Fast Source-Room-Receiver Acoustics Modeling

1st Yuancheng Luo
Amazon Inc.
mikeluo@amazon.com

2st Wontak Kim
Amazon Inc.
wontakk@amazon.com

***Abstract*—Advances in speaker, room, and device acoustic modeling have given rise to large scale simulations of their spatial-frequency responses suitable for tasks such as rapid hardware prototyping, audio front-end algorithm validation, and back-end data-set augmentation for machine learning. Joint modeling of sources, rooms, and receivers is computationally prohibitive due to the large combinatorial space, coupling between models, and overhead cost of data exchange. To address these issues, we introduce the complex spherical harmonics as a separable set of basis functions for representing each of these models and their first-order interactions. We then present a partitioned frequency-dependent image-source model expanded into the spherical harmonics for efficient impulse response synthesis. Results are validated against real-world measurements.**

***Index Terms*—Spherical harmonics, image-source, image-receiver, room acoustics, impulse response**

I. Introduction

Recent progress in speaker, room, and device acoustic modeling have found renewed applications in hardware microphone array design [3], audio front-end algorithm development [4], and back-end automatic speech recognition (ASR) [12], [13]. As computing costs decrease, these applications may find realizable benefits to joint or coupled modeling between sources-rooms-receivers in the spatial time-frequency domain. We propose a separable first-order approximation of the joint system response using the complex Spherical Harmonics (SH) [16] that is agnostic to the underlying source-room-receiver models. The separable format reduces the computational complexity by allowing for independent modeling and permutation of items from each category. Acoustic free-field responses of varied speaker designs and device form factors can be independently simulated using techniques such as finite and boundary element methods in commercial packages [5]. Their results are then decomposed along the SH basis functions and stored before efficiently permuting through all combinations of joint system responses realized into impulse responses (IR).

The SH basis functions have found varied applications in fitting both empirical and simulated measurements (see Fig. 1). Early applications saw use in representing head-related impulse responses [1], [7], [22] for spatial audio and binaural rendering. Plane-waves decomposition for microphone arrays recordings use the SH in its formulation [6], [18]. Acoustic beamforming filters in the SH domain can be designed [15]. To demonstrate an example of the separable SH format, we modify and then extend the well-known image-source model (ISM) [2], [8], [9], [14] traditionally used in room acoustics simulation and microphone array validation [21] via a cross-modal formulation [20]. Section II presents a new derivation of the SH formulation to source-room-receiver models in terms of matrix operators. Section III introduces image-receivers to ISM and derives the partitioned frequency-dependent image-source model (FDISM) with SH extension. Section IV shows experimental results and validation with measurement data.

II. Separable Spherical Harmonic Modeling

Spherical harmonics are orthogonal basis functions over spherical coordinates at degree l and order m given by

$$Y_l^m(\theta,\phi) = \sqrt{\frac{(2l+1)}{4\pi}\frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta) e^{im\phi}, \quad (1)$$

where (θ,ϕ) are the physics convention spherical coordinates for colatitude and azimuth respectively, and $P_l^m(\cos\theta)$ are the associated Legendre polynomials. Orthogonality or the integration of any two basis functions over spherical coordinates is given by

$$\int_\Omega Y_l^m(\theta,\phi) Y_{\bar{l}}^{\bar{m}*}(\theta,\phi) d\Omega = \delta_{l\bar{l}}\delta_{m\bar{m}}, \quad (2)$$

where the $*$ is the complex conjugate and δ is the Kronecker delta function. An arbitrary function $f(\theta,\phi)$ over the spherical coordinates can be decomposed into SH given by

$$f(\theta,\phi) \approx \sum_{l=0}^{P}\sum_{m=-l}^{l} C_l^m Y_l^m(\theta,\phi) = \mathbf{C}^T\mathbf{Y}(\theta,\phi), \quad (3)$$

