Arend *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:37
https://doi.org/10.1186/s13636-021-00224-5

EURASIP Journal on Audio,
Speech, and Music Processing

RESEARCH Open Access

Efficient binaural rendering of spherical microphone array data by linear filtering

Johannes M. Arend[1,2*†], Tim Lübeck[1,2†] and Christoph Pörschmann[1†]

Abstract

High-quality rendering of spatial sound fields in real-time is becoming increasingly important with the steadily growing interest in virtual and augmented reality technologies. Typically, a spherical microphone array (SMA) is used to capture a spatial sound field. The captured sound field can be reproduced over headphones in real-time using binaural rendering, virtually placing a single listener in the sound field. Common methods for binaural rendering first spatially encode the sound field by transforming it to the spherical harmonics domain and then decode the sound field binaurally by combining it with head-related transfer functions (HRTFs). However, these rendering methods are computationally demanding, especially for high-order SMAs, and require implementing quite sophisticated real-time signal processing. This paper presents a computationally more efficient method for real-time binaural rendering of SMA signals by linear filtering. The proposed method allows representing any common rendering chain as a set of precomputed finite impulse response filters, which are then applied to the SMA signals in real-time using fast convolution to produce the binaural signals. Results of the technical evaluation show that the presented approach is equivalent to conventional rendering methods while being computationally less demanding and easier to implement using any real-time convolution system. However, the lower computational complexity goes along with lower flexibility. On the one hand, encoding and decoding are no longer decoupled, and on the other hand, sound field transformations in the SH domain can no longer be performed. Consequently, in the proposed method, a filter set must be precomputed and stored for each possible head orientation of the listener, leading to higher memory requirements than the conventional methods. As such, the approach is particularly well suited for efficient real-time binaural rendering of SMA signals in a fixed setup where usually a limited range of head orientations is sufficient, such as live concert streaming or VR teleconferencing.

Keywords: Spherical microphone arrays, Binaural rendering, Spatial audio reproduction, Virtual acoustics

1 Introduction

Headphone-based binaural rendering of spatial sound fields is of great importance in the consumer sector for virtual reality (VR) and augmented reality (AR) applications as well as in research areas such as hearing science. Using a spherical microphone array (SMA) is a flexible method to capture a spatial sound field and render it for a single listener over headphones. One possibility is to measure spatial room impulse responses (SRIRs) with an SMA, which can then be used to generate binaural room impulse responses (BRIRs) [1–5]. To auralize the captured sound field, dynamic binaural synthesis is employed, i.e., the generated BRIRs are convolved (in real-time) with anechoic audio material. However, the major advantage of SMAs is that they can be used for real-time rendering of a spatial sound scene, such as a musical performance in a concert hall. In this case, the captured sound field is processed in real-time to generate ear signals that, when presented over headphones, virtually place the listener in the sound

*Correspondence: Johannes.Arend@th-koeln.de
†Johannes M. Arend, Tim Lübeck and Christoph Pörschmann contributed equally to this work.
[1]Institute of Communications Engineering, TH Köln - University of Applied Sciences, Betzdorfer Str. 2, 50679, Cologne, Germany
[2]Audio Communication Group, Technical University of Berlin, Einsteinufer 17c, 10587, Berlin, Germany

field [6–8]. Both methods allow binaural rendering with head tracking, i.e., rendering for arbitrary head orientations of the listener. Furthermore, individual head-related transfer functions (HRTFs) can be employed for binaural rendering of SMA data.

Recent advances in research yielded several solutions for binaural real-time rendering of SMA signals, such as the IEM Plug-in Suite [6, 9], SPARTA [8], and ReTiSAR [7, 10]. The overall concept of these toolboxes is similar. The sound field captured with an SMA is first spatially encoded in real-time, i.e., it is transformed to the spherical harmonics (SH) domain using the discrete SH transform (SHT) [11]. The resulting SH signals are then processed with radial filters, which are array-specific filter functions that compensate for the spatial extent and, in the case of a rigid sphere array, the scattering properties of the array body [3, 4]. A classical approach for binaural decoding of SH signals (also referred to as Ambisonics signals) is the use of so-called virtual loudspeakers [2, 6, 12–14]. By applying the inverse SH transform (ISHT) to the SH signals, spatially uniformly distributed plane waves are generated, which are then weighted with HRTFs of the corresponding directions. More recent methods perform binaural rendering directly in the SH domain, i.e., the HRTF set is transformed to the SH domain and then multiplied with the SH signals of the array [6, 7]. Both rendering methods are usually combined with further pre- or postprocessing methods, such as max-$\mathbf{r}_{\mathrm{E}}$ weighting [15], SH tapering [16], spherical head filters [17], or MagLS [6], to mitigate spatial aliasing and truncation errors caused by spatial discretization of the sound field in SMA capturing (see [5] for an overview of different mitigation approaches).

Real-time binaural rendering of SMA signals in the manner described above is computationally demanding, in particular because of the time-consuming SHT. Due to these performance requirements, the most recent implementation of ReTiSAR, for example, can only render SMA data up to a maximum spatial order of $N = 12$ on a standard laptop [10]. This spatial order corresponds to an SMA with a minimum number of $Q = 169$ microphones ($Q = (N + 1)^2$) and is thus sufficient for most common SMAs available in commercial or scientific contexts, which mostly do not exceed an SH order of $N = 7$ (e.g., em32 Eigenmike [18], Zylia ZM-1 [19], HØSMA [20]). However, content based on sequentially measured higher-order SMA data (see, for example [21] or [22], providing SMA data with $N = 29$ or $N = 44$, respectively) cannot be rendered in real-time with current implementations. Furthermore, approaches that perform spatial upsampling of real-world SMA signals before the SHT to enhance the rendering [23] significantly increase the spatial order and thus the number of audio channels, making real-time rendering of upsampled SMA signals impossible with currently available implementations.

In this paper, we present a simpler and computationally more efficient approach for real-time binaural rendering of SMA signals. As the entire encoding and decoding chain represents a linear time-invariant (LTI) system, it can also be described with a set of finite impulse response (FIR) filters. More precisely, the transmission from each microphone input to the decoded ear signal for the left or right ear can be described by one FIR filter each, resulting in a set of $Q \times 2$ FIR filters required for binaural decoding of SMA signals for one specific head orientation. Those filters can be precomputed (for any desired number of head orientations) for the specific SMA and HRTF configuration and applied to the SMA signals in real-time by fast convolution, similar to dynamic binaural synthesis. Superimposing the output of all filtered SMA signals yields exactly the ear signals produced by any of the conventional encoding and decoding chains described above, given that the settings are the same as for the FIR filter precomputation. According to their functionality to describe the transmission from the array microphones to the ears, we have named these filters SMATBIN (Spherical Microphone Array To Binaural) filters.

