Room acoustics rendering for immersive audio applications

...

Dr. Orchisama Das

DAFx 24

KING'S College LONDON

CCRMA

PhD student, 2016-2021

REALITY LABS

TESLA

Research intern

IoSR

Postdoc, 2021-2022

SONOS

Senior research scientist, 2022-2024

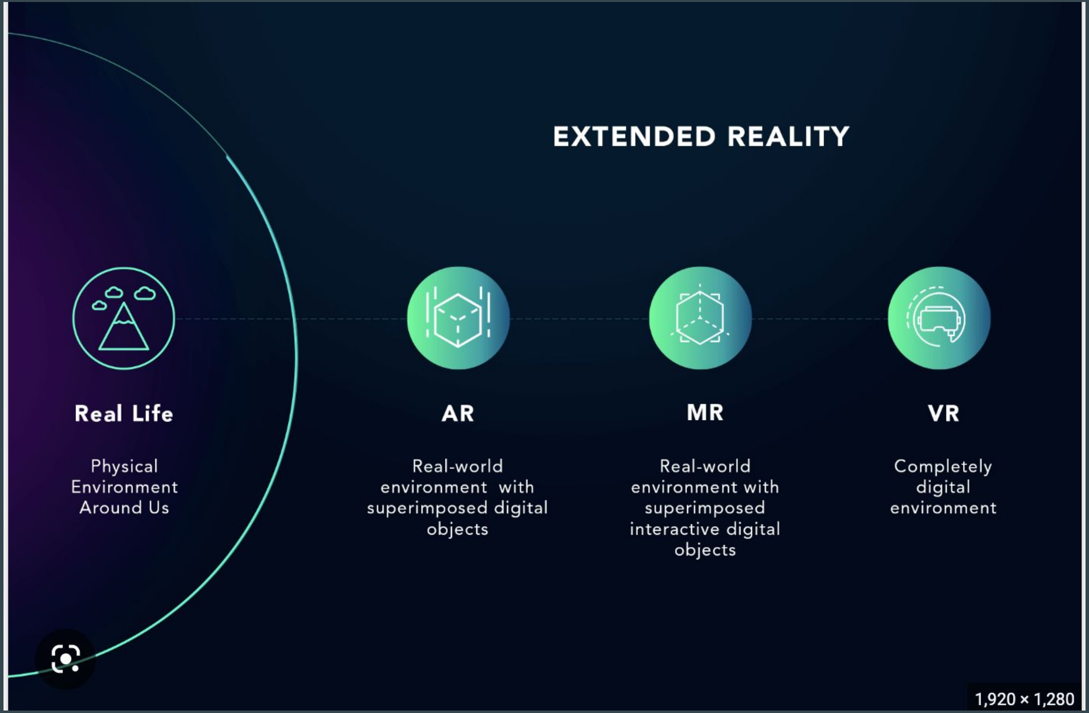
Outline

- Why is room acoustics modelling important for audio reproduction in eXtended Reality?
- How can we achieve spatial audio over headphones?
- Fundamentals of Room Impulse Responses
- Room acoustics rendering with convolution vs parametric delay networks
- 3DoF Binaural Room Impulse Response generation:
  - From simulations: Image-source method
  - From measurements : Spatial Decomposition Method and Spatial Impulse Response Redndering
- Parametric delay networks:
  - Feedback delay networks
  - Scattering delay networks
- Open questions
- Software demo

Link to Github repo :
https://github.com/orchidas/DAFx24-room-acoustics-tutorial/

Scan me!

eXtended Reality (XR)

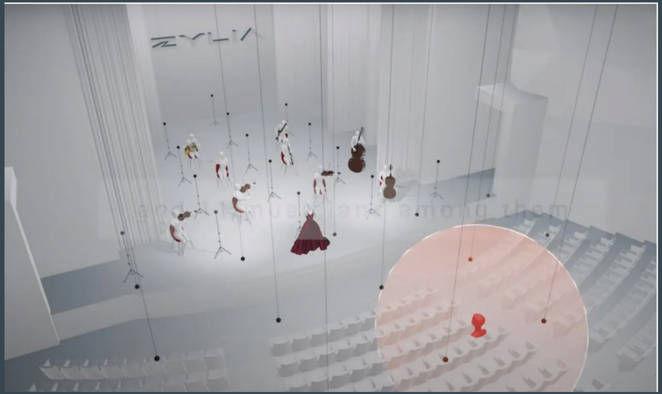


CC : Softengi

Audio in XR

- Audio in XR is spatial and typically delivered over headphones.
    - Re-create an out-loud listening experience over headphones.
- Must be rendered in 6DoF
    - Adaptive to user's head rotation and position translation.
- Must adapt the content being delivered to the user's listening environment.
    - Plausibility is most important in AR and MR.
    - 'Acoustic transparency' – the headphone listening experience is identical to an out-loud listening experience.

CC : Resonance Audio



CC : Zylia

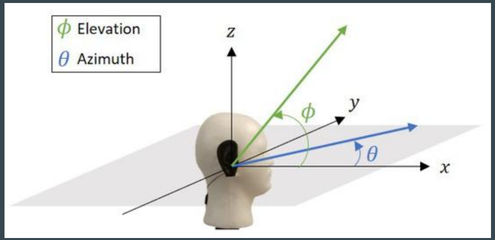
Binaural rendering

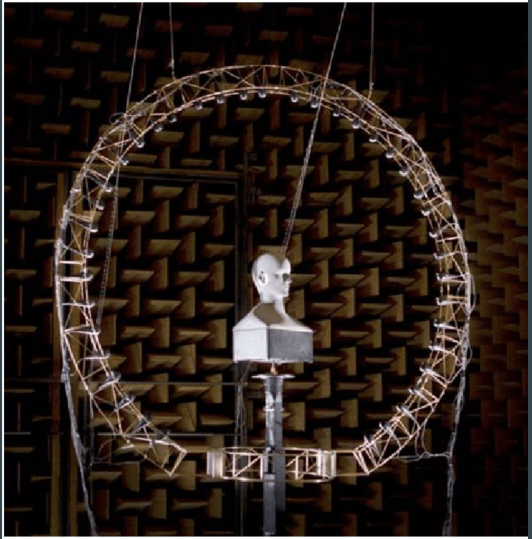
(Sound) objects can be placed in 3D and rendered binaurally (over headphones) using Head-Related Transfer Functions.

- The HRTF describes the transfer function between a point source in free field to the listener's ears.
- A unique HRTF maps each point in the elevation-azimuth plane to the listener's ears.
- Head-tracking used updates to interpolate between different HRTFs during rendering.
- For translation, a simple distance based weighting can be applied to the HRTF.

HRTF measurement is an expensive procedure.