where $C_l^m \in \mathbb{C}$ are complex weights for each basis function of degree and order l, m, and P the max degree or number of truncation terms of the set of spherical harmonics. The number of basis functions for degree l totals $2l+1$ and number of basis functions upto max truncation order P is $(P+1)^2$. Vectorizing across basis function order followed by the degree gives the column vector of coefficients $\mathbf{C} \in \mathbb{C}^{(P+1)^2\times 1}$ and SH evaluations $\mathbf{Y}(\theta,\phi) \in \mathbb{C}^{(P+1)^2\times 1}$. Moreover, system responses can be expanded along SH and efficiently convolved in the frequency domain. Let system responses $F(z,\theta,\phi)$ and $G(z,\theta,\phi)$ at frequency z be decomposed along SHs given by Eq. 3. By orthogonality of Eq. 2, frequency-domain spherical convolution [19] is the inner product given by

$$\begin{aligned} H(z) &= \int_\Omega F(z,\theta,\phi)\, G^*(z,\theta,\phi) d\Omega \\ &= \sum_{l=0}^{P}\sum_{m=-l}^{l} C_l^m(z,F)\, C_l^{m*}(z,G) \\ &= \mathbf{C}^H(z,G)\mathbf{C}(z,F), \end{aligned} \quad (4)$$

where $\mathbf{C}(z, F)$ and $\mathbf{C}(z, G)$ vectorizes the weights $C_l^m(z, F)$ and $C_l^m(z, G)$ belonging to degree l and order m SH basis functions of system responses F and G respectively.

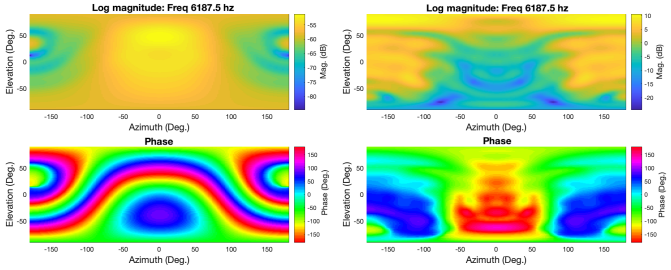

Fig. 1. Free-field responses of COMSOL [5] simulated speaker source (**left**) and mic receiver on a device (**right**) are fitted to 12^{th} max-order SHs.

Source-Room-Receiver Decomposition: The cross-modal formulation [20] can be simplified via the following arrangement. For a speaker-source and device-receiver, let their SH fitted free-field responses at 1 meter be given by

$$\begin{aligned} S(z,\,\theta,\phi) &= \mathbf{C}^T(z, S)\mathbf{Y}(\theta,\phi), \\ R(z,\,\theta,\phi) &= \mathbf{C}^T(z, R)\mathbf{Y}(\theta,\phi). \end{aligned} \tag{5}$$

For a source with SH free-field response $Y_l^m(\theta,\phi)$, let its SH expansion from source location to a receiver location with room effect be given by

$$D_l^m(z,\theta,\phi) = \mathbf{C}^T(z, D_l^m)\mathbf{Y}(\theta,\phi), \tag{6}$$

where $\mathbf{C}(z, D_l^m) \in \mathbb{C}^{(P+1)^2\times 1}$ are room-model dependent weights along SH bases. The speaker-source response in Eq. 5 can be expanded to Eq. 6 by substituting $D_l^m(z,\theta,\phi)$ for $Y_l^m(\theta,\phi)$ given by

$$\begin{aligned} \bar{S}(z,\,\theta,\phi) &= \sum_{l=0}^{P}\sum_{m=-l}^{l} C_l^m(z,\,S) D_l^m(z,\theta,\phi), \\ &= \mathbf{C}^T(z, S)\mathbf{C}(z, D)\mathbf{Y}(\theta,\phi), \end{aligned} \tag{7}$$

where $\mathbf{C}(z, D) \in \mathbb{C}^{(P+1)^2\times(P+1)^2}$ is the row matrix of $\mathbf{C}^T(z, D_l^m)$ weights. Applying SH convolution in Eq. 4 to the expanded speaker-source response in Eq. 7 and the device-receiver response in Eq. 5 results in the overall frequency response given by

$$\begin{aligned} G(z) &= \int_\Omega \bar{S}(z,\theta,\phi)\, R^*(z,\theta,\phi) d\Omega \\ &= \mathbf{C}^T(z,\,S)\,\mathbf{C}(z,\,D)\,\mathbf{C}^*(z,\,R), \end{aligned} \tag{8}$$

where the speaker-source, room, and device-receiver SH expansions are separated.

