Meyer-Kahlen, Nils; Schlecht, Sebastian J.; Lokki, Tapio

Parametric Late Reverberation from Broadband Directional Estimates

Published in:
2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)

DOI:
10.1109/I3DA48870.2021.9610928

Published: 10/09/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Meyer-Kahlen, N., Schlecht, S. J., & Lokki, T. (2021). Parametric Late Reverberation from Broadband Directional Estimates. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)* (pp. 1-10). Article 9610928 IEEE. https://doi.org/10.1109/I3DA48870.2021.9610928

Parametric Late Reverberation from Broadband Directional Estimates

Nils Meyer-Kahlen
Dpt. Signal Processing and Acoustics
Aalto University
Espoo, Finland
nils.meyer-kahlen@aalto.fi

Sebastian J. Schlecht
Dpt. Signal Processing and Acoustics and Dpt. of Media
Aalto University, Espoo, Finland
sebastian.schlecht@aalto.fi

Tapio Lokki
Dpt. Signal Processing and Acoustics
Aalto University
Espoo, Finland
tapio.lokki@aalto.fi

***Abstract*—Several parametric spatial room impulse response rendering methods use broadband directional estimates, whereby based on sample-by-sample direction-of-arrival estimation, a single channel room impulse response is distributed to multiple loudspeakers. To this end, it has been unclear how such simple parametric processing behaves in the late part of the response. To assess this question, we use simulations and a measurement to show that the commonly applied estimation methods based on the pseudo intensity vector and time difference of arrival estimation do preserve the directional information in the late response. Also, we show that estimated directional differences can be audible under best case conditions. As broadband rendering can sound 'rough' or 'grainy' for transient input signals due to insufficient pulse density in individual reproduction channels, we use a method to synthesize smooth sounding spatial reverberation. For this, the broadband estimates are used to calculate directional energy envelopes, which are applied to filtered noise sequences. The findings presented here help assessing and improving spatial room impulse response processing methods.**

***Index Terms*—SRIR, SDM, roughness, auralization**

I. Introduction

Spatial room impulse response (SRIR) processing techniques aim at reproducing the acoustics of a measured space on an arbitrary reproduction setup. Parametric methods achieve this by estimating directional parameters and synthesizing loudspeaker or headphone responses. Many variants of such methods have been introduced, such as Spatial Impulse Response Rendering (SIRR) [1][1], Higher Order SIRR (HO-SIRR) [2], the Spatial Decomposition Method (SDM) [3], a variant applied to first-order Ambisonics responses (ASDM) [4], a framework referred to as Reverberant Spatial Audio Object [5], and a parametric synthesis method applied in [6]. The methods may employ different microphone arrays for measurement (e.g. rigid spherical arrays, open arrays). They conduct directional analysis and rendering either broadband, i.e. in a sample-wise manner (SDM), or in the time-frequency domain, based on a short-time Fourier transform (SIRR, HO-SIRR). Commonly, methods for broadband estimation rely on the intensity vector [1], [4] or on Time Difference of Arrival (TDoA) estimation [3]. Both of these principles were already applied in early work on directional analysis of room impulse responses [7], but many more directional analysis methods could be applied [8]. After directional estimation, several panning methods can be used to synthesize loudspeaker responses by spatializing the measured RIR using the directional estimates, such as VBAP, nearest loudspeaker synthesis (NLS) or higher-order Ambisonics (HOA). Headphone rendering can be achieved by subsequent convolution of the loudspeaker responses with head-related impulse responses (HRIR) or rather direct binaural decoding in the case of Ambisonics.

Generally SRIR processing algorithms have been evaluated by assessing the quality of the finally rendered signal [2]–[4], [9], but as perceptual sensitivity to certain directional errors appears to be low [10], [11], it is hard to deduce directional estimation performance and technical limitations of the methods in this way.

In the case of broadband methods, it is clear that exact directional information can be extracted for the direct sound and the first few reflections found in a SRIR. Successful estimation in this time range is demonstrated for example by visualizations [12] or by successful combination of early reflection estimates from several receivers [13]. In the late part of the SRIR, where the reflection density is so high that several reflections arrive within each analysis window, it is tempting to assume that meaningful directional estimates cannot be obtained, implying that the directional estimates are random [2], i.e. uniformly distributed over the sphere.

However, there are observations that seem to contradict this view. In [14] for example, the distribution of the late energy is related to perceived envelopment in different concert halls. Although one should consider that the paper shows cumulative directional energy distributions, it does seem like a general shape is maintained in the late part. Historically, this property has even been one important reason to favor SDM in concert hall research over an early SIRR implementation, as legacy SIRR did not allow for direction-dependent diffuse rendering. This is opposed to the more recent HO-SIRR [2], which uses the directional information contained in a higher-order SRIR to preserve directionality also in the diffuse part of the response. Clearly, these observations raise the question of the exact functioning principle of single-direction estimation in the late

This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 812719.

[1]Please note the difference between the abbreviations *SRIR* and *SIRR*, which can easily be overlooked.

part of the response.

To approach this question, we first describe a simple stochastic model of an anisotropic diffuse field in Section III and examine the distribution of the normalized pseudo intensity vector (PIV) in such a scenario. Then, we show a simulation example comparing normalized PIV, and TDoA estimation for simulated open microphone arrays. Moreover, we show directional estimation in a SRIR measured in a room with strongly non-uniformly distributed absorption.

Using a listening test, we confirm that the maintained directional features obtained from broadband parametric processing can be audible under best case conditions, when compared to random directions for the late part, see Section V. For the test, we do not only apply standard Nearest Loudspeaker Synthesis (NLS), but instead employ a simple parametric method based on spatio-temporal shaping of filtered noise, which also maintains the directional information from the broadband estimates, see Section IV.

II. Background

A. Limited Spatio-Temporal Resolution

Single-direction broadband estimation and rendering is the most straightforward way of processing SRIRs. In workflows employing such estimation, one directional estimate $\hat{\boldsymbol{\theta}}(t)$ is computed for every sample of the SRIR, which is later used to synthesize loudspeaker responses.

For all broadband single-direction estimation, it is important to note that sample-wise computation does not imply that the time interval between two sound events arriving from different directions is as low as one sampling interval. In fact, the so called spatio-temporal limit bounds the time difference under which two ideal, specular reflections can still be estimated separately to

$$\Delta T_{\text{min}} = 2\frac{d_{\text{max}}}{c}, \tag{1}$$

where c is the speed of sound [3]. This limit corresponds to twice the propagation time along the largest distance between two microphones in the employed microphone array d_{max}.

