AUTHENTIC AURALIZATION OF ACOUSTIC SPACES BASED ON SPHERICAL MICROPHONE ARRAY RECORDINGS

J. Ahrens Chalmers University of Technology, Gothenburg, Sweden
H. Helmholz Chalmers University of Technology, Gothenburg, Sweden
C. Andersson Chalmers University of Technology, Gothenburg, Sweden

http://www.ta.chalmers.se/research/audio-technology-group/

1 INTRODUCTION

Ever since the invention of technology to capture and playback audio recordings, researchers and practitioners have been investigating ways to capture are re-create the sound of acoustic spaces[1,2]. New waves of enthusiasm tended to break out with every milestone in the development of new hardware. Historic reports on demonstrations of live-versus-recorded sound using rather basic equipment, such as the 1910 road show of Thomas Edison showcasing his Diamond Disk Phonograph, confirm that the audience did not notice when the switch between presentation modes occurred. The famous quote "*the ear could not tell when it was listening to the phonograph alone, and when to actual voice and reproduction together. Only the eye could discover the truth by noting when the singer's mouth was open or closed*" stems from a 1916 article in New York Evening Mail on an according demonstration at Carnegie Hall New York[3].

Many modern authors have expressed their doubt on the success of modern repetitions of such demonstrations using the original equipment[4]. The general audience nowadays is more accustom to both electroacoustic equipment as well as live performances so that their judgement may be expected to be more critical.

A system that aims at re-creating an audio experience can be designed in two ways: One can (I) aim at controlling the signals that arise at the user's ears directly – for example by means of headphones –, so called *head-related* reproduction. Or, one can (II) aim at producing a sound field that evokes the desired ear signals when impinging on a listener, which may be termed *room-related* reproduction[5].

Up until today, loudspeaker systems have been created that comprise up to 800 individual channels and use advanced processing methods to synthesize the physical structure of a desired sound field[6]. Even such systems have not been successful in producing an *authentic* experience, i.e., an experience that is indistinguishable from the original. Still, these systems can create a highly *plausible* experience, which is an experience that might not be indistinguishable from the original but that is highly believable[7]. Laymen might also use the term *realistic* in this context. The reasons why authenticity cannot be achieved with room-related systems are manifold and include limitations in the recording methods as well as the influence of the room in which the loudspeaker system is installed.

Recordings with dummy heads – i.e., a manikin that is equipped with microphones at the locations of the eardrums – have been considered very useful ever since the first explorations at Bell Laboratories and Philipps in the 1930s[8]. This concept directly avoids both main drawbacks of room-related systems: the limitations of the recording methods as well as the response of the reproduction room.

However, recordings performed with a dummy head can suffer from a lack of spatial fidelity. This is mainly because a fixed head orientation is encoded in the ear signals so that head movements of the listener upon playback cannot be taken into account. Hence, one speaks of *static* binaural playback. A circumvention of this drawback has been achieved by not recording a scenario live but measure the ear impulse responses of static sound sources (i.e., loudspeakers) in a given venue for different head orientations. Upon playback, the instantaneous orientation of the listener's head can be tracked with a sensor, and a given source signal can be convolved with the impulse responses that

correspond to the instantaneous orientation. One then speaks of *dynamic* or *head-tracked* binaural reproduction. The result is a tremendous improvement in terms of spatial fidelity[9,10] as far as the binaural playback being almost or fully indistinguishable from a real sound source[11].

Note that the signals from a dummy head, no matter whether recorded live or measured, have the acoustic response of the dummy encoded in them. This acoustic response is termed head-related transfer function (HRTF) and can be significantly different from the acoustic response of the individual listener. Discussions in how far this affects the result and the reproduction quality are complicated and on-going. A slightly simplified guideline informally supported by some researchers might read as follows: If all sound sources in the scene are located inside the horizontal plane and careful equalization is being applied, then the impairment that has to be expected is usually low or very low. Other researchers disagree.

The major limitation of data-based (i.e. measurement based) binaural re-synthesis is the fact that no live recordings are possible since all data needs to be measured under time-invariant conditions. One may also mention the fact that it is very impractical to perform the measurement with a real human to invoke individual HRTFs. The employed measured data brings along the requirement for recording program material without any spatial information. Firstly, this can be tedious for certain kinds of sources or for an extended number of sound sources. And secondly, it captures the signal only at one distance and incidence direction of the source (if only one microphone is being used), whose reproduction will evoke a different emulated room response than the real source.

