Regularized Spherical Fourier Transform for Room Impulse Response Interpolation

1st Julio Alarcón
Facultad de Ing. Electrónica y Eléctrica
Universidad Nacional Mayor de San Marcos
Lima, Perú
alarcon.ganoza@gmail.com

2nd Javier Solis-Lastra
Facultad de Ing. Electrónica y Eléctrica
Universidad Nacional Mayor de San Marcos
Lima, Perú
jsolisl@unmsm.edu.pe

3rd César D. Salvador
Perception Research
Lima, Perú
salvador@perception3d.com

***Abstract*—Room impulse responses (RIRs) are a key tool in architectural acoustics and spatial sound for virtual reality. They characterize the room response between a sound source and a receiver. Directional RIRs at the receiver position can be measured with spherical microphone arrays or calculated with 3D models and numerical methods. Either because of a low number of microphones or limited computational capacity, there is a need to interpolate the RIRs to obtain higher directional resolutions. The spherical Fourier transform (SFT) enables a promising physics-based interpolation approach. However, existing SFT algorithms for acoustic purposes highly depend on the spherical distribution of microphones. This paper presents a SFT algorithm based on Tikhonov regularization that is suitable for random spherical distributions. Results show that the regularized SFT maintains the interpolation error bounded at high-energy values in time and up to a maximum frequency determined by the number of microphones. An open-source implementation is made publicly available to foster the reproducibility of this research.**

***Index Terms*—Architectural acoustics, spatial sound technology, room impulse response, spherical Fourier transform, Tikhonov regularization.**

I. Introduction

Room impulse responses (RIRs) are a key tool in architectural acoustics [1] and spatial sound for virtual reality [2]. Traditionally, RIRs characterize the room response to sound transmission from an omnidirectional sound source to an omnidirectional receiver. Nowadays, there is a growing interest in directional receivers to discriminate among the different directions in which sounds may reach a listener. In this regard, directional RIRs can be measured with spherical microphone arrays [3] or calculated with 3D models and numerical methods [4], [5]. Either because of the low number of microphones or limited computational capacity, there is a need to interpolate the RIRs to obtain higher directional resolutions.

Geometric approaches such as baricentric interpolation have been used for acoustic data [6], [7]. Physics-based approaches are a more convenient choice because they consider the spatial nature of sound propagation. Physics-based interpolation has been used for RIRs captured with microphones distributed in rectangular arrays [8] and microphones uniformly distributed on rigid spherical baffles [9], [10]. Rigid spherical arrays, in particular, can leverage the angular solutions to the acoustic wave equation, in the form of the spherical Fourier transform (SFT) and the inverse spherical Fourier transform (ISFT), for interpolation purposes [11], [12]. However, existing SFT algorithms highly depend on the spherical distribution of microphones [13], [14].

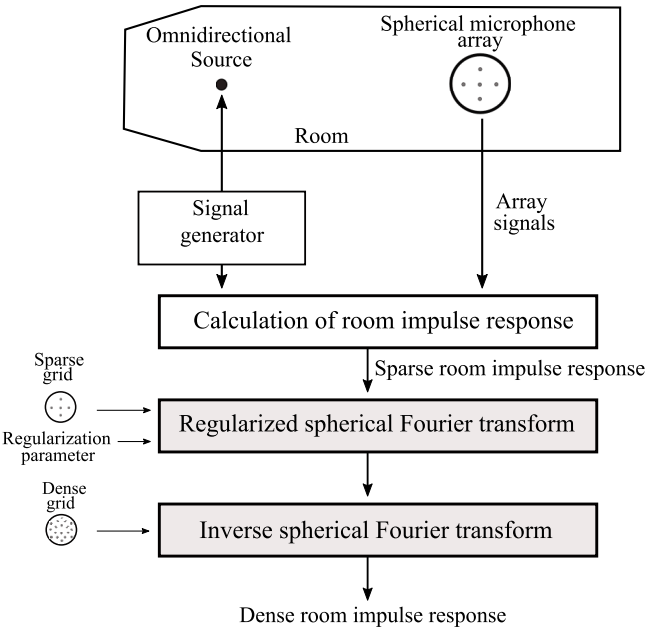


Fig. 1. Interpolation of room impulse responses.