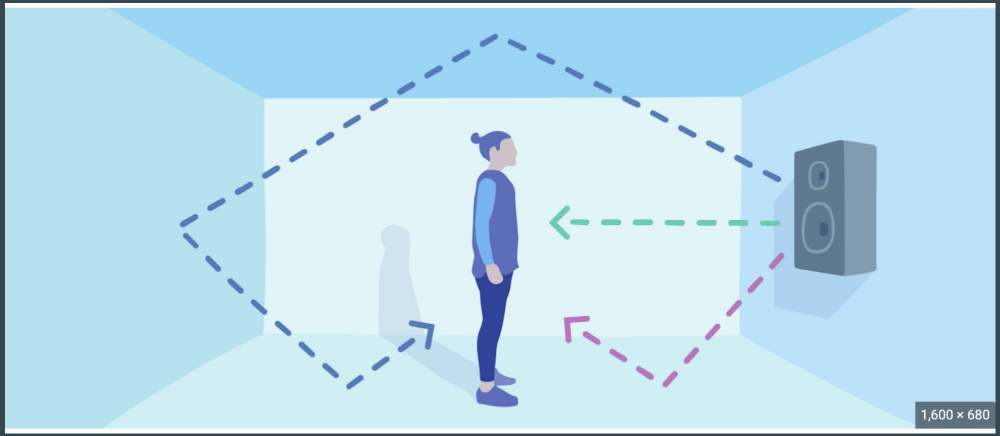
- Databases are used in commercial applications.
- Measurements need to be done in anechoic chambers with circular loudspeaker arrays.
- HRTFs vary according to the head-ear-torso shape of each individual.





Adapting to listening environment

- Recreating the acoustic fingerprint of a listening space is a necessary step in plausibility and acoustic transparency.
  - A concert hall needs to sound like a concert hall and a living room must sound like a living room.
- HRTFs render the direct sound component (in green), **but what about the room reflections?**



Room acoustics breakdown

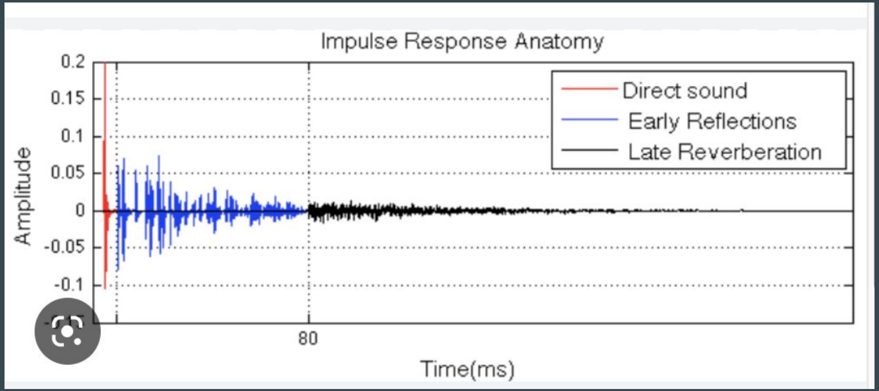
- A room impulse response (RIR) is the impulse response of a room (assuming a linear and time invariant system).
  - To make something sound 'roomy', we need to convolve dry audio with the RIR.
  - Measured by exciting the room with a broadband signal - sine sweep / balloon pop
- Room impulse responses can be decomposed into:
  - Direct path
  - Sparse early reflections
  - Diffuse late field
- Room impulse responses measured for each ear separately is a binaural room impulse response (BRIR).

*Depends on source-listener position and head orientation.

*Static in nature.

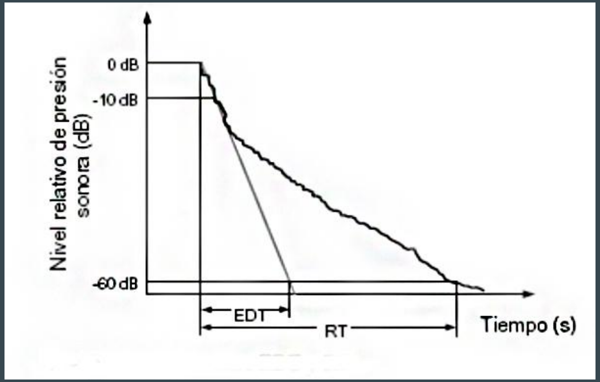
Dry speech

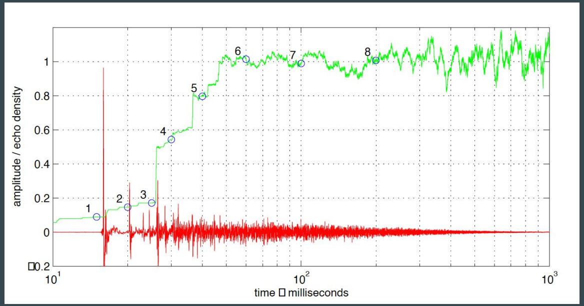
Convolved with RIR



Perceptual attributes

- Reverberation time:
  - Late reverb is modeled as an exponentially decaying white noise in each subband.
  - Time taken for the tail to decay to -60dB : **T60**.
  - Relates to perception of spaciousness and envelopment.
- Early decay time:
  - Time taken by the early reflections to decay after the direct sound.
  - Determines clarity.
- Echo density:
  - Sparse early reflections morph into dense echoes with time. Number of echoes increase polynomially with time.
  - Echo density determines how ‘lush’ or ‘fluttery’ the reverb sounds.
- Direct-to-reverberant ratio:
  - Ratio of the energy of the direct path to the reverberation tail.
  - Determines how ‘dry’ or ‘wet’ the output sounds.

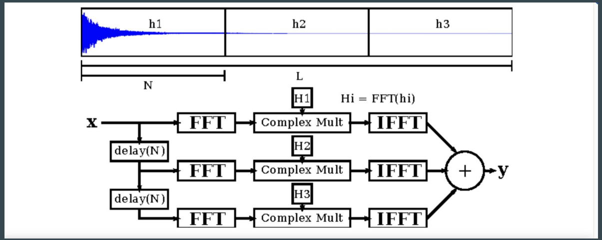




Normalised echo density

Room acoustics rendering techniques

- Via real-time convolution
  - BRIRs are usually a few hundred ms to a few seconds long (depending on T60 of room).
  - Partitioned convolution is implemented in the frequency domain in real-time.[1]
  - BRIRs are updated from a database based on head-tracking updates.



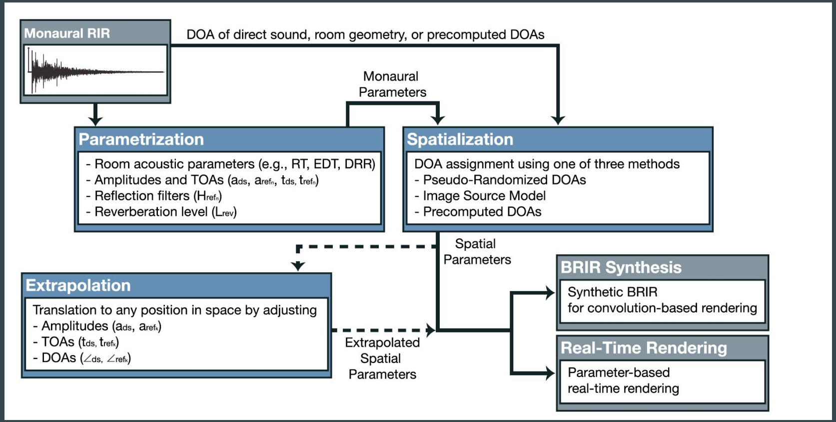
Real-time partition convolution
CC:Battenberg, DAFx-11

- Via parametric delay-networks
  - Uses a network of delay lines connected via a feedback loop.
  - Architecture ensures echo density increases with time, producing dense reverberation.
  - Losses are introduced via absorption filters.
  - Source-receiver position, head orientation updated in real-time by modulating the delay lines.