The proposed approach significantly reduces the complexity of implementing real-time binaural rendering of SMA signals while also being less computationally demanding. Thus, any existing software or hardware structure for efficient and fast real-time convolution can be used for binaural rendering of SMA signals of a very high order, i.e., with many audio channels. In the following, we first explain the common encoding and decoding chains briefly discussed above in greater detail. We then provide further details on the SMATBIN filter method and explain how the filters can be generated. Next, we compare BRIRs resulting from applying the SMATBIN filters with those resulting from common binaural rendering in two working examples. Finally, we compare the computational complexity and the memory requirements of the proposed approach to that of the common rendering methods and discuss the advantages and disadvantages of using the proposed filters for binaural rendering of SMA signals.

2 Binaural rendering methods

2.1 Virtual loudspeaker approach

The block diagrams in Fig. 1 show two common methods for binaural rendering of SMA data (top and middle) and the proposed approach using the SMATBIN filters (bottom). The block diagram on the top illustrates the classical virtual loudspeaker approach [2, 6, 12–14]. The Q microphone signals captured with an SMA are transformed to the SH domain employing the SHT. The resulting SH

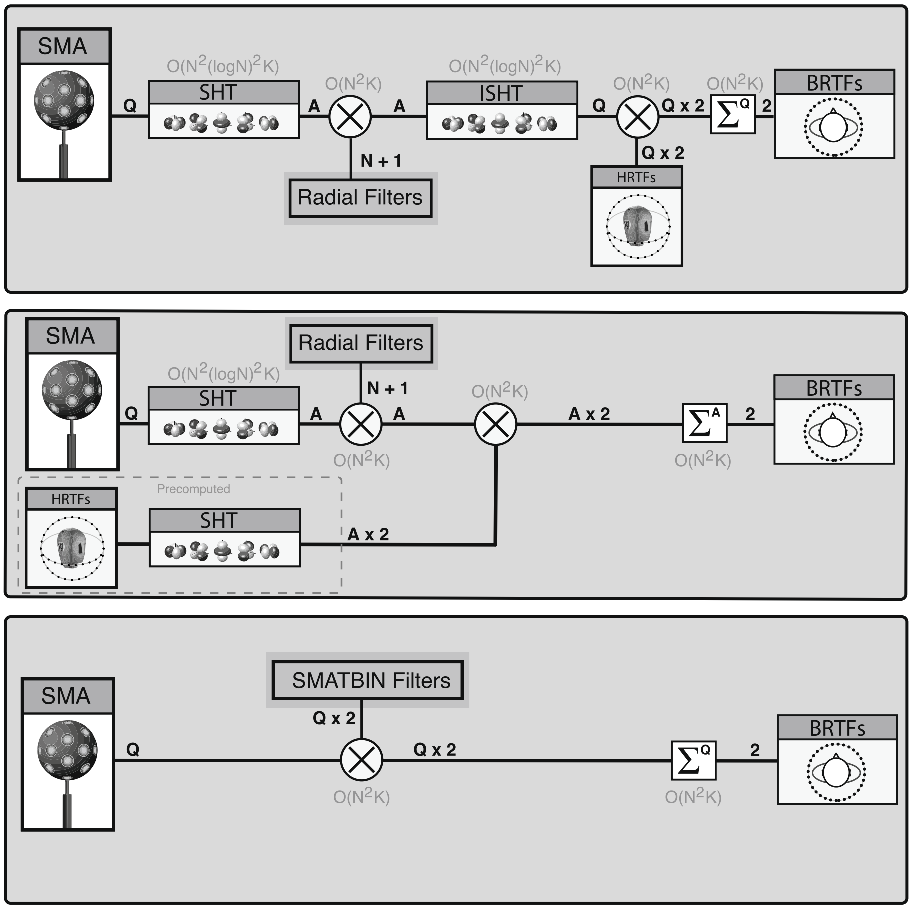

Fig. 1 Block diagrams illustrating the signal processing in the temporal frequency domain for binaural rendering of SMA data using the virtual loudspeaker approach (top), the SH domain approach (middle), and the proposed SMATBIN filter approach (bottom). *N*: spatial order, *Q*: number of microphones, *A*: number of SH channels, *O*: Landau's symbol, *K*: FFT size (top, middle) or SMATBIN filter length (bottom)

signals with A channels are multiplied with $N + 1$ order-dependent radial filters, and then plane waves for specific directions are generated by applying the ISHT. This procedure, known as plane wave decomposition, is usually performed on a spatial sampling grid of the same spatial order N as that of the SMA, but different sampling schemes (e.g., Lebedev or Gaussian schemes with the same order) yield different results [2, 3, 6, 24]. For simplicity, we assume in the present case that the sound field is decomposed to Q plane waves with the directions of the SMA sampling scheme. The Q plane waves are then multiplied with Q HRTFs for the corresponding directions, resulting in the virtual loudspeakers. Finally, the Q spatially weighted plane waves are summed up, yielding the two-channel binaural signal.

The block diagram shows the processing for a single head orientation. Binaural room transfer functions (BRTFs) for arbitrary head orientations can be generated in two different ways. One way is to use HRTFs for directions corresponding to the relative angles between head orientation and the plane wave directions [2, 3, 25]. Alternatively, the sound field can be rotated in the SH domain according to the head orientation before the ISHT [8, 14]. When using complex SH basis functions, the sound field rotation can be performed by Wigner-D weighting [26], whereas for real SH basis functions, a computationally more efficient rotation matrix obtained by recursion relations can be applied [27]. The processing, including switching the HRTFs or rotating the sound field depending on the head orientation, can be performed in real-time so that the spatial sound scene captured with the SMA can be reproduced binaurally in real-time [8].

2.2 Spherical harmonics domain approach

Alternatively, binaural decoding can be directly performed in the SH domain [6, 7, 10], as illustrated in Fig. 1 (middle). As with the virtual loudspeaker approach, the SMA signals are transformed to the SH domain, and radial filters are applied. For binaural decoding, the SH signals of the array with A channels are multiplied with the HRTF set, which was also transformed to the SH domain at the same spatial order N, resulting in an HRTF set with A SH channels. The final BRTF is obtained by summing up all A SH channels. BRTFs for arbitrary head orientations can be achieved by rotating the sound field in the SH domain applying a rotation matrix to the SH signals [26, 27]. All processing can also be done in real-time, enabling dynamic binaural auralization of SMA data [6, 7, 10]. Compared to the virtual loudspeaker approach, calculating the ear signals directly in the SH domain is less computationally expensive because the multiplication with the SH basis functions, which is part of the ISHT, is omitted.

