OPEN ACCESS Freely available online
S. V. A. Garí, J. M. Arend, 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 December).
DOI: https://doi.org/10.17743/jaes.2020.0063

Optimizations of the Spatial Decomposition Method for Binaural Reproduction

SEBASTIÀ V. AMENGUAL GARÍ,[1] *AES Member*, **JOHANNES M. AREND,**[1,2] *AES Student Member*,
(samengual@fb.com) (johannes.arend@th-koeln.de)

PAUL T. CALAMIA,[1] **AND PHILIP W. ROBINSON,**[1] *AES Member*
(pcalamia@fb.com) (philrob22@fb.com)

[1] *Facebook Reality Labs Research, Redmond, WA*
[2] *TH Köln - University of Applied Sciences, Cologne, Germany*

The spatial decomposition method (SDM) can be used to parameterize and reproduce a sound field based on measured multichannel room impulse responses (RIRs). In this paper we propose optimizations of SDM to address the following questions and issues that have recently emerged in the development of the method: (a) accuracy in direction-of-arrival (DOA) estimation with open microphone arrays utilizing time differences of arrival as well as with B-format arrays using pseudo-intensity vectors; (b) optimal array size and temporal processing window size for broadband DOA estimation based on open microphone arrays; (c) spatial and spectral distortion of single events caused by unstable DOA estimation; and (d) spectral whitening of late reverberation as a consequence of rapidly varying DOA estimates. Through simulations we analyze DOA estimation accuracy (a) and explore processing parameters (b) in search of optimal settings. To overcome the unnatural DOA spread (c), we introduce spatial quantization of the DOA as a post-processing step at the expense of spatial distortion for successive reflections. To address the spectral whitening (d), we propose an equalization approach specifically designed for rendering SDM data directly to binaural signals with a spatially dense HRTF dataset. Finally, through perceptual experiments, we evaluate the proposed equalization and investigate the consequences of quantizing the spatial information of SDM auralizations by directly comparing binaural renderings with real loudspeakers. The proposed improvements for binaural rendering are released in an open source repository.*

0 INTRODUCTION

The spatial decomposition method (SDM) [1] can be used to parameterize and reproduce a sound field based on measured multichannel room impulse responses (RIRs). The direction of arrival (DOA) of each sample of the RIR is first estimated and then the instantaneous energy is mapped to its corresponding direction using any loudspeaker or headphone-based reproduction method.

Since its initial conception the method has been used for multiple applications, such as concert hall analysis and auralization [2–4], stage acoustics [5, 6], car cabin acoustics [7, 8], acoustic preference research in small rooms [9, 10], speech intelligibility [11], or audio-visual perception in virtual reality (VR) [12]. Additionally it has been used in conjunction with multiple spatial audio reproduction methods, such as Vector-Base Amplitude Panning (VBAP) [13, 14], Nearest Loudspeaker Synthesis (NLS) [7, 15, 4], Ambisonics [16, 17], and binaural synthesis [16, 6]. The increasing popularity of the method can be partially attributed to the publicly available SDM Toolbox [18], released by the original developers of the method.

The extended usage of the method resulted in a series of questions and open issues. In this manuscript we focus on the case of the analysis of sound fields composed of broadband single sources in small and medium rooms and reproduced via binaural synthesis. In particular we address four particular topics in this work: (a) SDM can be used with various DOA estimation methods, such as time differences of arrival (TDOA) using open microphone arrays [1] or pseudo-intensity vectors (PIV) using B-format arrays [7].

*https://github.com/facebookresearch/BinauralSDM

However it is not yet clear which of the two estimation methods provides more accurate DOAs and under which conditions one of the two methods performs better.

(b) Various parameters such as array size and temporal processing window size for DOA estimation with an open microphone array have not been investigated systematically, and the question remains as to what are the optimal parameters for best possible DOA detection using broadband RIRs. (c) As the RIR progresses into the late reverberation, the DOA estimation becomes unreliable [1], leading to spatial spread of single events—a problem not yet addressed when using SDM for binaural reproduction.

(d) As a result of rapidly varying DOA estimates in the late part of the RIR, the late reverberation becomes spectrally white [7, 17]. However current solutions to overcome this issue [7] are computationally inefficient when rendering SDM data directly to binaural signals, requiring a suitable alternative for binaural reproduction. In the following paragraphs we introduce the above-mentioned points in more detail and outline how we examine the technical questions. We then briefly describe our proposed optimizations of SDM for the generation of binaural renderings, covering the three stages of processing—measurement, analysis, and rendering.

The original SDM method, developed for open microphone arrays and performing the DOA analysis using TDOA, was validated numerically and perceptually with reference to an image source model [1]. Later the same authors released an open implementation of the algorithm [18], including an alternative analysis approach based on broadband PIVs of B-format RIRs (similar to [19]), which has lately been popularized, and enabling the usage of the method with a greater variety of array configurations. Recent evaluations suggest that the DOAs are not reliably estimated when analyzing broadband B-format signals [20], although perceptually satisfactory results can be obtained with appropriate bandpass filtering of the raw RIRs and subsequent smoothing of the DOA estimates [16].

In this paper, we explore the analysis requirements and compare simulation and measurement results from PIV analysis to those of TDOA. Furthermore we investigate optimal parameters for the analysis using TDOA and open arrays. Note that we focus our investigations on the analysis of broadband events. Analysis using multiple bands, such as in [7], results in additional degrees of freedom in the search for optimal parameter values, which could be different in each analysis band.

In regard to headphone reproductions of SDM RIRs, the use of dense Head-Related Transfer Function (HRTF) datasets results in a higher degree of spatial resolution at the cost of potential timbral degradations. As the RIR progresses into the late reverberation, the DOA estimates become unstable and less reliable [1]. This causes consecutive samples of the RIR to be mapped to disparate locations—an effect that is accentuated by the fact that small fluctuations will result in reflections being mapped onto several adjacent HRTFs. To address this we discuss approaches for the post-processing of DOAs based on the spatial quantization and clustering of reflections, reducing the DOA spread significantly at the expense of clustering consecutive reflections onto the same directions. In particular we focus on the implications of using regular grids for quantization of the spatial information.

Rapidly varying DOA estimates cause a spectral whitening in certain portions of the rendered responses, as consecutive samples corresponding to the same band-limited event are mapped onto disparate locations as broadband events. This is especially relevant in small spaces with high reflection density [7] or at the late reverberation tail [16, 5, 6], where the DOA cannot be reliably estimated. The presence of this artifact has been reported with multiple reproduction approaches, such as NLS [7], Ambisonics [17], or binaural synthesis [6], and in typical rooms it generally results in an increase of the reverberation time at high frequencies.

Tervo et al. proposed a time-frequency equalization to address this problem and validated it in the application of car cabin acoustics [7]. This equalization method was designed for loudspeaker reproduction, and it generates time-varying filters for each of the loudspeaker (rendered) RIRs by comparing the average magnitude response of the rendered RIRs and original pressure RIR. Applying this approach to binaural rendering is possible by using a virtual loudspeaker approach, where each loudspeaker feed is convolved with the Head-Related Impulse Response (HRIR) corresponding to the loudspeaker location. However, with a spatially dense HRTF dataset, this approach becomes impractical due to computing limitations. In this paper we propose an alternative equalization approach comprising a reverberation correction process (RTMod) and the processing of the resulting Binaural Room Impulse Responses (BRIR) with a cascade of allpass filters (RTMod+AP).

Finally, as has been suggested previously [21], we hypothesize that the spatial resolution of the SDM auralizations can be reduced without perceivable degradations. We investigate the minimum required spatial resolution in perceptual experiments employing SDM auralizations by directly comparing binaural renderings with real loudspeakers.

The paper is structured as follows. Sec. 1 reviews the two approaches (TDOA and PIV) used in SDM for the DOA analysis. Sec. 2 evaluates the performance of the directional analysis for various common array and parameter configurations using simulations. Sec. 3 compares the results of the directional analysis conducted with TDOA and PIV on the same set of measurements from a tetrahedral array. Sec. 4 describes our proposed rendering approach to resynthesize binaural RIRs, including DOA post-processing, a novel equalization method for the reverberation and instrumental validation. Sec. 5 presents a perceptual evaluation on the plausibility of BRIRs with quantized spatial resolution. Secs. 6 and 7 present a discussion and conclusions, respectively.

1 DOA ESTIMATION

The basic paradigm of SDM involves assigning one DOA to each of the samples of a pressure RIR, implicitly assuming that the sound field is composed of a succession of

broadband specular events. Once this information is available it can be used for the directional analysis of an RIR or to re-synthesize the sound field using any loudspeaker or headphone-based method. Two main approaches are currently widespread to perform the DOA analysis, depending on the nature of the microphone array and available signals.

1.1 Time Differences of Arrival (TDOA) Method

In this section we review the method introduced by Tervo et al. in [1], which estimates DOA data from a multichannel RIR by exploiting the TDOAs between microphones. The estimation requires an open array of $M \geq 4$ microphones arranged in a 3D space. Although the authors recommended the use of omnidirectional microphones, accurate results have been obtained with arrays of cardioid microphones as well [15], suggesting that the requirements regarding directivity are somewhat flexible. However, if the data are intended to be used for auralization, at least one of the microphones must be omnidirectional or encoded to present an omnidirectional response. Alternatives to encode directional responses are proposed in [7] but are beyond the scope of this paper.

A sliding Hanning window of size L is applied to the RIR and at each time step the DOA is resolved for one single acoustic event. The window is moved in 1-sample steps and thus one DOA is estimated for each sample of the RIR. The size of the sliding window must be equal to or greater than the time needed for a plane wave to travel between the most distant microphones in the array. The available data regarding optimal array and window sizes are limited and one of the objectives of this paper is to find appropriate parameters.

Defining $\mathbf{h}_i$ and $\mathbf{h}_j$ as the windowed RIRs of microphones i and j at an arbitrary time instant, the TDOA of an event $\tau_{i,j}$ between microphones i, j can be estimated by finding the delay that maximizes the cross-correlation $\mathbf{r}_{\mathbf{h}_i,\mathbf{h}_j}$

$$\tau_{i,j} = \arg\max \{\mathbf{r}_{\mathrm{h_i,h_j}}\}. \tag{1}$$

Assuming a sound field model in which only one broadband sound event arrives within the windowed responses, $\tau_{i,j}$ can be related to the geometrical properties of the array and direction of propagation of the sound event.

$$\tau_{i,j} = (\mathbf{m}_i - \mathbf{m}_j)^{\mathrm{T}} \frac{\mathbf{d}_p}{c}, \tag{2}$$

where $\mathbf{m}$ [3 × 1] refers to the position of the microphones in cartesian coordinates, $\mathbf{d}_p$ [3 × 1] refers to the direction of propagation of a single event in the windowed response, T denotes the transpose operation, and c refers to the speed of sound. This operation is repeated for each of the $N_{mp} = \frac{M(M-1)}{2}$ microphone pairs.