B. TDoA and PIV Estimation

Different estimation methods have been applied in broadband SRIR processing. SDM uses TDoA-based directional estimation method [3]. For the method, the responses measured using a microphone array with at least $M = 4$, ideally omnidirectional capsules are analyzed in blocks $\boldsymbol{h}_m = \left[h_m(t - D/2 + 1), ..., h_m(t + D/2)\right]^T$ of D samples, with a hopsize of one sample. In each block, the TDoAs $\tau_{i,j}$ between all pairs of microphones i, j are computed by finding the maximum in their crosscorrelation function. Then, for each block, the direction of arrival is determined from the TDoAs and the microphone position vectors $\boldsymbol{r}_m$ by means of least-squares, where $(.)^\dagger$ symbolized the Moore-Penrose pseudo-inverse

$$\begin{aligned}
\boldsymbol{\tau} &= \left[\hat{\tau}_{1,2}, \hat{\tau}_{1,3}, ..., \hat{\tau}_{M-1,M}\right]^{\text{T}} \\
\boldsymbol{V} &= \left[\boldsymbol{r}_1 - \boldsymbol{r}_2, \boldsymbol{r}_1 - \boldsymbol{r}_3, ..., \boldsymbol{r}_{M-1,M}\right]^{\text{T}} \\
\boldsymbol{k} &= \boldsymbol{V}^\dagger \boldsymbol{\tau} \\
\underline{\hat{\boldsymbol{\theta}}} &= \frac{\boldsymbol{k}}{||\boldsymbol{k}||}.
\end{aligned}$$

For a more detailed description of the estimation algorithms, see [12]. When implementing TDoA analysis, the spatio-temporal limit (1), becomes especially apparent, as in the used block processing, the block size should not fall below $D_{\text{min}} = \Delta T_{\text{min}} f_{\text{s}}$ for the estimator to compute meaningful results, where f_{s} is the sampling rate. If plane-waves from several directions occur in such a window, the individual directions cannot be resolved reliably. As all microphones used for TDoA analysis should be omnidirectional, the omnidirectional response h_{w}, which is required for rendering, can be extracted from either one.

Another common analysis method used for both narrowband [1], [2] and broadband processing is intensity vector estimation. When calibration and constants are omitted, the intensity vector is also referred to as pseudo intensity vector [15]. Applying the PIV is especially straightforward if first-order Ambisonics room impulse responses (ARIR) are available. In broadband methods employing PIV estimation [4], [16], [17], the omnidirectional pressure is simply multiplied with the three first-order directional components sample-by-sample. The resulting vector's length can be interpreted as an estimate of diffuseness, which in turn is used in narrowband processing of SRIR [1] and in the parametric time-frequency spatial audio method DirAC [18], which operates on running signals. To obtain a unit direction vector from the intensity vector estimate, the vector is normalized.

The functioning principle of PIV estimation can be explained from several different perspectives. An especially compact and useful derivation of an already normalized PIV estimator [19], [20] starts from SH theory. In the SH domain impulse response $\breve{\boldsymbol{h}}$, a plane-wave is ideally represented by the SHs of maximal order $\boldsymbol{y}_N(\boldsymbol{\theta}) = \left[Y_0^0(\boldsymbol{\theta}), Y_1^{-1}(\boldsymbol{\theta}), ..., Y_N^N(\boldsymbol{\theta})\right]^{\text{T}}$, evaluated at the incidence angle $\boldsymbol{\theta}(t)$

$$\breve{\boldsymbol{h}}(t) = \boldsymbol{y}_N(\boldsymbol{\theta}(t))s(t), \tag{2}$$

where $s(t)$ is the signal of the associated plane-wave. By inserting the definition of spherical harmonics in Cartesian coordinates and solving for $\boldsymbol{\theta}(t)$, one immediately obtains the normalized PIV

$$\hat{\boldsymbol{\theta}}(t) = \begin{bmatrix} \hat{x}(t) \\ \hat{y}(t) \\ \hat{z}(t) \end{bmatrix} = \frac{1}{\sqrt{3}\breve{h}_{\text{w}}(t)} \begin{bmatrix} \breve{h}_{\text{x}}(t) \\ \breve{h}_{\text{y}}(t) \\ \breve{h}_{\text{z}}(t) \end{bmatrix}, \tag{3}$$

where $\breve{h}_{\text{w}}(t)$ is the omnidirectional component and $\left[\breve{h}_{\text{x}}(t), \quad \breve{h}_{\text{y}}(t), \quad \breve{h}_{\text{z}}(t)\right]^{\text{T}}$ are the first-order SH components.

Note that in comparison to the more common formulation of the PIV, the directional components are divided by the omnidirectional component instead of multiplied by it. Both operations fulfill the same purpose: They make the estimate independent of the sign of the signal to avoid flipped estimates. The result of (3) should ideally have unit length. However, the normalization is only correct if the assumption of one plane-wave per time instance holds. For the case of two reflections, the equations can also be solved using two independent observations, as shown in [20]. The probability density function (PDF) of the estimated vector in cases where there are many more directions is the subject of the late field estimation properties discussed below.

It is important to note that broadband PIV estimation would be ideal in the case of a perfectly coincident array. For a real, non-coincident array, it should only be applied after low-pass filtering the response, with the cut-off frequency of the filter set to the spatial aliasing frequency of the array as in [4]. This implies that the impulse response of such a low-pass filter distributes energy in time, effectively acting as an analysis window as well. Sometimes, PIV estimation has also been done blockwise [21].

With either one of the methods, TDoA or PIV, the limited spatio-temporal resolution implies that accurate analysis is only possible in the early part of the response, where one plane-wave arrives in each analysis window.

C. Broadband rendering

Once the directional estimates are obtained, arguably the simplest rendering stage for synthesizing loudspeaker responses g_l based on the directional estimates is Nearest Loudspeaker Synthesis, as used in [3]. In the broadband case, the omnidirectional response is combined with the instantaneous directional estimate by assigning it to the nearest loudspeaker position out of the set of available loudspeakers $\boldsymbol{\Theta}_L = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_L\}$, such that

$$\begin{aligned} l_{\mathrm{NL}}(t) &= \arg\min_l ||\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}_l(t)|| \\ g_l(t) &= \begin{cases} h_{\mathrm{w}}(t) & l = l_{\mathrm{NL}}(t) \\ 0 & l \neq l_{\mathrm{NL}}(t). \end{cases} \end{aligned} \tag{4}$$

Two problems have been observed with such simple broadband rendering, which are discussed in more detail in Section IV.