Part 1 - Generating BRIRs for rendering via convolution

- Via measurements[2]
  - Just like HRTFs, we can measure a full set of BRIRs and interpolate between them in real-time with head tracking.
  - Late tail can be rendered statically (no need to update with tracking).
  - Infeasible because most people do not have a multi-speaker set-up in their listening environments.
- Via simulations
  - Geometric methods such as
    - Image-source model[3,4] in shoebox rooms.
    - Ray-tracing (eq: CATT acoustics)
  - Recently, wave based solvers are becoming commercially available (eg: Treble technologies).
  - Infeasible unless we know exact geometry and room materials.

- Via parameterisation
  - Spatial Decomposition Method[5,6]
    - Detect the direction of arrival of each reflection in a room impulse response (RIR) using a microphone array.
    - Binaural generation using a weighted sum of HRTFs.
    - Spectral whitening due to incorrect DoA estimation in late tail needs to be corrected for.
  - Paraspax[7]
    - Only needs one omni measurement.
    - Parameter extraction of psychoacoustically important features, followed by extrapolation.



Paraspax breakdown
CC : Journal of the AES

BRIR generation with Image-source model

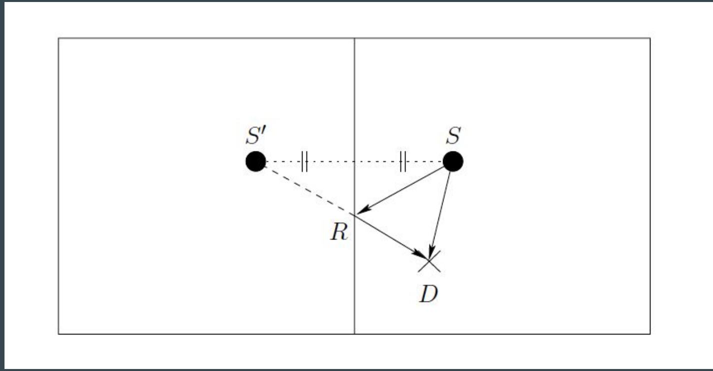
- Exact solution of the wave equation for shoebox rooms with rigid walls.

$$\nabla^2 p(\boldsymbol{r},t) - \frac{1}{c^2}\frac{\partial^2 p(\boldsymbol{r},t)}{\partial^2} = -s(\boldsymbol{r},t),$$

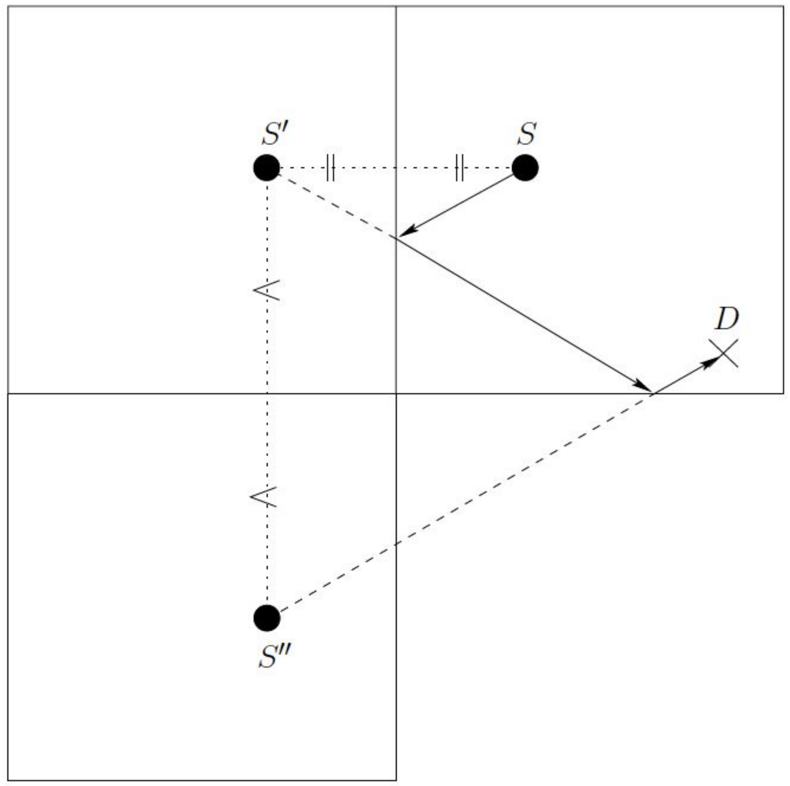
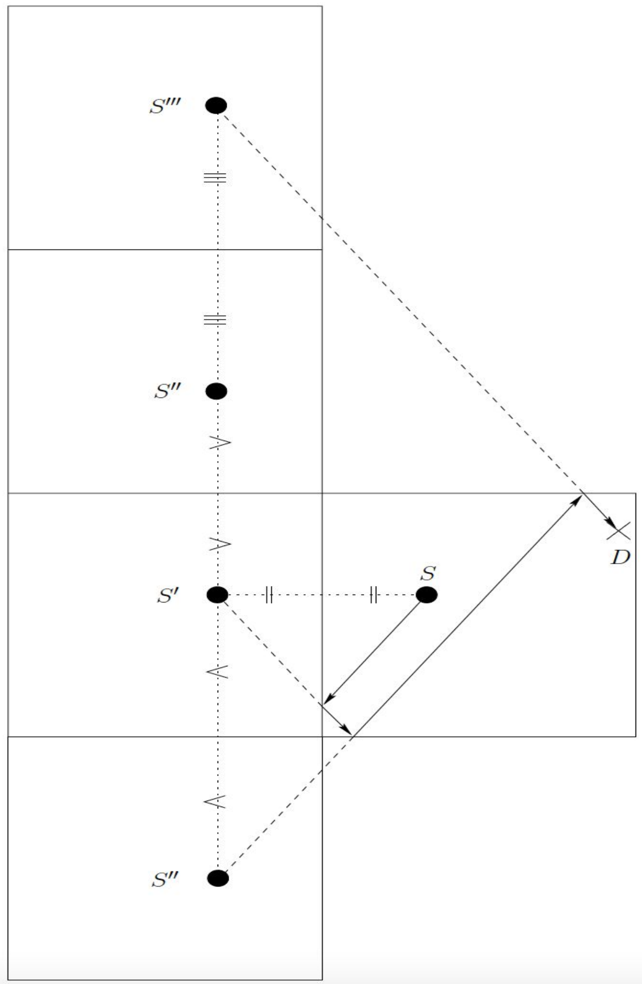
$$\nabla^2 P(\boldsymbol{r},\omega) + k^2 P(\boldsymbol{r},\omega) = -S(\boldsymbol{r},\omega)$$

Helmholtz equation

- Virtual image source associated with each wall in the room.
- RIR is a scaled sum of delta functions, each corresponding to a reflection from a unique image source.



