기하음향 기반 확산 음장 시뮬레이션을 위한 앰비소닉 렌더링 기법

*유필선, **Franz Zotter, ***유재현, *최정우
*한국과학기술원, **그라츠국립음악대학교, ***한국전자통신연구원
pilz@kaist.ac.kr, zotter@iem.at, jh0079@etri.re.kr, jwoo@kaist.ac.kr

Ambisonic Rendering for Diffuse Sound Field Simulations based on Geometrical Acoustics

*Pilsun Eu, **Franz Zotter, ***Jae-hyoun Yoo, *Jung-Woo Choi
*Korea Advanced Institute of Science and Technology
**University of Music and Performing Arts Graz
***Electronics and Telecommunications Research Institute

ABSTRACT

The diffuse sound field plays a crucial role in the perceptual quality of the auralization of virtual scenes. Diffuse Rain is a geometrical scattering model which enables the simulation of diffuse fields that is compatible with acoustic ray tracing, but is often computationally expensive. We develop a novel method that can reduce this cost by rendering the large number of Diffuse Rain data in Ambisonics format. The proposed method is evaluated in a shoebox scene simulation run on MATLAB, in reference to a more faithful method of rendering the Diffuse Rain data ray-by-ray. The EDC and IACC of the binaural output show that the simulated diffuse field can be rendered in Ambisonics with only minimal deviations in energy decay and spatial quality, even with 1st-order Ambisonics.

1. Introduction

Regarding the recent trend and the rapid development of virtual reality (VR) media platforms that aim to provide users with immersive virtual experiences, there is a rising demand for fast, high-quality acoustic simulations.

Ray tracing is a geometrical simulation method that is increasingly becoming an attractive model for the real-time auralization in VR, due to the rise of graphical processing units (GPU) that can handle the vast number of rays required for perceptually accurate results. However, unlike wave-based numerical methods, traditional ray tracing cannot emulate the wave phenomena for realistic acoustics, such as scattering and diffusion of sound from material surfaces.

Diffuse Rain [1] is an acoustic model that incorporates sound scattering into ray tracing by deterministically calculating the energy transfers from scattering point candidates visible to listener, avoiding the issue of potential computational explosions in scattering models such as ray splitting. However, Diffuse Rain can lead to an immense number of data and rendering cost when the listener is exposed to a large number of scattering points.

To reduce this cost, we propose a technique of rendering Diffuse Rain results in Ambisonics. We introduce the formulation and necessary assumptions for the encoding of energetic rays in Ambisonics and the estimation of the diffuse sound field covariance matrix, which stores the spatial sound information for the binaural output synthesis. We compare the Ambisonics method to a brute-force approach that renders each individual ray energy impulse with a Head-Related Impulse Response (HRIR), which we take as the most faithful representation of simulation results.

2. Geometrical Acoustics

A. Acoustic Ray Tracing

Acoustic ray tracing, specifically *stochastic* acoustic ray tracing, statistically models the acoustics of a virtual scene by emitting discrete sound rays uniformly in random directions and propagating them within the scene geometry. The individual ray energies decay over time as they get absorbed into the air and the reflecting surfaces. Intersections between ray paths and a finite-volume detector are logged as energy values in discrete time-bin histograms. This temporal energy data is converted to an audio output with a rendering method of choice, in the form of an impulse response (IR), or the room IR (RIR) of the scene.

Although ray tracing is a high-frequency approximation that considers sound as particles instead of waves, frequency-dependent energy decay is still calculated with air and material absorption coefficients defined for octave frequency bands within the audible range. Therefore, ray energies will not only be stored in time-bins, but also frequency-bins. For binaural audio rendering, ray energies are stored in angle-dependent histograms used in the synthesis of binaural RIRs (BRIRs).