3 SMATBIN filter method

As both above-mentioned encoding and decoding chains represent LTI systems for which the principle of superposition holds, the transmission path from each microphone of the SMA to the left and right ear can be described by a pair of FIR filters. Such a pair of filters can be calculated by applying a unit impulse (Dirac delta) to the respective channel of the SMA, while assigning zeros to the other channels, and performing the usual encoding and decoding as described above. Applying unit impulses to each channel of the SMA successively, while always assigning zeros to the other channels, yields a set of $Q \times 2$ FIR filters – the SMATBIN filters. Algorithm 1 shows the pseudocode for generating SMATBIN filters for one head orientation using either the virtual loudspeaker approach or the SH domain approach for binaural decoding. To generate SMATBIN filters for arbitrary head orientations, rotation must be integrated at the appropriate point in the algorithm. The sound field rotation in the SH domain is implemented after the radial filtering in step 5, whereas the HRTF switching is integrated in step 8 (see also Sections 2.1 and 2.2). Notably, the proposed principle can be used to convert not only the discussed common methods, but any approach for binaural rendering of SMA data, including any of the popular mitigation approaches implemented in the rendering [5], into a set of FIR filters.

The block diagram on the bottom of Fig. 1 shows the simple structure for binaural rendering when using the SMATBIN filters. Each of the Q microphone signals is convolved with the corresponding two-channel filter and then, all Q filtered microphone signals are summed up, yielding the two-channel binaural signal. The approach thus omits the computationally expensive SHT, and real-time binaural rendering can be achieved by an efficient and fast convolution of the SMA signals with the SMATBIN filters. For dynamic binaural auralization, the filter sets are precomputed for a suitably large number of head orientations, resulting in $Q \times 2 \times M$ filters, with M the number of head orientations. In real-time rendering, the SMATBIN filters are selected and switched according to the head orientation, just as any common binaural renderer does.

Algorithm 1 Pseudocode for generating the SMATBIN filters for one head orientation using either the virtual loudspeaker approach or the SH domain approach for binaural decoding. See text for more information on how to integrate head orientations.

```
1:  for $q = 1{:}Q$ do
2:      Apply unit impulse to the $q$-th microphone
3:      Perform FFT
4:      Perform SHT
5:      Perform radial filtering
6:      if Virtual Loudspeaker Approach then
7:          Perform ISHT
8:          Multiply plane waves with HRTFs for same directions
9:          Sum over $Q$ weighted plane waves
10:     else if SH Domain Approach then
11:         Weight SH signals with HRTFs
12:         Sum over $A$ SH channels
13:     end if
14:     Perform IFFT
15:     Export $q$-th SMATBIN filter
16: end for
```

4 Results

4.1 Working examples

To evaluate the proposed method, we implemented two working examples comparing binaural rendering of SMA data using the SMATBIN filters with the rendering chain implemented in the SOFiA toolbox [28]. As all spherical microphone array processing in the present work was performed using SOFiA, and the SMATBIN filters for the two examples were based on the SOFiA rendering chain, the BRTFs/BRIRs produced by the two methods should ideally be identical.

For binaural decoding, we used the `sofia_binauralX` function, which employs the virtual loudspeaker approach in combination with HRTF switching to account for arbitrary head orientations [2, 3]. The HRTFs used in SOFiA are from a Neumann KU100 dummy head measured on a Lebedev grid with 2702 sampling points [29]. The HRTFs are transformed to the SH domain at a sufficiently high order of $N = 35$, allowing artifact-free SH interpolation to obtain HRTFs for any direction corresponding to the directions of the plane waves [3].

For both working examples, the radius of the rigid sphere array was $r = 8.75$ cm, and the radial filter gain was soft-limited to 20 dB [30]. The SMATBIN filter length was defined as $K = 2048$ taps at a sampling rate of $f_s = 48$ kHz. Figure S1 in the supplementary material (Additional file 1) shows an example of SMATBIN filters with

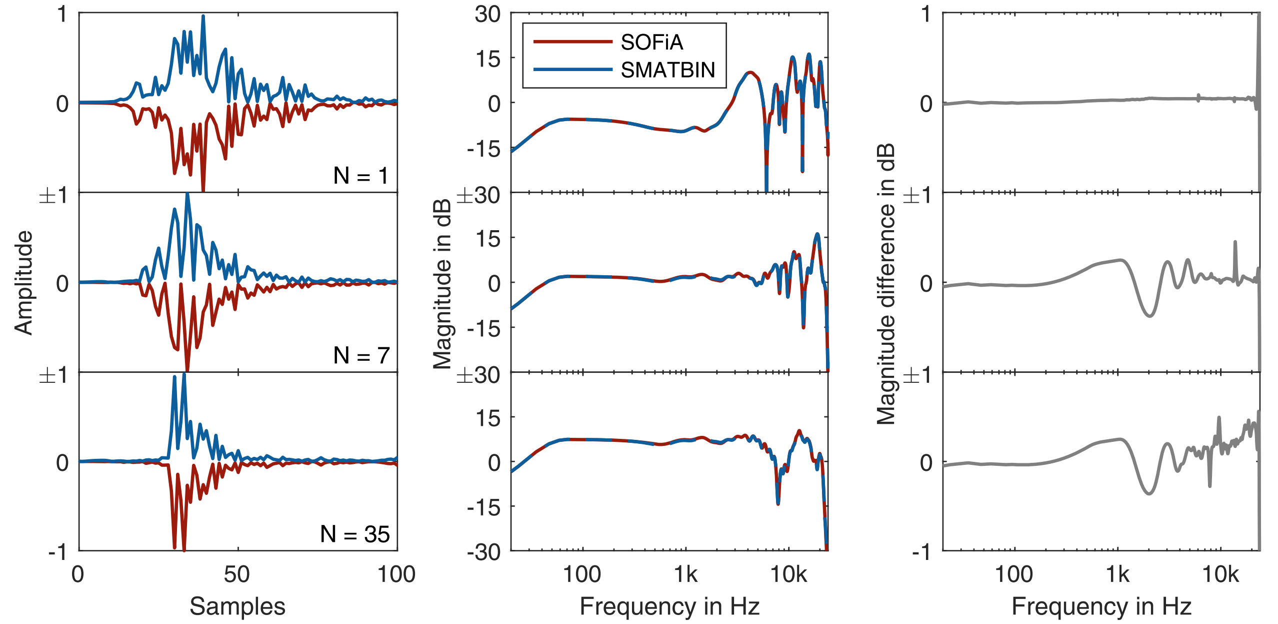

Fig. 2 Working Example 1: Left-ear BRIRs/BRTFs resulting from the SOFiA (reference) and the SMATBIN filter processing ($K = 2048$ taps) for frontal head orientation ($\phi = 0°$, $\theta = 90°$) and a single broadband plane wave incident from the front ($\phi = 0°$, $\theta = 90°$) on three different rigid sphere arrays with Lebedev sampling schemes of orders $N = \{1, 7, 35\}$. For better illustration, the absolute amplitudes of the SOFiA BRIRs in the left column are inverted. The right column shows the magnitude differences between the SOFiA and the SMATBIN BRTFs

the above-mentioned array and filter parameters for a Lebedev sampling scheme of order $N = 1$. The described implementations with functions to calculate the SMATBIN filters and generate results plots, as well as various demo implementations, are available online[1].