III. Frequency Dependent Image Source Model

Image-source modeling follows from a subset of ray-tracing models restricted to specular type reflection over planar surfaces. The law of reflections equates the angle of incidence of a wave to a surface to the angle of reflection away from the surface. Given a point source v and receiver r, the angle of incidence w.r.t. a surface can be found constructing rays to source and receivers imaged or reflected across the boundary. The intersection of rays coincident to an image-source and image-receiver shown in Fig. 2 form the point of reflection and the wave propagation path. Placing a virtual source at v_* in free-field simulates the system response of source v reflecting off the boundary before reaching receiver r.

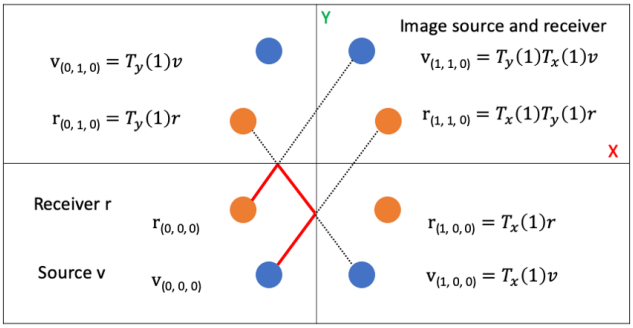

Fig. 2. Higher-order image-source v_{110} applies affine matrix transforms to source v. Image-receiver r_{110} reverses the matrix transforms to receiver r.

A. Extended Geometry Model

An image-source's coordinates for a reflection about a wall can be specified as a series of affine transformations given by translation, reflection, and inverse-translation operations. For successive reflections about either parallel or orthogonal walls belonging to shoe-box like rooms, the set of transformations can be indexed by the first reflection about the $(\pm x, \pm y, \pm z)$ axis and the subsequent number of reflections along each axis. This follows from the observation that reflections about different axes commute but reflections about the same axis do not, allowing for image-sources to be arrayed into cells indexed along a Cartesian grid as shown in Fig. 2. For source vertex $v \in R^{4\times 1}$ with components $(x,\, y,\, z,\, 1)$ within a rectangle room of dimensions $(l_x,\, l_y,\, l_z)$ with origin $(0,\, 0,\, 0)$ at the room's center, its image-source vertex v_{ijk} at cell index $(i,\, j,\, k)$ has matrix recurrence relations given by

$$\begin{aligned} T_x(i) &= \begin{bmatrix} -1 & 0 & 0 & \text{sgn}(i)l_x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} T_x(-i + \text{sgn}(i)), \\ T_y(j) &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & \text{sgn}(j)l_y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} T_y(-j + \text{sgn}(j)), \\ T_z(k) &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & \text{sgn}(k)l_z \\ 0 & 0 & 0 & 1 \end{bmatrix} T_z(-k + \text{sgn}(k)), \\ v_{ijk} &= T_x(i)T_y(j)T_z(k)v, \end{aligned} \tag{9}$$

where for $\circ \in \{x,\, y,\, z\}$, the base case is the identity transform $T_\circ(0) = I$ and the transforms are involutory as $T_\circ^{-1}(\pm 1) = T_\circ(\pm 1)$. Successive reflections about orthogonal walls along

x, y, z axes commute whereas successive reflections about parallel walls intersecting the $\pm$ axes do not commute via inductive proof following the base cases:

$$\begin{aligned} T_\circ(\pm 1)T_{\circ'}(\pm 1) &= T_{\circ'}(\pm 1)T_\circ(\pm 1), \quad \circ \neq \circ', \\ T_\circ(1)T_{\circ'}(-1) &\neq T_{\circ'}(-1)T_\circ(1), \quad \circ = \circ'. \end{aligned} \tag{10}$$

Thus, Eq. 10 maps transformations of v to v_{ijk} in Eq. 9 by their multiplicity and first reflection along the x, y, z axes.