CC:Habets, RIRGenerator


S′
S
D
S″

S‴
S″
S′
S
D
S″

$$\begin{aligned} \text{Source location } &:= (x_s, y_s, z_s) \\ \text{Mic location } &:= (x, y, z) \\ \text{Room dimensions } &:= (L_x, L_y, L_z) \end{aligned}$$

$$\mathcal{M} = \{(m_x, m_y, m_z) : m_x, m_y, m_z \in (-N, \ldots, N)\};$$

$$\mathcal{P} = \{(q, j, k) : q, j, k \in (0, 1)\}$$

$$\mathbf{R_p} = [(1 - 2q)x_s - x, (1 - 2j)y_s - y, (1 - 2k)z_s - z]$$

$$\mathbf{R_m} = [2m_x L_x,\ 2m_y L_y,\ 2m_z L_z]$$

Distance between each image source and mic

$$h(\mathbf{r}, \mathbf{r}_s, t) = \sum_{\mathbf{p} \in \mathcal{P}} \sum_{\mathbf{m} \in \mathcal{M}} \frac{\delta(t - \|\mathbf{R_p} + \mathbf{R_m}\|/c)}{4\pi\|\mathbf{R_p} + \mathbf{R_m}\|}$$

$$= \sum_{\mathbf{p} \in \mathcal{P}} \sum_{\mathbf{m} \in \mathcal{M}} \frac{\delta(t - \tau)}{4\pi d}$$

$$\boxed{H(\mathbf{r},\mathbf{r}_s;\omega)=\sum_{\mathbf{p}\in\mathcal{P}}\sum_{\mathbf{m}\in\mathcal{M}}\frac{\exp(ik\|\mathbf{R_p}+\mathbf{R_m}\|)}{4\pi\|\mathbf{R_p}+\mathbf{R_m}\|}}$$

Solution to Helmholtz equation for a point source

Adding wall absorption

Reflection coefficients of 6 walls

$$h(\mathbf{r},\mathbf{r}_s,t)=\sum_{\mathbf{n}\in\mathcal{P}}\sum_{\mathbf{m}\in\mathcal{M}}\beta_{x_1}^{|m_x-q|}\beta_{x_2}^{|m_x|}\beta_{y_1}^{|m_y-j|}\beta_{y_2}^{|m_y|}\beta_{z_1}^{|m_z-k|}\beta_{z_2}^{|m_z|}\frac{\delta(t-\tau)}{4\pi d},$$

Binauralising the Image method

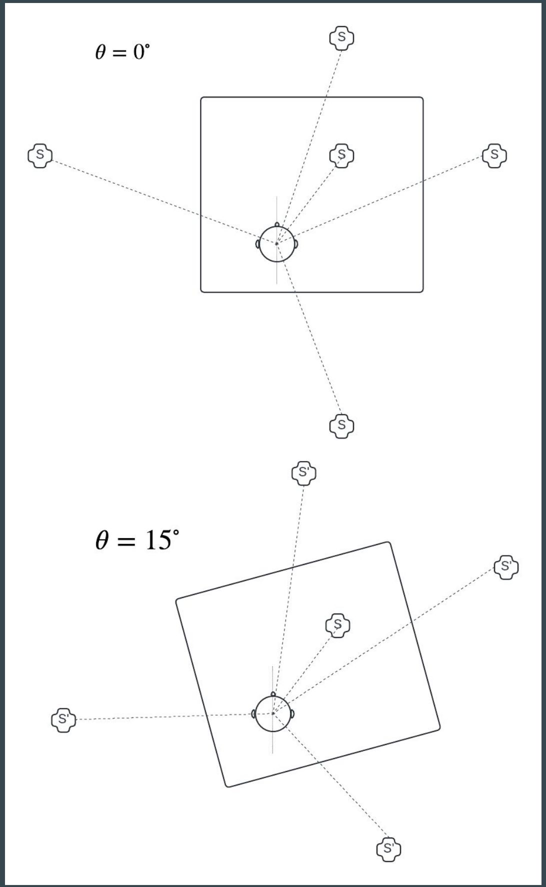
Binauralising the RIR for with head rotation

$$h_{L,R}(\mathbf{r},\mathbf{r}_s,t)=\sum_{\mathbf{p}\in\mathcal{P}}\sum_{\mathbf{m}\in\mathcal{M}}\frac{\delta(t-\|\mathbf{R_p}+\mathbf{R_m}\|/c)}{4\pi\|\mathbf{R_p}+\mathbf{R_m}\|}*\mathrm{hrir}_{L,R}(t,\theta_{p,m},\phi_{p,m})$$

$\theta_{p,m}$, $\phi_{p,m}$ are the angles between the image source and listener

For a head rotation of (θ, ϕ), the room rotates in the opposite direction by $(-\theta, -\phi)$, so does the position of the image sources, and hence, the DoAs

$$h_{L,R}(\mathbf{r},\mathbf{r}_s,t,\theta,\phi)=\sum_{\mathbf{p}\in\mathcal{P}}\sum_{\mathbf{m}\in\mathcal{M}}\frac{\delta(t-\|\mathbf{R_p}+\mathbf{R_m}\|/c)}{4\pi\|\mathbf{R_p}+\mathbf{R_m}\|}*\mathrm{hrir}_{L,R}(t,\theta_{p,m}-\theta,\phi_{p,m}-\phi)$$



BRIR generation from measurements - Spatial Decomposition Method[5,6]

- Image method / ray tracing work if the properties of the room - eg: geometry, absorption coefficients are known.
- How to parameterise a measured space for 3DoF rendering?
- Spatial Decomposition Method:
    - DoA of each reflection is estimated from a multichannel RIR
    - DoA is mapped to corresponding direction using loudspeaker or headphone based reproduction method.
    - DoA quantisation ensures that reflections spread out in time are assigned the same direction.
    - Spectral whitening of late reverberation is avoided by altering the late decay time in subbands and passing through a cascade of allpass filters.