B. Diffuse Rain based Scattering

Diffuse Rain [1] models the scattering of sound by geometrically calculating the amount of scattered energy transferred to a finite-volume detector, with the ray reflection points visible to the detector as the scattering points. The transferred energy from each point is determined by a scattering distribution, a function of the angle against surface normal that is independent of the incident ray angle. The conventional Lambertian distribution $w(\Theta) = A\cos(\Theta)$ is used in this work, where A is a normalization factor and Θ is the angle between the detector position and the surface normal.

When a ray with energy E_p is incident at the scattering point, a fraction α (absorption coefficient) of the energy is absorbed by the surface material, from which a fraction s (scattering coefficient) is scattered into the half-sphere in the direction of the normal. The portion from this scattered energy transferred to a detector is the integration of $w(\Theta)$ over the solid angle Ω subtended by the detector's volume. With A determined by normalizing $w(\Theta)$ over the half-sphere, the energy transferred to a *spherical* detector is given as

$$\mathrm{E} = \mathrm{E_p} \cdot (1-\alpha) \cdot \mathrm{s} \cdot 2 \cdot (1-\cos\gamma) \cdot \cos\Theta, \qquad (1)$$

where γ is half of the open angle from the scattering point to the sphere.

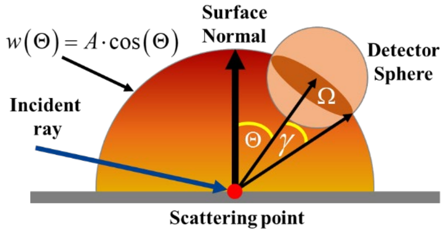

Figure 1. An instance of Diffuse Rain energy transfer.

This energy is treated in the same way as a detected ray from specular reflections, and is logged in an energy histogram at the time-bin for the detection time of the scattering event. Since coefficients α and s are frequency-dependent, energies are also logged in the histogram at frequency-bins.

Although not essential for the acoustic rendering process, the incident ray *energy, time and direction* data from both specular reflections and Diffuse Rain are stored as image source data for the sake of explicit representation and ease of handling. From this point on, we refer to an instant of incident energy detection as an *image source*.

3. Acoustic Rendering

A. Ambisonic Rendering

Since Diffuse Rain generates an image source for every point of reflection visible to the detector, rendering its results often requires a high computation cost. We propose a method for reducing this rendering cost by encoding the large number of image sources in an Ambisonics format with much fewer channels, as a trade-off between spatial resolution and cost, variable with the encoding spherical harmonic (SH) order.

In the usual sense, N^{th}-order Ambisonic encoding [2] of a multi-channel audio signal involves transforming each i^{th} microphone channel sound pressure data using an encoding vector with SH weights for orders $n = 1, 2, \dots, N$ and degrees $m = -n, -n+1, \dots, n-1, n$

$$[\vec{\mathrm{y}}_{\mathrm{i}}]_{n,m} = Y_n^m(\Omega_i)\,, \qquad (2)$$

$$Y_n^m(\theta,\phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} \cdot P_n^m(\cos\theta) \cdot e^{im\phi}\,, \qquad (3)$$

where $\Omega_i = (\theta_i, \phi_i)$ is the i$^{\text{th}}$ channel direction with elevation θ_i and azimuth ϕ_i, $P_n^m(\cdot)$ are the associated Legendre polynomials. Since our goal is to encode discrete image source energies and not signal amplitudes, we devise a new appropriate formulation.

1) *Ray Encoding and Covariance Construction*

Denoting the diffuse field IR in N^{th}-order SH domain at the position and time of interest as $\vec{\mathbf{p}} = [p_1(t) \quad \cdots \quad p_{l_{max}}(t)]^{\mathrm{T}}$, $l_{max} = (N+1)^2$, the covariance between its harmonics is expressed as the expectation value of the vector product