III. Estimation Properties in the Late Field

To assess the properties of the estimator in the late field, we first introduce a model for an anisotropic diffuse field and show the distribution of PIV estimates for such a field. Next, we run simulations based on this model, assessing the directional estimator's distribution and also the resulting level distribution of the rendered output for both PIV and TDoA-based estimators. Then, we apply the estimators to the late field of a real room with non-uniformly distributed absorption.

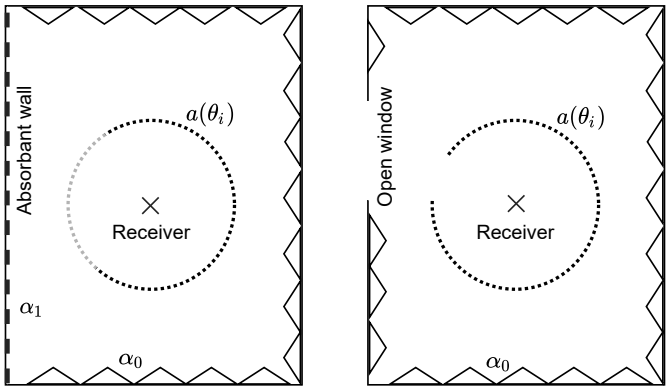


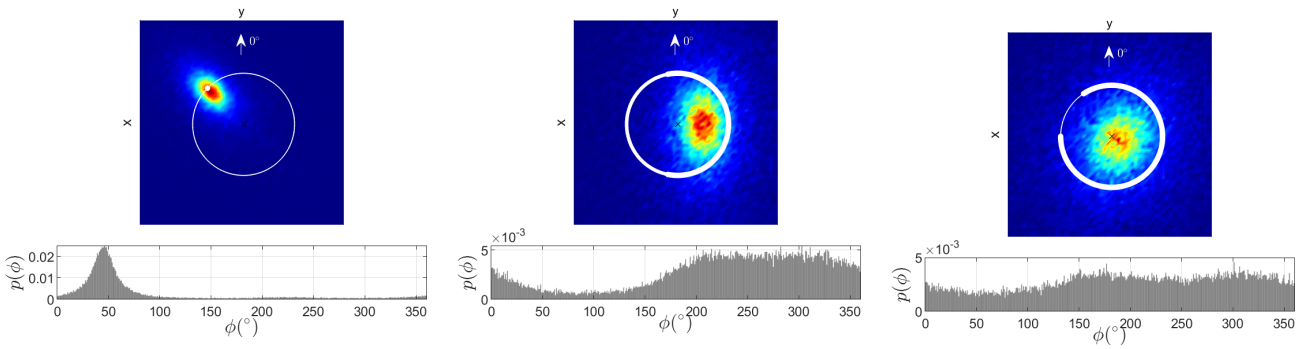
Fig. 1: Two scenarios that may cause an anisotropic diffuse field: A room with one more strongly absorbing wall (absorption coefficients $\alpha_1 > \alpha_0$) and one room with an open window. The dots represent the Gaussian sources scaled by $a(\boldsymbol{\theta}_i)$ that are used to model the resulting late, diffuse field. For such a model, we must assume a room with irregular geometry or a shoebox with sufficient scattering to decouple its three perpendicular wall pairs; otherwise a lower level would be expected from the right side as well.

A. Diffuse sound field model

To assess the behavior of PIV and TDoA estimation in the diffuse field, we first assume that the field is constructed from I uncorrelated Gaussian noise signals scaled by the level function $a(\boldsymbol{\theta})$, such that

$$\begin{aligned} s_i = a_i U_i \ \forall i = 1, ..., I, \\ U_i \sim \mathcal{N}(0, 1), \end{aligned} \tag{5}$$

where the uncorrelated components arrive from a (quasi-) uniformly distributed set of directions $\boldsymbol{\Theta}_I = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_I\}$ and $a_i = a(\boldsymbol{\theta}_i)$ is a continuous level function evaluated at these directions. In general, the Gaussian assumption for diffuse late reverberation is well established [22], [23]. Simulations of diffuse fields with a similar model are described in [24] as well. However, our model is different in that it assumes Gaussian noise with different levels from different directions rather than a combination of plane-waves and isotropic Gaussian noise. This represents a rather strict modelling assumption for evaluating the estimators. In the physical world, such a field may potentially be approximated by a closed room, which comprises sufficient scattering, but where the absorption is not distributed uniformly, such that some surfaces are less reflective than others. Then, extreme anisotropy is found for example in the direction of an open window or in a room with irregularly distributed absorption, illustrated in Figure 1. In real rooms with limited scattering, sound from nearby directions would be correlated more strongly. Additionally, in a shoebox shaped room, increased absorption on one side would also decrease the level of incoming sound from the opposite side.



(a) A test case, in which sound from $\phi_i = 50°$ has 40 dB more energy than the diffuse sound from all other directions.

(b) A scenario with a more strongly absorbing wall from which diffuse sound is attenuated by -12 dB.

(c) The scenario resembling and open window, with no diffuse sound from this direction.

Fig. 2: Diffuse field simulation of normalized 2D PIV estimation using Gaussian noise from 1000 directions, weighted by the directional level distribution $a(\boldsymbol{\theta})$ indicated by the white circles. In the upper half of each plot, a 2D histogram is shown. On the bottom, the resulting distribution of the estimates angle is presented.

B. PIV in the late field

For PIV estimation, we show the PDF of the resulting vector $p(\hat{\boldsymbol{\theta}}; a(\boldsymbol{\theta}))$, which depends on the present level distribution of the diffuse field. To do so, we plug in the diffuse sound model, (5) into the estimator (3) to obtain

$$\hat{\boldsymbol{\theta}} = \frac{1}{\sqrt{3}\sum_i s_i} \sum_i \begin{bmatrix} x_i s_i \\ y_i s_i \\ z_i s_i \end{bmatrix}, \tag{6}$$

$$= \frac{1}{\sqrt{3}} \frac{\sum_i \boldsymbol{\theta}_i a_i U_i}{\sum_i a_i U_i} = \frac{\boldsymbol{B}}{W}, \tag{7}$$

where $\boldsymbol{B} : \Omega \to \mathbb{R}^3$ and $W : \Omega \to \mathbb{R}$ are random variables.