SDM for binaural reproduction

DoA matrix estimated by time-difference of arrival method.

$$\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N] \in \mathbb{R}^{3\times N}$$

Central mic captures omni pressure response

$$\boldsymbol{p} = [p_1, \ldots, p_N] \in \mathbb{R}^{1\times N}$$

Rotated DoA matrix corresponding to head rotation of $\boldsymbol{\theta}$ (azimuth) and $\boldsymbol{\Phi}$ (elevation)

$$\boldsymbol{D}^u = \boldsymbol{R}_z(-\theta_u)\boldsymbol{R}_y(-\phi_u)\boldsymbol{D}$$

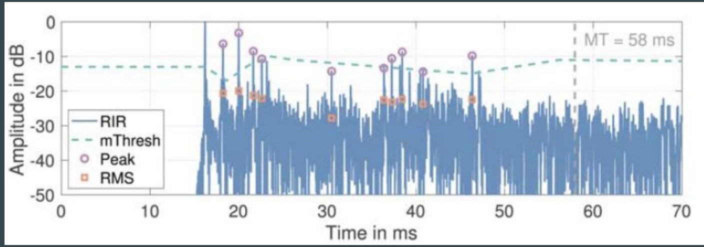
Indices of closest HRIRs for each sound event in each head orientation are selected by Euclidean distance minimisation

$$\hat{k}_n^u = \arg \min_{n\in 1,\ldots,N} d\left[\boldsymbol{D}_n^u, \hat{\boldsymbol{D}}\right],$$

$\hat{\boldsymbol{D}} \in \mathbb{R}^{3\times K}$ are DoAs from HRTF set

BRIRs at any head orientation constructed by delaying HRIRs corresponding to the right DoA and weighting them by the pressure

$$\mathrm{BRIR}^u(t) = \sum_{n=1}^{N} p_n \, \mathrm{HRIR}_{\hat{k}_n^u} \circledast \delta(t-n), \qquad \mathrm{HRIR} \in \mathbb{R}^{L\times K\times 2}$$



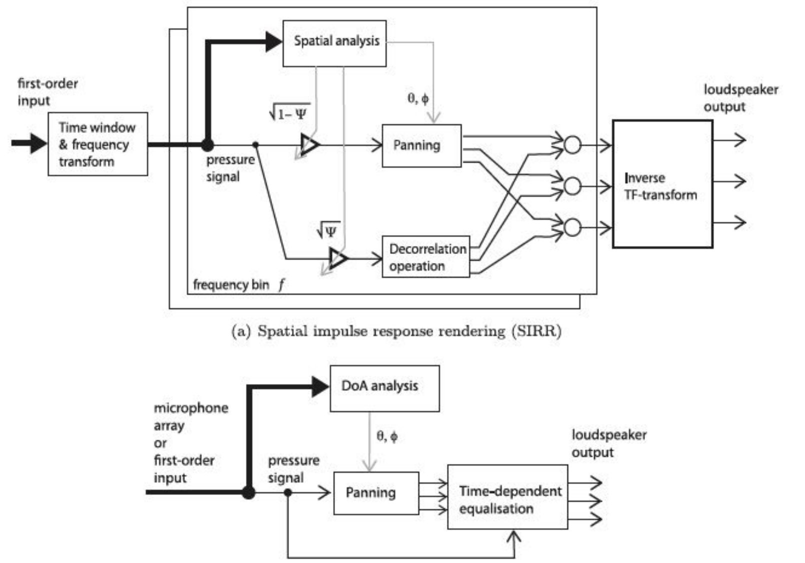
CC: Gari et al., Journal of the Audio Engineering Society

BRIR generation from measurements - Spatial Impulse Response Rendering Method[8]

- Uses first-order ambisonic microphones to record RIR.
- Analysis of each time-frequency component with STFT.
- Pseudo intensity vectors used for estimation of DoAs in each TF bin.
- Each TF bin split into diffuse and non-diffuse parts based on a diffuseness metric.
    - Non-diffuse part of the omni response is reproduced from estimated DoA using vector-based amplitude panning (VBAP).
    - Diffuse part reproduced by decorrelation and distribution uniformly around listener.
- Rotations imposed by modifying the estimated DoAs.

CC: Rode

SIRR vs SDM[9]



(a) Spatial impulse response rendering (SIRR)

(b) Spatial decomposition method (SDM)

CC: McCormack et al, Journal of the Audio Engineering Society

Advantages and Drawbacks

- Once a full BRIR dataset has been generated, convolution, interpolation and rendering is simple.
- SDM and SIRR provide a good perceptual match to real rooms.
- Image method is simple to understand and implement and is the choice for dataset generation for ML applications.
- 6DoF BRIR dataset storage and lookup for online rendering is expensive. Data for a single room can be a few GB.
- All methods are approximate, not physically accurate.

Part 2 - Rendering with parametric delay networks

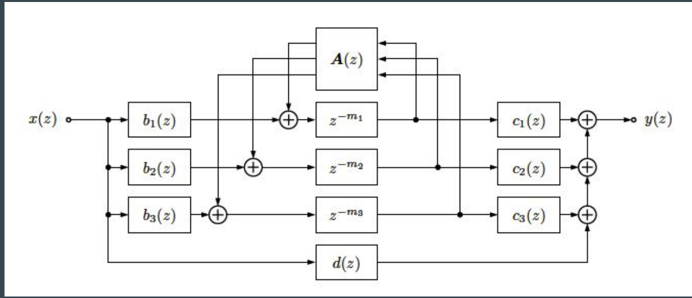
- Takes the dry audio signal as input and gives a mono/binaural/multichannel reverberated output.
- No pregenerated BRIR set required for 6DoF audio.
    - Instead we tune their parameters to produce a BRIR-like impulse response.
    - The parameters are updated in real-time to model 6 DoF movement.
    - More efficient than convolution with long impulse responses.
- General architecture:
    - Network of delay lines that are interconnected with feedback.
    - Operation in the time domain.
    - Number of echoes build up over time.

Feedback Delay Networks[10,11]

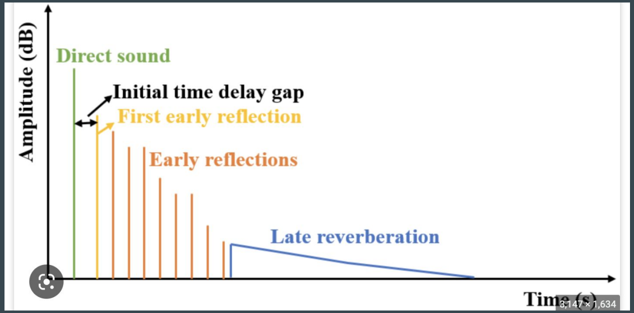
- Highly parameterisable network used for rendering **late reverberation.**
- Delay lines connected via a lossless feedback matrix.
- Losses are introduced via absorption filters.
    - These are tuned based on the frequency-dependent T60s in a room.
- Number of delay lines and feedback matrix controls build-up of echo-density.
- Input-output filters (b,c) and delay line lengths control source-listener positions.
    - Input-output filters also control equalisation.
    - Delay line lengths are co-prime to prevent comb filtering.
- Direct path filter d(z) is required to reproduce the direct path accurately

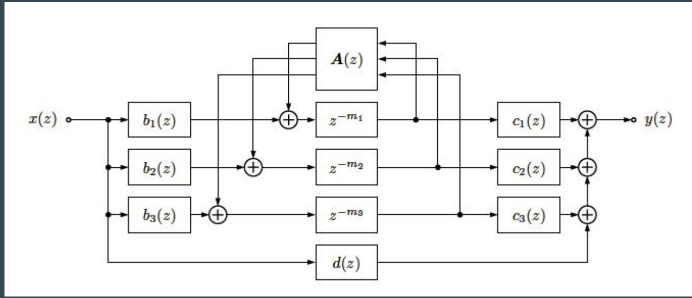
CONS:

- Parameter tuning is an art.
- Replicating a space is tricky.
- Model is not physical.



