AURALIZATION OF OMNIDIRECTIONAL ROOM IMPULSE RESPONSES BASED ON THE SPATIAL DECOMPOSITION METHOD AND SYNTHETIC SPATIAL DATA

Jens Ahrens

Audio Technology Group
Division of Applied Acoustics
Chalmers University of Technology
412 96 Gothenburg, Sweden

ABSTRACT

The spatial decomposition method decomposes acoustic room impulse responses into a pressure signal and a direction of arrival for each time instant of the pressure signal. An acoustic space can be auralized by distributing the pressure signal over the available loudspeakers or head-related transfer functions so that the required instantaneous propagation direction is recreated. We present a user study that demonstrates based on binaural auralization that the arrival directions can be synthesized from random data such that the auralization is nearly indistinguishable from the auralization of the original data. The presented concept constitutes the fundament of a highly scalable spatialization method for omnidirectional room impulse responses.

Index Terms— Reverberation, spatial decomposition method, binaural rendering, head-related transfer functions

1. INTRODUCTION

A large variety of algorithms for the creation of artificial reverberation has been proposed for classical loudspeaker setups such as stereo and surround [1]. This includes methods based on filter networks or based on convolution with measured or simulated data. Unfortunately, these methods do no scale well to be directly applicable with more advanced modern setups, especially those employing an object-based paradigm, and a standard solution is not available. Systems employing audio objects typically render a certain number of reverberation channels as virtual sources [2, 3].

We propose and evaluate here the use of the auralization approach associated with the spatial decomposition method (SDM) [4] as basis for the creation of artificial reverberation that is scalable with respect to the degrees of freedom of the spatial information.

SDM uses room impulse responses that were captured with a typically compact microphone array to obtain the omnidirectional pressure signal as well as the instantaneous arrival direction for each of the digital samples that the pressure signal is composed of. The geometry of the array is flexible as long as the required pressure signal as well as the instantaneous intensity can be obtained from the data. SDM has been primarily used for analysis and visualisation of the directional properties of room impulse responses and is particularly popular in concert hall acoustics [5].

SDM has also successfully been applied for auralizations of spaces. Slightly simplified, SDM data are auralized by distributing the digital samples of the pressure signal to the available loudspeakers such that the instantaneous arrival directions of the signal are maintained as precisely as possible. Example works are [6, 7, 8], all of which used loudspeaker systems in auralization. Binaural auralization of SDM data is essentially similar to loudspeaker rendering as the available head-related transfer function (HRTF) measurement points are used as virtual loudspeakers. The systems presented in [6, 8], for example, play the obtained signals directly from the available loudspeakers while [7] uses Ambisonics encoding of the components.

SDM constitutes a simple and intuitive means of representing room impulse responses including the directional information. It is therefore an excellent starting point for the design of an artificial reverberator.

Many of the distinct characteristics of an acoustic space are encoded in the omnidirectional room impulse response, and a plethora of measurements of such room impulse responses of a large range of venues are available on the internet. Unfortunately, the use of these data in auralization is limited without the directional information. We therefore evaluate the perceptual properties of omnidirectional room impulse responses auralized with synthetic directional information in this paper. We chose binaural rendering for this because it constitutes the most controlled and reproducible scenario. We apply head tracking in the rendering to mitigate risks for impairment of the spatial fidelity of the rendering [9]. A further advantage is that the binaural auralizations can be directly compared to dummy head auralizations of the same scenario so that their authenticity can be evaluated. This strategy has been successfully applied with spherical microphone array data, for example, in [10, 11].

An approach to binaural auralization that may be considered an advancement of SDM-based rendering was presented recently in [12]. We chose to use the core SDM-based rendering as it can be more straightforwardly applied to loudspeaker setups, too.

An alternative approach to binauralization of omnidirectional room impulse responses is presented in [13], which uses a more involved decomposition of the room impulse response into direct sound, early reflections, and late reverberation. This decomposition is combined with an mirror-source model of a suitable room geometry to account for listener movements. A perceptual evaluation is not available as to our awareness.

In the following, we briefly review SDM analysis and synthesis of room impulse responses. We then describe the employed approach to generating synthetic spatial data and then present the user study, which is the main contribution of this paper.

2. SPATIAL DECOMPOSITION AND SYNTHESIS

SDM estimates the direction of arrival (DOA) of the sound pressure signal of a room impulse response in short time windows along the

entire impulse response. These data are obtained from compact microphone arrays the geometry of which is not important as long as the desired information can be deduced. The microphone arrays do typically not comprise a scattering object.