Figure 2 shows the distribution of the estimates (in 2D, as it allows for better visualization), as well as the finally obtained directional distribution for three cases: in the theoretical case of one noise source that is 30 dB louder than the others, and in the two scenarios sketched out in Figure 1, with an absorbing wall or an open window. On the bottom of each plot, the distribution of the angular estimate $\hat{\phi} = \operatorname{atan}(\hat{y}/\hat{x})$ is presented. In all cases, the estimated vector is distributed according to a multivariate Cauchy distribution, which stems from (7) constituting a ratio of dependent Gaussian variables. The emergence of the Cauchy distribution in intensity estimation has also been shown in [25] for a scenario of one source and one interferer. A full derivation of the resulting estimates is the subject of future work.

Even though the estimation results in these examples show directional trends, it is also clear that the distribution of the estimates is not capable of closely following arbitrary directional functions $a(\boldsymbol{\theta})$. The levels may only scale the covariance and the mean of the PDF $p(\hat{\boldsymbol{\theta}}; a(\boldsymbol{\theta}))$, giving rise to a certain set of possible angular distributions. Such scaling and shifting can be seen in the 2D histograms presented in Figure 2.

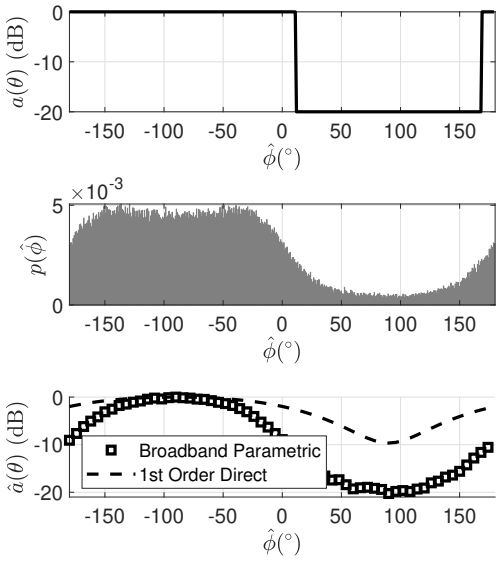
In the case of one dominant noise source shown in Figure 2a, the mean is shifted almost all the way to the unit circle and the variance becomes small. In the absorbing wall example shown in Figure 2b, the mean of the distribution is shifted to the right and the general trend of $a(\boldsymbol{\theta})$ is well maintained. In the open window example (Figure 2c), where no sound at all arrives from the small angular range on the left, the reduction in probability for estimating the corresponding directions is visible, but weaker.

On the one hand, this means that the kinds of level distributions, which the PIV is capable of estimating in the late field are severely limited. On the other hand, it is clear that some directional information is maintained, thus the assumption that the intensity vector generates uniformly random directions in the diffuse field is not true.

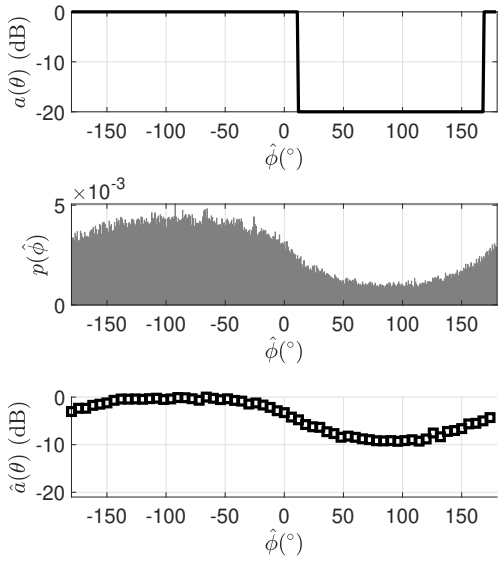
C. Microphone array simulations

As a next step, simulations are conducted for PIV and TDoA utilizing the same diffuse field model, with plane-waves impinging from $I = 360$ directions on two different microphone arrays. In this simulation, the estimated directions $\hat{\boldsymbol{\theta}}(t)$ are then utilized to perform simple NLS rendering as shown in (4). By assigning the omnidirectional response's sample to a set of $L = 60$ virtual loudspeakers at directions $\boldsymbol{\Theta}_L$, equidistantly placed in a circle, the directional estimates are converted back to a level distribution $\hat{a}(\boldsymbol{\theta}_l)$, which can be compared with the original level distribution before the estimation and rendering process.

For simulation, microphone arrays were used, which can be considered typical choices for the applied estimators. PIV estimation was simulated for a tetrahedral array of ideal cardioid capsules, and with a radius of $r = 1.5$ cm, reminiscent of the Soundfield microphone used below. TDoA estimation was simulated for an open array of six omnidirectional capsules, on the faces of a cube, as in the intensity probe commonly used for SDM [3]. The same radius was used. The results are shown in Figures 3a and 3b. The first row represents the original level distribution $a(\boldsymbol{\theta})$. The second row shows the distribution of the estimate $p(\hat{\boldsymbol{\theta}}; a(\boldsymbol{\theta}))$. The last row shows the resulting RMS after NLS rendering.



(a) PIV estimation in a simulated, anisotropic diffuse field using an open array of four cardioid capsules in tetrahedral arrangement.



(b) TDoA estimation in the same simulated, anisotropic diffuse field using an open array of six omnidirectional capsules.

Fig. 3: Estimation in an anisotropic diffuse field, where one wall is more absorbing than the others. For comparison, linear processing using a first-order beam-former in case of the tetrehedral array. The black squares represent a set of 60 rendering points (virtual loudspeakers) used for NLS.

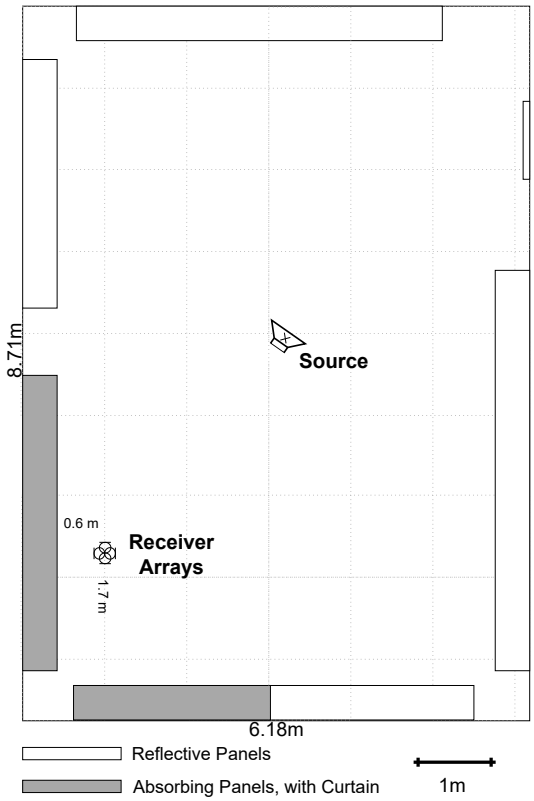


Fig. 4: Variable acoustics room 'Arni' used for the measurements presented in Figure 5 and the listening test. The receiver was placed close to an absorbing wall and the source was directed towards the room corner in order to excite horizontal directions with equal energy.