The formulation is extended to receiver vertex $r \in R^{4\times 1}$ and corresponding image-receiver vertex r_{ijk} that belongs to the ray-traced path of image-source vertex v_{ijk}. We observe that the affine transforms in Eq. 9 are applied in reverse order via backtracking the ray-traced path. For odd number of reflections between parallel walls, the transformations are palindrome and thus identical. For even number of reflections between parallel walls, the first reflection occurs about the opposing wall. Thus an image-receiver vertex is specified by

$$r_{ijk} = T_x(\otimes(i))T_y(\otimes(j))T_z(\otimes(k))r, \tag{11}$$

where $\otimes(n) = -n$ for even n and $\otimes(n) = n$ for odd n.

B. Extended Transfer Function Model

The image-source transfer function $G_{ijk}(z)$ at cell (i, j, k) is extended to model separable air $A_{ijk}(z)$, wall material $M_{ijk}(z)$, source $S_{ijk}(z)$, receiver $R_{ijk}(z)$ responses given by

$$\begin{aligned} G_{ijk}(z) &= A_{ijk}(z)M_{ijk}(z)S_{ijk}(z)R_{ijk}(z), \\ G(z) &= \sum_i \sum_j \sum_k G_{ijk}(z), \end{aligned} \tag{12}$$

where $G(z)$ is the overall transfer function summed over image-source indices $|i|, |j|, |k| \leq N$.

Air: The image-source's propagation through air can be modeled by a distance component equivalent to the ray-traced path length, and an attenuation component with variable high-frequency roll-off. A first-order approximation can be modeled using a simple inverse-square law gain component and an integer sample-delay $s_{ijk} = \left\lfloor \frac{D(v_{ijk},r)}{c} F_s \right\rfloor$ at a fixed sample-rate F_s given by

$$A_{ijk}(z) = \frac{B_{ijk} z^{-s_{ijk}}}{D(v_{ijk}, r)}, \quad D(v_{ijk}, r) = ||v_{ijk} - r||_2, \tag{13}$$

where $D(v_{ijk}, r)$ is the distance of image-source to receiver in meters, c the speed of sound (meters/sec), and $B_{ijk} = \pm 1$ is a stochastic phase-inversion term. Randomizing the sign of the phase per image-source zeros the DC offset, removes chirp-like artifacts in empty rooms, and whitens the spectra prior to room material modeling. The fractional component of the delay can then be modeled by convolving with a filter sampled from a windowed sinc kernel [9].

Wall Material: The transfer functions for wall reflections along x, y, z dimensions are given by $M_x(z)$, $M_y(z)$, $M_z(z)$ respectively. Back-tracing the image source's ray from receiver to source verifies that the total number of reflections w.r.t. walls aligned with x, y, z axes is equivalent to the image-source index (i, j, k). By the commutative property of transfer functions in series, we model the overall wall reflection response as exponentiated transfer functions given by

$$M_{ijk}(z) = M_x^{|i|}(z)M_y^{|j|}(z)M_z^{|k|}(z), \tag{14}$$

where the material transfer functions $M_x(z)$, $M_y(z)$, $M_z(z)$ have magnitude response that can be fitted to wall-absorption coefficients. It is useful to constrain the filters to be in minimum phase and have frequency response less than or equal to unity everywhere so that the IR is compact in time and the frequency response is bounded above after exponentiation. To achieve a desired $T_{60}(z)$ decay time (sec) given any room dimensions, one approach is to realize $M_x(z)$, $M_y(z)$, $M_z(z)$ as common minimum-phase FIR filters from the real cepstrum [17] upto gain adjustment. The cepstral method requires log-magnitude targets over uniform sampled frequency z along each room dimension, which can be specified by

$$20 \log_{10} \left| M_{\{x,y,z\}}(z) \right| = \frac{-60\, l_{\{x,y,z\}}}{T_{60}(z) c \sqrt{\xi}}, \tag{15}$$

where $\xi = \mathrm{E}\left[(B + B')^2\right] = 2$ is the expected power of the sum of two stochastic phase-inversion processes.