The room impulse response is analyzed in segments. The time-difference of arrival is determined for each time window for each of the microphone pairs in the array through cross-correlation. Subsequently, a minimum mean square error problem is solved to obtain the final estimate of the average DOA for the time window under consideration [4]. The analysis window advances in steps of 1 sample so that one DOA estimate is obtained for each digital sample of the impulse response.

The pressure signal can be obtained in different ways depending on the microphone array. If the array comprises omnidirectional microphones and the array is compact, the signal of any of the microphones may be employed directly. Arrays that do not employ omnidirectional microphones such as tetradedral Ambisonics microphones [14] require different dedicated solutions.

Auralization of the obtained pressure and DOA signals is performed by either distributing the signal samples over the available loudspeakers via nearest neighbor interpolation [6], vector-base amplitude panning [4], Ambisonics encoding [7], and the like, with compensation for potentially varying loudspeaker distance from the listening location applied. The loudspeaker signals finally need to be equalized to achieve the correct signal spectrum at the listening location [4]. The resulting impulse responses for each of the loudspeakers is then convolved with the source signal.

In this paper, we employ auralization based on nearest neighbor interpolation because of its simplicity and excellent quality [6].

3. METHOD

For the present study, we used the data of the rooms "another living room" (ALR) from [15] with a reverberation decay time of 0.25 s and "Promenadikeskus concert hall" (PCH) from [16] with a reverberation decay time of 2.4 s. The impulse response of ALR was acquired using a microphone array composed of 6 omnidirectional capsules equally distributed over a notional spherical surface with a radius of 12.5 cm (i.e., at the corners of a regular octahedron). The impulse response of PCH was acquired using a tetrahedral microphone producing a B-format signal.

The standard SDM according to [4] was employed at a sampling frequency of 48 kHz to obtain the spatial data, i.e., the instantaneous azimuth and elevation of the arrival direction of each sample of the pressure impulse response, for the two rooms. Fig. 1 exemplarily depicts the first 20 ms of the data of room ALR. The direct sound occurring between approx. 0.5 and 1 ms is apparent as well as a few reflections might be discernable at approx. 3 ms and 4.5 ms. The direct sound and the mentioned reflections are also apparent in the azimuth: The determined azimuth appears to be rather stable for the duration of the related event. The unwrapped azimuth turns out to be approx. 723° for the duration of the direct sound, which is equivalent to 3° and is therefore sane. The direct sound is also apparent in the computed elevation (approx. 0°), but the arrival elevations of the reflections are less obvious.

Fig. 2 depicts the spatial data for the same room for a later time interval. No obvious structure of the spatial data is apparent anywhere through the entire duration of the impulse response apart from the first few milliseconds as discussed above. Our attempts to analyze the data using standard methods in spherical statistics have not been fruitful.

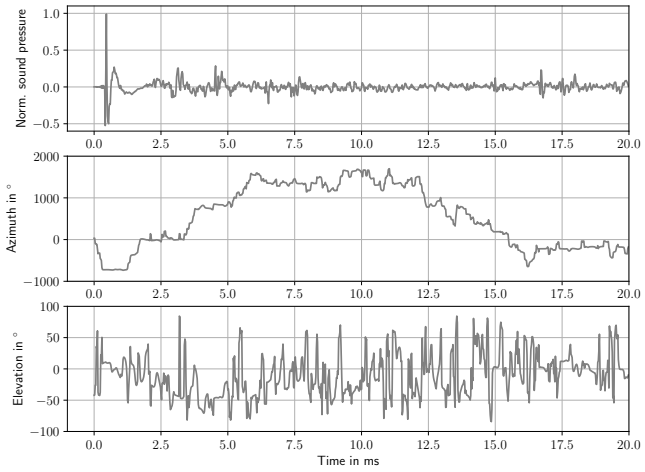

Fig. 1. Sound pressure signal (top) as well as unwrapped azimuth (middle) and elevation (bottom) as determined by SDM

We therefore started off with uniformly distributed random DOAs. The multivariate standard normal distribution is spherically symmetric so that we can randomly distribute points on a spherical surface by generating a set of Cartesian coordinates x, y, z using a Gaussian distribution and projecting each point onto a sphere by normalizing with $1/\sqrt{x^2+y^2+z^2}$. In terms of SDM, the angular position the obtained points may be assumed to represent the spatial data of a diffuse signal. The black curves in Fig. 3 are example data obtained through the procedure described above.