The time differences for each pair and the difference vectors for their positions are collected into the vector $\boldsymbol{\tau}$ [N_{mp} × 1] and matrix $\mathbf{V}$ [3 × N_{mp}], respectively. Then Eq. (2) can be rewritten as

$$\boldsymbol{\tau} = \mathbf{V}^{\mathrm{T}} \frac{\mathbf{d}_p}{c}. \tag{3}$$

By calculating the Moore-Penrose pseudoinverse $(\cdot)^+$ of $\mathbf{V}^{\mathrm{T}}$ the least-squares solution is obtained, resolving the direction of propagation of the event.

$$\mathbf{d}_p = (\mathbf{V}^{\mathrm{T}})^+ \boldsymbol{\tau} c. \tag{4}$$

Finally, the DOA vector $\mathbf{d}$ [3 × 1] is the opposite vector of the direction of propagation

$$\mathbf{d} = -\mathbf{d}_p. \tag{5}$$

The previous process is repeated for each sample of the measured RIRs, resulting in a matrix $\mathbf{D}$ containing the DOA for each sample. The reader is referred to [1] for further details regarding the algorithm. Throughout this work we used the implementation provided in the SDM Toolbox [18] for Matlab to conduct the presented investigations.

1.2 Pseudo-Intensity Vectors (PIV) Method

The DOA estimation can also be done using alternative approaches, provided that one DOA is assigned to each sample in the RIR. The Spatial Impulse Response Rendering (SIRR) method [19, 22] introduced the use of pseudo-intensity vectors for the estimation of narrow band directional information from B-format (First Order Ambisonics—FOA) RIRs. As opposed to SDM, SIRR further aims at dividing the RIR into a directional and diffuse component. More recently a higher order variant (HO-SIRR) was introduced [23], introducing the capability of identifying the direction of arrival of multiple events arriving simultaneously.

The original conception of SDM [1] only contemplated DOA analysis using the TDOA method. However, in the SDM Toolbox [18], Tervo et al. included the PIV analysis approach to generate DOA estimates to be used with SDM. The method is largely based on that used for the characterization of the directional sound field component in SIRR. However the SDM Toolbox only included analysis of broadband responses. While PIV analysis using multiple bands could be used in conjunction with SDM, to the best knowledge of the authors this has not been evaluated in the past. Additionally, in spite of the growing popularity of this analysis approach with SDM, only a few recent studies have analyzed its objective and perceptual performance [20, 24, 25], and the topic warrants further attention.

As with the TDOA method, the goal is to obtain one directional estimate for each sample in the RIR.

$$\mathbf{D}(n) = \begin{bmatrix} \hat{x}(n) \\ \hat{y}(n) \\ \hat{z}(n) \end{bmatrix} = h_w(n) \begin{bmatrix} h_x(n) \\ h_y(n) \\ h_z(n) \end{bmatrix} * \mathbf{w}(k) \tag{6}$$

where h_w(n) is the omnidirectional channel (W) of the B-format signal, which approximates the pressure RIR. The three components of the pseudo-intensity vectors are represented by h_x, h_y, and h_z and correspond to the figure-of-eight virtual microphones of the B-format signal aligned with the X, Y, and Z axes, respectively. The DOA estimates are convolved with a Hanning window**w**for smoothing. This convolution is effectively a low-pass filter on the DOA data, and the optimal size of the window is currently unknown.

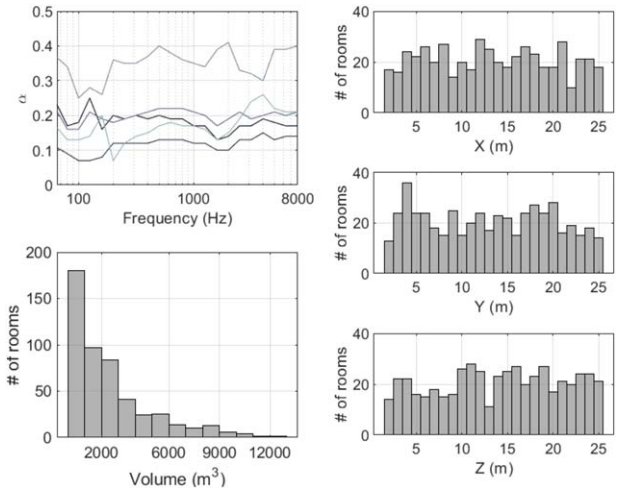

Fig. 1. Absorption data (top left), volume histogram (bottom left), and side-length histograms of the simulated rooms (right).

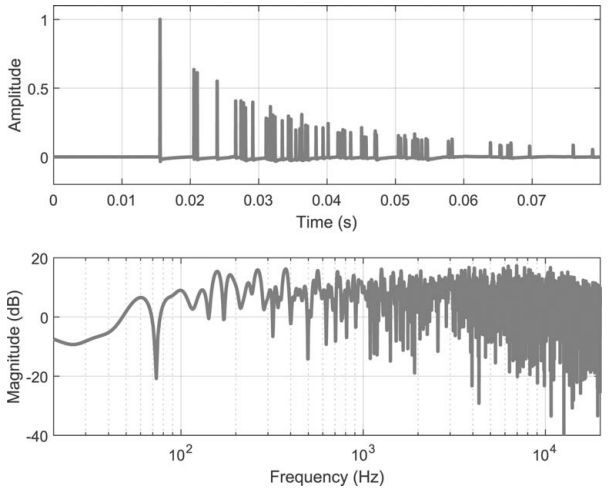

Fig. 2. ISM simulated RIR (top) and magnitude spectrum (bottom) for a room of dimensions $8 \times 6 \times 9$ m.

2 SIMULATIONS

In order to investigate the performance of the two presented approaches in idealized conditions, we simulated multichannel RIRs of 500 shoebox rooms. The wall lengths were randomly chosen using a uniform distribution containing lengths between 2 and 25 m, in order to cover a meaningful range of room sizes. The source and receiver locations were randomized as well in each simulation. Image Source Method (ISM) [26] simulations including frequency-dependent material absorption and air absorption were conducted using AKtools [27]. The materials for each wall are different and kept constant for all the room configurations. Room size distributions and absorption properties are shown in Fig. 1.

In order to allow for the evaluation of SDM with B-format signals, we expanded the functionality of the simulator to include ideal first-order microphones. The simulated RIRs contain 64 sound events, corresponding to the direct sound and specular reflections up to third order. Although the analysis we present in this section is not generalizable to the entire RIR, we decided to focus only on strong specular events for two reasons: it is known that the spatial analysis performed by SDM does not provide accurate results when multiple sound events start overlapping, as is the case in the late reverb [1], and the directionality of the late reverb is of limited perceptual relevance in common rooms [28]. An exemplary simulated RIR is presented in Fig. 2.

2.1 Evaluation Metric

We propose an objective metric ε_{DOA} to evaluate the performance of the DOA estimations. For each of the samples in the RIR we compute the angular distance between the ground truth DOA $\mathbf{D}_{ISM}(n)$ and estimated direction $\mathbf{D}(n)$. These are then weighted by the energy of each sample and normalized by the total energy of the RIR.

$$\epsilon_{DOA} = \frac{\sum_{n=1}^{N} \arccos\{\mathbf{D}(n)^{\mathrm{T}}\mathbf{D}_{ISM}(n)\}p(n)^2}{\sum_{n=1}^{N} p(n)^2} \tag{7}$$

where p_n represents the instantaneous pressure of sample n and N is the number of samples in the RIR. Note that the DOA vectors contained in the matrices $\mathbf{D}_{ISM}$ and $\mathbf{D}$ are expressed in cartesian coordinates and normalized to define unit vectors.

It is worth noting that the proposed metric relies on comparing the estimated DOA of each sample in the RIR. Thus it is only suitable for the case in which sound events in the RIR are not overlapping and each sample in the RIR has only one associated DOA.

2.2 Time Differences of Arrival (TDOA) Evaluation

As discussed previously, the requirements for estimation with time difference of arrival consist of a compact microphone array with at least four microphones arranged in a 3D space. If the data are intended for auralization, one of the microphones must be omnidirectional—for analysis only, multiple directivities are acceptable. These somewhat relaxed requirements resulted in a variety of experimental works using various array configurations, including microphone arrays arranged in orthogonal directions of various sizes—with or without a center microphone [7, 5, 9, 20, 6], a tetrahedral array with a physical or virtual omnidirectional microphone [15, 20], or a 12-element star-shaped array [20].

2.2.1 Array Size

Arrays composed of 6 or 7 omnidirectional microphones (3 orthogonal pairs, with or without a center microphone) seem to be among the most commonly used topologies [7, 20, 6, 4, 9]. However, to the best of our knowledge, no formal comparison between array topologies and dimensions has been completed to date.

Given the extended use of this topology, we chose to investigate the optimal size of this geometry. To that end we performed a DOA analysis using the function `SDMpar` from the SDM Toolbox [18] on the 500 simulated ISM

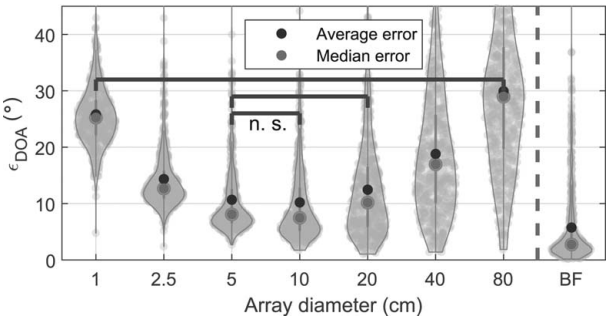

Fig. 3. DOA error as a function of microphone array diameter (500 ISM simulations at $f_s = 48$ kHz). The open array is a 7-microphone array with a center capsule and 6 capsules arranged as pairs in orthogonal directions. BF refers to ideal B-format signals. Brackets refer to statistically non-significant differences between groups ($p > 0.01$). The statistical analysis was conducted using a Kruskal-Wallis test with Tukey's range post-hoc correction (balanced dataset with non-normal distributions).

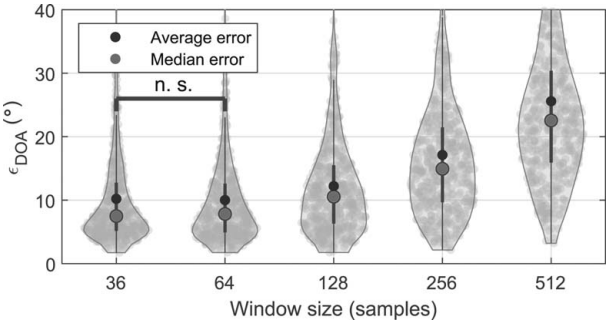

Fig. 4. DOA error as a function of window size (500 ISM simulations at $f_s = 48$ kHz for a 7-microphone open array of 10-cm diameter with a center capsule and 6 capsules arranged as pairs in orthogonal directions). Brackets refer to statistically non-significant differences between groups ($p > 0.01$). The statistical analysis was conducted using a Kruskal-Wallis test with Tukey's range post-hoc correction (balanced dataset with non-normal distributions).

RIRs. The results in Fig. 3 show that for a sampling rate of 48 kHz, an array with diameters of 5 and 10 cm results in the smallest DOA estimation errors among the evaluated dimensions. Smaller and larger sizes yield statistically significant increases in error.

The presented results partially confirm the findings of perceptual validations by Ahrens [20]. They found that at this sampling rate ($f_s = 48$ kHz), when comparing auralizations against a reference BRIR, arrays of 6 sensors with diameters of 10 and 20 cm result in smaller perceptual differences than arrays of 4-cm diameter or other configurations such as a tetrahedral array of 4.8-cm diameter or a 12-element array of 10-cm diameter.

It is worth highlighting the substantial difference between the average and median errors in all cases reported in Fig. 3. This suggests that the directional estimation error is especially high in some cases, leading to long tailed distributions.