$$\mathbf{C} = E\{\vec{\mathbf{p}}\vec{\mathbf{p}}^{\mathrm{H}}\}\,. \tag{4}$$

We assume that the diffuse sound field modeled with Diffuse Rain image sources is approximately *uniform in time and direction*. Then, we can regard the i$^{\text{th}}$ image source, or ray, within a time-bin as an uncorrelated and random noise sample recorded at a direction $\boldsymbol{\Omega_i}$, with the variance $\boldsymbol{\sigma_i^2} = \boldsymbol{E_i}$ equal to ray energy. Then, image source data can effectively be encoded in $\boldsymbol{N^{th}}$-order SH domain by estimating $\mathbf{C}$ as

$$\mathbf{C} \approx \tilde{\mathbf{C}} = \textstyle\sum_i \vec{\mathbf{y}}_i^* \sigma_i^2 \vec{\mathbf{y}}_i^{\mathrm{T}}\,, \tag{5}$$

where vectors $\vec{\mathbf{y}}_i$ have the same definition as (2), but with indices i representing the i$^{\text{th}}$ image source, not channels.

2) *IR synthesis and Binaural Decoding*

Spatial information is extracted from the estimated covariance matrix to reproduce the diffuse field. To this end, the eigenvectors $\tilde{\mathbf{U}} = [\vec{\mathbf{u}}_1 \quad \cdots \quad \vec{\mathbf{u}}_{l_{max}}]$ and the eigenvalues $\widetilde{\boldsymbol{\Lambda}} = diag(\lambda_1, \ldots, \lambda_{l_{max}})$ are calculated from the eigendecomposition $\tilde{\mathbf{C}} = \tilde{\mathbf{U}}\widetilde{\boldsymbol{\Lambda}}^2\tilde{\mathbf{U}}^{\mathrm{H}}$. Then, the estimated diffuse field IR can be synthesized as a linear combination of noise samples

$$\vec{\tilde{\mathbf{p}}} = \textstyle\sum_j (\lambda_j \vec{\mathbf{u}}_j) r_j(t) = \tilde{\mathbf{U}}\widetilde{\boldsymbol{\Lambda}}\vec{\mathbf{r}}\,, \tag{6}$$

where $\vec{\mathbf{r}} = [r_1(t) \quad \cdots \quad r_{l_{max}}(t)]^{\mathrm{T}}$ are random Gaussian noise generated for each eigenvector. Incidentally, the covariance matrix of the estimated diffuse field

$$E\{\vec{\tilde{\mathbf{p}}}\vec{\tilde{\mathbf{p}}}^{\mathrm{H}}\} = \tilde{\mathbf{U}}\widetilde{\boldsymbol{\Lambda}}E\{\vec{\mathbf{r}}\vec{\mathbf{r}}^{\mathrm{H}}\}\widetilde{\boldsymbol{\Lambda}}\tilde{\mathbf{U}}^{\mathrm{H}} = \tilde{\mathbf{U}}\widetilde{\boldsymbol{\Lambda}}^2\tilde{\mathbf{U}}^{\mathrm{H}} = \tilde{\mathbf{C}} \tag{7}$$

is equal to the original estimated covariance, given that $r_j(t)$ are uncorrelated to each other, such that $E\{\vec{\mathbf{r}}\vec{\mathbf{r}}^{\mathrm{H}}\} = \mathbf{I}$.

To obtain the left and right BRIR signals $p_L(t)$ and $p_R(t)$, the estimated diffuse sound field is first decoded into a virtual speaker configuration, using a spherical t-design appropriate for the maximum SH order [3]. The decoded signals are convolved with the HRIRs at the speaker directions and summed up to obtain the BRIR

$$p_{L,R}(t) = \textstyle\sum_{k=1}^{K} \vec{\mathbf{H}}_{L,R}^{k}(t) * \vec{\tilde{\mathbf{p}}}_{L,R}^{k}(t)\;. \tag{8}$$

Here, $\vec{\mathbf{H}}_{L,R}^{k}(t) = \left[H_{L,R}^{1}(t) \quad \cdots \quad H_{L,R}^{K}(t)\right]^{\mathrm{T}}$ and K is the number of virtual speakers. The synthesis of the diffuse part for the BRIR in a virtual scene with Ambisonic rendering is summarized in figure 2.