Diffuse-like reverberation tails can also be simulated by avoiding flutter effects in the wall material design. Given any room size and set of material transfer functions, it is desirable to compensate for the anisotropy of the distribution of image-source locations by attenuating strong reflections along the major axes. This can be formulated in terms of solving for the material transfer function's gain correction w.r.t. the material transfer function's cross-power ratio and inverse cross-room-dimension ratios. The system of equations is given by

$$\frac{p_x g_x}{p_y g_y} = \frac{l_y}{l_x}, \quad \frac{p_x g_x}{p_z g_z} = \frac{l_z}{l_x}, \quad (g_x\, g_y\, g_z)^{\frac{1}{3}} = g_c, \tag{16}$$

where $\{p_x, p_y, p_z\} = \int \{M_x(\omega), M_y(\omega), M_z(\omega)\} d\omega$ are the transfer function powers integrated over frequency $\omega \in [-\pi, \pi]$, $\boldsymbol{\lambda}^T = (g_x \leq 1, g_y \leq 1, g_z \leq 1)$ the unknown but constrained gain correction terms to be solved for, and g_c a free-parameter constraining the geometric mean of the gain correction terms. Log-transformation of Eq. 16 reduces to a linear system of equations where $\log \boldsymbol{\lambda}$ is both non-positive and preserves total power when $g_c = 1$. Satisfying both constraints on $\boldsymbol{\lambda}$ may be infeasible but a non-negative least squares (NNLS) solution to the matrix system given by

$$\begin{aligned} \mathbf{A}\mathbf{x} = \mathbf{b}, \text{ s.t. } \mathbf{x} \geq 0, \quad \mathbf{A} &= \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix}, \\ \mathbf{b} &= -\begin{bmatrix} \log l_y - \log l_x - \log p_x + \log p_y \\ \log l_z - \log l_x - \log p_x + \log p_z \\ 3 \log g_c \end{bmatrix}, \end{aligned} \tag{17}$$

is feasible s.t. the desired gain correction terms are $\boldsymbol{\lambda} = e^{-\mathbf{x}}$.

Source Receiver: The exact acoustic path between an image-source and receiver can be realized by back-tracing or reversing the affine transformations in Eq. 9. The process follows from testing the intersection between the image-source to receiver ray and the first wall within the room

boundaries, reflecting the image-source and receiver across the wall to obtain a new pair of image-source and image-receiver, and repeating until the image-source vertex is coincident to that of source v in cell $(0, 0, 0)$. Applying the same set of transformations to receiver vertex r gives the image-receiver r_{ijk} in Eq. 11 as shown in Fig. 2. If the free-field responses of the source and receiver along unit vector direction $p, q \in R^{3\times1}$ are given by $S(z, p)$, $R(z, q) \in \mathbb{C}$ respectively, then the transfer functions between image-receiver to source and image-source to receiver are given by

$$\begin{aligned} S_{ijk}(z) &= S\left(z,\, p_{ijk}\right), \quad R_{ijk}(z) = R\left(z,\, q_{ijk}\right), \\ p_{ijk} &= \frac{r_{ijk} - v}{||r_{ijk} - v||_2}, \quad q_{ijk} = \frac{v_{ijk} - r}{||v_{ijk} - r||_2}. \end{aligned} \tag{18}$$

For reference, it is possible to sample the free-field responses of either source or receiver via a lookup table computed from anechoic measurements or from simulation. A separable decomposition is possible by fitting the free-field source and receiver responses along SHs and grouping the cross-modal SH terms into the room component.

Separable SH Extension: We give an alternative derivation of the cross-modal ISM formulation [20] via two observations: First, the free-field speaker-source and device-receiver responses are separated into direction and weight components following the SH decomposition in Eq. 3 and 5 given by

$$\begin{aligned} S_{ijk}(z) &= \mathbf{C}^T(z,\, S)\mathbf{Y}(p_{ijk}),\ \mathbf{Y}(p_{ijk}) \in \mathbb{C}^{(P_S+1)^2\times1}, \\ R_{ijk}(z) &= \mathbf{C}^T(z,\, R)\mathbf{Y}(q_{ijk}),\ \mathbf{Y}(q_{ijk}) \in \mathbb{C}^{(P_R+1)^2\times1}, \end{aligned} \tag{19}$$

where $\mathbf{Y}(p_{ijk})$, $\mathbf{Y}(q_{ijk})$ are the SH basis functions evaluated at directions p_{ijk}, q_{ijk} respectively, and $\mathbf{C}(z, S)$, $\mathbf{C}(z, R)$ are the SH weights of the free-field source and receiver responses. Second, substituting the source and receiver terms $S_{ijk}(z)$, $R_{ijk}(z)$ into the overall transfer function in Eq. 12 yields the separable form in Eq. 8 given by