The generalization, and currently the most general approach available according to the present authors' opinion, is the dynamic binaural auralization of spherical microphone array recordings, which has the potential to overcome all of the limitations and drawbacks that were mentioned above. The remainder of this paper outlines the underlying concept and presents an overview of recent advances as well as the remaining challenges.

2 SPHERICAL MICROPHONE ARRAYS

2.1 Concept

Spherical microphone arrays like the one depicted in Fig. 1 are able to capture the spatial structure of a sound field, i.e., they capture the time signal that a sound field carries as well as geometric information such as the propagation direction and curvature of the wave fronts. In other words, one obtains an acoustic fingerprint of the sound field. HRTFs sets, particularly when acquired under anechoic conditions, represent the response of the ear to sound incidence as a function of the direction of incidence. In other words, one obtains an acoustic fingerprint of the human ear.

Putting it in simple words, one has available all geometric information of a sound field as well as the transfer function from any sound field to the ears of a human. This makes it possible to compute the signals that would arise at the ears of the listener – the person whose HRTFs are available – when the listener is exposed to the sound field that was captured by the microphone array. The convenience of the underlying concept is that both the sound field as well as the HRTFs can be rotated so that the listener's ear signals can be computed for arbitrary head orientations. This may be performed in real-time so that head-tracking can be applied and – under ideal conditions – the signals at the ears of the listener are identical to the signals that would arise when listening to the original sound field no matter what head rotations are being performed. The complete processing pipeline then represent a virtual head that is placed in the sound field with arbitrary orientation.

2.2 Mathematics

The mathematical foundation of the signal processing pipeline is the concept of spherical harmonics decomposition[12]. Spherical harmonics form an orthogonal set of basis functions for any function of

finite energy that is defined on the surface of a sphere. The according coefficients then represent the function under consideration independent of the position on the sphere.

Figure 1: An experimental 64-channel microphone array with a size similar to that of a bowling ball (photograph by Karim Helwani)

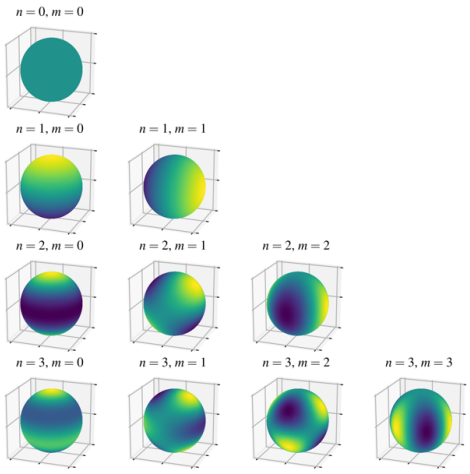

Figure 2: Visualizations of spherical harmonics up to order 3

Let's look at the Fourier transform along time as an analogy: The Fourier transform converts a time-domain signal into a set of coefficients that constitute the complex amplitudes that need to be applied to a set of given basis functions in order to reconstruct the signal. These coefficients are termed *spectrum* and are independent of time. The basis functions are oscillations of different frequencies.

The concept of spherical harmonics decomposition is exactly the same: The function of interest, for example the sound pressure along the surface of a sphere, is decomposed into oscillations, and the according coefficients are independent of space and represent the strength of the oscillation in the signal (cf. Fig. 2). While time-frequency spectra exhibit a certain amount of intuitive information, this is generally not the case with such space-frequency spectra. The spatial frequency in this case is termed *order* and the meaning of the spatial spectra is rather abstract. A simple way of putting it is, lower orders represent slow variations of the pressure on the sphere, higher order represent faster variations on the sphere and more details on the spatial structure of the sound field.

Mathematically speaking, one needs to integrate the sound pressure over the surface of a sphere to obtain the spherical harmonics coefficients, i.e., to perform the transform. This is equivalent to using a continuous distribution of microphones, i.e., an infinite number of infinitesimal microphones. This is obviously not possible, and one needs use a finite set of discrete microphones. This limits the maximum order – and in other words, the amount of spatial detail – that can be obtained from the microphone signals. The rule of thumb is that a maximum obtainable order of N requires at least $(N+1)^2$ microphones. The operation that is an integration in the continuous domain is a summation of the microphone signals in the discrete domain.