When analyzing small spaces, such as a car cabin [7], it might be desirable to use smaller arrays to enable the use of smaller analysis windows. In these cases higher sampling rates might be necessary in order to avoid quantization in the resolved DOA estimates caused by insufficient time resolution. However a comparison between 48 and 96 kHz carried out using a compact tetrahedral array of approximately 4.8 cm of diameter (Core Sound TetraMic) found no significant benefit of increasing the sample rate when analyzing larger halls [15]. A formal comparison of array sizes at multiple sampling rates warrants more investigation. However the data presented here serve to lay a foundation and provide formal validation of the suitability of a relatively popular array topology used for SDM.

2.2.2 Window Size

The size of the sliding window used for DOA estimation theoretically governs the compromise between temporal and spatial resolution. While a larger window length would enable a more robust estimation of single events, it also increases the probability of multiple events arriving within the same window and consequently violating the soundfield model based on a succession of specular events. Tervo et al. [1] recommend the use of a window that is slightly longer than the time that it takes for one acoustic event to travel along the longest array dimension. In order to formally validate this recommendation we completed the DOA analysis and computed the estimation error using the 10-cm array configuration.

In Fig. 4 the DOA estimation errors [computed with Eq. (7)] for various window sizes are reported. As can be seen, shorter windows yield smaller errors that increase steadily with increasing window length. For the evaluated case of a 10-cm diameter array at 48 kHz, a window size of 36 samples seems most appropriate, although differences to the 64-sample window do not seem obvious. Thus we conclude that sizes between 36 and 64 samples are appropriate for this configuration. We hypothesize that fine tuning might provide a practical benefit depending on the structure of the specific analyzed RIR.

Theoretically, for the studied case, smaller windows could be used, as long as the window length is larger than the time needed for a plane wave to travel between the two most distant microphones. However the minimum value allowed in the SDM Toolbox is slightly higher and selected by default and thus for practical reasons we decided to limit the smallest size.

2.3 Pseudo-Intensity Vectors (PIV) Evaluation

The use of coincident array configurations (B-format arrays) has recently become more common, as it does not require specific open array topologies and the signals can be obtained in a variety of ways, either by using B-format arrays or taking subsets of Ambisonic signals in higher-order spherical arrays. However the rendering results generated with SDM and PIV DOA estimation have often been found to be unsatisfactory [24, 25, 23, 20]. For instance, in direct comparisons against a reference, SDM renderings from B-format arrays and PIV DOA estimation presented lower perceptual ratings than those from open array configura-

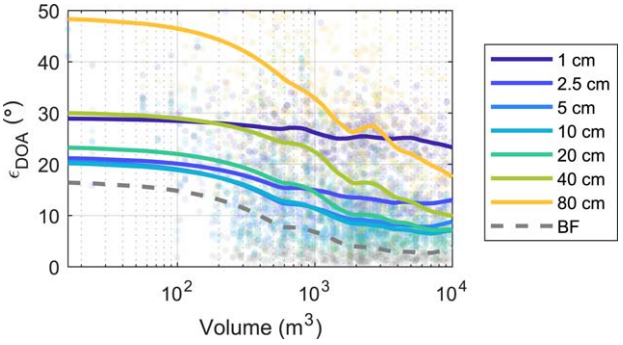

Fig. 5. DOA estimation error as a function of room volume for various array configurations. Circular markers represent individual observations; lines show a moving average (200 samples) of the observations.

tions and TDOA DOA estimation [20] or first-order SIRR renderings [23]. It is shown in [20] that the use of measured B-format RIRs results in poor directional estimates, even for samples of the direct sound. That could be partially attributed to the imperfect directional properties and spatial aliasing exhibited by B-format microphones above the spatial aliasing frequency. However it has also been suggested that the equalization process provided in the SDM Toolbox could be partially responsible for the generation of audible artifacts [24], as it suffers from time aliasing [20]. In addition, unlike with open microphone arrays, in the B-format estimation the DOAs are obtained from single-sample snapshots of the pseudo-intensity vectors and the estimates are prone to very fast variability. These effects can be partially mitigated by band-pass filtering and smoothing the DOA estimates, and some studies have reported that SDM auralizations from B-format arrays can indeed be perceptually very close to a reference [16]. Given the mixed results, we aim at evaluating the performance of broadband PIV estimation, as it is generally used in SDM.

Similar to our analysis with the open microphone arrays, we simulated 500 RIRs corresponding to shoebox rooms. In this case we defined ideal B-format microphones and used the function `SDMbf` (without windowing) to obtain the DOA estimates. As reported in Fig. 3, in an ideal simulation the results from a B-format array are approximately half an order of magnitude better than the best tested open array configuration. These results suggest that the B-format analysis might be preferable to TDOA. However in practice the quality of the results strongly depends on the A-to-B encoding and the quality of the reconstructed first-order directional patterns. This is further explored in Sec. 3.

2.4 Room Size

The DOA estimation error is presented as a function of room volume for different array sizes in Fig. 5. It can be seen that the estimation error tends to decrease with increasing room volume, with bigger arrays presenting more accentuated reductions in error. This is expected, as in larger rooms the average time between consecutive reflections is greater, thus reducing the probability of two reflections arriving simultaneously or within the same analysis window.

It is observed that the B-format array performs better than any open array counterparts at all room volumes. However, note that in this case the simulated B-format signals exhibit ideal directivities and are free from spatial aliasing, which is generally not the case in measured signals. For open arrays, very small and big arrays perform consistently worse than medium-sized arrays. This confirms our findings suggesting that medium-sized arrays (5, 10, and 20-cm diameter) are preferred at a sampling rate of $fs = 48$ kHz (see SEC. 2.2.1).

It is important to note once again that this analysis is based on limited-order ISM simulations without diffuse energy, thus representing the best case scenario for the analysis. We hypothesize that the same behavior would generalize to measured rooms, given that prominent reflections are already more spaced in time in larger spaces, but it is not possible to confidently generalize these findings from the presented results.

3 MEASUREMENTS

In addition to the aforementioned sound field assumptions, in simulations the array sensors exhibit ideal characteristics and the RIRs are free of noise. The PIV method resolves the DOA by providing an exact solution, while the TDOA method uses a pseudo-inverse, thus providing a least squares solution. We hypothesize that measurement noise, non-ideal microphone directivity limitations, and imperfections in the A-to-B format conversion might result in a noticeable analysis degradation, especially for the PIV method.

In order to compare the results of the TDOA and PIV approaches in a practical scenario, we conducted RIR measurements using a tetrahedral microphone (CoreSound Tetramic) in an apartment-like scene with a tall absorptive ceiling (see *FRL Apartment* in the *Replica* dataset [29]).

We used the A-format signals (four cardioid microphones at the vertices of the tetrahedron) to conduct the DOA analysis based on TDOA and the B-format signals for the PIV method. The A-to-B conversion is performed following the manufacturer's recommendations—using individually calibrated encoding matrices and the software *VVMic*. Note that the array is relatively small (∼2 cm diameter) and thus not the optimal choice for the open array case. In addition the microphones are not omnidirectional but cardioid, potentially even further compromising the performance of the TDOA algorithm. However this enables a more accurate and direct comparison than repeated measurements with different arrays placed at the same position. Given that both analysis methods can be used with this array, a direct comparison aids in establishing guidelines for algorithm choice in practical scenarios.

Given the relatively complex scene geometry and large amount of furniture (see Fig. 6), reliable DOA ground truth data are not available. Thus we focus on a qualitative comparison of the results obtained using various window sizes.

Fig. 6. Top view of the room used for the measurements in the PIV and TDOA comparison. The visualization is part of the Replica Dataset [29] and the furniture was in a different configuration during the acoustic measurements. Approximate source (S) and receiver (R) locations are marked as black squares.

Fig. 7 contains a collection of spatial energy maps corresponding to the analyzed measurement. We focus the analysis on the first 20 ms of the RIR, as they contain several prominent reflections. Both methods present considerable agreement for the DOA of more energetic samples, with longer windows providing more stable estimates that result in energy clusters at discrete locations. A slight angular offset is apparent when comparing the two methods. Provided that the signals for the analysis come from the same measurement set, a possible explanation for this offset is a non-ideal encoding of the first-order directional patterns of the B-format array.

Given that the ground-truth data for the DOAs are not available, it is not straightforward to assess which of the estimations is closer to the actual DOAs. However it seems reasonable to conclude that raw DOA estimates for the PIV method (without a smoothing window) present significantly worse performance than when they are smoothed, with estimated DOAs of multiple highly energetic samples scattered around the entire sphere—including samples of the direct sound. This could result in noticeable localization artifacts if these data were to be used directly for auralization. Thus, in a practical scenario, a smoothing window should be used when using the PIV method. Note that in the PIV case the windowing acts as a low-pass filter on the pseudo-intensity vectors, as a Hanning window is convolved with the product of omnidirectional and velocity RIRs. Although the stability of the PIV analysis increases with longer windows, the DOA of the strongest events is not significantly affected by the window size.

For the TDOA approach, longer windows result in clustering many low-energy DOA samples into larger clusters, resulting in cleaner energy maps. However, note how when increasing the window size from 16 to 64 samples, the estimated DOA of one specific reflection changes drastically. Specifically, in the left column of Fig. 7, the reflection represented by the light green icons presents most of its energy around $[-160^\circ, 20^\circ]$ in the 16-sample plot, then moves to $[55^\circ, -40^\circ]$ in the 64-sample plot.

Another identifiable difference between the two methods is that while the overall stability improves for both methods with longer windows, in the PIV analysis the estimates follow continuous traces, creating trailing patterns between DOAs of strong events due to the effects of the window convolution. This is also somewhat present in TDOA estimates, although to a much smaller extent.

4 BINAURAL RENDERING

Binaural room impulse responses (BRIRs) can be synthesized as a weighted sum of HRTFs corresponding to each DOA, appropriately delayed and weighted by the amplitude of the instantaneous pressure of the omnidirectional RIR. We presented this method previously in [6]. Alternative implementations based on binaural Ambisonics and virtual loudspeaker layouts can be found in [16] and the SDM Toolbox [18], respectively.

The SDM sound field is defined by a [1 × N] vector $\mathbf{p} = [p_1, p_2, ..., p_N]$ containing the pressure RIR and a [3 × N] matrix $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_N]$ indicating the DOA for each of the samples in cartesian coordinates. The DOA matrix, **D**, can be rotated to render BRIRs corresponding to arbitrary head orientations,

$$\mathbf{D}^u = \mathbf{R}_z(-\theta^u)\,\mathbf{R}_y(-\phi^u)\,\mathbf{D} \tag{8}$$

where $\mathbf{R_y}$ and $\mathbf{R_z}$ represent rotation matrices corresponding to the head orientation (θ^u, ϕ^u). Here a right-hand coordinate system is used with positive Y corresponding to left and positive Z to up.[1]

The indices $\hat{k}_n^u$ of the closest HRIRs for each sound event in each head orientation, u, are selected by finding the nearest HRIR for each sample, n, in the rotated DOA matrices $\mathbf{D}^u$.

$$\hat{k}_n^u = \underset{n\in 1,...,N}{\arg\min}\{d(\mathbf{D}_n^u, \hat{\mathbf{D}})\} \tag{9}$$

where $\hat{\mathbf{D}}$ is a [3 × K] matrix containing the source/receiver relative orientations of the HRIR dataset in cartesian coordinates and $d(\cdot, \cdot)$ is the Euclidean distance.