Together with the result after PIV estimation and rendering in Figure 3a, we have also included the spatial distribution that would be obtained by linear SH processing for determining the shape of the late part response. One might refer to this as discrete plane-wave decomposition (PWD) or, equivalently, directing a first-order hyper-cardioid beam-former to each of the rendering directions [26].

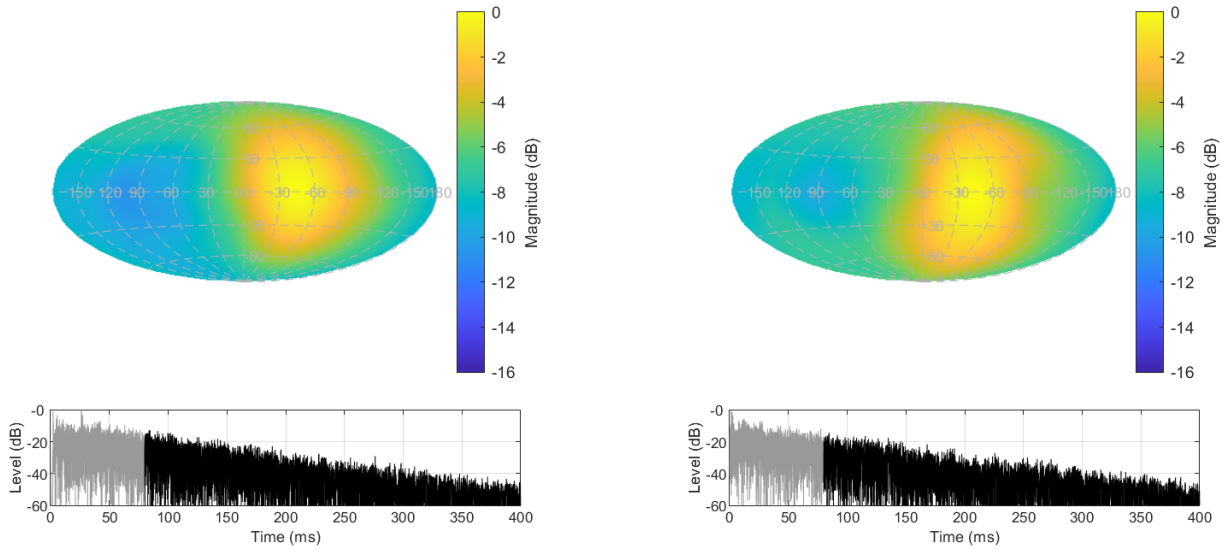
For higher-order input, PWD yields high directional resolution; it is used for example in [27] to analyse the late part of room impulse responses. For first-order, this simulation suggests that such linear processing follows the particular level distribution $a(\boldsymbol{\theta})$ less closely than PIV estimation.

Clearly, for the TDoA-based estimation, the statistics of the estimation result is different from that of PIV estimation. In the present simulation, the TDoA-based estimator follows the dip in the level distribution less closely than the PIV, but is actually closer to the distribution obtained by first-order PWD. However, also here, the directional estimates are far from uniformly distributed over the sphere.

D. Analysis of a real space

As the next step, a measurement was conducted in the variable acoustics room "Arni" at the Aalto Acoustics Lab [28], in order to check if directional information would be maintained in the late part of a measured response as well. The chamber has dimensions 8.71 m $\times$ 6.81 m $\times$ 3.6 m. All of its variable panels were set to the reflective setting, except those on one half-wall on the left of the microphone and one section behind the microphone, see Figure 4.

In this configuration, the room had a reverberation time of approximately $T_{30} = 0.62$ s. A Genelec 8331AP loudspeaker was used for the measurement.



(a) Analysis of a real space using the Soundfield ST-350 microphone and PIV processing.

(b) Analysis of a real space using the GRAS 50VI intensity probe and TDoA-based directional information. Also here, directional information is maintained.

Fig. 5: The map shows the estimated energy distribution in the time range 80–400 ms (black area of the omnidirectional RIR shown on the bottom), measured next to an absorbing wall using PIV and TDoA estimation respectively. The contained directional information is clearly visible even within the dynamic range of 16 dB: The level on the left and the rear of the array are attenuated.

The loudspeaker was directed towards the front right corner of the room to excite horizontal directions equally. Two microphone arrays - a Soundfield ST350 as a tetrahedral array of cardioids, and a GRAS 50VI intensity probe as an open array of omnidirectional capsules were placed 60 cm away from the left absorbing wall. The energy distributions was estimated using PIV for the ST-350 and TDoA estimation for the GRAS 50VI and are shown in Figure 5, integrated in the time range $80 - 400$ ms. In this extreme example, directional differences in the resulting distribution reach up to 12 dB for both estimators. The directional differences seem to be slightly larger in the case of PIV estimation.

The extreme case was selected, such that differences would potentially be audible in the subsequent listening test, where perceptual consequences were generally expected to be small. In order to have access to a reference for the test, also a BRIR was measured using the KEMAR head and torso simulator.

IV. Rendering

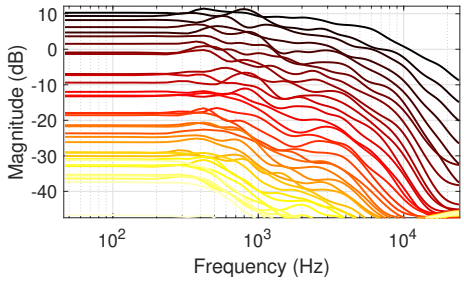
Now, we would like to check if, under best-case conditions, the estimated directional distributions lead to audible differences when compared to replacing them with uniformly distributed directions. As mentioned above, two problems have been stated with respect to standard broadband rendering. Firstly, it has been observed that the rendered response may contain more high frequencies than the omnidirectional response, which can be referred to as whitening. The effect has first been reported in [29] and the typical solution is to apply spectrogram or filterbank-based time-varying equalization to the response [17], [29]. If such equalization is not done carefully, it can lead to audible artefacts of its own. The second problem, roughness, can be perceived when transient signals are rendered. Since such critical signals have been rarely used in practice, this issue of roughness has been discussed less frequently. However in [10], [20], compensation strategies based on introducing all-pass filters have been employed.