Comparing the random data to the real one from Fig. 2 suggests that the random data are more erratic. We therefore applied different amounts of smoothing by means of a moving mean filter of orders 5 and 9 for the unwrapped azimuths as well as a moving median filter of orders 5 for the elevations. The moving mean filter tended to push away the elevations from the poles, which was considered unfavorable, and which was not so pronounced with the moving median filter. We waived smoothing the elevation data with order 9 to avoid squashing the arrival directions too much towards the horizontal plane. Exemplary data are included in Fig. 3.

We divided the room impulse response into the segments direct sound, early reflections, and late reverberation to apply different amounts of smoothing as detailed in Sec. 4. The duration of the direct sound was identified manually for simplicity, the start of the late reverberation was identified using the method from [17]. We observed in a pilot study that it does not seem favorable to smooth the spatial data of the late reverberation. The smoothing was only applied to the early reflections in those cases where we were using the original data for the direct sound and only to the direct sound and early reflections when the entire spatial data was synthesized.

4. PERCEPTUAL EVALUATION

We created binaural auralizations of the data of the two rooms ALR and PCH using the HRTFs of a Neumann KU 100 dummy head from [18]. The data have a very high angular resolution of 2° so that the ear impulse responses can be computed for an according grid of head orientations without an intermediate encoding in spherical harmonics or the like, and head tracking can be applied straightforwardly.

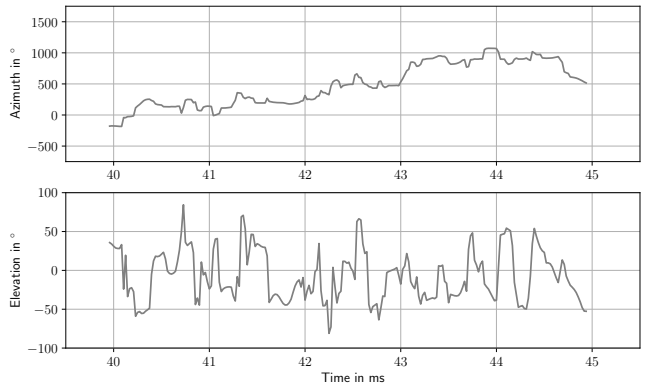

Fig. 2. Spatial data of room ALR for a different time interval than in Fig. 1; top: unwrapped azimuth; bottom: elevation

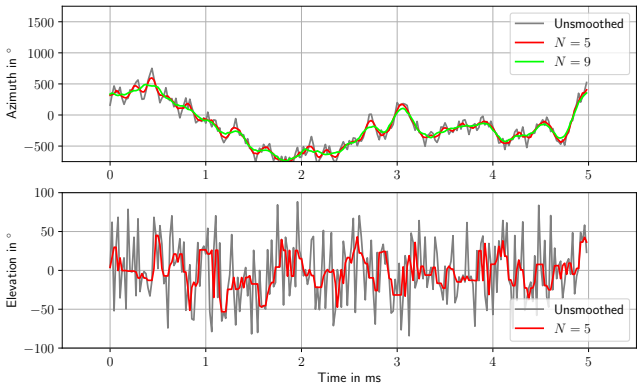

Fig. 3. Random spatial data with different amounts of smoothing; top: unwrapped azimuth; bottom: elevation

The primary goal of the experiment is to perceptually evaluate the synthetic spatial data compared to the spatial data obtained from the measurements.

We did not observe considerable differences in the timbre between different experimental conditions when conducting a pre-study. The perceptual differences that arise are primarily related to spatial attributes such as source distance, locatedness of the source, diffuseness of the reverberation and the like. We therefore chose the subjects' task to be a rating of the overall perceptual distance between stimuli using a slider with a continuous scale ranging from "no difference" via "small difference", "moderate difference", "significant difference" to "large difference". We opted for a pairwise comparison rather than a MUSHRA-like paradigm as the observed perceptual differences are mostly small in magnitude.

In most conditions, the manipulated stimulus was compared to the auralization of the original SDM data. This was performed separately for the two rooms. The manipulations that were applied were randomization of the spatial data, i.e., the DOAs, of certain parts of the room impulse response. The tested conditions are listed in Tab. 1 and comprise randomization of the late reverberation (LR), randomization of the late reverberation as well as the early reflections (LR + ER), as well as randomizing the spatial data of the entire room impulse response whereby different amounts of smoothing were applied.