The BRIR for an arbitrary head orientation, $\mathbf{BRIR}^u$, is then constructed by delaying the HRIRs corresponding to indices $\hat{k}_n^u$ at the nth position by n samples and multiplying them by the instantaneous pressure p_n contained in the pressure RIR:

$$\mathbf{BRIR}^u(t) = \sum_{n=1}^{N} p_n\,\mathbf{HRIR}_{\hat{k}_n^u} \circledast \delta(t-n), \tag{10}$$

where **HRIR** is a three-dimensional [H × K × 2] matrix containing an HRIR dataset of H samples (per channel) and K source/receiver relative orientations. Samples in the BRIR are indicated by t.

To improve the timbral fidelity of the binaural reproduction, these rendered BRIRs can be further perceptually optimized by processing the DOA matrix **D** prior to the binaural rendering and performing reverberation equalization

[1]Note that the the DOA must be rotated in a reversed order to achieve a correct rotation. Roll rotation is excluded from the equation.

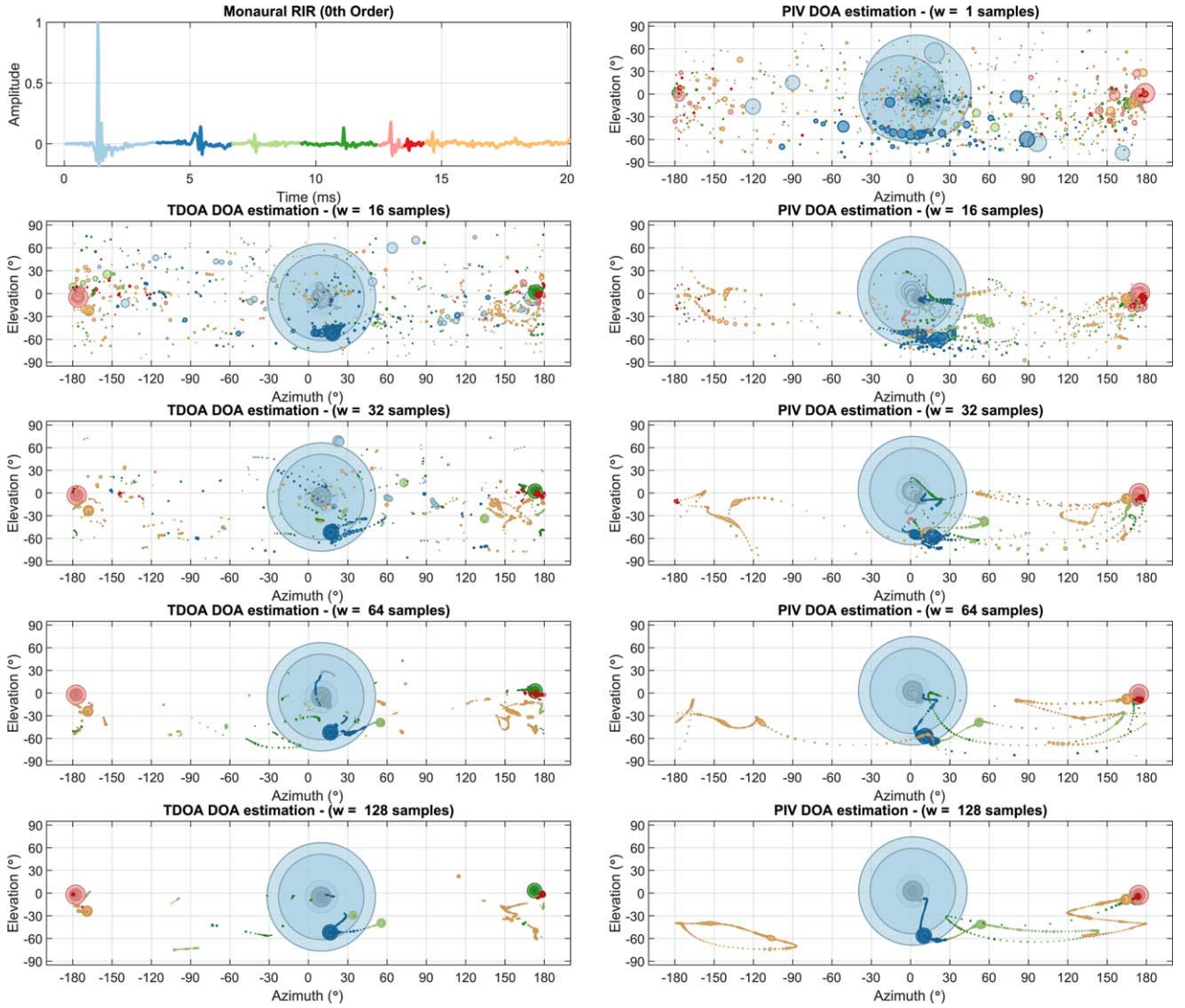

Fig. 7. Comparison of the DOA estimation results obtained from TDOA (left) and PIV (right) analysis at various window sizes ($f_s = 48$ kHz) using a tetrahedral array (Tetramic). Each circular marker represents a sample of the monaural RIR, with area proportional to the instantaneous energy of each sample. Note that in PIV windows are convolved with the signal, effectively low-passing the RIR.

on the rendered BRIRs. These processes are described in the subsequent sections.

4.1 DOA Postprocessing

In measured RIRs, sound events usually span multiple samples, as opposed to simulations, where events are usually very compact in time. As demonstrated in Sec. 3, when analyzing measured RIRs, it is common to obtain DOA estimates that fluctuate over the course of a single event. This can potentially result in spatial spread and spectral distortions of these events. Thus it is desirable to post-process the DOA estimates to minimize potentially audible artifacts.

In loudspeaker rendering, a common approach is the use of Nearest Loudspeaker Synthesis [7, 9], which assigns the DOA to the closest loudspeaker. While this reduces the spatial spread of single events by collapsing nearby DOA values to a single location, it might result in noticeable localization shifts, especially if the distance between the direction of the direct sound and the closest loudspeaker is larger than the minimum audible angle. An optimization method of the loudspeaker layout is available in [30].

When dealing with binaural synthesis, previous studies suggest that using a moving-average filter to smooth the DOA estimates is an effective post-processing approach [6, 16]. When using synthetic spatial data to auralize an omnidirectional RIR, it is desirable to use a certain degree of smoothing on the spatial data [21], resulting in smaller perceptual differences than when using random unfiltered data. Here we discuss potential alternatives to the post-processing of the DOA based on clustering of reflections and spatial quantization.

4.1.1 Direct Sound

As demonstrated in Fig. 7, the DOA of the most energetic samples in the RIR seem to be reliably estimated, although it is not uncommon to observe trailing patterns between those. Considering that in practice each acoustic event has a specific spectral shape, each event spans several

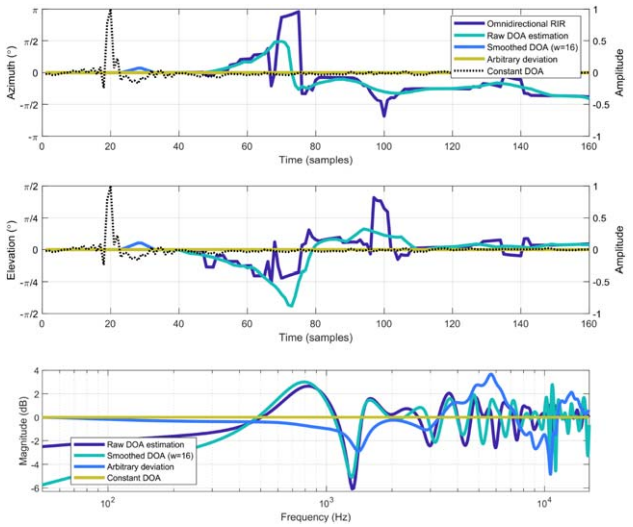

Fig. 8. DOA of the direct sound (top two panels) and spectral deviations resulting from mapping samples to multiple locations (bottom panel).

samples of the RIR. Thus mapping consecutive samples to disparate locations could potentially lead to spectral artifacts. To mitigate this and preserve the spectral properties of the direct sound we propose a post-processing step on the DOA matrix, enforcing a stable DOA for the direct sound.

We first locate the index ds of the sample with the largest amplitude in the RIR (direct sound)

$$ds = \arg\max\{|\mathbf{p}|\} \tag{11}$$

and then enforce the direction of the direct sound on the first i samples of the RIR

$$\mathbf{D}_p(n) = \begin{cases} \mathbf{D}(ds) & n \leq i \\ \mathbf{D}(n) & n > i \end{cases} \tag{12}$$

where $\mathbf{D}_p$ is the DOA matrix with constant DOAs for the first initial i samples and the original directions for the rest of the RIR.

Adjusting i in a case-by-case basis allows optimizing the trade-off between spectral and spatial fidelity. Theoretically the maximum allowed value of i is equal to the initial time delay gap (ITDG) of the RIR. Due to pre-ringing of each sound event, in practice i will be slightly lower. The goal is to maintain a constant DOA for the direct sound for as long as possible without distorting the DOA of the first reflection. Note that as low frequency components extend longer in time, in RIRs with longer ITDG the spectral distortion of the direct sound can be corrected to a greater extent.

To demonstrate the effect of this post-processing step in a practical scenario we auralized an RIR measured with a 7-microphone array (10-cm diameter, one central microphone) and compared the magnitude spectrum of the direct sound when auralized using the original DOA as estimated using the TDOA approach or when enforcing a stable DOA on first $i = 160$ samples.

The data are reported in Fig. 8, showing the DOA estimates in four different cases: raw DOA estimates as produced by the SDM algorithm (using minimum window size—36 samples), smoothed DOA estimates with a moving average filter of 16 samples, stable estimates with an arbitrary deviation for illustrative purposes, and a reference case with perfectly stable DOA. It is observed that even when the DOAs of the most energetic samples are appropriately estimated, spectral deviations of up to 6 dB are present. Note that in the example presented here the responses correspond to the left channel of a BRIR rendered with a very dense HRIR dataset (20,624 directions). The effects are likely dependent on the rendering configuration (loudspeaker layout or HRIR grid) and thus not easily generalizable. However it is expected that rendering setups with more spatial resolution will suffer from higher spectral distortions, as small fluctuations in the DOA result in samples being mapped at different locations.

4.1.2 DOA Quantization

The same spatial-timbral trade-off discussed for direct sound is present for early reflections and late reverberation. This becomes especially severe when low-passed and overlapped reflections appear in the RIR, rendering the high temporal resolution of the DOAs unusable. This is typically the case in common rooms, where air absorption and surface absorption tend to attenuate high frequencies more than low frequencies. In this case using all the available information results in spreading specular reflections onto multiple directions, leading to spatial and timbral degradations. We thus suggest a straightforward approach based on clustering of early reflections to reduce the spatial spread of early reflections.

There are a number of suitable methods for the spatial clustering:

- Virtual layout optimization as in [30]: The selected DOAs are chosen from weighted spatial energy maps derived using the original DOA data and pressure RIR. The advantage of this approach is that an adaptive grid allows for graceful downsampling of the DOA.
- Density Based Spatial Clustering (DBSCAN) [31]: This method can be used to identify portions of the RIR with meaningful DOA data and cluster them. The parts of the RIR in which there are no reliable DOA data can be rendered separately as diffuse components or arbitrary directions can be enforced to preserve spectral information. At the time of writing we obtained preliminary results related to a post-processing algorithm using DBSCAN—although these are out of the scope of this manuscript.
- Quantization using an arbitrary grid: Using sparse spatial grids for quantization is the equivalent of using finite fixed virtual loudspeaker layouts. Although they might not provide an optimal layout, the implementation is straightforward. Below we include the rendering steps for an arbitrary grid. We further ex-

plored the minimum grid resolution for a Lebedev grid in perceptual tests (see Sec. 5).