One could technically install the microphones along a spherical surface in mid-air, whereby the microphones would need to be acoustically transparent. It is preferable for many reasons, incl. favorable mechanical setup as well as requiring less aggressive signal processing, to mount pressure microphones flush with the surface of a acoustically rigid spherical object. The presence of this scattering object obviously alters the signals that are captured by the individual microphones of the array due to reflection, diffraction, and the like. Fortunately, it is possible to completely remove this effect by signal processing in the spherical harmonics domain without any knowledge on the impinging sound field by the so-called *radial filter*[12].

The same transform is applied to the HRTFs. I.e., one sums up the HRTFs for the different directions of incidence to obtain a spherical harmonics representation of them. Here, no radial filter is necessary. The “wedding” between the sound field and the HRTFs takes place inside the spherical harmonics domain[13]. This means that there is no such thing as a one-to-one mapping between microphones and HRTFs or anything similarly intuitive. All microphones end up in a given HRTF in one or the other way, if one insists on using this simple picture. The “magic” happens in the spherical harmonics domain in which the information is typically not very tangible.

This operation is performed for the left and right ear separately, and the result is the left and right ear signals that are caused by the sound field captured by the microphone array for a given head orientation. The center of the head coincides with the center of the microphone array. To compute the ear signals for a different head orientation, either the sound field or the HRTF set needs to be rotated, which can be performed for arbitrary angles inside the spherical harmonics domain. Rotations about the vertical axis are straightforward and efficient to compute. Other rotations like nodding or tilting of the head are mathematically well defined but computationally expensive. Fig. 3 summarizes the processing pipeline.

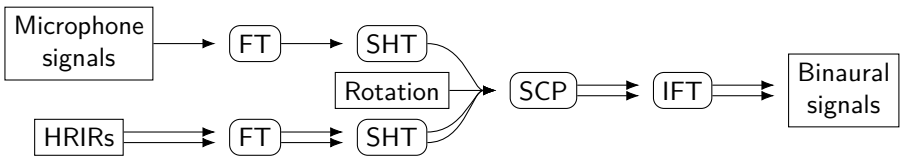

Figure 3: Processing pipeline; FT: Fourier transform; SHT: Spherical harmonics transform; SCP: Spherical coefficients processing; IFT: Inverse Fourier transform

2.3 Practical considerations

2.3.1 Upper frequency limit

The most significant departure from the theoretical requirements is the employment of a finite set of discrete microphones instead of a continuous distribution. This discretization causes *spatial aliasing*. Spatial aliasing constitutes ambiguities in the spatial information that is being captured. This means that information like the instantaneous propagation direction of the captured sound field can be represented incorrectly.
Theoretically, spatial aliasing is apparent at all frequencies f [12]. However, it is marginal at lower frequencies f. There is a limiting frequency f_A, the so-called spatial aliasing frequency, above which the energy of the aliasing is significant and the spatial information is incorrect. The spatial aliasing frequency is proportional to the number of microphones squared and inversely proportional to the radius of the scattering object. Table 1 lists the spatial aliasing frequencies f_A for an array of radius of 8.75 cm, which is similar to that of a human head.

The placement of the microphones on the spherical surface is not straightforward since uniform layouts exist only based on the five platonic solids and are available only up to 20 microphones. Quasi-uniform layouts exist for arbitrary numbers of microphones. We refer the reader to[14] for a more detailed treatment.

In summary, spatial aliasing defines the upper frequency limit up to which spherical microphone arrays are precise.

Max. order	No. of microphones	f_A
5	50	3.1 kHz
8	110	5.0 kHz
29	1202	18 kHz

Table 1: Spatial aliasing frequencies for an array of radius of 8.75 cm with different required numbers of microphones when using a Lebedev sampling grid

2.3.2 Lower frequency limit

The lower frequency limit is determined by the radius of the scattering object. A larger radius means higher spatial accuracy at a given frequency. Since a larger radius also reduces the spatial aliasing frequency, practical arrays are a compromise between these two limitations.
It is intuitive that a small array has trouble deducing detailed spatial information at low frequencies where the wavelength is much lower than the size of the array. The difference between the microphone signals is minuscule. Mathematically speaking, the problem is *ill-posed*. This means that massive amplifications have to be applied to the signals to extract the desired information. This makes the process vulnerable to any sort if inaccuracy be it in the microphone placement, mismatches in, for example, gain between the microphones, or microphone self-noise.