4.1.1 *Working example 1*

In the first working example, we simulated a single broadband plane wave incident from the front ($\phi = 0°$, $\theta = 90°$, with ϕ the horizontal angle ranging from 0° to 360° and θ the vertical angle ranging from 0° to 180°) on three different rigid sphere arrays with Lebedev sampling schemes of orders $N = \{1, 7, 35\}$, corresponding to 6, 86, and 1730 sampling points respectively. Besides the more common orders $N = 1$ and $N = 7$, we decided to show the implementation with the rather high order $N = 35$ to verify that no artifacts or instabilities occur even when processing with a very high number of SMATBIN filters. From these SMA signals, we calculated BRIRs using the SOFiA implementation employing plane wave decomposition and virtual loudspeaker rendering (see Fig. 1 (top)) as well as using the proposed SMATBIN filter method where the SMA signals are simply filtered and then superimposed to achieve a BRIR (see Fig. 1 (bottom)).

Figure 2 compares the left-ear BRIRs/BRTFs resulting from the SOFiA and the SMATBIN filter processing, taking frontal head orientation ($\phi = 0°$, $\theta = 90°$) as an example. The absolute amplitudes of the broadband pressure BRIRs (left column) are nearly identical in their overall time-energy structure with matching amplitude and time events. Accordingly, the magnitude frequency responses of the respective BRTFs (middle column) show no considerable differences and are almost identical at all examined spatial orders. Consistent with that, the magnitude differences (right column) are minimal over the entire audible frequency range from 20 Hz to 20 kHz for all examined spatial orders, with a maximum of about ±0.5 dB at higher frequencies.

In further analysis, we compared BRIRs for 360 head orientations in the horizontal plane (1° steps from 0° to 360°), generated based on the SMA signals for a single plane wave incident from the front as described above. For a perception-related evaluation of the spectral deviations, we calculated for each head orientation the absolute energetic difference ΔG between SOFiA and SMATBIN BRIRs in 40 auditory gammatone filter bands between 50 Hz and 20 kHz [31, 32], as implemented in the Auditory Toolbox [33]. Figure 3 shows the so determined left-ear differences on the example of $N = 7$ for all 360 head orientations (gray lines) and averaged over all head orientations (blue line). In general, the differences are minimal and well below an assumed just-noticeable difference (JND) of 1 dB

[1] Available: https://github.com/AudioGroupCologne/SMATBIN

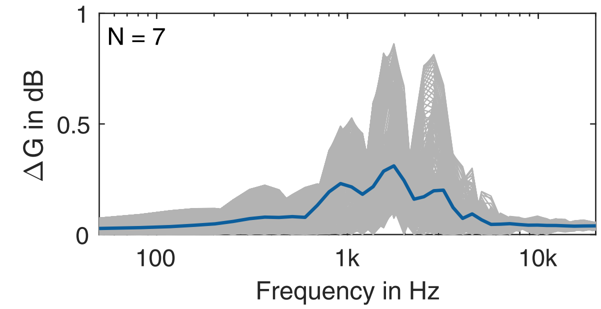

Fig. 3 Working Example 1: Left-ear absolute energetic difference ΔG in auditory gammatone filter bands between BRTFs resulting from the SOFiA (reference) and the SMATBIN filter processing ($K = 2048$ taps) for 360 head orientations in the horizontal plane (1° steps from 0° to 360°) and a single broadband plane wave incident from the front ($\phi = 0°$, $\theta = 90°$) on a rigid sphere array with a Lebedev sampling scheme of order $N = 7$. The gray lines show the difference for each individual head orientation, the blue line shows the difference averaged across all head orientations

[34] and thus can be considered perceptually uncritical. For certain head orientations, the differences reach a maximum of approximately 0.8 dB in the frequency range of about 2-3 kHz. These larger differences occur mainly for lateral sound incidence, i.e., for head orientations in the range of 90° and 270°. Smaller differences with a maximum of approximately 0.3 dB in the range of 2-3 kHz occur for frontal and rear sound incidence, i.e., for head orientations in the range of 0° and 180°. The average difference across head orientations is generally very small, but increases slightly towards mid frequencies, reaching a maximum of approximately 0.3 dB at about 2 kHz.

4.1.2 Working example 2

In the second working example, we evaluated the proposed method using measured SMA data of a real, more complex sound field. Specifically, we employed data captured with the VariSphear measurement system [35] on a Lebedev grid of order $N = 44$ in a classroom at TH Köln [22]. The shoebox-shaped classroom has a volume of 459m^3 and a mean reverberation time of about 0.9 s (0.5 - 8 kHz). The sound source was a Neumann KH420 loudspeaker, placed at a distance of about 4.50 m and a height of 1.40 m in front of the VariSphear array. We spatially resampled the measurements to Lebedev grids of orders $N = \{1, 7, 35\}$ using SH interpolation. From these (resampled) SMA data, we calculated BRIRs using the SOFiA rendering chain as well as the SMATBIN filter method.