$$\begin{aligned} G(z) &= \mathbf{C}^T(z,\, S)\, \mathbf{C}(z,\, D)\, \mathbf{C}(z,\, R), \\ \mathbf{C}(z,\, D) &= \sum_i \sum_j \sum_k \mathbf{Q}_{ijk}(z), \\ \mathbf{Q}_{ijk}(z) &= A_{ijk}(z) M_{ijk}(z) \mathbf{Y}(p_{ijk}) \mathbf{Y}^T(q_{ijk}) \end{aligned} \tag{20}$$

where $\mathbf{Q}_{ijk}(z) \in C^{(P_S+1)^2\times(P_R+1)^2}$ are the contributions of each image-source of the room to the summation of the cross-modal SH weights $\mathbf{C}(z, D)$. Last, the source and receiver responses are oriented along SO(3) via SH rotations [11].

Time-Frequency Partitioning: The room component matrix $\mathbf{C}(z, D)$ can be reconstructed in the time-domain via the inverse Fast Fourier Transform (IFFT) at uniform sampled frequency $z = e^{j\omega}$ over each image-source's contribution in Eq. 20 given by $\mathbf{c}(t, D) = \sum_{ijk} \mathbf{q}_{ijk}(t)$ where $\mathbf{q}_{ijk}(t) = \text{IFFT}\left(\mathbf{Q}_{ijk}(\omega)\right)$. For uniform block-size L and $N \geq L$ point FFT, the n^{th} time-frequency partition is given by $\mathbf{C}_n(\omega, D) = \text{FFT}\left(\mathbf{c}(nL + [1:L], D)\right)$. Last, the impulse response is reconstructed via overlap-add method:

$$\begin{aligned} G_n(\omega) &= \mathbf{C}^T(\omega,\, S)\, \mathbf{C}_n(\omega,\, D)\, \mathbf{C}(\omega,\, R), \\ \mathbf{g}_n(t) &= \text{IFFT}\left(G_n(\omega)\right), \quad g(t) = \sum_n \mathbf{g}_n(t - nL). \end{aligned} \tag{21}$$

IV. Experiments

Truncation Order Study: Limiting the max number of basis functions P in the SH decomposition of Eq. 3 introduces reconstruction error when fitting to speaker and device's free-field responses in Eq. 5. The overall reconstruction error after regularized least-squares fitting [10] increases for higher frequencies and decreases for large max order P in wide-band as shown in Fig. 3 due to presence of spatial details or rough features in high frequency.

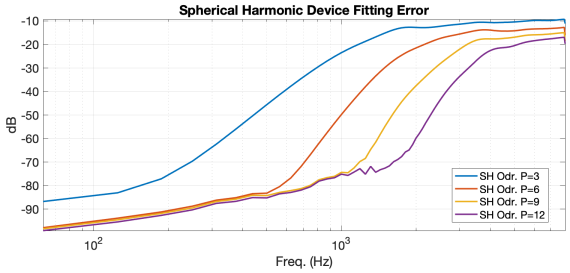

Fig. 3. Device free-field reconstruction error increases in frequency due to prevalence of rough features. SH fitting by truncated singular value decomposition reduces reconstruction error for larger max truncation P.

SH Partitioned FDISM Study: For IRs realized via the separable SH decomposition in Eq. 20 and the reference transfer function in Eq. 12, the reconstruction error depends on both truncation order P and the set of acoustic paths between source and receiver in a room. The IR shown in Fig. 4 was simulated in a cuboid room of size 9 meters at source location $(9.1,\, 0.9,\, 5.4)$, receiver location $(8.1,\, 1.9,\, 5.2)$, equal wall-reflection FIR coefficients $[0.6954,\, 0.2487]$, and upto 8^{th} order reflections. The SH time-frequency partitioning in Eq. 21 occurs between early and late portions of the IR and expanded to $P_S = P_R = 12$ max-truncation order SHs. Reconstruction error is within 2 dB below the Nyquist rate. The early partition is dominated by the direct acoustic path whereas the late partition has more uniformly distributed reflections over the spherical coordinates.