The direct sound of the signal has a strong influence on the overall perception [19]. We chose not specifically synthesize the spatial data of the direct sound for the following reasons: 1) Synthesizing spatial data for the direct sound is straightforward. 2) SDM data for the direct sound is sometimes impaired, especially when B-format signals are used. We want to avoid making our subjects compare clean synthetic data with potentially impaired original data as this may lead to incorrect conclusions. We did include conditions where the entire spatial data were synthetic without applying specific treatment to the direct sound.

We also added the condition of the pressure room impulse response auralized as a virtual source (pressure + HRTF). The motivation was two-fold: 1) A pre-study showed that the perceptual differences are mostly very small, which makes the task challenging for the subjects. The present condition is significantly different from the reference, which makes the subjects gain confidence in their performance. 2) This condition can be easily reproduced in other user studies so that it constitutes an anchor.

Room	Auralization method	Randomized component(s)	Smoothing order	Label in Fig. 4
ALR / PCH	SDM			S
ALR / PCH	SDM	LR		$S^{(L)}$
ALR / PCH	SDM	ER + LR		$S^{(E)}$
ALR / PCH	SDM	All		$S^{(A)}$
ALR / PCH	SDM	ER + LR	5	$S^{(E,5)}$
ALR / PCH	SDM	All	5	$S^{(A,5)}$
ALR / PCH	SDM	ER + LR	9	$S^{(E,9)}$
ALR / PCH	Pressure + HRTF			P
PCH	DH (no HT)			DH

Table 1. List of stimuli evaluated in the user study; pressure: sound pressure signal of the room impulse response; pressure + HRTF: sound pressure signal of the room impulse response auralized as virtual source; SDM: standard SDM rendering; LR: late reverberation; ER: early reflections; DH: dummy head; HT: head tracking

The data for PCH comprise the binaural room impulse responses (BRIRs) of a B&K HATS dummy head at the same location like the microphone array so that the auralization based on the microphone data can be compared directly to the dummy head ear signals. This situation comprises two caveats: 1) We were required to use the HRTFs of a different dummy head (Neumann KU 100) than the one for which we had the measured BRIRs available. This may cause slight differences with respect to the timbre, which we minimized through manual equalization. 2) The BRIRs are only available for one head orientation so that these data cannot be auralized with head tracking. To achieve a meaningful comparison, we also removed the head tracking for all stimuli while being compared against the BRIRs.

The complete ethics protocol including information on how and how long the collected data are going to be stored was explained to the subjects before the start of the experiment. They were then provided with written instructions on their task and performed a training of 6 manually selected comparisons. The experiment comprised the evaluation of 17 stimulus pairs (8 for ALR and 9 for PCH) each of which was presented twice and in random order and with random button assignment yielding a total number of 34 ratings performed by each subject. The subjects were allowed to proceed at their own pace. A loop of male speech with a duration of 2 min was used as source signal. The signal was playing continuously while the subjects were switching between the stimuli.

The presentation of the stimuli was performed using the software SoundScape Renderer (SSR) [20, 21] running in binaural room synthesis mode. SSR convolves a given input signal with that pair of impulse responses that corresponds to the instantaneous head orientation as provided by a head tracker. The use of head tracking is essential in such studies in order to avoid distortion of the spatial perception [9, 22]. We employed a Polhemus Patriot and AKG K702 open-design headphones. The experiment was conducted in an acoustically treated laboratory room.

When a change in head orientation occurs, then SSR convolves the current signal block with the current as well as with the previous set of filters and crossfades with a cosine ramp between the signals. The block size was set to 256 samples at a sampling frequency of 48 kHz with 2 blocks of buffering. The overall latency of the system is composed of the latency of the tracker (18.5 ms), 2 blocks of buffering (2 x 5 ms), 1 block delay due to signal routing and processing (5 ms), and approximately a half block delay due to the crossfade after the convolution (2.5 ms). This amounts to 36 ms, which is well below audibility [23].

Each stimulus condition is represented by one virtual sound source in SSR to which the corresponding set of impulse responses was assigned. SSR comprises a TCP/IP interface for remote control of all its functionality. The graphical user interface (GUI) of the experiment was created in Matlab and communicated with SSR over TCP/IP. Switching to a given stimulus in the GUI makes SSR unmute the corresponding source (and mute all others). Unmuting and muting are implemented in SSR with cosine fade-in and fade-out ramps. This produces a smooth transition between stimuli.