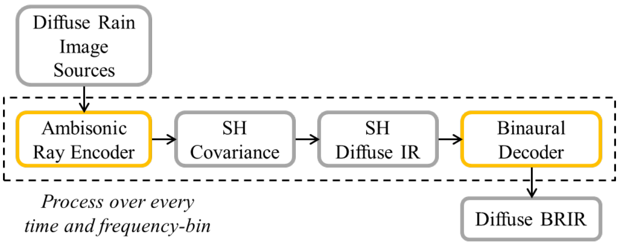

Figure 2. Diffuse field BRIR synthesis with Ambisonic rendering.

B. Ray-by-Ray Rendering

The ray-by-ray method is the direct synthesis of the BRIR as a linear combination of individual ray contributions. For each image source direction and frequency-bin energies, the respective band-pass filtered HRIR is amplitude-scaled, time-shifted, then summed onto the final BRIR after a convolution with a Gaussian noise segment. In the evaluation of the Ambisonic rendering of diffuse sound fields, we shall consider this approach as the "ground truth" that is most true to the generated Diffuse Rain data. The ray-by-ray method is also used for the rendering of the specular reflection part of the BRIR, where relatively few image sources are involved.

4. Evaluation

A. Simulation Settings

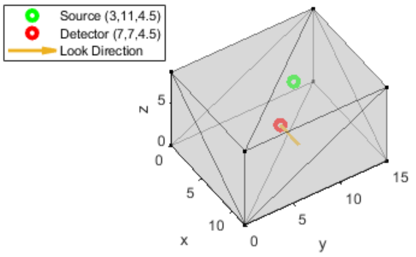

Figure 3. Simulation shoebox scene (dimensions in m).

The two rendering methods are implemented in a proof-of-concept acoustic simulation run on MATLAB. 20,000 rays are traced over 20 reflection orders within a virtual shoebox scene, with a spherical detector of radius 0.25 m and sampling frequency of 48 kHz. Histograms are built with 10 ms time-bins and 6 frequency-bins corresponding to octave bands of center frequencies 125 Hz to 4 kHz. The average absorption and scattering coefficients for the scene are 0.15 and 0.8, respectively. *Gypsilab* MATLAB library's *openRay* toolbox [4] is used for ray tracing, and the HUTUBS HRTF database [5] with 220 sampling points is used for binaural rendering.

B. Energy Decay Curve

The energy decay curve (EDC) of an IR is defined as the normalized tail integral [6] or the energy decay from the total signal power up to time t:

$$\mathrm{EDC}(t) = \frac{\int_t^\infty h^2(\tau)d\tau}{\int_0^\infty h^2(\tau)d\tau}. \qquad (9)$$

Figure 4 shows the EDCs from 1st and 3rd-order Ambisonic rendering, and the ray-by-ray method. The energy deviations of Ambisonic rendering from the ray-by-ray method are small near the beginning, but increases with propagation time. Nevertheless, the mean deviations are 0.71 dB for the 1st-order Ambisonics, and 1.03 dB for 3rd-order, showing that the diffuse field reproduced by the proposed method is a good estimate in terms of acoustic energy decay.

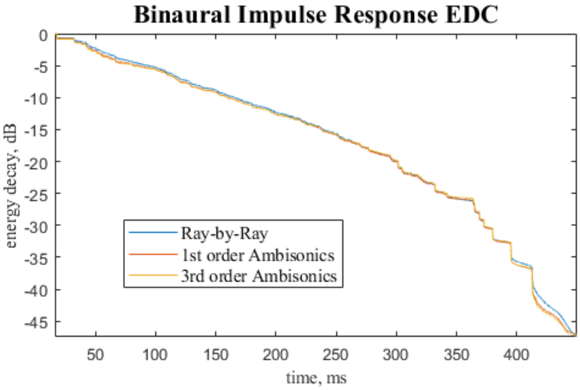

Figure 4. EDC of BRIRs. 0 dB is the total BRIR energy.