As for testing the audibility of the directional information estimated in the late reverberation tail, we would like to use a transient signal. Therefore, we use a simple parametric rendering approach, which is described next.

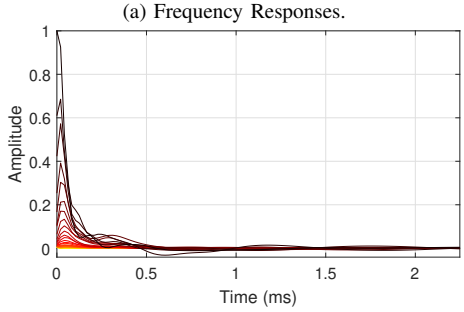
A. Short-time Filters

The applied parametric rendering is based on a model that uses white, Gaussian noise, filtered using a global short-term filter per time instance, modulated by one energy envelope per reproduction channel. The method has similarity to the parametric model introduced in [30], which may use directional information from a measured BRIR to shape binaural noise in separate bands. The model applied here is capable of incorporating time-frequency dependence of the omnidirectional response and the detected broadband time-direction dependency, but it cannot reproduce frequency-dependent directional differences.

As a first step, the omnidirectional response is analyzed in order to create filters that mimic its short-time spectral content. It is one of the advantages of the method that these filters can be designed in order to have a well controlled impulse response, which avoids time-aliasing and artefacts.

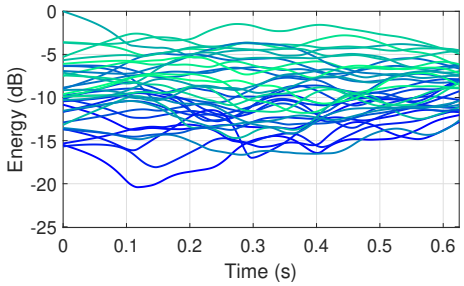


(a) Frequency Responses.

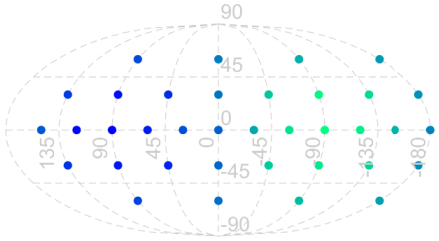


(b) Impulse Responses f_b.

Fig. 6: Short-time coloration filters obtained from block-wise analysis of the omnidirectional RIR. The color indicates the starting time of the block from $t = 0$ s (black) to $t = 1.2$ s (yellow)



(a) Short-time directional modulation functions d_l.



(b) Loudspeaker positions $\mathbf{\Theta}_L$ used for rendering.

Fig. 7: Directional decay for different analysis positions. One color corresponds to one loudspeaker position. Note the lower level of the decay curves on the left side, where the absorbing wall is located (darker blue).

To create the filters, the omnidirectional room response is first separated into blocks of $B = 1024$ samples (at 48 kHz) and minimum phase filters are designed for each block, with impulse response $f_b(t)\ \forall b = 1, ..., \lceil L_\mathrm{s}/B \rceil$, where L_s is the length of the response.

The filters are set to a constant gain below $f_\mathrm{lim} = 500$ Hz, since standard NLS rendering is used for low frequencies. The filters and their short impulse responses are shown in Figure 6.

B. Directional Modulation Functions

Next, the directional information is extracted. For this, NLS is performed and the energy contained in each of the rendering channels is compared to that of the omnidirectional RIR. Like this, directional modulation functions are obtained for each channel l. These modulation functions need to be smoothed, for example using a hann window $w(t)$ in order to mitigate amplitude modulation that would otherwise be perceivable when rendering transient sounds

$$d_l(t) = \frac{\sum_\tau g_l^2(t-\tau)w(\tau)}{\sum_l \sum_\tau g_l^2(t-\tau)w(\tau)}. \tag{8}$$

The smoothed directional modulation functions are shown in Figure 7. If the aim is to render to a dense set of loudspeakers, the directional modulation functions can also be interpolated spatially.

C. Synthesis

When finally synthesizing the response, standard NLS rendering is used for the first 30 ms – a range in which early reflections are still identified properly. Then, a 5 ms crossfade is used to transit to the parametrically rendered part. Standard NLS rendering is kept below 500 Hz, as at low frequencies, roughness is not perceivable; a perfectly reconstructing Linkwitz-Riley crossover filter is used to separate the two bands. For high frequencies, the fully parametric rendering is done by applying the filter determined for the current time range to an independent sequence of Gaussian noise $n(t)$ for each channel. Then, the directional modulation function is applied to each reproduction channel, such that a synthesized response $\hat{g}_l$ is obtained as

$$\hat{g}_l(t) = d_l(t) \sum_\tau n(t-\tau) f_{\lceil \frac{t}{B} \rceil}(\tau), \tag{9}$$

where $\lceil . \rceil$ denotes rounding to the next larger integer, which is used to select the correct short-time filter f_b for the current time range.

Now, it is possible to use the obtained responses for loudspeaker rendering. For the listening test, the loudspeaker channels are then convolved with diffuse-field equalized HRIRs of the KEMAR HATS conducted in the loudspeaker array, such that comparison against the BRIR reference measurement becomes possible.

V. Perceptual Evaluation

A short listening test was conducted so that participants could complete the test in their home office due to the COVID-19 restrictions. The nine experienced participants used different headphone models, however, they were all high quality headphones.

Two stimuli were rendered using the obtained responses: A transient, synthetic kick-drum, and a short anechoic speech excerpt.

For each stimulus, participants were first asked to rate similarity in terms of spatial impression alone, and then to judge general similarity applying their own criteria, which were collected after the test. The full set of stimuli included the binaural reference (ref), rendering using standard NLS without any compensation based on TDoA estimation (TDoA - NLS), as well as the parametrically rendered responses based on TDoA and PIV estimation (TDoA - Param and PIV - Param). As an additional stimulus, the directional estimates of the parametrically rendered TDoA-based response where replaced with random directions, uniformly distributed over the sphere, starting at 80 ms (Late Random). An anchor was created by using random directions for the complete response (also modifying the direct sound) and applying a highpass filter at 200 Hz, to induce a noticeable spectral difference.