Figure 4 compares the left-ear BRIRs/BRTFs for frontal head orientation generated using SOFiA or the SMATBIN filter method. Also for the complex sound field, the time-energy structure of the two broadband pressure BRIRs (left column) is almost identical. Consequently, the 1/6-octave smoothed magnitude responses (middle column) are largely identical for all spatial orders examined, and the magnitude differences (right column) are minimal, with a

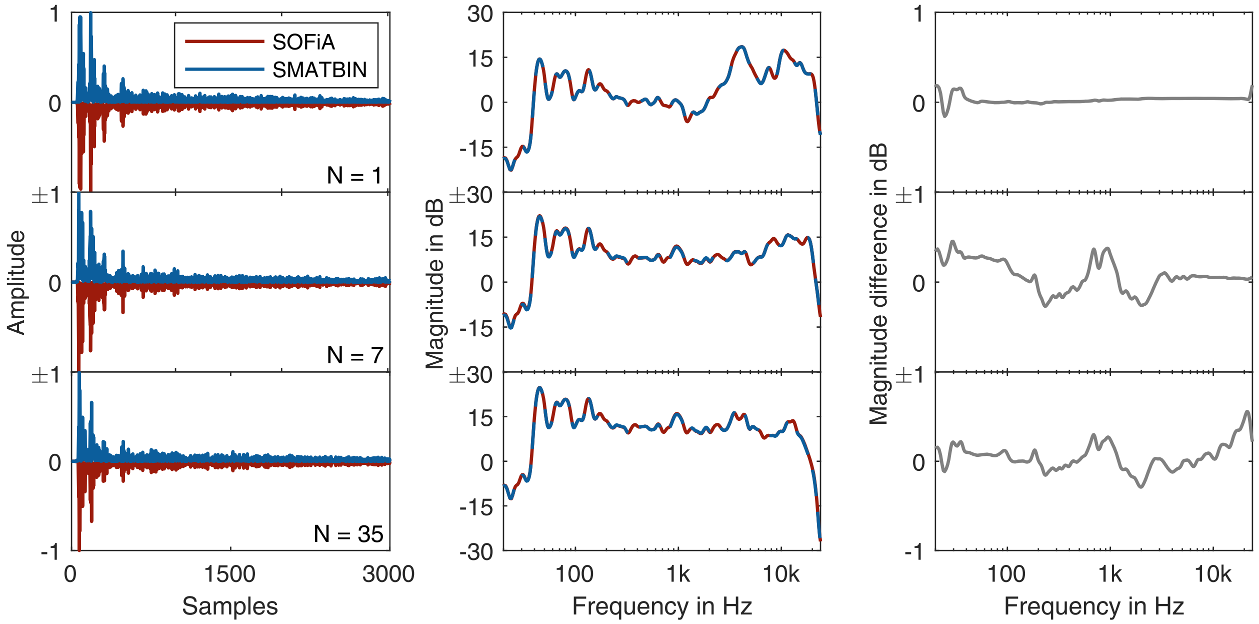

Fig. 4 Working Example 2: Left-ear BRIRs/BRTFs resulting from the SOFiA (reference) and the SMATBIN filter processing ($K = 2048$ taps) for frontal head orientation ($\phi = 0°$, $\theta = 90°$) and impulse responses of a classroom for three different rigid sphere arrays with Lebedev sampling schemes of orders $N = \{1, 7, 35\}$. For better illustration, the absolute amplitudes of the SOFiA BRIRs in the left column are inverted and the magnitudes of the BRTFs in the middle column are 1/6-octave smoothed. The right column shows the magnitude differences between the SOFiA and the SMATBIN BRTFs

maximum range of about ±0.5 dB over the entire audible frequency range.

The analysis of the absolute energetic difference ΔG across 360 head orientations in the horizontal plane and selected SH order $N = 7$ revealed differences that should be perceptually uncritical as they are clearly below the assumed JND of 1 dB (see Fig. 5). At frequencies below 100 Hz and in the range between 500 Hz and 3 kHz, the differences for certain head orientations reach a maximum of about 0.4 dB, but decrease again above 3 kHz. The average difference across head orientations does not exceed 0.2 dB in the entire audible frequency range and even tends towards 0 dB at frequencies above 3 kHz.

4.1.3 Interim summary

The results of the two working examples clearly show that the presented approach can be used equivalently to the established but much more complex virtual loudspeaker approach for binaural rendering of SMA data or for generating BRIRs from SMA measurements. Theoretically, the result of the two compared methods should even be completely identical. In practice, however, minimal differences between the binaural signals can occur because of the filter design, i.e., because of the necessary further processing of the filters after sampling the rendering chain with unit pulses, such as windowing, truncation, or delay compensation.

The supplementary material (Additional file 1) contains further BRIR/BRTF plots for Working example 1 for the (more application-oriented range of) orders $N = \{1, 3, 7\}$, selected SMATBIN filter lengths, and selected head orientations. Similar to Fig. 2, the results of the SOFiA and SMATBIN renderings are nearly identical as long as the SMATBIN filters have a sufficient number of filter taps. If the number of FIR filter taps is too small (approximately below 512 taps), deviations from the reference occur in the low-frequency range (<100 Hz) because of insufficient frequency resolution. The SMATBIN filter length can thus be used to adjust the accuracy of the binaural reproduction (compared to the reference) in the low-frequency range, but also the required computing power and memory requirements, as the computing effort for the real-time convolution as well as the required memory space depends on the number of filter taps.

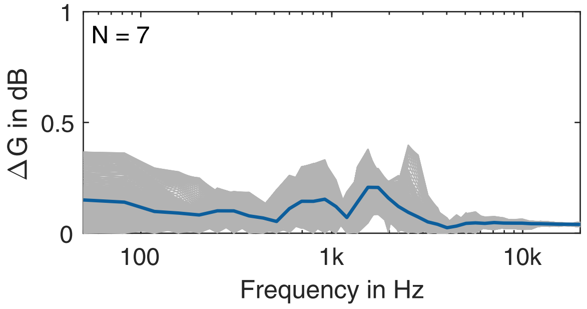

Fig. 5 Working Example 2: Left-ear absolute energetic difference ΔG in auditory gammatone filter bands between BRTFs resulting from the SOFiA (reference) and the SMATBIN filter processing ($K = 2048$ taps) for 360 head orientations in the horizontal plane (1° steps from 0° to 360°) and impulse responses of a classroom for a rigid sphere array with a Lebedev sampling scheme of order $N = 7$. The gray lines show the difference for each individual head orientation, the blue line shows the difference averaged across all head orientations

4.2 Computational complexity

In particular, towards higher orders N, the SHT dominates the computational complexity[2] of the common virtual loudspeaker and SH domain approaches. As the SHT must be performed for each frequency bin, it scales linearly with the filter length or FFT size K. The SMATBIN filter approach omits the SHT and reduces the entire encoding and decoding chain to linear filtering and summation (see Fig. 1), thereby decreasing the complexity for binaural rendering of SMA data, as detailed in the following.

The conventional SHT has a complexity of $O(N^4 K)$ and thus, the calculation effort increases rapidly as a function of spatial order N [36]. Optimized methods for performing the SHT with reduced complexity still require $O(N^2 (logN)^2 K)$ or $O(N^{\frac{5}{2}} (logN) K)$ steps, depending on the optimization [36, 37].