The self-noise in the microphones is uncorrelated between different microphones and has therefore no physical relation. (Note that the relation between the individual microphone signals evoked by a sound field is represented by the wave equation.) While the large gains that need to be applied at low frequencies to extract the acoustic information yield a meaningful result for all physically related signals, they simply boost the self-noise, which can then be very prominent in the output signals.

The theoretically required gains can be as high as hundreds of decibels for an order of, say, 8 or higher, which is prohibitive in practice. It has therefore been proposed to limit the gain, which is equivalent to reducing the order at low frequencies, or in other words, the level of spatial detail.

It is certainly not a coincidence that arrays of a size similarly to that of a human head have been identified as an excellent compromise that does not require excessive gains to extract a sufficient amount of information at lower frequencies. We refer the interested reader to[15] for a detailed treatment.

3 RESEARCH

3.1.1 Past and current research

Despite the circumstance that practically feasible arrays are precise only within a few octaves, it has turned out that the output can be perceptually pleasing. A large part of this may be attributed to the circumstance that arrays can be designed such that the frequency range in which strong auditory spatial localization cues are apparent is rendered precisely. (Note that an array of the size of a human head equipped with 50 microphones is accurate only between, say, 200 Hz and 3.1 kHz.) These findings have led to a significant research activity in the past few years, which is summarized in the following.

The mathematical concept of spherical harmonics has been known since a long time. Research on the general concept of sound field analysis based on spherical microphone arrays started to become active in the early 2000s[16,17]. One of the early works on the binaural rendering of spherical microphone array data is[18] and research activity on this matter has picked up significantly in the recent years.
Technical and physical aspects have been known comprehensively by now. Most of the current research investigates the perceptual properties of binaural renderings. The studies presented in[13,19-22] investigated – and some of them also predict – the perception with respect to overall quality or with respect to higher-level attributes that were either elicited from the subjects themselves or prescribed by the experimenter. Array captures with different parameters were compared against each other.

We will discuss the studies performed in[15,20,23,24] in slightly more detail here. These studies performed perceptual comparisons of head-tracked headphone renderings of array recordings to head-tracked headphone renderings of dummy head recordings of the same scenarios. Particularly, the amount of studies presented in[15] is extensive.

All stimuli in the mentioned studies were produced based on the data from[25,26] which are freely available. These data comprise impulse response measurements in different rooms from a loudspeaker to the microphones of different rigid-sphere arrays as well as to the ears of a dummy head in the same location like the microphone array. The measurements were performed such that time invariance of the rooms may be assumed. The signal processing was performed using[27] or[28]. Both toolboxes are freely available, too. The output of the processing pipeline is then a pair of ear impulse responses that represents the transfer function from the measurement loudspeaker through the room, the array, and the processing pipeline. The output is computed for different head orientations so that head tracking can be applied. The auralization is performed by a convolution engine like SoundScape Renderer[29].

The different spherical microphone arrays in the measurement data were emulated through the VariSphear single-microphone scanning array[30]. It a robotic arm that is equipped with a measurement microphone that is flush mounted in a rigid spherical scattering object of a radius of R = 8.75 cm. The construction rotates such that the microphone can be moved to arbitrary positions on the surface of the spherical scattering object while keeping the center of the scattering object still. This way, arbitrary sampling grids can be emulated and data for up to 1202 microphones are available, which corresponds to order 27.

The direct dummy head measurement data may be interpreted as ground truth against which the output of the processing pipeline is being compared if the processing pipeline uses the dummy head's HRTFs. Recall that the output of the processing pipeline represents the ear signals of a virtual head in the captured sound field. The experiments allow for drawing conclusions on the authenticity of the array renderings as any departure of the output of the processing pipeline from the measured dummy head signals, the ground truth, may be interpreted as an artifact of the pipeline.

The results show that renderings with orders below 8 can sound significantly different from the direct dummy head data. No further reduction in the perceived difference is expected for order above 12.