Most importantly, in terms of spatial impression, the results obtained for the kickdrum sample show a significant difference ($p = 0.0078$, Wilcoxon signed rank test) between the renderings using the estimated directions and the random directions in the late part, see Figure 8. When listening to this transient example, the reverberant tail can be perceived as part of the sound itself, and the transition to the random directions is clearly audible as a shift of energy to the left side during the response. Also, in the general similarity rating, the late random directions received worse results. However, it was also expected that the standard NLS rendering would receive worse results, given the audible roughness, but it seems that spectral differences between all rendered sounds were as important as these artefacts when judging general similarity. Only one participants indicated that roughness was used as a criterion, while all except one mentioned attributes like coloration, spectral content or timbre.

Interestingly, the speech sample did not evoke any significant differences, see Figure 9. Here, the reverberant tail is not heard as clearly as part of the sound, and it seems that other differences, for example due to coloration, that occur between the renderings where as important as the shift caused by the late part.

VI. Discussion and Future Work

The listening test results show that using a transient input signal, directional differences in the range detectable using parametric estimation methods can be perceived. In [11], it has been shown that randomizing late directions only caused very minor differences in terms of similarity when compared to rendering using the original data. Also in our continuous speech sample, the differences in the late part did not lead to significant differences between the renderings. Also, it should be noted that we deliberately chose an extremely anisotropic response, and that for a more regular room, audibility is expected to decrease even for transient signals.

A closer examination of different signals and directional distributions and the parametric rendering is beyond the scope of this paper and would require a more controlled listening test. Dedicated work on the perceptual detectability of directional information in the late part of the responses as such has been published recently as well [27].

Although not the main focus of this work, it should also be mentioned is that as in earlier tests comparing parametric SRIR renderings to a binaural references [31], a clear difference to the reference was always notable. It is probably mostly due to coloration differences, which are hard to avoid. While it is unlikely that such differences are problematic when comparing different rooms with each other, or creating plausible or transfer-plausible renderings for extended realities [32], further effort should go into making fully authentic reproductions using SRIR methods.

In future work, the statistics of the PIV estimator will be derived for the presented sound field model. Moreover, a formal description of the TDoA estimator in such a field will be formulated.

With respect to the rendering approach using filtered noise, there is a similarity to single channel reverberators based on shaped velvet noise [33]. The simple parametric model used here would be easily modified to use velvet noise for each reproduction channel as well, in order to create more efficient spatial reverberators in the future.

VII. Conclusion

All in all, we can conclude that single-direction parametric estimators are in principle able to detect directional distributions in the late field. This could be seen in simulations and a SRIR measurement that was taken close to an absorbing wall in a real room. However, we have also described that such estimation yields results in a statistical sense, and that the possible kinds of directional distributions are therefor limited.

In this work, a simple parametric rendering approach has been applied, which uses the obtained directional estimation, but does not suffer from roughness or whitening as the standard broadband NLS rendering. Even using this method, the rendered sound is noticeably different from a binaural reference. However, we have shown that the detected distributions are in principle noticeably, but only under best-case conditions, when rendering a transient sounds.

References

[1] J. Merimaa, "Spatial Impulse Response Rendering I: Analysis and Synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, Dec. 2005.

[2] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall, "Higher-order spatial impulse response rendering: investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution," *J. Audio Eng. Soc.*, vol. 68, no. 5, pp. 368–354, May 2020.

[3] S. Tervo, J. P. Tynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc.*, vol. 61, no. 1, pp. 17–28, Mar. 2013.

[4] M. Zaunschirm, M. Frank, and F. Zotter, "Binaural Rendering with measured room responses: first-order ambisonic microphone vs. dummy head," *Appl. Sci.*, vol. 10, no. 5, p. 1631, Feb. 2020.

[5] P. Coleman, P. J. B. Jackson, L. Remaggi, and F. Melchior, "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc*, vol. 65, no. 1, pp. 66–77, Feb. 2017.

[6] P. Stade, J. Arend, and C. Pörschmann, "A parametric model for the synthesis of binaural room impulse responses," in *173rd Meeting of Acoustical Society of America and 8th Forum Acusticum*, Boston, Massachusetts, Aug. 2017, p. 015006.

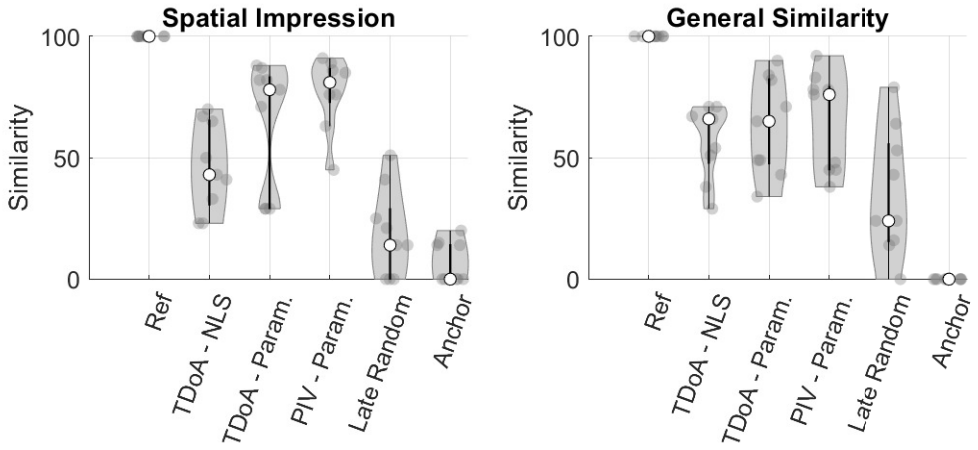


Fig. 8: Listening test results using a synthetic kick drum as a stimulus. Observe the difference between the parametric rendering using the TDoA or PIV result and the random directions.