All other processing steps for binaural rendering of SMA data depend on N only with $O(N^2)$. The FFT and IFFT, required in all rendering methods to transform the SMA signals to frequency domain and the binaural signals to time domain, respectively, both have a complexity of $O(N^2 K\, logK)$. Linear filtering in frequency domain, which in the present case corresponds to either applying the radial filters to the SH signals or the SMATBIN filters to the SMA signals, has a complexity of $O(N^2 K)$, and summing up all channels also has a complexity of $O(N^2 K)$.

Thus, the SHT has the highest complexity depending on N in the entire rendering chain, and especially for large N, its calculation effort significantly exceeds that of all other processing steps. As a result, by omitting the SHT, the SMATBIN filter method allows a more efficient binaural rendering of SMA data than the conventional methods.

4.3 Memory requirements

The lower computational complexity of the SMATBIN filter method comes at the cost of higher memory requirements, as a set of filters must be precomputed and stored for each required head orientation. To estimate by example how much more memory the SMATBIN filter approach requires compared to the virtual loudspeaker or

[2] With the term computational complexity, we refer in the following to running time or time complexity.

SH domain approach, we assume in the following an SMA with a Lebedev sampling scheme of order $N = 12$, i.e., $Q = 230$ microphones and $A = 169$ SH channels, a bit depth of 32 bit, i.e., $P = 4$ bytes per filter tap, and a filter length of $K = 2048$ taps.

For the virtual loudspeaker approach, $N+1$ radial filters with a length of K taps and $2\times D$ HRTF filters with a length of L taps must be stored, with D the number of directions of the HRTF set. The total memory requirement calculates as $[\,((N+1)\times K + 2\times D\times L)\times P]$. Assuming an HRTF set with $D = 2702$ directions and $L = 128$ taps, the virtual loudspeaker approach requires 2.9 MB.

With the SH domain approach, only $2\times A$ HRTF filters in the SH domain need to be stored in addition to the radial filters. Here, the total memory requirement calculates as $[\,((N+1) + A\times 2)\times K\times P]$, which also results in 2.9 MB.

In the case of the SMATBIN filter method, the required memory scales with the number of microphones Q and the number of head orientations M. The total memory requirement calculates as $[\,Q\times 2\times M\times K\times P]$. Assuming that, as is often the case, only head orientations in the horizontal plane with a sufficiently high resolution of 2° are rendered [38], yields $M = 180$ head orientations and a total memory requirement of 678 MB. Thus, the SMATBIN filter method requires significantly more memory space than the other two methods, but is computationally less demanding. Accordingly, it must be decided on a case-by-case basis, depending on the technical requirements of a rendering system, whether memory space can be sacrificed for a lower computational load.

5 Discussion

Real-time binaural rendering of SMA data is currently being intensively researched and is becoming increasingly important for various VR and AR applications. However, common rendering methods are extremely computationally demanding, especially for high-order SMAs, and require quite sophisticated real-time signal processing. With the SMATBIN filter method, we presented in this paper a less computationally demanding approach for real-time binaural rendering of SMA data. The presented method allows representing any common rendering chain as a set of precomputed FIR filters, which are then applied to the SMA signals in real-time using fast convolution to generate the binaural signals. As the rendering process is reduced to simple linear filtering with a two-channel FIR filter per SMA channel, it is easier-to-implement using any existing hardware and software solution for fast convolution. Established binaural renderers, such as the SoundScape Renderer [39] or PyBinSim [40], are well suited for this purpose, as they already have implemented methods for optimal filter switching according to the listener's head orientation (see the demo implementation using the SoundScape Renderer in the SMATBIN repository[1]).

The technical evaluation results clearly show that the SMATBIN filer method can be used equivalently to the conventional methods. Thus, BRIRs generated with SMATBIN filtering were almost identical to BRIRs generated with the common virtual loudspeaker method [2, 3]. Furthermore, we showed that by omitting the costly SHT, rendering with SMATBIN filters has significantly lower computational complexity and is thus less computationally demanding than, for example, the virtual loudspeaker or SH domain approach [7, 10]. However, example calculations showed that the lower computational cost of the SMATBIN filter method comes along with considerably higher memory requirements than those for the virtual loudspeaker or SH domain approaches.

The advantages of lower computational complexity are not only accompanied by higher memory requirements, but also by less flexibility. As the SMATBIN filters are always precomputed for a specific SMA configuration with specific HRTFs, neither the SMA nor the HRTFs can be exchanged quickly and flexibly within an application without recalculating the filters or loading a complete precomputed filter set for the changed configuration. Moreover, the en- and decoding are no longer decoupled, and basic SH domain processing such as beamforming, sound field rotation, or spatial effects applied to the sound field in the SH domain, as available in the IEM Plug-in Suite [6, 9], are not possible at all.

Apart from our proposed method, there are alternative filtering methods for binaural rendering of microphone array captures that also omit to transform the sound field to the SH domain. One example is the virtual artificial head [41, 42], which is a filter-and-sum beamformer based on a planar microphone array with 24 microphones used to generate BRIRs. Another recent approach is beamforming-based binaural reproduction [43], with the concept of generating BRIRs directly from signals of arbitrary array structures (spherical or planar) by applying beamforming filter structures. Interestingly, depending on the parameterization of the beamformer, the results are equivalent to SH processing. For example, when using an SMA, the array output of a maximum directivity beamformer corresponds to a plane wave decomposition for the look-direction [11, 43]. Unlike the proposed filter method, however, to the best of our knowledge none of the beamformer methods have been implemented for real-time rendering of array streams, but only for generating BRIRs that are then used for auralization using dynamic binaural synthesis. That said, comparing the performance and computational demands of different beamforming-based methods with the SMATBIN filter approach in a real-time framework would be an interesting study for future research.

Although the proposed method and beamforming-based methods share some similarities in terms of using a specific filter structure for binaural rendering, sampling SH-based rendering chains as performed with the SMATBIN filter approach has some advantages. For one thing, many aspects of SH processing are well understood, both technically and perceptually, such as the required grid resolution, the frequency characteristics of the beams, or the behaviour of spatial aliasing [11], to name a few. These findings can be used to create optimized rendering chains, which can than be sampled and stored as FIR filters for more efficient binaural rendering. Furthermore, there are several approaches to mitigate undersampling errors when using real-world SMAs (e.g., max-$\mathbf{r}_\mathrm{E}$ weighting [15], SH tapering [16], spherical head filters [17], or MagLS [6]), which can also be sampled as part of the rendering chain using the SMATBIN filter method and integrated into a real-time implementation. Thus, using SMATBIN filters makes it possible to integrate any mitigation approach (under development) into a real-time framework without extensive modifications of the real-time processing chain. More specifically, any rendering chain, no matter how complex, which may be difficult to implement in real-time, can be sampled using the presented method and used for real-time rendering using a standard convolution engine.