CC: Schlecht, DAFx-20





State-space representation of SISO FDN:

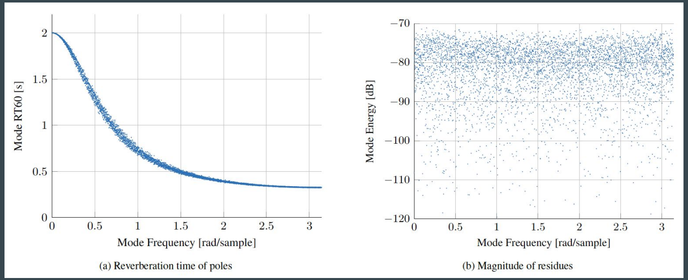
$$\begin{aligned}
\boldsymbol{s}(n) &= \boldsymbol{A}\boldsymbol{s}(n-m) + \boldsymbol{b}x(n) \\
y(n) &= \boldsymbol{c}^T\boldsymbol{s}(n) + dx(n), \\
\boldsymbol{s}(n) &= [s_1(n), s_2(n), \ldots, s_N(n)]^T \\
\boldsymbol{s}(n-m) &= [s_1(n-m_1), s_2(n-m_2), \ldots, s_N(n-m_N)]^T
\end{aligned}$$

Transfer function:

$$\begin{aligned}
H(z) = \frac{Y(z)}{X(z)} &= \boldsymbol{c}^T(z)\left[\boldsymbol{D}_m\left(z^{-1}\right) - \boldsymbol{A}(z)\right]^{-1}\boldsymbol{b}(z) + d(z) \\
&= \boldsymbol{c}^T(z)\boldsymbol{P}(z)^{-1}\boldsymbol{b}(z) + d(z) \\
&= d + \sum_{i=1}^{\mathcal{R}} \frac{\sigma_i}{1-\lambda_i z^{-1}} \\
\boldsymbol{D}_m\left(z^{-1}\right) &= \mathrm{Diag}\left[z^{-m_1}, \ldots, z^{-m_N}\right]
\end{aligned}$$

- Modes, $\boldsymbol{\lambda}_i$ are roots of the characteristic polynomial[12], $\boldsymbol{P(z)}$
- Number of modes is the sum of the delay line lengths, $\mathcal{R} = \sum_{i=1}^{N} m_i$
- Attenuation is introduced by replacing each delay element, z^{-1} ,with a lossy filter, $\gamma_i(z)z^{-1}$, $\gamma_{dB}(e^{j\omega}) = \dfrac{-60}{f_s T_{60}(\omega)}$
- With attenuation, the characteristic polynomial is $\boldsymbol{P'(z)}$ and the modes are attenuated to be $\boldsymbol{\lambda'}_i$

$$\boldsymbol{P}'(z) = \boldsymbol{D}_m\left(z^{-1}\right)\boldsymbol{\Gamma}_m(z) - \boldsymbol{A}(z)$$
$$\boldsymbol{\Gamma}_m(z) = \mathrm{Diag}[\gamma_1(z), \ldots, \gamma_N(z)]$$
$$\lambda_i = \lambda'_i \left|\Gamma\left(\Lambda'_i\right)\right|^{-1}$$

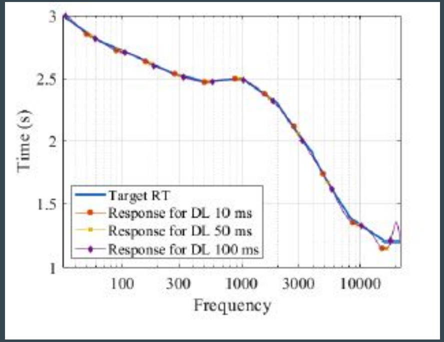


CC: Schlecht, DAFx-20

CC: Prawda, DAFx-19

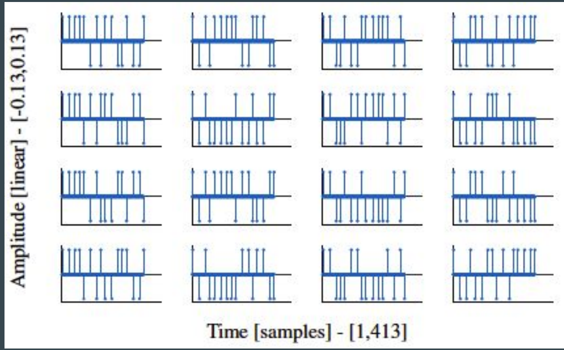
Tuning the FDN

- Attenuation filter design to match a measured space[13,14]
  - Graphic equaliser to match measured subband T60.
  - Alternately, an IIR fitting method such as warped Prony's method can be used.
- Feedback matrix optimisation[15]
  - Feedback matrices are typically designed to be unitary (lossless) with the property $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$
  - The feedback matrix directly controls the echo density
    - To increase the echo density, scalar matrices can be replaced with filter feedback matrices (FFM) with FIR filters in each element of the matrix.
    - FFMs need to be paraunitary to preserve energy, ie., $\boldsymbol{A}^T(z^{-1})\,\boldsymbol{A}(z) = \boldsymbol{I}$
    - Velvet FFMs consist of a sparse set of $\mp$1s and are shown to increase echo density with minimal computation



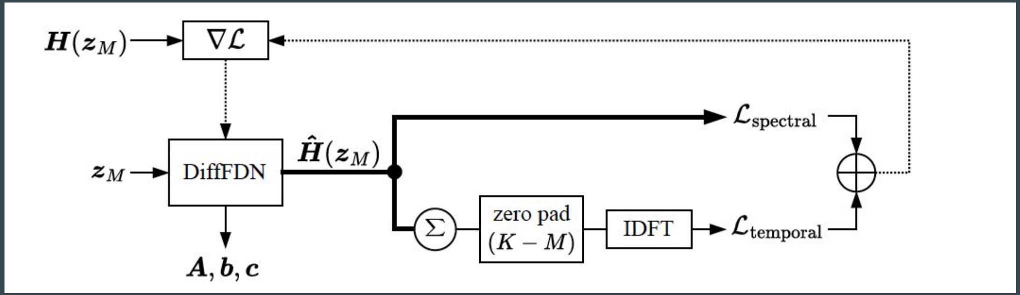
GEQ filter fitting for a measured concert hall

CC: Schlecht, IEEE TASLP



Velvet feedback matrix example for 4 delay line FDN

- Tuning input and output gains[16,17]
  - FDN allpass completion finds b's and c's for fixed delay line lengths and fixed feedback matrix to get an allpass magnitude response.
  - Differentiable FDN tunes the input-output gains as well as the feedback matrix to achieve maximally flat magnitude response whilst maintaining a dense distribution of echoes.
    - Samples the FDN response as an FIR filter at M frequency bins
    - Learns the parameters, $\boldsymbol{A}$, $\mathbf{b}$ and $\mathbf{c}$ for fixed delay line lengths and a homogeneous decay.