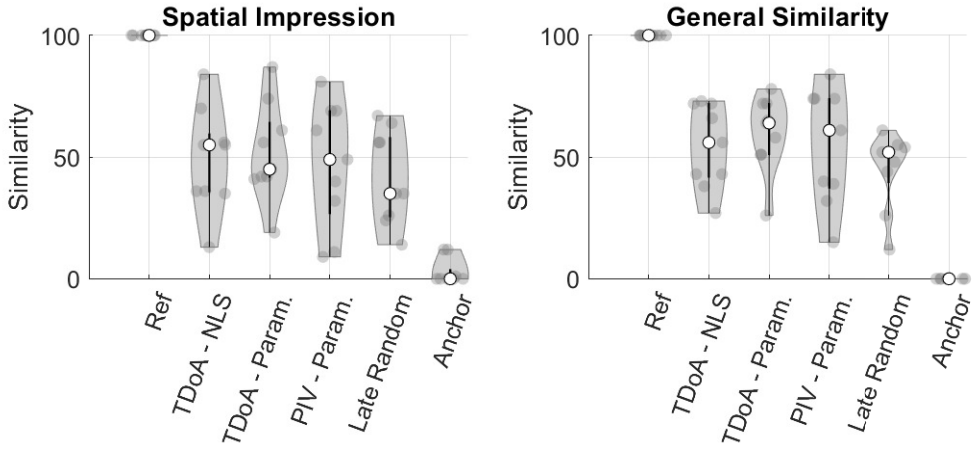


Fig. 9: Listening test results using a short speech sample. Here, there are no significant differences.

[7] Y. Yamasaki and T. Itow, "Measurement of spatial information in sound fields by closely located four point microphone method." *J.Acoust.Soc.Jpn (E)*, vol. 10, no. 2, pp. 101–110, 1989.

[8] S. Tervo and A. Politis, "Direction of Arrival Estimation of Reflections from Room Impulse Responses Using a Spherical Microphone Array," *IEEE Trans. Audio Speech Lang. Process.*, vol. 23, no. 10, pp. 1539–1551, Oct. 2015.

[9] V. Pulkki and J. Merimaa, "Spatial impulse response rendering II: reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc.*, vol. 54, no. 1/2, pp. 3–20, Jan. 2006.

[10] S. V. A. Gari, P. T. Calamia, and P. W. Robinson, "Optimizations of the spatial decomposition method for binaural reproduction," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 959–976, 2020.

[11] J. Ahrens, "Auralization of Omnidirectional Room Impulse Responses Based on the Spatial Decomposition Method and Synthetic Spatial Data," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, May 2019, pp. 146–150.

[12] J. Pätynen, S. Tervo, and T. Lokki, "Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses," *J. Acoust. Soc. Am.*, vol. 133, no. 2, pp. 842–857, Feb. 2013.

[13] O. Puomio, N. Meyer-Kahlen, and T. Lokki, "Locating Image Sources from Multiple Spatial Room Impulse Responses," *Appl. Sci.*, vol. 11, no. 6, p. 2485, Mar. 2021.

[14] T. Lokki, L. McLeod, and A. Kuusinen, "Perception of loudness and envelopment for different orchestral dynamics," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2137–2145, 2020.

[15] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *European Signal Processing Conference (EUSIPCO)*, Aug. 2010.

[16] M. Frank and F. Zotter, "Spatial impression and directional resolution in the reproduction of reverberation," in *DAGA - Fortschritte der Akustik*, Mar. 2016.

[17] M. Zaunschirm, F. Zagala, and F. Zotter, "Auralization of High-Order Directional Sources from First-Order RIR Measurements," *Appl. Sci.*, vol. 10, no. 11, May 2020.

[18] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric Time-Frequency Domain Spatial Audio*. Wiley, IEEE Press, 2017.

[19] J. Daniel and S. Kitic, "Time Domain Velocity Vector for Retracing the Multipath Propagation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 421–425.

[20] L. Gölles and F. Zotter, "Directional enhancement of first-order ambisonic room impulse responses by the 2+2 directional signal estimator," in *15th International Conference on Audio Mostly*, Sep. 2020, pp. 38–45.

[21] D. Protheroe and B. Guillemin, "3D Impulse Response Measurements of Spaces Using an Inexpensive Microphone Array," in *International Conference on Room Acoustics*, 2013.

[22] J. A. Moorer, "About this reverberation business," *Computer Music J.*, vol. 3, no. 2, pp. 13–28, 1979.

[23] R. Badeau, "Common mathematical framework for stochastic reverberation models," *J. Acoust. Soc. Am*, vol. 145, no. 4, pp. 2733–2745, Apr. 2019.

[24] N. Epain and C. T. Jin, "Spherical Harmonic Signal Covariance and Sound Field Diffuseness," *IEEE Trans. Speech Audio Process.*, vol. 24, no. 10, p. 12, 2016.

[25] B. Günel, "On the statistical distributions of active intensity directions," *Journal of Sound and Vibration*, vol. 332, no. 20, pp. 5207–5216, Sep. 2013.
[26] D. Zotkin, R. Duraiswami, and N. Gumerov, "Plane-Wave Decomposition of Acoustical Scenes Via Spherical and Cylindrical Microphone Arrays," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 1, pp. 2–16, Jan. 2010.
[27] B. Alary, P. Massé, S. J. Schlecht, M. Noisternig, and V. Välimäki, "Perceptual analysis of directional late reverberation," *J. Acoust. Soc. Am.*, vol. 149, no. 5, pp. 3189–3199, May 2021.
[28] K. Prawda, S. J. Schlecht, and V. Välimäki, "Evaluation of Reverberation Time Models with Variable Acoustics," in *Sound and Music Computing Conference (SMC)*, 2020.
[29] S. Tervo, J. Pätynen, N. Kaplanis, and e. al, "Spatial analysis and synthesis of car audio system and car cabin acoustics with a compact microphone array," *J. Audio Eng. Soc.*, vol. 63, no. 11, pp. 914–925, Feb. 2015.
[30] P. Stade and J. M. Arend, "Perceptual Evaluation of Synthetic Late Binaural Reverberation Based on a Parametric Model," in *AES Conference on Headphone Technology*, 2016.
[31] J. Ahrens, "Perceptual Evaluation of Binaural Auralization of Data Obtained from the Spatial Decomposition Method," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, Oct. 2019, pp. 65–69.
[32] S. A. Wirler, N. Meyer-Kahlen, and S. J. Schlecht, "Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes," in *AES Conference on Audio for Virtual and Augmented Reality*, 2020.
[33] V. Välimäki, B. Holm-Rasmussen, B. Alary, and H.-M. Lehtonen, "Late Reverberation Synthesis Using Filtered Velvet Noise," *Appl. Sci.*, vol. 7, no. 5, p. 483, May 2017.