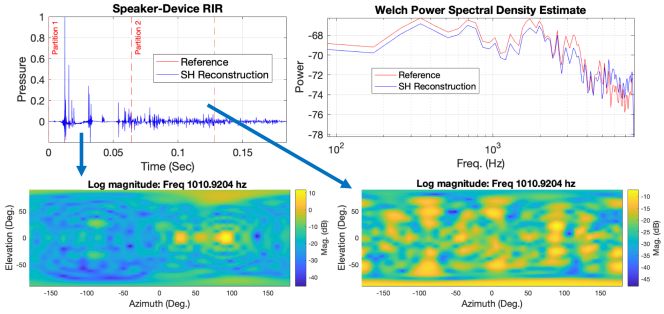

Fig. 4. **Top:** Source-to-receiver reference IR (sample rate 16 kHz, partition size 1024 samples, free-field resolution 62.5 Hz) and SH reconstruction. **Bottom:** dB response at 1 kHz of a mono-pole source expanded upto 12^{th} order SHs at the device location in a room for early and late partitions.

Room Fitting Validation: A sample IR is measured in a room of dimensions $(6.35,\, 4.01,\, 2.54)$ using a pink-noise sequence from a loudspeaker at source position

$(5.365, 1.2, 1.25)$ and recorded at a microphone on a cylindrical solid body at receiver position $(5.365, 3.655, 1.32)$ in meters. To specify room reflection filters in Eq. 14, we estimate the frequency-dependent RT60 of the measured IR for Eq. 15 by fitting an exponential model to the dB spectrogram: $g_{dB}(t, \omega) = \log\left(\alpha e^{-\beta t} + \gamma\right)$ where t is time (sec), $g_{dB}(t, \omega)$ the dB power of spectrogram at time-frequency t, ω, and α, β, γ are the unknown onset gain, T60 decay rate, and noise floor respectively. The IR is generated from Eq. 21 and then equalized against the measured IR's power spectral density where the direct portion, early reflections, and overall RT60 profile align as shown in Fig. 5.

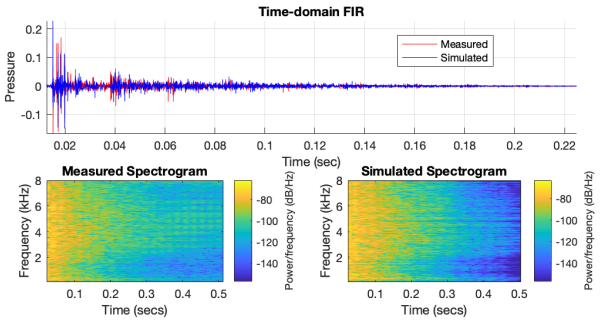

Fig. 5. **Top:** Time-domain responses of measured and FDISM IRs align direct and early reflections. **Left:** Measured IR spectrogram with high noise floor. **Right:** FDISM IR preserves spectrogram's RT60.

Room Flutter Correction: In musical applications, it is desirable that artificial room IRs have a smooth energy decay and increasing echo density. Flutter in the reverberation tail appear as distinct echos over time which can be corrected by decreasing the wall-reflection gain following Eqs. 16, 17, compensating for strong reflections between the longer room axis. For illustration, the sample IR generated from a room of size $(12.35, 4.01, 2.54)$ meters with wall reflection FIR coefficients $[0.9716, 0.0056]$, $[0.9387, 0.0054]$, and $[0.9029, 0.0117]$ along the (x, y, z) dimensions in Fig. 6 exhibits strong flutter in the tail. The NNLS solution $\boldsymbol{\lambda}^T = (0.3776, 1, 1)$ reduces the gain of the transfer function along the longest dimension.

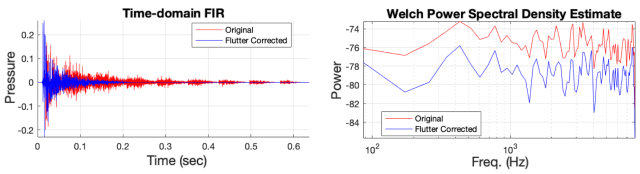

Fig. 6. Flutter correction adjusts overall gain for each wall material filter and preserves overall spectra envelope shape of the IR.