Differentiable FDN training architecture

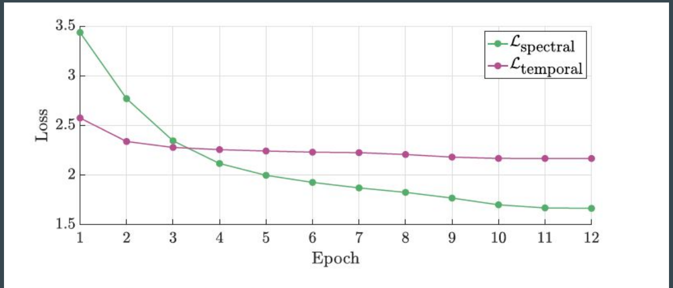


Figure 5: *Progression of spectral and temporal components of the loss function during optimization.*

Differentiable FDN training results

CC: Dal Santo, DAFx-23

Binaural late reverb generation with FDN

- Interaural coherence matching[18,19]

Uncorrelated output from FDN

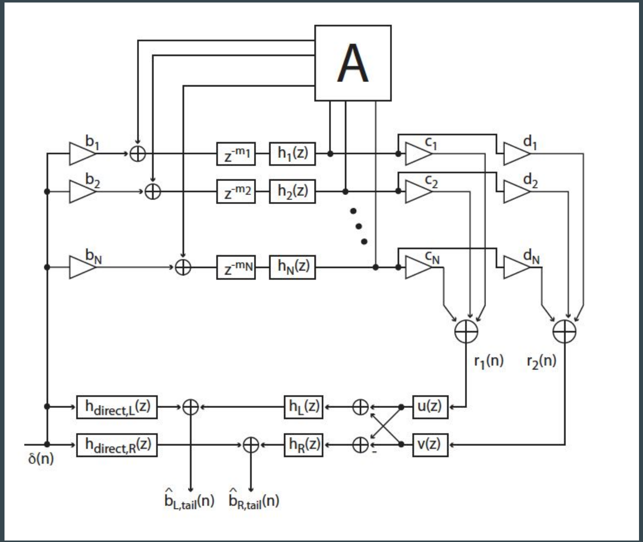
$$\hat{b}_{L,tail}(n) = (u * r_1 + v * r_2)(n)$$
$$\hat{b}_{R,tail}(n) = (u * r_1 - v * r_2)(n)$$

Interaural coherence

Derived from IAC of HRTF/BRIR set

$$\Phi(\omega) = \frac{\left|\sum_{k=1}^{K} H_L(\omega,k) H_R^*(\omega,k)\right|}{\sqrt{\sum_{k=1}^{K} |H_L(\omega,k)|^2 \sum_{k=1}^{K} |H_R(\omega,k)|^2}}$$

$$U(\omega) = \sqrt{\frac{1+\phi(\omega)}{2}}, \quad V(\omega) = \sqrt{\frac{1-\phi(\omega)}{2}}$$

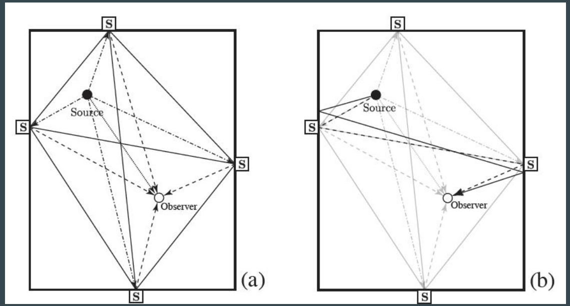


CC: Menzer, Journal of the Audio Engineering Society

Scattering Delay Networks[20]

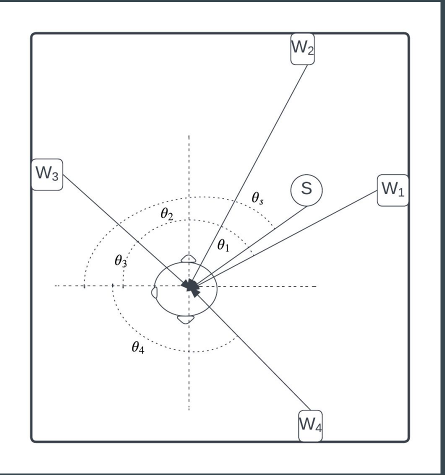
- Unlike the FDN, Scattering delay networks can model the entire room response accurately - direct path, early reflections and late reverberation.
- Combines digital waveguide mesh with FDNs.
  - Deals with pressure variables directly.
  - Each reflection point on a wall is a node.
  - Bi-directional delay lines connect different nodes.
  - Uni-directional delay lines connect source and nodes to listener.
  - Lossless scattering matrix at each node scatters the incoming pressure variables to the other nodes.
  - Filters in delay lines introduce frequency-dependent losses.
  - 1/r attenuation modeled by input and output gains.

CC: De Sena, IEEE TASLP



SDN for 2D room

- Renders first-order reflections exactly and higher-order reflections approximately.
- Delay line lengths and node positions updated in real-time as the listener moves around the room.
- Parameterisable - source and listener directivities can easily be modeled
- Binauralisation is simple and elegant



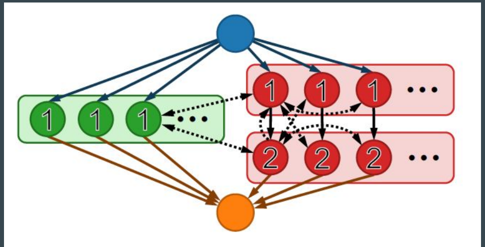
Output of delay line between source and mic

$$y_L = s * h_l(\theta_s, \phi_s) + \sum_{i=1}^{6} w_i * h_L(\theta_i, \phi_i)$$

$$y_R = s * h_R(\theta_s, \phi_s) + \sum_{i=1}^{6} w_i * h_R(\theta_i, \phi_i)$$

Output of delay line between ith node and mic

Higher-order SDN[21,22]



2nd order SDN topology

- Additional placement of nodes where higher-order image sources would be.
- For N^{th} order SDN,
  - First $(N-1)^{th}$ order nodes are connected unidirectionally to the source and the receiver.
  - N^{th} order nodes are grouped according to their "bounce" number:
    - Source connected to the first bounce nodes.
    - Receiver connected to the last bounce nodes.
    - 1^{st} bounce nodes -> 2^{nd} bounce nodes -> .... -> N^{th} bounce nodes
    - Recursive (bidirectional) connections added randomly between nodes.
    - A low-complexity network architecture places the a single N^{th} order node on each wall (instead of $6 * 5^{N-1}$ nodes)
      - Position of single node determined by centroid, first-bounce or wall-centre

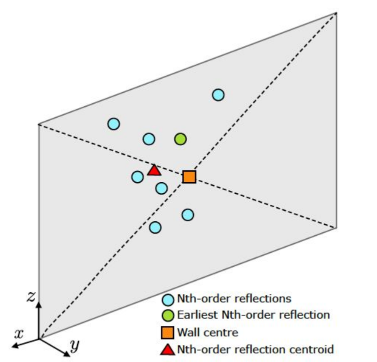


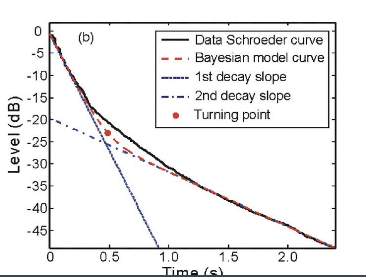
Figure 3: Diagram of one of the room walls explaining the proposed node placement strategies.