Defining $\mathbf{D}_Q$ as a matrix containing the directions of an arbitrary grid in cartesian coordinates, the original DOA matrix can be quantized by finding the closest directions in the quantized grid

$$q(n) = \underset{n\in i,\ldots,N}{\arg\min}\{d(\mathbf{D}(n), \mathbf{D}_Q)\} \tag{13}$$

where q refers to the indices corresponding to the closest directions. Then, following a similar approach to Eq. (12), a final matrix of quantized DOAs $\mathbf{D}_{pq}$ can be defined.

$$\mathbf{D}_{pq}(n) = \begin{cases} \mathbf{D}_{ds}(n) & n \le i \\ \mathbf{D}_Q(q(n)) & n > i \end{cases} \tag{14}$$

Note that in the matrix $\mathbf{D}_{pq}$ the direction of the direct sound is enforced to be constant, as described in Sec. 4.1.1. At this point, using $\mathbf{D}_{pq}$ as the input variable in Eq. (8) and solving Eqs. (9) and (10) results in a re-synthesized BRIR with the post-processed DOAs as explained in this section.

4.2 Reverb Equalization

Direct auralization of an RIR using DOA data to either map the energy to discrete loudspeakers or generate BRIRs [as in Eq. (10)] results in a perceivable spectral whitening of those parts of the RIR with unreliable DOA estimation. When DOA estimates fluctuate randomly, single-band limited sound events are mapped onto disparate locations, resulting in broadband sound events. This is especially important in the late reverberation tail, resulting in an increase of the reverberation at high frequencies [5] or in environments with high echo density, such as small rooms or a car cabin [7]. An analysis of this rendering artifact and a time-frequency equalization to compensate for it were introduced in [7]. This equalization approach uses the pressure RIR p as a reference to generate a time-varying filter for each of the rendered directions. This is especially useful when SDM is used for loudspeaker-based auralization, as only a relatively low number of directional streams need to be equalized. However in binaural rendering with dense HRTF datasets this approach becomes impractical from a computing and memory perspective.

In this section we introduce the RTMod and RTMod+AP methods, which correct the reverberation time by acting on the BRIRs directly without using directional feeds as an intermediate step. The main idea of RTMod is to decompose the BRIR into fractional octave bands, modify the energy envelope of each subband separately, and finally reconstruct the broadband BRIR. The RTMod+AP variant is based on the same concept, but it processes the output signals through a cascade of 3 Schroeder Allpass filters to increase the echo density of the late reverberation.

Note that this approach is specifically designed for rendering directly into binaural signals. When using Ambisonics as an intermediate format, time-frequency equalization can be done in the spherical harmonics domain [16].

4.2.1 RTMod Equalization

To generate the band limited components of the BRIR we use the same implementation of a perfect reconstruction filter bank [32] found in the SDM Toolbox [18]. Assuming that the time-frequency deviations of the rendered BRIRs with regard to the original pressure RIR are not time dependent we can manipulate the energy envelope of each subband by using exponential functions.

$$\mathbf{BRIR}^u_{\text{corr}}(t) = \sum_{f=1}^{F} \mathbf{BRIR}_{\text{corr}^u,f}(t) \tag{15}$$

$$\mathbf{BRIR}^u_{\text{corr},f}(t) = \mathbf{BRIR}^u_f(t) e^{-t(d_{1,f}-d_{0,f})} \tag{16}$$

where $\mathbf{BRIR}_{\text{corr}^u}$ is the corrected BRIR and f refers to each frequency band. The constants $d_{1,f}$ and $d_{0,f}$ determine the amount of correction of each subband envelope and are determined using the RT_{60} of the pressure RIR and the BRIR.

$$d_{0,f} = \frac{\ln(10^6)}{2\,RT_{60,\text{resynth},f}} \tag{17}$$

$$d_{1,f} = \frac{\ln(10^6)}{2\,RT_{60,\text{orig},f}} \tag{18}$$

where $\text{RT}_{60,\ \text{resynth},\ f}$ and $\text{RT}_{60,\ \text{orig},\ f}$ refer to the reverberation time of band f of the uncorrected BRIR and pressure RIR, respectively.

After applying RTMod equalization the reverberation time of the resulting BRIRs are within one JND unit, which is defined as 5% of the RT_{60} according to ISO 3382, at most frequencies (see and Sec. 4.3 and [6] for a more detailed analysis of the equalization). However, due to the violations of the sound-field model, the late reverberation presents a more coarse fine envelope, likely due to consecutive events interfering constructively and destructively. Through informal listening we concluded that these artifacts are largely negligible when auralizing continuous signals but audible when rendering highly impulsive sounds.

4.2.2 RTMod+AP Equalization

One way to reduce the signal-dependent quality of the late reverb is by increasing the echo density of the late reverberation to achieve a more smooth decay. The goal is to break up strong specular reflections formed by constructive interference of multiple reflections due to incorrectly estimated DOA into multiple reflections.

Allpass (AP) filters have been extensively used in audio for decorrelation [33–35] or artificial reverberation [36–39] for their ability to act as impulse expanders. As such, when AP filters are designed as Schroeder Allpass sections [36, 37], they can be effectively used to increase the echo density of the late reverberation without affecting its spectral properties. This results in an RIR with a smoother time envelope. Additionally, if both left and right channels are processed with the same filters, the Inter-Aural Cross Correlation (IACC) is left unaffected. We propose the use of a cascade of 3 Schroeder Allpass filters (see Fig. 9) to process the late reverb of the broadband corrected BRIRs (RTMod+AP).

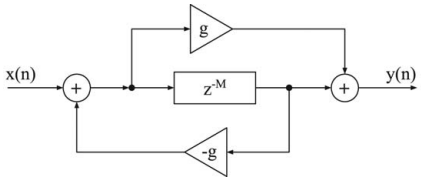

Fig. 9. Block diagram of a Schroeder Allpass filter.

The design of the Schroeder Allpass sections is based on two parameters—a delay (M) and gain (g). The values of these parameters can be largely based on artificial reverberation design. As such, delays must be coprime to minimize strong modulation effects. In the present paper we have obtained satisfactory results using delays of 37, 113, and 215 (on experiments run at a sampling rate of $f_s = 48$ kHz). The gain g can be set by using a desired reverberation time for the filters (RT_{60filt}).

$$g_{dB} = -60\frac{M}{f_s RT_{60\,filt}} \tag{19}$$

$$g = 10^{g_{dB}/20} \tag{20}$$

By setting relatively short reverberation times (in the order of 0.1 s) the RT60 of the BRIRs is largely unaffected by the Schroeder Allpass filters.

In order to process only the late reverberation, where spatial information is largely incorrect, we split the BRIRs at the mixing time [28] and the late reverb is processed using the Schroeder AP filter cascade. This effectively increases the diffuseness of the late reverberation without significantly changing its energy or IACC. Finally, the early reflections and processed late reverberation are summed back together using cosine ramps in the cross-fading region.

The choice of optimal number of filters and their parameters might be application dependent. Additionally the use of dynamic filter parameters could simplify the implementation and avoid explicitly dividing the BRIRs into early response and late reverberation. Adaptive filtering is used in SIRR to generate the time-varying diffuse component of the RIR. A similar application to the present approach warrants further research.

4.3 Instrumental Validation

To compare the performance of the equalization methods we compared a dummy-head reference measurement (KEMAR) with BRIR renderings generated using HRTF measurements of the same mannequin—source to the left (70° azimuth, 4° elevation). We used a microphone array of 10-cm diameter with a central microphone and 6 microphones arranged in pairs on orthogonal axes, an analysis window of 62 samples and a moving average window of 16 samples to smooth the DOA estimates. In the renderings we quantized the DOAs to 50 directions using a Lebedev grid while keeping the first 160 samples fixed to the original direct sound direction.

For the RTMod+AP equalization we used a mixing time of 3,800 samples (80 ms) and crossfade ramps of 1,024 samples. The filter delays were fixed to 37, 113, and 215

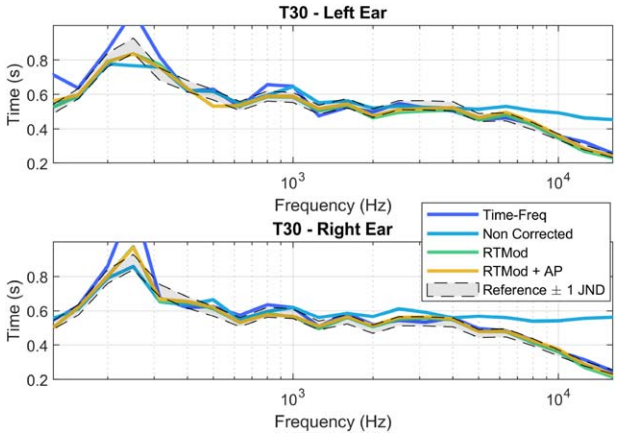

Fig. 10. Reverberation time (T30) of various reverberation equalization techniques.

samples and their reverberation time was 0.1 s. The entire rendering process, from pressure RIR and DOA data to equalized RTMod+AP BRIRs, takes approximately 0.15 s on a laptop PC (Intel Core i7, 7th gen) running Matlab 2018 and Windows 10. In comparison the rendering using time-frequency equalization from [7] takes 8.9 s. Note that these refer to the rendering of one BRIR, corresponding to one arbitrary head orientation.

Examples of estimated T_{30} for a BRIR processed with the presented methods are shown in Fig. 10. In this case, the reference T_{30} is obtained from the pressure RIR. It is clear that the non-corrected case, implemented as in Eq. (10), yields an excessive reverberation time above 4 kHz. The time-frequency equalization from [7] presents T30 results closer to the reference, although overestimated at low frequencies (approx. 250 Hz) and around 1 kHz. Finally, both RTMod [as in Eqs. (15) and (16)] and RTMod+AP methods present the closest results to the reference. Both RTMod and RTMod+AP present T_{30} errors smaller than the strictest accepted value of reverberation time JND (5% per ISO 3382-1:2009) over nearly the entire frequency range.

IACC has been linked to the perceived spatial quality of concert hall acoustics [40]. We computed the IACC for all the equalization methods on the full BRIR as well as early (0 to 80 ms) and late (80 ms to end) portions (see Fig. 11). Although discrimination thresholds and perceptual interpretation of IACC are topics of current research, we utilize a JND value of 0.075 as defined in ISO 3382-1:2009 for reference. The greatest deviations for all three methods are at low and mid frequencies, below 1 kHz. The time-frequency method presents the highest error at all ranges and portions of the RIR. Both RTMod and RTMod+AP methods provide a significant improvement, with deviations within ± 1 JND across the entire spectrum except at one band for the early part of the BRIR. At the late reverb deviations increase slightly at low frequencies. Note that the RTMod and RTMod+AP methods are almost equivalent in the early part of the BRIR, as the allpass filter cascade is only applied to the late reverberation (with a fade-in ramp). Additionally the negligible differences when comparing RTMod and

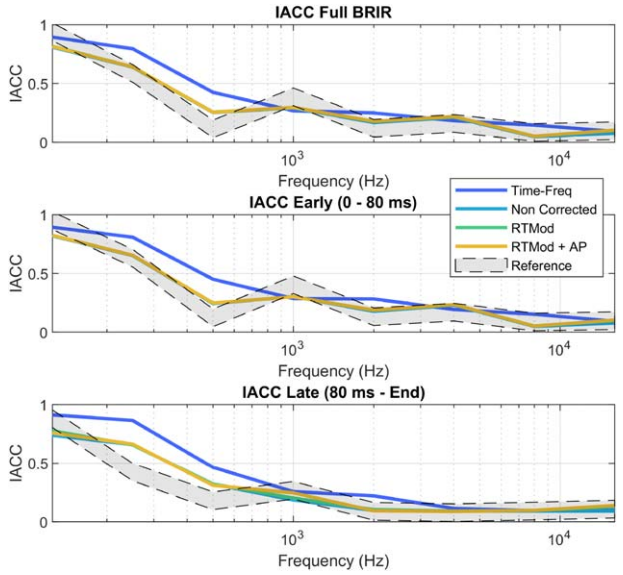

Fig. 11. Inter-Aural Cross Correlation of the reference and rendered BRIRs with various equalization methods.