The presented method offers advantages, particularly for fixed SMA to binaural chains that should be rendered efficiently. Due to its lower computational complexity, the SMATBIN filter method enables real-time rendering of high-order SMA signals ($N > 12$, which is currently the maximum feasible order with the real-time renderer ReTiSAR [10] on a standard laptop). In an informal pilot study, we implemented dynamic binaural rendering of 12th-order SMA signals on a standard laptop (Apple MacBook Pro 15 Mid 2018, Intel Core i7 2,6 GHz) using the SMATBIN filter approach in combination with the SoundScape Renderer for fast convolution and Cockos REAPER for audio playback of the multi-channel stream, resulting in a CPU load of only about 26% on average. Thus, using SMATBIN filters should also enable real-time rendering of SMA signals that are first spatially upsampled to improve the quality [23], which significantly increases the spatial order and thus the number of audio channels. However, a direct objective comparison of the SMATBIN filter approach with other real-time rendering chains such as ReTiSAR [10] or SPARTA [8] in terms of required computational power is not easily possible, as the implementations as well as the frameworks in which the renderers run differ too much to obtain meaningful results.

Overall, the SMATBIN filter method is particularly well suited for real-time binaural rendering of SMA signals in VR or AR applications where the setup is fixed and the focus is on computationally efficient and pristine binaural reproduction of the sound field. Similar to recordings with a dummy head, the SMATBIN filter approach does not allow any further processing of the sound field. Thus, live concert streaming or VR teleconferencing, for example, which require no further processing and typically only a limited range of head orientations, could benefit from the presented method. For related consumer applications, which often do not require any flexible change in setup, the SMATBIN filter approach could even be embedded in a hardware system, enabling highly efficient binaural rendering of SMA signals in real-time.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13636-021-00224-5.

Additional file 1: The supplementary material provides additional results figures.

Acknowledgements

Not applicable.

Authors' contributions

All authors contributed equally to this work. JMA, TL, and CP implemented the method in Matlab and drafted the manuscript. JMA and TL created all the figures. TL and JMA implemented the Matlab and SoundScape Renderer demos. CP conceptualized the work. All authors read and approved the final manuscript.

Funding

The authors acknowledge funding by the German Federal Ministry of Education and Research (BMBF) under project reference 03FH014IX5-NarDasS and by the by European Regional Development Fund (ERDF) under project reference EFRE-0801444. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The supplementary material (Additional file 1) provides further results plots. The Matlab implementation of the proposed method is available on https://github.com/AudioGroupCologne/SMATBIN. The repository provides functions to calculate SMATBIN filters for arbitrary rigid sphere array configurations and head orientations as well as functions to generate results plots. Furthermore, the repository includes demo implementations for binaural rendering of simulated and measured SMA data using the proposed SMATBIN filter approach as well as an integration example demonstrating real-time binaural rendering of a commercially available SMA using SMATBIN filters and the SoundScape Renderer [39].

Declarations

Consent for publication

All authors agree to the publication in this journal.

Competing interests

The authors declare that they have no competing interests.

Received: 18 May 2021 Accepted: 30 September 2021
Published online: 06 November 2021

References

1. A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, B. Rafaely, Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution. J. Acoust. Soc. Am. **133**(5), 2711–2721 (2013). https://doi.org/10.1121/1.4795780