Advantages and Drawbacks

- Delay networks are efficient to implement and can be run on resource constrained devices.
- Full generation of 6DoF BRIR dataset and online lookup is not required.
- Position translation is much more intuitive in delay networks compared to data-based convolution reverberators.
- Tuning them to sound be perceptually indistinguishable from a measured room is still a challenge.
- They do not model wave effects like scattering and diffraction.

Open challenges

- Perceptually indistinguishable is a high bar that has not yet been achieved. At best, we aim for plausibility.
  - Whilst 3DoF rendering with head-tracking is simple, 6DoF rendering with position translation is more complicated.
  - Data-driven RIR interpolation methods are promising[23,24] since they don't rely on simple geometric assumptions.
- Dynamic rendering in coupled spaces
  - Anisotropic two stage decay is observed in coupled spaces.
  - Late reverberation models need to account for this.
  - Recent surge in literature on the topic.[14,25,26]
- HRTF personalisation
  - HRTF varies among individuals, thus we localise sounds differently from each other.
  - To get the best immersive experience, personalised HRTFs* should be used.



CC: Xiang, JASA

* Apple introduces personalised HRTFs

References

1. Battenberg, Eric, and Rimas Avizienis. "Implementing real-time partitioned convolution algorithms on conventional operating systems." In *Proceedings of the 14th International Conference on Digital Audio Effects. Paris, France*, pp. 248-235. 2011.
2. Satongar, Darius, Yiu W. Lam, and Chris Pike. "Measurement and analysis of a spatially sampled binaural room impulse response dataset." In *21st International Congress on Sound and Vibration*, pp. 1-8. 2014.
3. Allen, Jont B., and David A. Berkley. "Image method for efficiently simulating small‐room acoustics." *The Journal of the Acoustical Society of America* 65, no. 4 (1979): 943-950.
4. Habets, Emanuel AP. "Room impulse response generator." *Technische Universiteit Eindhoven, Tech. Rep* 2, no. 2.4 (2006): 1.
5. Amengual Garí, Sebastià V., Johannes M. Arend, Paul T. Calamia, and Philip W. Robinson. "Optimizations of the spatial decomposition method for binaural reproduction." *Journal of the Audio Engineering Society* 68, no. 12 (2021): 959-976.
6. Tervo, S., Pätynen, J., Kuusinen, A. and Lokki, T., 2013. Spatial decomposition method for room impulse responses. *Journal of the Audio Engineering Society*, *61*(1/2), pp.17-28.
7. Arend, Johannes M., Sebastià V. Amengual Garí, Carl Schissler, Florian Klein, and Philip W. Robinson. "Six-degrees-of-freedom parametric spatial audio based on one monaural room impulse response." *Journal of the Audio Engineering Society* 69, no. 7/8 (2021): 557-575.
8. Merimaa, Juha, and Ville Pulkki. "Spatial impulse response rendering I: Analysis and synthesis." *Journal of the Audio Engineering Society* 53, no. 12 (2005): 1115-1127.
9. McCormack, L., Pulkki, V., Politis, A., Scheuregger, O. and Marschall, M., 2020. Higher-order spatial impulse response rendering: Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution. *Journal of the Audio Engineering Society*, *68*(5), pp.338-354.
10. Jot, J.M. and Chaigne, A., 1991, February. Digital delay networks for designing artificial reverberators. In *Audio Engineering Society Convention 90*. Audio Engineering Society.

11. Schlecht, S., 2020. FDNTB: The feedback delay network toolbox. In *International Conference on Digital Audio Effects* (pp. 211-218). DAFx.
12. Schlecht, S.J. and Habets, E.A., 2019. Modal decomposition of feedback delay networks. *IEEE Transactions on Signal Processing*, *67*(20), pp.5340-5351.
13. Prawda, K., Välimäki, V. and Schlecht, S., 2019, September. Improved reverberation time control for feedback delay networks. In *International Conference on Digital Audio Effects*. University of Birmingham.
14. Das, O. and Abel, J.S., 2021. Grouped feedback delay networks for modeling of coupled spaces. *Journal of the Audio Engineering Society*, *69*(7/8), pp.486-496.
15. Schlecht, S.J. and Habets, E.A., 2020. Scattering in feedback delay networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, pp.1915-1924.
16. Schlecht, S.J., 2021. Allpass feedback delay networks. *IEEE Transactions on Signal Processing*, *69*, pp.1028-1038.
17. Dal Santo, G., Prawda, K., Schlecht, S. and Välimäki, V., 2023, September. Differentiable feedback delay network for colorless reverberation. In *International Conference on Digital Audio Effects* (pp. 244-251). Aalborg University.
18. Menzer, F. and Faller, C., 2010. Investigations on an early-reflection-free model for BRIRs. *Journal of the Audio Engineering Society*, *58*(9), pp.709-723.
19. Menzer, F. and Faller, C., 2009. Binaural reverberation using a modified Jot reverberator with frequency-dependent interaural coherence matching. In *Proceedings of the 126th AES Convention*.
20. De Sena, E., Hacıhabiboğlu, H., Cvetković, Z. and Smith, J.O., 2015. Efficient synthesis of room acoustics via scattering delay networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(9), pp.1478-1492.
21. Scerbo, M., Das, O., Friend, P. and De Sena, E., 2022. Higher-order scattering delay networks for artificial reverberation. In *Proceedings of the 25th International Conference on Digital Audio Effects (DAFx-22)*.
22. Vinceslas, L., Scerbo, M., Hacıhabiboğlu, H., Cvetković, Z. and De Sena, E., 2023, October. Low-Complexity Higher Order Scattering Delay Networks. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 1-5). IEEE.
23. Antonello, N., De Sena, E., Moonen, M., Naylor, P.A. and Van Waterschoot, T., 2017. Room impulse response interpolation using a sparse spatio-temporal representation of the sound field. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(10), pp.1929-1941.
24. Das, O., Calamia, P. and Gari, S.V.A., 2021, June. Room impulse response interpolation from a sparse set of measurements using a modal architecture. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 960-964). IEEE.
25. Atalay, T.B., Gül, Z.S., De Sena, E., Cvetković, Z. and Hacıhabiboğlu, H., 2022. Scattering delay network simulator of coupled volume acoustics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, pp.582-593.
26. Hold, C., McKenzie, T., Götz, G., Schlecht, S. and Pulkki, V., 2022. Resynthesis of spatial room impulse response tails with anisotropic multi-slope decays. *journal of the audio engineering society*, *70*(6), pp.526-538.