings of the 43rd Annual Meeting on Acoustics, DAGA, Kiel, Germany, 6–9 March 2017.
17. Neidhardt, A.; Tommy, A.I.; Pereppadan, A.D. Plausibility of an interactive approaching motion towards a virtual sound source. In Proceedings of the 144th International AES Convention, Milan, Italy, 23–26 May 2018.
18. Zotter, F.; Frank, M.; Schörkhuber, C.; Höldrich, R. Signal-independent approach to variable-perspective (6DoF) audio rendering from simultaneous surround recordings taken at multiple perspectives. In Proceedings of the 46th Annual Meeting on Acoustics, DAGA, Hannover, Germany, 16–19 March 2020.
19. Jekosch, U. *Voice and Speech Quality Perception—Assessment and Evaluation*; Springer Series in Signals and Communications Technology; Springer: Berlin/Heidelberg, Germay, 2005.
20. Raake, A. *Speech Quality of VoIP—Assessment and Prediction*; John Wiley and Sons: Chichester, West Sussex, UK, 2006.
21. Plenge, G. Über das Problem der Im-Kopf-Lokalisation. *Acustica* **1972**, *26*, 241–252.
22. Werner, S.; Klein, F.; Mayenfels, T.; Brandenburg, K. A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events. In Proceedings of the 8th International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, 6–8 June 2016. [CrossRef]
23. Keen, R.; Freyman, R.L. Release and re-build of the listern's models of auditory space. *J. Acoust. Soc. Am.* **2009**, 3243–3251. [CrossRef] [PubMed]
24. ITU-R. *BS.1116-3 Methods for the Subjective Assessment of Small Impairments in Audio Systems*; Recommendation, International Telecommunication Union, Radiocommunication Sector: 2015. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1116-3-201502-I!!PDF-E.pdf (accessed on 21 December 2020).
25. ITU-R. *BS.1534-3 Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*; Recommendation, International Telecommunication Union, Radiocommunication Sector: 2015. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-I!!PDF-E.pdf (accessed on 21 December 2020).
26. Neidhardt, A.; Reif, B. Minimum BRIR grid resolution for interactive position changes in dynamic binaural synthesis. In Proceedings of the 148th International AES Convention, Online, 2–5 June 2020.
27. Mills, A.W. On the Minimum Audible Angle. *J. Acoust. Soc. Am.* **1958**, *30*, 237–246. [CrossRef]
28. Simpson, W.E.; Lee, D.S. Head movement does not facilitate perception of the distance of a source of sound. *Am. J. Psychol.* **1973**, *1*, 151–159. [CrossRef]
29. Ashmead, D.H.; Leroy, D.; Odom, R.D. Perception of the relative distances of nearby sound sources. *Percept. Psychophys.* **1990**, *4*, 326–331. [CrossRef] [PubMed]

30. Götz, G.; Kamandi, S. *Optimization of the Number and Spatial Distribution of Binaural Room Impulse Responses in an Augmented Auditory Reality Application*; Media Project; Technische Universität Ilmenau, Electronic Media Technology Group: Ilmenau, Germany, 2018.
31. Lindau, A.; Kosanke, L.; Weinzierl, S. Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses. In Proceedings of the 128th Int AES Convention, London, UK, 22–25 May 2010.
32. Kamandi, S. Perception of Simplifications of the Room Acoustics in a Dynamic Binaural Synthesis for Listener Translation. Master's Thesis, Technische Universität Ilmenau, Ilmenau, Germany, 2019.
33. Sloma, U.; Klein, F.; Werner, S.; Pappachan Kannookadan, T. Synthesis of Binaural Room Impulse Responses for Different Listening Positions Considering the Source Directivity. In Proceedings of the 147th International AES Convention, New York, NY, USA, 16–19 October 2019.
34. Lindau, A. The Perception of System Latency in Dynamic Binaural Synthesis. In Proceedings of the 35th Annual Meeting on Acoustics, DAGA, Rotterdam, The Netherlands, 23–26 March 2009.
35. Slutsky, D.A.; Recanzone, G.H. Temporal and spatial dependency of the ventriloquism effect. *Neuroreport* **2001**, *1*, 7–10. [CrossRef] [PubMed]
36. Wefers, F. Partitioned Convolution Algorithms for Real-Time Auralization. Ph.D. Thesis, RWTH Aachen, Aachen, Germany, 2014.
37. Meesawat, K.; Hammershø i, D. The time when the reverberation tail in a binaural room impulse response begins. In Proceedings of the 115th International AES Convention, New York, NY, USA, 10–13 October 2003.
38. Neidhardt, A.; Florian Klein, Niklas Knoop, T.K. Flexible Python Tool for Dynamic Binaural Synthesis Applications. In Proceedings of the 142th International AES Convention, Berlin, Germany, 20–23 May 2017.
39. Kuhn-Rahloff, C. *Realitätstreue, Natürlichkeit, Plausibilität-Perzeptive Beurteilungen in der Elektroakustik*; Springer: Heidelberg, Germany; London, UK; New York, NY, USA, 2012.
40. Lindau, A.; Weinzierl, S. Assessing the plausibility of virtual environments. *Acta Acust. United Acust* **2012**, *98*, 804–810. [CrossRef]
41. Meyer, D. Einfluss des Kopfhörermodells auf Reale und Virtuelle Schallquellen in Augmented Acoustics Anwendungen (Influence of the Type of Headphone on Real and Virtual Sound Sources). Bachelor Thesis, Technische Universität Ilmenau, Ilmenau, Germany, 2020.
42. Köster, F.; Guse, D.; Wältermann, M.; Möller, S. Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech. In Proceedings of the 41st Annual Meeting on Acoustics, DAGA, Nuremberg, Germany, 16–19 March 2015.
43. Werner, S.; Neidhardt, A.; Klein, F.; Brandenburg, K. Comparison of Different Methods to Create an Interactive Augmented Auditory Reality Scenario Using Sparse Binaural Room Impulse Response Measurements. In Proceedings of the 44th Annual Meeting on Acoustics, DAGA, Garching/Munich, Germany, 19–23 March 2018.
44. VIVE. HTC VIVE Tracker. Available online: https://www.vive.com/en/accessory/vive-tracker/ (accessed on 17 December 2020).
45. Erbes, V.; Schultz, F.; Lindau, A.; Weinzierl, S. An extraaural headphone system for optimized binaural reproduction. In Proceedings of the 38th Annual Meeting on Acoustics, DAGA, Oldenburg, Germany, 7–11 July 2012; pp. 313–314.
46. Kimmo Saunisto. VituixCAD—Simulation Software for Passive and Active Multi-Way Loudspeakers. Available online: https://kimmosaunisto.net/Software/Software.html (accessed on 17 December 2020).
47. Coleman, P.; Franck, A.; Jackson, P.J.B.; Hughes, R.J.; Remaggi, L.; Melchior, F. Object-Based Reverberation for Spatial Audio. *J. Audio Eng. Soc.* **2017**, *65*, 66–77. [CrossRef]
48. Brinkmann, F.; Gamper, H.; Raghuvanshi, N.; Tashev, I. Towards encoding perceptually salient early reflections for parametric spatial audio rendering. In Proceedings of the 148th International AES Convention, Online, 2–5 June 2020.