V. Conclusions

We proposed a separable format for first-order-approximation of joint speaker-room-device modeling using SH basis functions. We then extended the image method to model frequency-dependent characteristics before generalizing into the SH domain. Last, we validated the model using both simulated data and measurements. Future work includes SH based beamforming design, data synthesis for acoustic echo cancellation, barge-in validation, and augmentation for sound-source localization and ASR training.

Acknowledgment

We would like to thank Sanjay Yengul, Mrudula Athi, Guandong Pan, and Jim Sun for their discussions and assistance in data collection and experimental validation.

References

[1] Ahrens, Jens, Mark RP Thomas, and Ivan Tashev. "HRTF magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data." Proceedings of IEEE APSIPA, 2012.

[2] Allen, Jont B., and David A. Berkley. "Image method for efficiently simulating small-room acoustics." The Journal of the Acoustical Society of America 65.4 (1979): 943-950.

[3] Chhetri, Amit, et al. "On Acoustic Modeling for Broadband Beamforming." 27th European Signal Processing Conference (EUSIPCO). IEEE, 2019.

[4] Chhetri, Amit, et al. "Multichannel Audio Front-End for Far-Field Automatic Speech Recognition." 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018.

[5] COMSOL Multiphysics, "Acoustic module–user guide," 2017.

[6] Duraiswami, Ramani, et al. "Plane-wave decomposition analysis for spherical microphone arrays." IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.

[7] Evans, Michael J., James AS Angus, and Anthony I. Tew. "Analyzing head-related transfer function measurements using surface spherical harmonics." The Journal of the Acoustical Society of America 104.4 (1998): 2400-2411.

[8] Funkhouser, Thomas, et al. "A beam tracing method for interactive architectural acoustics." The Journal of the acoustical society of America 115.2 (2004): 739-756.

[9] Habets, Emanuel AP. "Room impulse response generator." Technische Universiteit Eindhoven, Tech. Rep 2.2.4 (2006): 1.

[10] Hansen, Per Christian. "The truncated SVD as a method for regularization." BIT Numerical Mathematics 27.4 (1987): 534-553.

[11] Ivanic, Joseph, and Klaus Ruedenberg. "Rotation matrices for real spherical harmonics. Direct determination by recursion." The Journal of Physical Chemistry 100.15 (1996): 6342-6347.

[12] Kim, Chanwoo, et al. "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home." (2017).

[13] Ko, Tom, et al. "A study on data augmentation of reverberant speech for robust speech recognition." IEEE ICASSP, 2017.

[14] Lehmann, Eric A., and Anders M. Johansson. "Diffuse reverberation model for efficient image-source simulation of room impulse responses." IEEE TASLP 18.6 (2009): 1429-1439.

[15] Li, Zhiyun, and Ramani Duraiswami. "Flexible and optimal design of spherical microphone arrays for beamforming." IEEE Transactions on Audio, Speech, and Language Processing 15.2 (2007): 702-714.

[16] Müller, Claus. Spherical harmonics. Vol. 17. Springer, 2006.

[17] Oppenheim, Alan V., and Ronald W. Schafer. Digital Signal Processing, Englewood Cliffs, NJ, Prentice-Hall, 1975

[18] Park, Munhum, and Boaz Rafaely. "Sound-field analysis by plane-wave decomposition using spherical microphone array." The Journal of the Acoustical Society of America 118.5 (2005): 3094-3103.

[19] Rafaely, Boaz. "Plane-wave decomposition of the sound field on a sphere by spherical convolution." The Journal of the Acoustical Society of America 116.4 (2004): 2149-2157.

[20] Samarasinghe, Prasanga N., et al. "Spherical harmonics based generalized image source method for simulating room acoustics." The Journal of the Acoustical Society of America 144.3 (2018): 1381-1391.

[21] Wabnitz, Andrew, et al. "Room acoustics simulation for multichannel microphone arrays." Proceedings of the International Symposium on Room Acoustics. 2010.

[22] Zotkin, Dmitry, Ramani Duraiswami, and Nail A. Gumerov. "Regularized HRTF fitting using spherical harmonics." IEEE WASPAA, 2009.