C. Interaural Cross-Correlation

The interaural cross-correlation (IACC) is the maximum value of the normalized interaural cross-correlation function between the left and right IRs:

$$\mathrm{IACC} = \max[\Phi_{LR}(\tau)], \ |\tau| \leq 1\, ms \ ,$$

$$\Phi_{LR}(\tau) = \frac{\int_{t_1}^{t_2} p_L(t+\tau)p_R(t)dt}{\sqrt{\int_{t_1}^{t_2} p_L^2(t)dt \int_{t_1}^{t_2} p_R^2(t)dt}}. \qquad (10)$$

Psychoacoustic studies show a close relationship between the IACC and the quality of the spatial sound, and it has been observed that lower values of the late IACC ($t_1 = 80$ ms and $t_2 = 1$ s), imply higher degrees of sound field diffusion [7].

Figure 5 shows that the IACCs from the three rendering cases are very similar, especially for 3rd-order. This indicates a good preservation of the degree of diffusion within the scene.

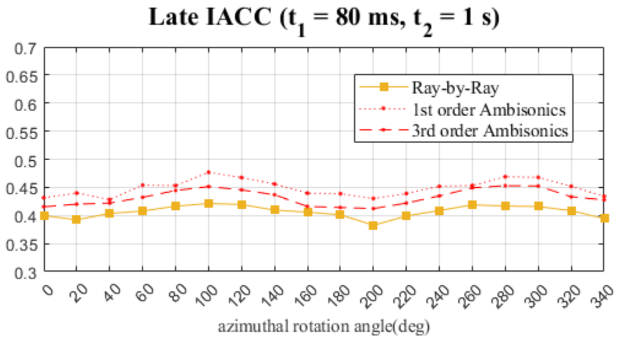

Figure 5. Late IACC of BRIRs, averages of 10 simulation results. Rotation is anti-clockwise from look direction in figure 3.

5. Conclusion

The proposed method reduces the rendering cost for diffuse field simulations by encoding the large Diffuse Rain data in Ambisonics. Instead of individually processing an arbitrary number of image sources, the diffuse field is synthesized in N^{th}-order Ambisonics for just $(N+1)^2$ eigenvectors. Compared with the ray-by-ray method using EDCs and IACCs of the results, it is shown that Ambisonic rendering is able to accurately estimate the energy decay and the degree of sound diffusion. Higher-order Ambisonic rendering yields a closer match of IACCs with the ray-by-ray method, but no other striking differences are observed, suggesting that even 1st-order Ambisonics may be sufficient for perceptually acceptable diffuse sound field simulations.

6. Acknowledgements

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [22ZH1200, The research of the basic media · contents technologies]

7. References

[1] Schröder, D. and Vorländer, M., "*RAVEN: A real-time framework for the auralization of interactive virtual environments*" (2011).

[2] Zotter, F. and Matthias F., "*Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*", Springer Nature. (2019).

[3] Hardin, R.H. and Sloane, N.J.A., "Spherical Designs", (http://neilsloane.com/sphdesigns/), Retrieved October 31, 2022.

[4] Aussal, M., *gypsilab*, (github.com/matthieuaussal/gypsilab), GitHub. Retrieved October 3, 2022.

[5] Brinkmann, F., Dinakaran, M., Pelzer, R., Wohlgemuth, J.J., Seipel, F., Voss, D., Grosche, P. and Weinzierl, S., "*The HUTUBS head-related transfer function (HRTF) database*" (2019).

[6] Smith, J.O., "*Physical Audio Signal Processing*", W3K Publishing. (2010).

[7] Xie, B., "*Head-related transfer function and virtual auditory display*". J. Ross Publishing. 2013.