RTMod+AP to the non-corrected BRIRs demonstrate that the spatial properties of the equalized BRIRs are preserved through the equalization method, which acts only on the time-energy properties.

Through this section we demonstrated the benefits of both RTMod and RTMod+AP over the uncorrected BRIRs and those with time-frequency equalization as in [7]. However neither T_{30} nor IACC analysis demonstrate the benefits of the AP addition to the RTMod equalization. As discussed extensively through the manuscript, a sound field description based on a succession of specular events is violated in the late reverberation. This results in a noisy time envelope, as the DOA estimates of the late reverberation are less reliable and multiple events interfere with each other.

Since allpass filters maintain the magnitude response of a signal while introducing changes in the phase, they can be effectively used to modify the fine time structure of a BRIR. In Fig. 12 we present the late reverberation amplitude envelopes for the left ear of the reference re-synthesized BRIRs. Although all the re-synthesized BRIRs present a noisier envelope than the reference, the use of allpass filters in the RTMod+AP method contributes to a smoothing of the envelope. In the studied case we used a cascade of three filters and we hypothesize that parameter tuning (number of filters, delays, decay time) could provide further gains. Informal listening revealed a significant improvement in the perceived similarity between the reference and RTMod+AP, as compared to the other methods. Refinements of the process and a formal perceptual evaluation are left for future work.

5 PERCEPTUAL EVALUATION OF SPATIAL QUANTIZATION

In previous sections we have objectively evaluated the effect of microphone array topology and proposed DOA post-processing and BRIR rendering alternatives to the original SDM implementation that allow rendering of BRIRs with high spatial density HRTF datasets. In this section we present a perceptual study in which we investigate the perceived plausibility of auralizations using DOA post-processing, as introduced in SEC 5.1 and RTMod equalization.

5.1 Implementation

In the experiment we conduct pairwise comparisons of real loudspeakers and renderings with various degrees of spatial resolution in the early reflections and late reverberation. The experiment was conducted in the same space used for objective comparisons (see Figs. 10–12 for acoustical parameters). We generated renderings using the RTMod equalization method (without AP cascade filtering), based on quantized DOA matrices using 7 grids of increasing resolution (1, 2, 6, 14, 26, and 50 points). The lowest resolution (1 point) collapses all the energy to the direction of direct sound. Grids with 2 and 6 points quantize the energy to left/right and left/right, top/bottom, and front/back, respectively. Larger grids (14, 26, and 60 points) are based on Lebedev grids. Spatial energy maps of each variant are shown in Fig. 13. The direct sound was fixed to the original direction for the first 128 samples in all the cases. The HRTF dataset used for rendering was from a KEMAR mannequin, obtained from boundary element method (BEM) simulations with 20,624 directions. The BRIRs were rendered with a resolution of 1° azimuth and 5° elevation.

The real-time rendering was done using a custom Max/MSP patch enabling dynamic rendering of 2 DOF (yaw and pitch) BRIRs. To save computational resources and memory, the reverberation was rendered separately and statically after a conservative mixing time (80 ms). Previous studies have shown that in typical rooms the dynamic rendering of late reverberation is not audible [28].

Tracking was implemented using an OptiTrack system with markers on both the listener and loudspeaker. This effectively enables pseudo-6DOF rendering, i.e., the relative angle between the loudspeaker and listener was always correctly tracked, thus rendering the correct direction for the direct sound. This ensures that small unintended translations of the subjects during the listening test do not result in perceivable localization shifts.

To enable direct comparisons between loudspeakers and binaural renders we used non-occluding headphones (AKG K1000). Although the occlusion from these headphones is arguably smaller than with generic on-ear or over-the-ear headphones, a comparison showed that they exhibit differences of 6 dB at 10 kHz when comparing HRTFs with and without headphones [41]. As the occlusion effect of the headphones is direction dependent, one possible way to compensate for them would be to render the BRIRs using HRIR datasets measured on subjects wearing the headphones. However as the stimuli were generated using generic HRTFs and including the headphone occlusion would not remove its effect from the real loudspeakers we decided not to include it. Informal listening revealed that

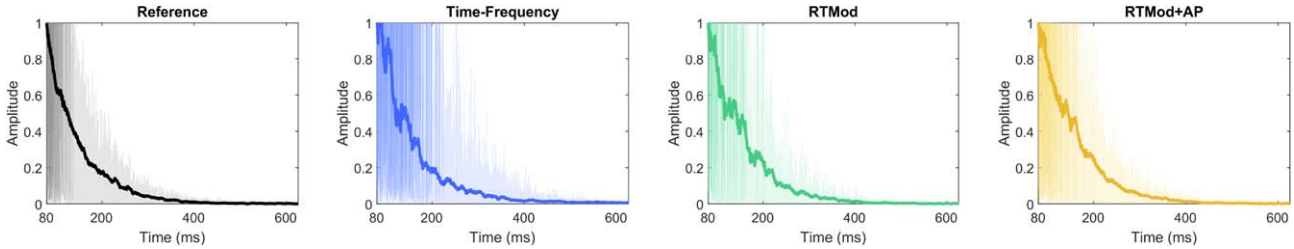

Fig. 12. Left ear absolute pressure signals (thin curves) and envelopes (thick curves) for the Reference BRIR (measured with a KEMAR dummy head) and various equalization methods applied to re-synthesized BRIRs.

the headphone occlusion did not affect the perceived location or spatial properties of the real loudspeakers and was only causing small coloration effects at high frequencies.

All the presented stimuli were band-passed between 200 Hz and 8 kHz using 15th-order Butterworth filters, which largely mitigated the effects of headphone occlusion at high frequencies. Generic headphone compensation filters based on KEMAR measurements using the same pair of headphones were used for all subjects. The filters were generated following the technique from [42].

5.2 Procedure

An increasingly relevant application of binaural rendering of room acoustics is the presentation of virtual sources in augmented reality scenarios along with real sources. Besides traditional similarity metrics, in the recent past dynamic binaural rendering has been evaluated in terms of plausibility [43] or authenticity [44]. While the definition of authenticity is unequivocal, i.e., the evaluated stimulus is perceptually indistinguishable from a reference under all listening conditions, plausibility experiments can be implemented in various degrees of strictness. For instance, in [43] and [45], the plausibility was assessed by using a yes/no test, in which listeners were asked to identify whether the stimulus was presented from a loudspeaker or binaurally using headphones. In this case listeners were hearing renderings corresponding to the room in which the experiments were carried out and were presented explicit versions of the real and virtual audio during the training phase. In comparison, in [46] the criterion is somewhat less strict, as listeners were presented with either simulations or measurement-based auralizations of real spaces and asked whether the stimuli corresponded to real or simulated rooms. In this case the listeners relied entirely on internal references related to plausibility of room acoustics and simulation artifacts that would differentiate real from simulated rooms. When utilizing only internal references, effects such as room acoustical divergence [47] or listener adaptation [48] could lead to changes in perceived externalization, thus affecting the plausibility ratings.

In order to account for plausibility at a stricter degree, we designed a 2-Alternative Forced Choice (2AFC) test in which in each trial subject was presented with two stimuli from the same location and asked to identify which of the two stimuli sounded more plausible. The concept of plausibility was discussed with the subjects and in this task it was equivalent to choosing which of the two simuli they thought corresponded to the sound generated by a real loudspeaker in the room (see Fig. 14). In each of the trials either both stimuli would be virtual or one of them would correspond to a real loudspeaker. This results in a test paradigm that combines plausibility based on the comparison to an internal reference (when two virtual sources are presented) or an explicit reference (when one of the sources is the real loudspeaker).

In order to eliminate potential influences due to visual elements or localization mismatch due to the use of generic HRTFs, the loudspeaker was hidden behind a curtain. The audio content used in the test was a sequence of castanets and only one single source location was used. Although the subjects were not given feedback or explicitly presented the real and virtual stimuli for comparison, they underwent an introduction to the experiment and conversation with the experimenter that allowed them to get acquainted with the natural acoustics of the room.

A group of eight expert listeners without known hearing problems participated in the test. All the subjects had previously participated in listening tests involving binaural audio and were familiar with room acoustic terminology. The total number of trials per subject was 21, resulting from all the possible pair combinations of 7 stimuli without repetitions, with 6 stimuli corresponding to re-synthesized BRIRs and 1 corresponding to the real loudspeaker in the room. The decision of not including repetitions responded to the fact that listeners were highly trained and we aimed at avoiding fatigue during the test.

To ensure subject reliability we provided listeners with unlimited time and instructed them to make use of natural head rotations in order to fully explore the sound scenes before making a decision. Listeners could switch back and forth between the two presented stimuli, which were played in sync and in loop. All the listeners heard the same stimuli, although the order of presentation was randomized. The collection of the responses was done using a touchscreen and GUI (see Fig. 14), minimizing the interaction between the subjects and experimenters.

5.3 Results

The results of the test are a decision matrix for each subject, corresponding to all the comparisons in the pairwise test design. By adding the rows of the matrix we can obtain the total number of selections of each stimulus. We use this

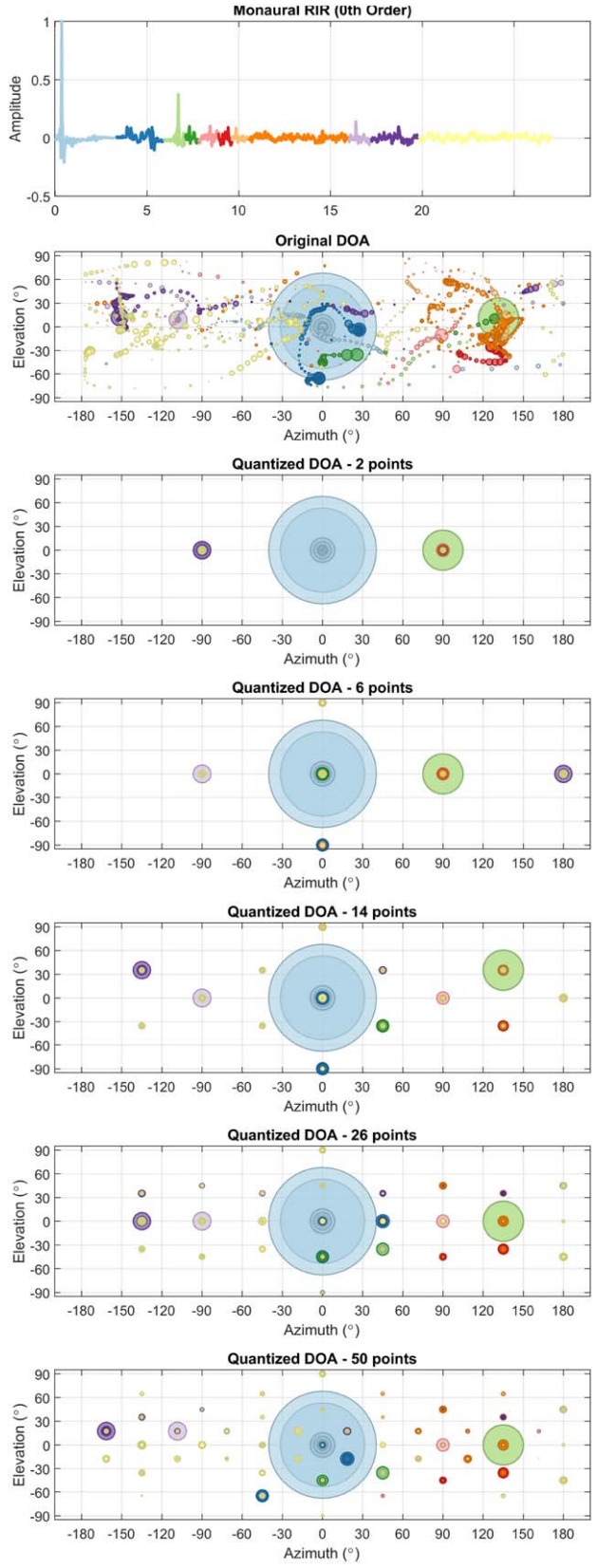

Fig. 13. Spatial energy maps of the renderings included in the listening test.