2. B. Bernschütz, A. V. Giner, C. Pörschmann, J. M. Arend, Binaural Reproduction of Plane Waves With Reduced Modal Order. Acta Acust. united Ac. **100**(5), 972–983 (2014). https://doi.org/10.3813/AAA.918777
3. B. Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*. (Doctoral dissertation, TU Berlin, 2016). https://doi.org/10.14279/depositonce-5082
4. J. Ahrens, C. Andersson, Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre. J. Acoust. Soc. Am. **145**(4), 2783–2794 (2019). https://doi.org/10.1121/1.5096164
5. T. Lübeck, H. Helmholz, J. M. Arend, C. Pörschmann, J. Ahrens, Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data. J. Audio Eng. Soc. **68**(6), 428–440 (2020). https://doi.org/10.17743/jaes.2020.0038
6. F. Zotter, M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording*. (Springer, Cham, Switzerland, 2019), p. 210. https://doi.org/10.1007/978-3-030-17207-7
7. H. Helmholz, C. Andersson, J. Ahrens, in *Proc. of the 45th DAGA*. Real-Time Implementation of Binaural Rendering of High-Order Spherical Microphone Array Signals (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2019), pp. 1462–1465
8. L. McCormack, A. Politis, in *Proc. of the AES International Conference on Immersive and Interactive Audio*. SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods (Audio Engineering Society, Inc., New York, 2019), pp. 1–12
9. IEM Plugin Suite. https://plugins.iem.at. Accessed 09 Sept 2021
10. H. Helmholz, T. Lübeck, J. Ahrens, S. V. A. Garí, D. L. Alon, R. Mehra, in *Proc. of the 46th DAGA*. Updates on the Real-Time Spherical Array Renderer (ReTiSAR) (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2020), pp. 1169–1172
11. B. Rafaely, *Fundamentals of Spherical Array Processing*. (Springer, Berlin, Germany, 2015), p. 193. https://doi.org/10.1007/978-3-662-45664-4
12. A. McKeag, D. McGrath, in *Proc. of the 6th AES Australian Regional Convention*. Sound Field Format to Binaural Decoder with Head Tracking (Audio Engineering Society, Inc., New York, 1996), pp. 1–9
13. J.-M. Jot, V. Larcher, J.-M. Pernaux, in *Proc. of the 16th International AES Conference on Spatial Sound Reproduction*. A Comparative Study of 3-D Audio Encoding and Rendering Techniques (Audio Engineering Society, Inc., New York, 1999), pp. 281–300
14. M. Noisternig, A. Sontacchi, T. Musil, Robert Höl, in *Proc. of the 24th AES International Conference on Multichannel Audio - The New Reality*. A 3D Ambisonics Based Binaural Sound Reproduction System (Audio Engineering Society, Inc., New York, 2003), pp. 1–5
15. F. Zotter, M. Frank, All-round ambisonic panning and decoding. J. Audio Eng. Soc. **60**(10), 807–820 (2012)
16. C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, I. J. Tashev, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Improving Binaural Ambisonics Decoding by Spherical Harmonics Domain Tapering and Coloration Compensation (IEEE, New York, 2019), pp. 261–265. https://doi.org/10.1109/ICASSP.2019.8683751
17. Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, Spectral equalization in binaural signals represented by order-truncated spherical harmonics. J. Acoust. Soc. Am. **141**(6), 4087–4096 (2017)
18. mh acoustics, em32 Eigenmike. https://mhacoustics.com/products. Accessed 09 Sept 2021
19. Zylia, ZM-1. https://www.zylia.co/zylia-zm-1-microphone. Accessed 09 Sept 2021
20. O. Moschner, D. T. Dziwis, T. Lübeck, C. Pörschmann, in *Proc. of the 148th AES Convention*. Development of an Open Source Customizable High Order Rigid Sphere Microphone Array (Audio Engineering Society, Inc., New York, 2020), pp. 1–5
21. P. Stade, B. Bernschütz, M. Rühl, in *Proc. of the 27th Tonmeistertagung - VDT International Convention*. A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios (Verband Deutscher Tonmeister e.V. (VDT), Cologne, 2012), pp. 1–17
22. T. Lübeck, J. M. Arend, C. Pörschmann, in *Proc. of the 47th DAGA*. A High-Resolution Spatial Room Impulse Response Database (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2021), pp. 1–4
23. C. D. Salvador, S. Sakamoto, J. Treviño, Y. Suzuki, in *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR)*. Enhancing binaural reconstruction from rigid circular microphone array recordings by using virtual microphones (Audio Engineering Society, Inc., New York, 2018), pp. 1–9
24. Z. Ben-Hur, J. Sheaffer, B. Rafaely, Joint sampling theory and subjective investigation of plane-wave and spherical harmonics formulations for binaural reproduction. Appl. Acoust. **134**, 138–144 (2018). https://doi.org/10.1016/j.apacoust.2018.01.016
25. G. Enzner, M. Weinert, S. Abeling, J.-M. Batke, P. Jax, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Advanced System Options for Binaural Rendering of Ambisonics Format (IEEE, New York, 2013), pp. 251–255. https://doi.org/10.1109/ICASSP.2013.6637647
26. B. Rafaely, M. Kleider, Spherical Microphone Array Beam Steering Using Wigner-D Weighting. IEEE Signal Process. Lett. **15**, 417–420 (2008). https://doi.org/10.1109/LSP.2008.922288
27. J. Ivanic, K. Ruedenberg, Rotation Matrices for Real Spherical Harmonics. Direct Determination by Recursion. J. Phys. Chem. **100**(15), 6342–6347 (1996). https://doi.org/10.1021/jp953350u
28. B. Bernschütz, C. Pörschmann, S. Spors, S. Weinzierl, in *Proc. of the International Conference on Spatial Audio (ICSA)*. SOFiA Sound Field Analysis Toolbox (Verband Deutscher Tonmeister e.V. (VDT), Cologne, 2011), pp. 8–16
29. B. Bernschütz, in *Proc. of the 39th DAGA*. A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100 (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2013), pp. 592–595
30. B. Bernschütz, C. Pörschmann, S. Spors, S. Weinzierl, in *Proc. of the 37th DAGA*. Soft-Limiting der modalen Amplitudenverstärkung bei sphärischen Mikrofonarrays im Plane Wave Decomposition Verfahren (Soft Limiting of the Modal Amplitude Gain for Spherical Microphone Arrays Using the Plane Wave Decomposition Method) (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2011), pp. 661–662
31. F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, S. Weinzierl, A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses. J. Audio Eng. Soc. **67**(9), 705–718 (2019). https://doi.org/10.17743/jaes.2019.0024
32. J. M. Arend, F. Brinkmann, C. Pörschmann, Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions. J. Audio Eng. Soc. **69**, 104–117 (2021). https://doi.org/10.17743/jaes.2020.0070
33. M. Slaney, Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work, Version 2. Technical Report #1998-010. Interval Research Corporation (1998). https://engineering.purdue.edu/~malcolm/interval/1998-010/
34. F. Brinkmann, S. Weinzierl, in *Proc. of the AES Conference on Audio for Virtual and Augmented Reality (AVAR)*. Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition (Audio Engineering Society, Inc., New York, 2018), pp. 1–10
35. B. Bernschütz, C. Pörschmann, S. Spors, S. Weinzierl, in *Proc. of the 36th DAGA*. Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio (Design and Construction of a Variable Spherical Microphone Array for Research in Room Acoustics and Virtual Audio) (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2010), pp. 717–718
36. J. R. Driscoll, D. M. Healy, Computing Fourier Transforms and Convolutions on the 2-Sphere. Adv. Appl. Math. **15**(2), 202–250 (1994). https://doi.org/10.1006/aama.1994.1008
37. M. J. Mohlenkamp, A Fast Transform for Spherical Harmonics. J. Fourier Anal. Appl. **5**, 158–184 (1999). https://doi.org/10.1007/bf01261607
38. A. Lindau, S. Weinzierl, in *Proc. of the EAA Symposium on Auralization*. On the Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical and Lateral Direction (European Acoustics Association (EAA), Espoo, 2009), pp. 1–6
39. M. Geier, J. Ahrens, S. Spors, in *Proc. of the 124th AES Convention*. The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods (Audio Engineering Society, Inc., New York, 2008), pp. 1–6
40. A. Neidhardt, F. Klein, N. Knoop, T. Köllmer, in *Proc. of the 142nd AES Convention*. Flexible Python tool for dynamic binaural synthesis applications (Audio Engineering Society, Inc., New York, 2017), pp. 1–5
41. E. Rasumow, M. Blau, S. Doclo, S. Van De Par, M. Hansen, D. Puschel, V. Mellert, Perceptual Evaluation of Individualized Binaural Reproduction Using a Virtual Artificial Head. J. Audio Eng. Soc. **65**(6), 448–459 (2017). https://doi.org/10.17743/jaes.2017.0012

42. M. Fallahi, M. Hansen, S. Doclo, S. Van De Par, D. Püschel, M. Blau, Evaluation of head-tracked binaural auralizations of speech signals generated with a virtual artificial head in anechoic and classroom environments. Acta Acust. **5**(30), 1–18 (2021). https://doi.org/10.1051/aacus/2021025
43. L. Madmoni, J. Donley, V. Tourbabin, B. Rafaely, in *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR)*. Beamforming-based Binaural Reproduction by Matching of Binaural Signals (Audio Engineering Society, Inc., New York, 2020), pp. 1–8

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.