5. RESULTS

14 voluntary subjects of different gender and in the age range of 26-40 years participated in the experiment. All have self-reported unimpaired hearing. Boxplots of the subjects' responses are depicted in Fig. 4. They show the median value of the data via the horizontal line, the 25th and 75th percentiles via the gray box, the whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually via circles. The notches represent the 95 % confidence interval of the median.

The main observations are the following:

- The hidden reference is reliably identified (S - S, room ALR).
- The perceptual differences are mostly smaller for room ALR, which is the dryer of the two.
- The perceived difference is mostly between "none" and "small" even for those conditions where all spatial data other than the direct sound was synthetic ($S^{(E)}$, $S^{(E,5)}$, $S^{(E,9)}$ for ALR; $S^{(E,9)}$ for PCH).
- Randomization of the entire spatial data including the direct sound clearly differentiates the auralization from that of the original data ($S^{(A)}$, $S^{(A,5)}$). Dedicated spatial data therefore have to be synthesized for the direct sound.
- Smoothing of the synthetic data (of the early reflections) reduces the perceptual distance to the original data ($S^{(E,5)}$ vs. $S^{(E)}$ and $S^{(E,9)}$ vs. $S^{(E)}$).
- The perceptual distance between the pressure signal auralized as a virtual source (P) is much larger than the distances of all synthetic data to the original data.
- The SDM-based auralization sounds significantly different from the dummy head auralization (room PCH, DH - S)

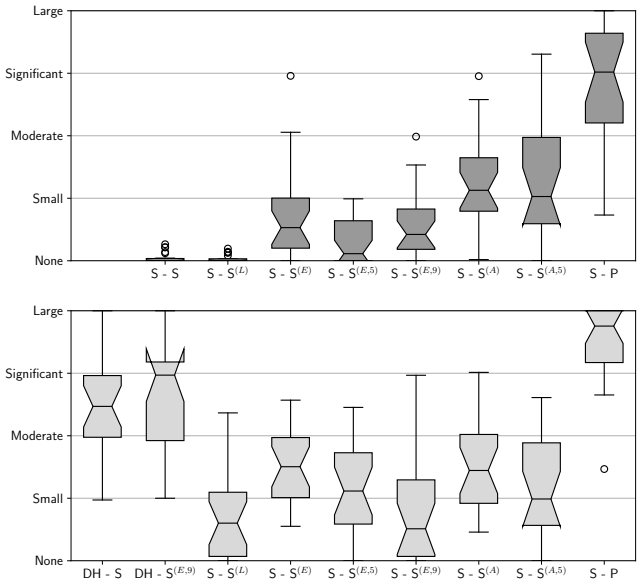

Fig. 4. Box plots; top: room ALR; bottom: room PCH

The observation that the SDM-based auralization is different from the auralization of the according dummy head data has already been reported in [12]. The observed differences mainly relate to the spectral balance of direct sound and reverberation as well as spatial attributes. Our presented approach provides the potential for mitigating the discrepancy as it provides considerable freedom in manipulating the spatial data.

Some stimuli that are composed of the original direct sound and synthetic data otherwise were rated as having a difference to the original data that is smaller than "small". This proves the effectiveness of the presented approach. The stimuli most similar to the original used smoothing of the early reflection data of order 5 with room ALR ($S^{(E,5)}$) and order 9 with room PCH ($S^{(E,9)}$).

The data indicate that dedicated spatial data has to be synthesized for the direct sound, which seems straightforward when knowing the location of the sound source. It is unclear at this stage why the observed differences tend to be smaller for the acoustically dryer room ALR than for the concert hall PCH. We have made the informal observation that the auralizations of SDM based on B-format signals, like it is the case for PCH, tend to sound less plausible than when arrays like the one with room ALR are employed. This circumstance might have affected the results.

6. CONCLUSIONS

We presented a first exploration of the auralization of room impulse responses based on the spatial decomposition method and synthetic spatial data. Our experiment showed that it is indeed possible to synthesize spatial data for early reflections and late reverberation such that the perceptual difference to the original data is negligible. Future work includes the synthesis of spatial data for the direct sound as well as treatment of scenarios that have more distinct properties than the tested ones such as a receiver that is located close to a reflecting surface.