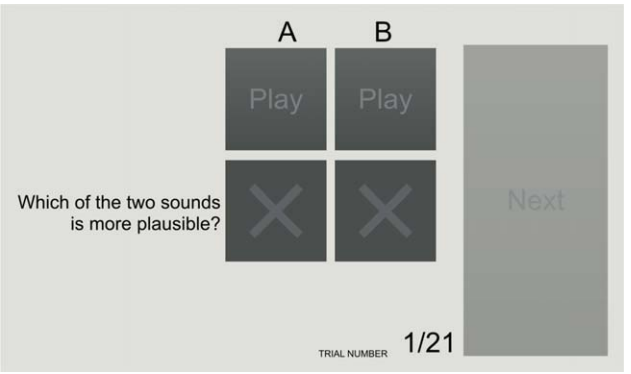

Fig. 14. GUI of the 2-AFC plausibility test.

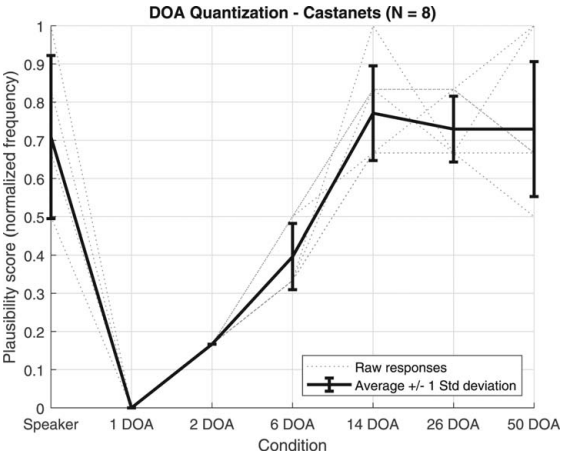

Fig. 15. Perceived plausibility of a real loudspeaker and SDM binaural renderings with various degrees of spatial resolution.

value and normalize it by the total number of presentations of each stimulus to obtain a 'Plausibility Score' P_i,

$$P_i = \frac{\sum_{j=1}^{N_s} a_{i,j}}{N_s - 1} \tag{21}$$

where $a_{i,j}$ is the response to comparison of stimuli $i$and j and has a value of 1 if i is selected as more plausible than j (and 0 if vice versa). N_s refer to the total number of stimuli (7 in this case). A P_i value of 1 would indicate that stimulus i is always selected as being more plausible than the rest in all cases. The results for each subject and a group average are shown in Fig. 15.

The results suggest that listeners perceive renderings with 14 or more DOAs as being as plausible as the real loudspeaker. The small spread at conditions 1, 2, and 6 DOA suggests that all listeners reliably discriminated between these conditions and the real loudspeaker or higher resolution renderings. In addition to showing that increased DOA resolution is not necessarily relevant in a practical scenario, the results suggest that the rendering improvements based on RTMod allow for the rendering of plausible virtual sources, even in the explicit comparison to real sources. Although the perceptual results in this specific scene are clear—there is no perceptual benefit in using more than 14 DOAs for rendering—it would be beneficial to conduct

similar tests in a variety of environments to ensure that they are generalizable.

6 DISCUSSION

The spatial decomposition method was initially developed for the research of concert hall acoustics [2, 3], and the public availability of a toolbox for Matlab caused a recent surge in its usage. It is nowadays applied in all kinds of room acoustics-related research, including car cabin acoustics [7, 8], stage acoustics [5, 14], audio-visual perception in virtual reality [12], speech intelligibility [11], dynamic binaural rendering [16, 6, 20], and acoustic preference in small rooms [9], among others. However it is a parametric method and despite its generalized use there is a lack of extensive perceptual validation and context-dependent optimization in the literature.

In an attempt to disentangle the effects of each step in the entire process and its application to binaural rendering we reviewed each stage separately. We focused on the case in which a sound field is generated in an enclosed space (small or medium room) by a single broadband source. The choice of microphone array and analysis parameters has a clear impact on the results of the DOA analysis. Although the B-format method can perform much better in simulations, the applicability to a real scenario might be case dependent and influenced by the encoding of the raw signals into B-format signals. Although we have not evaluated the possibility of performing a band-limited DOA estimation in several bands, this has been explored in the literature and applied in various studies [7, 9] using the TDOA approach. We want to note the fact that performing the PIV analysis in multiple bands would in fact converge to the same analysis procedure described in the (first order) SIRR method [19, 23] (although SIRR estimates a diffuse sound field component as well).

While using loudspeaker auralizations makes a comparison with a reference room difficult, binaural re-synthesis of an acoustic environment allows a direct comparison with dummy head recordings (or real spaces if non-occluding headphones are used). Recent studies have reported mixed results when using SDM auralizations. In various studies, authors presented satisfactory experimental results reporting perceptual ratings of SDM-based auralizations as being very similar to reference dummy head measurements [20, 16, 21, 17]. However all of these studies utilize custom implementations of the rendering part. In [20, 21], Ahrens explicitly mentions modifications to the time-frequency equalization to avoid perceivable time aliasing. In [16, 17], Zaunschirm et al. implement two variants, one based on a process similar to the one we describe in Eq. (10) and another based on Ambisonics upmixing. Studies using the original implementation found in the SDM Toolbox have reported more significant and case-dependent differences [24, 15, 23]. We thus want to draw attention to the equalization process as a critical factor in the quality of final renders.

We showed that both RTMod and RTMod+AP methods provide a substantial objective improvement as compared to the original time-frequency equalization. However they also increase the number of parameters in the rendering stage and may thus have a potential impact on the robustness of the method. The critical step is correctly estimating the reverberation time of both the original RIR and pre-corrected BRIRs. Although we have informally completed extensive perceptual validation of both variants of the proposed equalization we recognize the need for further formal evaluation including various acoustic spaces and content in order to evaluate the generalization of the observed improvements.

The fact that auralizations with spatial information quantized to 14 directions are perceived as equally plausible as a real loudspeaker suggests that the spatial resolution of the reverberation can be aggressively reduced without incurring perceptual degradations. Recent investigations led to similar conclusions when comparing Ambisonics renderings with full spatial resolution for the direct sound and reduced order for the reverberation [49]. A DOA clustering approach based on the perceptual relevance of prominent reflections could lead to further reduction of the needed number of directions used to render reflections.

7 CONCLUSIONS

In this paper we presented an optimization of SDM for the binaural auralization of multichannel RIRs, including the optimization of SDM analysis parameters, reduction of spatial resolution of the rendered BRIRs, and implementation of a new equalization approach for binaural renderings.

ISM simulations suggest that for a sampling rate of 48 kHz, an open array with a diameter close to 10 cm with an analysis window between 36 and 64 samples provides the lowest DOA error. For the case of B-format analysis, the results with simulations are significantly better than with an open array. However imperfections in A-to-B conversion in practical applications are not captured in simulations, and it is not possible to generalize the results from the B-format array to measurements.

Measurements with a Tetramic suggest that both TDOA and PIV methods are suitable to estimate the DOA of the strongest events in an RIR. When using PIV, longer convolution windows (lower low-pass cutoff frequencies) are preferred to obtain more stable DOA estimations. Choosing optimal window sizes might be case and array dependent and warrants more research.

We presented a reverberation equalization approach (RTMod+AP) composed of a Reverberation Time Modification (RTMod) step and an Allpass (AP) cascade filtering, yielding better objective results than the state of the art equalization for SDM at much lower computational cost.

Perceptual results suggest that equalization with RTMod provides perceptually plausible results when comparing dynamic binaural auralizations to real loudspeakers. Complete perceptual evaluation of RTMod+AP is left for future work.

The same perceptual experiments reveal that quantizing the early reflections and late reverberation DOA estimates to a Lebedev grid of 14 points does not result in perceptual

degradations when compared to denser grids, even with the use of a static late reverberation tail.

Future work includes the investigation of alternative methods to reduce the spatial resolution of DOA estimates and improve timbral preservation. Besides DOA quantization, time-varying or energy-informed clustering approaches could be explored to further reduce the spatial requirements without incurring perceptual impairments. A systematic analysis comparing both objective and perceptual performance of multiband directional analysis would help inform optimal parameters in a wider range of scenes. Finally full listening tests comparing the performance of RTMod+AP with other equalization approaches are also part of future work.

8 ACKNOWLEDGMENT

We want to thank Henrik Hassager, Nils Meyer-Kahlen, and Prof. Tapio Lokki for fruitful discussions on the work and valuable feedback on the manuscript. We are also indebted to the anonymous reviewers, whose contribution greatly improved the general quality of the paper.

9 REFERENCES

[1] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28 (2013 Jan.).

[2] 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 (2013 Jan.), https://doi.org/10.1121/1.4770260.

[3] T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo, "Concert Hall Acoustics: Repertoire, Listening Position, and Individual Taste of the Listeners Influence the Qualitative Attributes and Preferences," *J. Acoust. Soc. Am.*, vol. 140, no. 1, pp. 551–562 (2016 Jul.), https://doi.org/10.1121/1.4958686.

[4] S. V. Amengual Garí, J. Pätynen, and T. Lokki, "Physical and Perceptual Comparison of Real and Focused Sound Sources in a Concert Hall," *J. Audio Eng. Soc.*, vol. 64, no. 12, pp. 1014–1025 (2016 Dec.), https://doi.org/10.17743/jaes.2016.0035.

[5] S. V. Amengual Garì, D. Eddy, M. Kob, and T. Lokki, "Real-Time Auralization of Room Acoustics for the Study of Live Music Performance," presented at the *Fortschritte der Akustik - DAGA 2016* (2016 March).

[6] S. V. Amengual Garí, W. O. Brimijoin, H. G. Hassager, and P. W. Robinson, "Flexible Binaural Resynthesis of Room Impulse Responses for Augmented Reality Research," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 161–166 (2019 Sept.), https://doi.org/10.25836/sasp.2019.31.

[7] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki, "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 (2015 Nov.), https://doi.org/10.17743/jaes.2015.0080.