Fig. 4 depicts example results from a very recent study[31]. These demonstrate that the subjects did not perceive a difference between the head-tracked dummy head auralisation of the given scenario and the auralisation based on the spherical microphone array when the order of the array is 8 or higher. In other words, authentic reproduction of the scenario has been achieved.

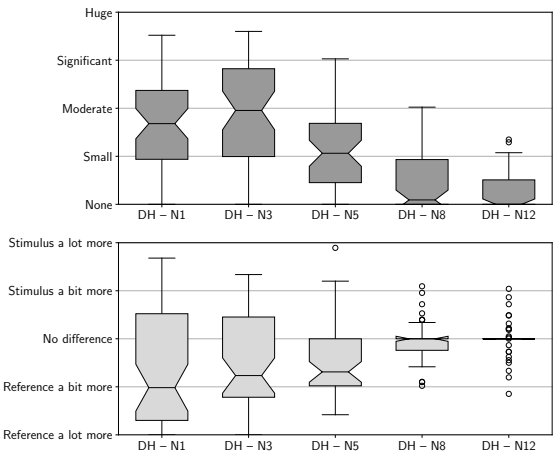

Figure 4: Boxplots of the responses of 20 listeners on the perceptual distance between the dummy head auralisation of a scenario (the "reference") and the output of the array processing pipeline (the "stimulus"). The top plot shows ratings of the difference with respect to *timbre* on a continuous scale from "None" to "Huge"; the bottom plot shows ratings of the difference with respect to *spaciousness* on a continuous scale from "Reference is a lot more spacious" to "Stimulus is a lot more spacious". DH refers to the dummy head; NX specifies the order of the microphone array. The scenario contained a single sound source at a lateral position in a recording studio control room. Data are from[31].

3.1.2 Remaining challenges

As explained in the previous section, research has shown that an order of 8 can be enough to achieve authenticity of the auralization. The 110 microphones required for this are on the limit but still feasible in practice. This is good news as physical accuracy would require several orders of magnitude more microphones, assuming that it is achievable at all.

The results have been achieved based on measured impulse responses and therefore under ideal conditions especially regarding the complete absence of microphone mismatch and self-noise. These are the two main effects that need to be studied to complete the basic investigation.

Microphone mismatch is easy to emulate as the measured impulse responses can be manipulated accordingly. Evaluating the effect of microphone self-noise requires an implementation based on streamed signals rather than based on impulse responses. According to our awareness, such an implementation has not been presented yet and is currently being created by the present authors. Our preliminary experiments have shown that the computational complexity is high but likely to be feasible in real time with general computing hardware.

Once the basic investigation has been completed, it will be interesting to know how much fidelity one loses when using simpler hardware, i.e., fewer microphones and smaller spheres, that is cheaper to manufacture, and how much of the quality loss can be mitigated by additional signal processing.

4 CONCLUSIONS

Binaural renderings of spherical microphone arrays are essentially recordings of audio scenes with a virtual head whose head orientation may be changed in real time upon playback. It is therefore possible to exploit natural mechanisms to impose binaural cues onto the signals while being able to apply head tracking. Research in the field is active and recent results confirm that authenticity can be achieved under certain circumstances. Remaining questions to be answered concern the employment of imperfect hardware as well as potential in simplifying the hardware.

5 REFERENCES

1. M.A. Gerzon, “Recording Concert Hall Acoustics for Posterity,” JAES 23(7) (1975)
2. A. Farina and R. Ayalon, “Recording Concert Hall Acoustics for Posterity,” in Proc. of the AES 24th International Conference on Multichannel Audio, Banff, Canada, June 26 – 28 (2003)
3. Harvith, J., and Harvith, S. Edison, Musicians and the Phonograph: A Century in Retrospect, Greenwood Press, N.Y (1987).
4. http://seanolive.blogspot.com/2010/07/why-live-versus-recorded-listening.html (accessed Aug. 31, 2018)
5. J. Blauert and R. Rabenstein, “Loudspeaker Methods for Surround Sound,” in Proc. of 57th Open Seminar on Acoustics, OSA, Gliwice, Poland (2010)
6. J. Ahrens, “Analytic Methods of Sound Field Synthesis,” Springer-Verlag, Berlin (2012)
7. H. Wierstorf, “Perceptual Assessment of Sound Field Synthesis,” Doctoral Thesis, University of Technology Berlin (2014)
8. S. Paul, “Binaural Recording Technology: A Historical Review and Possible Future Developments,” Acta Acustica United with Acustica, Vol. 95, pp. 767–788 (2009)
9. D.R. Begault, A.S. Lee, E.M. Wenzel, and M.R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source,” in 108th Conv. of the AES (2000)
10. A. Lindau, “Binaural resynthesis of acoustical environments,” Doctoral Thesis, University of Technology Berlin (2014)