7. REFERENCES

[1] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, July 2012.

[2] Martin Bürner, "A Realtime Multichannel Reverberation Framework for the SoundScape Renderer," MSc thesis, Technische Universität Berlin, 2017.

[3] Philip Coleman, Andreas Franck, Philip J B Jackson, Richard J. Hughes, Luca Remaggi, and Frank Melchior, "Object-based reverberation for spatial audio," *AES: Journal of the Audio Engineering Society*, vol. 65, no. 1-2, pp. 66–77, 2017.

[4] Sakari Tervo, Jukka Pätynen, Antti Kuusinen, and Tapio Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc*, vol. 61, no. 1/2, pp. 17–28, 2013.

[5] Jukka Pätynen, Sakari Tervo, and Tapio Lokki, "Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 842–857, 2013.

[6] Sakari Tervo, Jukka Pätynen, Neofytos Kaplanis, Morten Lydolf, Søren Bech, and Tapio Lokki, "Spatial Analysis and Synthesis of Car Audio System and Car-Cabin Acoustics with a Compact Microphone Array," *Journal of the Audio Engineering Society*, vol. 63, no. 11, pp. 914–925, 2015.

[7] Matthias Frank and Franz Zotter, "Spatial impression and directional resolution in the reproduction of reverberation," in *Proc. of DAGA*, Aachen, Germany, 2016, pp. 1304–1307.

[8] Sebastià V. Amengual Garí, Winfried Lachenmayr, and Eckard Mommertz, "Spatial analysis and auralization of room acoustics using a tetrahedral microphone," *The Journal of the Acoustical Society of America*, vol. 141, no. 4, pp. EL369–EL374, 2017.

[9] Durand R. Begault, Alexandra S. Lee, Elizabeth M. Wenzel, and Mark 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 Convention of the AES*, May 2000.

[10] Benjamin Bernschütz, "Microphone arrays and sound field decomposition for dynamic binaural recording," PhD thesis, Technische Universität Berlin, 2016.

[11] Jens Ahrens and Carl Andersson, "Perceptual Evaluation of Headphone Auralization of Rooms Captured with Spherical Microphone Arrays with Respect to Spaciousness and Timbre," *The Journal of the Acoustical Society of America (accepted for publication)*, 2018.

[12] Markus Zaunschirm, Matthias Frank, and Franz Zotter, "BRIR synthesis using first-order microphone arrays," in *144th Convention of the AES*, 2018.

[13] Christoph Pörschmann, Philipp Stade, and Johannes M. Arend, "Binauralization of Omnidirectional Room Impulse Responses - Algorithm and Technical Evaluation," *Proceedings of the DAFx 2017*, pp. 345–352, 2017.

[14] Johann-Markus Batke, "The B-Format Microphone Revisited," in *Proceedings of the Ambisonics Symposium*, Graz, Austria, June 2009.

[15] Tapio Lokki, "A toolbox for spatial analysis and synthesis of room acoustic impulse responses," [Online]. Available: `https://users.aalto.fi/~ktlokki/sdm_tools.html`, 2018.

[16] Juha Merimaa, Timo Peltonen, and Tapio Lokki, "Concert hall impulse responses – Pori, Finland: Reference," Tech. Rep., [Online]. Available: `http://www.acoustics.hut.fi/projects/poririrs/`, 2005.

[17] Jonathan Abel and Patty Huang, "A Simple, Robust Measure of Reverberation Echo Density," in *121st Convention of the AES*, San Francisco, CA, USA, 206.

[18] Benjamin Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Proceedings of AIA/DAGA*, Meran, Italy, Mar. 2013, DEGA, pp. 592–595.

[19] Aki Haapaniemi and Tapio Lokki, "Identifying Concert Halls from Source Presence vs. Room Presence," *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. EL311–EL317, 2014.

[20] Matthias Geier, Sascha Spors, and Jens Ahrens, "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," in *124th Convention of the AES*, May 2008.

[21] SSR Team, "SoundScape Renderer," [Online] Available: `http://spatialaudio.net/ssr/`, 2019.

[22] Alexander Lindau, "Binaural resynthesis of acoustical environments," PhD thesis, Technische Universität Berlin, 2014.

[23] Alexander Lindau, "The perception of system latency in dynamic binaural synthesis," in *Proceedings of NAG/DAGA*, Rotterdam, The Netherlands, Mar. 2009, DEGA, pp. 1063–1066.