[8] N. Kaplanis, S. Bech, S. Tervo, J. Pätynen, T. Lokki, T. van Waterschoot, and S. H. Jensen, "A Method for Perceptual Assessment of Automotive Audio Systems and Cabin Acoustics," presented at the *AES 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)* (2016 Jan.), conference paper 6-3.

[9] N. Kaplanis, S. Bech, T. Lokki, T. van Waterschoot, and S. Holdt Jensen, "Perception and Preference of Reverberation in Small Listening Rooms for Multi-Loudspeaker Reproduction," *J. Acoust. Soc. Am.*, vol. 146, no. 5, pp. 3562–3576 (2019 Nov.), https://doi.org/10.1121/1.5135582.

[10] S. Tervo, P. Laukkanen, J. Pätynen, and T. Lokki, "Preferences of Critical Listening Environments Among Sound Engineers," *J. Audio Eng. Soc.*, vol. 62, no. 5, pp. 300–314 (2014 May), https://doi.org/10.17743/jaes.2014.0022.

[11] O. Kokabi, F. Brinkmann, and S. Weinzierl, "Prediction of Speech Intelligibility Using Pseudo-Binaural Room Impulse Responses," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. EL329–EL333 (2019 Apr.), https://doi.org/10.1121/1.5099169.

[12] A. Saurì Suárez, N. Kaplanis, S. Serafin, and S. Bech, "In-Virtualis: A Study on the Impact of Congruent Virtual Reality Environments in Perceptual Audio Evaluation of Loudspeakers," presented at the *2019 AES International Conference on Immersive and Interactive Audio* (2019 Mar.), conference paper 67.

[13] J. Pätynen, S. Tervo, and T. Lokki, "Amplitude Panning Decreases Spectral Brightness With Concert Hall Auralizations," in *Proceedings of the AES 55th International Conference: Spatial Audio* (2014 Aug.), conference paper P-13.

[14] S. V. Amengual Gari, M. Kob, and T. Lokki, "Analysis of Trumpet Performance Adjustments Due to Room Acoustics," in *Proceedings of the International Symposium on Room Acoustics (ISRA)*, pp. 65–73 (2019 Sept.).

[15] S. V. Amengual Garì, W. Lachenmayr, and E. Mommertz, "Spatial Analysis and Auralization of Room Acoustics Using a Tetrahedral Microphone," *J. Acoust. Soc. Am.*, vol. 141, no. 4, pp. EL369–EL374 (2017 Apr.), https://doi.org/10.1121/1.4979851.

[16] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR Synthesis Using First-Order Microphone Arrays," presented at the *144th Convention of the Audio Engineering Society* (2018 May), convention paper 9944.

[17] 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 (2020 Feb.), https://doi.org/10.3390/app10051631.

[18] S. Tervo and J. Pätynen, "SDM Toolbox for Matlab," https://www.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox, accessed: 2020-06-29.

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

[20] J. Ahrens, "Perceptual Evaluation of Binaural Auralization of Data Obtained From the Spatial Decomposition Method," in *Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 65–69 (2019 Oct.), https://doi.org/10.1109/WASPAA.2019.8937247.

[21] J. Ahrens, "Auralization of Omnidirectional Room Impulse Responses Based on the Spatial Decomposition Method and Synthetic Spatial Data," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 146–150 (2019 May), https://doi.org/10.1109/ICASSP.2019.8683661.

[22] 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 (2006 Feb.).

[23] 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. 338–354 (2020 May), https://doi.org/10.17743/jaes.2020.0026.

[24] C. Hold, *Spatial Decomposition Method on Non-Uniform Reproduction Layouts*, *Master's thesis*, TU Berlin (2019 Aug.).

[25] L. McCormack, A. Politis, O. Scheuregger, and V. Pulkki, "Higher-Order Processing of Spatial Impulse Responses," in *Proceedings of the 23rd International Congress on Acoustics*, pp. 4909–4916 (2019 Sept.).

[26] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950 (1979), https://doi.org/10.1121/1.382599.

[27] F. Brinkmann and S. Weinzierl, "AKtools—An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics," presented at the *142nd Convention of the Audio Engineering Society* (2017 May), convention paper 309.

[28] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 60, no. 11, pp. 887–898 (2012 Nov.).

[29] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. De Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The Replica Dataset: A Digital Replica of Indoor Spaces," *arXiv preprint arXiv:1906.05797* (2019).

[30] O. Puomio, J. Pätynen, and T. Lokki, "Optimization of Virtual Loudspeakers for Spatial Room Acoustics Reproduction With Headphones," *Appl. Sci.*, vol. 7, no. 12, p. 1282 (2017 Dec.), https://doi.org/10.3390/app7121282.

[31] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, vol. 34, pp. 226–231 (1996).

[32] J. Antoni, "Orthogonal-Like Fractional-Octave-Band Filters," *J. Acoust. Soc. Am.*, vol. 127, no. 2, pp. 884–895 (2010 Feb.), https://doi.org/10.1121/1.3273888.

[33] G. S. Kendall, "The Decorrelation of Audio Signals and Its Impact on Spatial Imagery," *Comp. Music J.*, vol. 19, no. 4, pp. 71–87 (1995).

[34] M. Bouéri and C. Kyriakakis, "Audio Signal Decorrelation Based on a Critical Band Approach," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), convention paper 6291.

[35] E. Kermit-Canfield and J. Abel, "Signal Decorrelation Using Perceptually Informed Allpass Filters," in *Proceedings of the 19th International Conference on Digital Audio Effects*, pp. 225–231 (2016 Sept.).

[36] M. R. Schroeder and B. F. Logan, "'Colorless' Artificial Reverberation," *J. Audio Eng. Soc.*, vol. 9, no. 3, pp. 192–197 (1961 Jul.).

[37] M. R. Schroeder, "Natural Sounding Artificial Reverberation," *J. Audio Eng. Soc.*, vol. 10, no. 3, pp. 219–223 (1962 Jul.).

[38] V. Välimäki, J. Parker, and J. S. Abel, "Parametric Spring Reverberation Effect," *J. Audio Eng. Soc.*, vol. 58, no. 7/8, pp. 547–562 (2010 Jul.).

[39] L. Dahl and J. -M. Jot, "A Reverberator Based on Absorbent All-Pass Filters," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)-00* (2000 Dec.).

[40] T. Okano, L. L. Beranek, and T. Hidaka, "Relations Among Interaural Cross-Correlation Coefficient (IACCE), Lateral Fraction (LFE), and Apparent Source Width (ASW) in Concert Halls," *J. Acoust. Soc. Am.*, vol. 104, no. 1, pp. 255–265 (1998 Jun.), https://doi.org/10.1121/1.423955.

[41] C. Pörschmann, J. M. Arend, and R. Gillioz, "How Wearing Headgear Affects Measured Head-Related Transfer Functions," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 49–54 (2019 Sept.), https://doi.org/10.25836/sasp.2019.27.

[42] J. G. Bolaños, A. Mäkivirta, and V. Pulkki, "Automatic Regularization Parameter for Headphone Transfer Function Inversion," *J. Audio Eng. Soc.*, vol. 64, no. 10, pp. 752–761 (2016 Oct.), http://doi.org/10.17743/jaes.2016.0030.

[43] A. Lindau and S. Weinzierl, "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acust. United Acust.*, vol. 98, no. 5, pp. 804–810 (2012 Sept./Oct.), https://doi.org/10.3813/AAA.918562.

[44] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the Authenticity of Individual Dynamic Binaural Synthesis," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1784–1795 (2017 Oct.), https://doi.org/10.1121/1.5005606.

[45] C. Pike, F. Melchior, and T. Tew, "Assessing the Plausibility of Non-Individualised Dynamic Binaural Synthesis in a Small Room," presented at the *AES 55th Interna-*

tional Conference: Spatial Audio (2014 Aug.), conference paper 6-1.

[46] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A Round Robin on Room Acoustical Simulation and Auralization," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2746–2760 (2019 Apr.), https://doi.org/10.1121/1.5096178.

[47] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A Summary on Acoustic Room Divergence and Its Effect on Externalization of Auditory Events," in *Proceedings of the Eighth International Conference on Quality of Multimedia Experience (QoMEX 2016)* (2016 Jun.), https://doi.org/10.1109/QoMEX.2016.7498973.

[48] F. Klein, S. Werner, and T. Mayenfels, "Influences of Training on Externalization of Binaural Synthesis in Situations of Room Divergence," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 178–187 (2017 Mar.), https://doi.org/10.17743/jaes.2016.0072.

[49] I. Engel, C. Henry, S. V. Amengual Garí, P. W. Robinson, D. Poirier-Quinot, and L. Picinali, "Perceptual Comparison of Ambisonics-Based Reverberation Methods in Binaural Listening," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 121–126 (2019 Sept.), https://doi.org/10.25836/sasp.2019.11.

THE AUTHORS


Sebastià V. Amengual


Johannes M. Arend


Paul Calamia


Philip Robinson

Sebastià V. Amengual is currently a research scientist at Facebook Reality Labs Research working on room acoustics, spatial audio, and auditory perception. He received a Diploma Degree in Telecommunications with major in Sound and Image in 2014 from the Polytechnic University of Catalonia (UPC) in 2014, completing his Master's Thesis at the Norwegian University of Science and Technology (NTNU). His doctoral work at the Detmold University of Music focused on investigating the interaction of room acoustics and live music performance using virtual acoustic environments. His research interests lie in the intersection of audio, perception, and music.

•

Johannes M. Arend received a B.Eng. degree in Media Technology from HS Düsseldorf, Düsseldorf, Germany, in 2011 and an M.Sc. degree in Media Technology from TH Köln, Cologne, Germany, in 2014. Since 2015, he has been a Research Fellow and working toward a Ph.D. degree at TH Köln and TU Berlin in the field of spatial audio with a focus on binaural technology, auditory perception, and audio signal processing. Between September 2019 and March 2020 he was a research intern at Facebook Reality Labs Research.

•

Paul Calamia is a research scientist on the Audio Team at Facebook Reality Labs Research, where he conducts research in room acoustics for augmented-reality applications. Previously he was a member of the Technical Staff at MIT Lincoln Laboratory in the Bioengineering Systems and Technologies Group and the Advanced Undersea Systems and Technology Group. His other prior positions include Assistant Professor in the Graduate Program in Architectural Acoustics at Rensselaer Polytechnic Institute in Troy, NY, Consultant and Head of R&D at Kirkegaard Associates in Chicago, IL, and Acoustical Engineer at Wyle Laboratories in Arlington, VA. He holds a bachelor's degree in mathematics from Duke University, a Master's degree in electrical and computer engineering from the Engineering Acoustics Program at the University of Texas at Austin, and a Ph.D. in computer science from Princeton University.

•

Philip Robinson is a research science manager in room acoustics and auditory perception at Facebook Reality Labs Research (FRL Research) in Redmond, WA. Prior to joining FRL Research, he incorporated virtual acoustics simulation and reproduction systems into building design processes at the architecture firm of Foster + Partners. He was a Fulbright Scholar and post-doctoral researcher at Aalto University in Finland, where he studied perception of concert hall acoustics, spatial auditory resolution, and echo thresholds. He has been a visiting researcher at EPFL in Switzerland and Hanyang University in South Korea. He received a Ph.D. from Rensselaer Polytechnic Institute in Troy, NY in 2012. In a previous life, he was a registered architect in his home state of New Mexico. He remains passionate about architecture, the study of which gave him a great interest in perception of environments, real or virtual.