11. F. Brinkmann, Al. Lindau, M. Vrhovnik, S. Weinzierl, “Assessing the authenticity of individual dynamic binaural synthesis,” in Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April (2014)
12. B. Rafaely, “Analysis and design of spherical microphone arrays,” IEEE Transactions on Speech and Audio Processing 31(1), pp. 135–143 (2005)
13. A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, “Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution,” JASA 133(5), pp. 2711–2721 (2013)
14. F. Zotter, “Sampling strategies for acoustic holography/holophony on the sphere,” in Proc. of NAG/DAGA (2009)
15. B. Bernschütz, “Microphone arrays and sound field decomposition for dynamic binaural recording,” PhD thesis, University of Technology Berlin (2016)
16. J. Meyer and G. Elko, “A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield,” in Proc. ICASSP, Orlando, FL, USA, pp. 1781–1784 (2002)
17. T.D. Abhayapala and D.B. Ward, “Theory and design of high order sound field microphones using spherical microphone array,” in Proc. ICASSP, Orlando, pp. 1949-1952 (2002)
18. R. Duraiswami, D.N. Zotkin, Z. Li, E. Grassi, N.A. Gumerov, and L.S. Davis, “High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues,” in 119th Convention of the AES (2005)
19. F. Melchior, O. Thiergart, G.D. Galdo, D. de Vries, and S. Brix, “Dual radius spherical cardioid microphone arrays for binaural auralization,” in Proc. of 127th Conv. of the AES (2009)
20. A. Neidhardt, “Untersuchungen zur räumlichen Genauigkeit bei der binauralen Auralisation von Kugelarraydaten [text in German]” MSc thesis, Graz University of Technology (2015)
21. J. Nowak, K.-P. Jurgeit, and J. Liebetrau, “Assessment of spherical microphone array auralizations using open-profiling of quality (opq),” in 8th Int. Conf. on QoMEX (2016)
22. J. Nowak, J., and S. Klockgether, “Perception and prediction of apparent source width and listener envelopment in binaural spherical microphone array auralizations,” JASA 142(3), pp. 1634–1645 (2017)
23. J. Ahrens, C. Hohnerlein, and C. Andersson, “Auralization of acoustic spaces based on spherical microphone array recordings,” in Proc. of Acoustics ’17, ASA/EAA (2017)
24. C. Andersson. “Headphone auralization of acoustic spaces recorded with spherical microphone arrays” MSc thesis, Chalmers University of Technology (2017)
25. P. Stade, B. Bernschütz, B., M. and Rühl, “A spatial audio impulse response compilation captured at the WDR broadcast studios,” in 27th VDT Int. Convention (2012) (http://audiogroup.web.th-koeln.de/wdr_irc.html)
26. B. Bernschütz, “A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100,” in Proc. of AIA/DAGA, DEGA, Meran, Italy, pp. 592–595 (2013) (http://audiogroup.web.th-koeln.de/ku100hrir.html)
27 C. Hohnerlein, J. Ahrens, “Spherical microphone array processing in Python with the sound_field_analysis-py toolbox,” in Proc. of DAGA, Kiel, Germany (2017) (https://github.com/AppliedAcousticsChalmers/sound_field_analysis-py)
28. B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, “SOFiA sound field analysis toolbox,” in Proc. of (ICSA) (2011)
29. M. Geier, S. Spors, and J. Ahrens, “The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods,” in 124th Conv. of the AES (2008) (http://spatialaudio.net/ssr/)
30. B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, “Entwurf und Aufbau eines sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio [text in German],” in Proc. of DAGA, Oldenburg, Germany, pp. 717– 718 (2009)
31. J. Ahrens, C. Andersson, “Perceptual Evaluation of Headphone Auralization of Rooms Captured with Spherical Microphone Arrays with Respect to Spaciousness and Timbre”, JASA (2018) (submitted)