This paper presents an open-source SFT algorithm that is suitable for interpolating RIRs captured with microphones randomly distributed on a spherical baffle. The SFT algorithm is implemented with the classic regularization method of Tikhonov that minimizes errors in the ℓ^2-norm [15], [16]. Figure 1 shows an overview of RIR interpolation. First, the SFT analyses initial RIRs captured with a spherical array that has a sparse number of microphones. Then, the ISFT synthesizes the RIRs for a dense number of directions. An open-source implementation[1] of this algorithm is made publicly available to foster the reproducibility of this research.

The reminder of this paper is organized as follows: Sec. II

978-1-6654-1221-6/21/$31.00 ©2021 IEEE

[1]https://github.com/AlarconGanoza/sphericalAcoustic

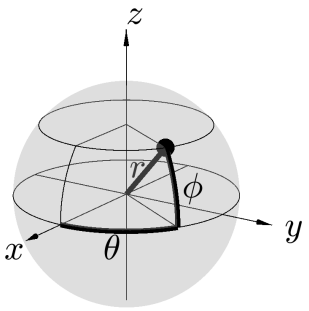


Fig. 2. Spherical coordinates. A point $\vec{r} = (r, \theta, \phi)$ is specified by its radial distance r, azimuth angle $\theta \in [-\pi, \pi]$, and elevation angle $\phi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

formulates the regularized SFT, Sec. III discusses interpolation errors, and Sec. IV states the conclusions.

II. Formulation of the regularized SFT

Spherical RIRs can be represented by a linear combination of spherical harmonic functions [11]. This representation defines the SFT and the ISFT [12], [14]. Figure 2 shows the spherical coordinate system used throughout this paper to describe these spherical transforms.

A. Continuous SFT

Let $\psi(\vec{r}, t)$ denote a finite-energy RIR describing the transmission of sound from a sound source position to a point $\vec{r} = (r, \theta, \phi)$ on the surface of a continuous recording sphere of radius r. The time domain is denoted by t.

The SFT of the spherical distribution ψ is defined by

$$\psi_{nm}(r,t) = \int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} \psi(r,\theta,\phi,t) Y_{nm}(\theta,\phi) \cos\phi d\phi d\theta. \quad (1)$$

Here, ψ_{nm} are SFT coefficients whereas Y_{nm} are real-valued spherical harmonic functions of order n and degree m:

$$Y_{nm}(\theta,\phi) = N_{nm} P_n^{|m|}(\sin\phi) \times \begin{cases} 1, & m = 0, \\ \sqrt{2}\cos(m\theta), & m > 0, \\ \sqrt{2}\sin(|m|\theta), & m < 0, \end{cases} \quad (2)$$

where P_n^m is the non-normalized associated Legendre function and the normalization factor is defined as

$$N_{nm} = (-1)^{|m|} \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}}. \quad (3)$$

The ISFT recovers ψ from ψ_{nm} in (1) as follows:

$$\psi(r,\theta,\phi,t) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \psi_{nm}(r,t) Y_{nm}(\theta,\phi). \quad (4)$$

B. Spherical discretization

In practice, (1) is approximated for a discrete distribution of L points on the recording sphere, denoted by $\vec{r}_\ell = (r, \theta_\ell, \phi_\ell)$, where $\ell = 1, ..., L$. These points are the positions of the L microphones on the spherical array. Let $\psi(\vec{r}_\ell, t)$ denote the RIR from the source in the room to the ℓ-th microphone. Since the number of microphones is finite, the sum in (4) can only be computed up to a maximum number of terms N determined by the following sampling theorem on the sphere:

$$(N+1)^2 \leq L. \quad (5)$$

Hence, the sum in (4) remains truncated as

$$\psi(\vec{r}_\ell, t) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \psi_{nm}(r,t) Y_{nm}(\theta_\ell, \phi_\ell) + \varepsilon_N(\vec{r}_\ell, t), \quad (6)$$

where ψ_{nm} are the SFT coefficients to be calculated and ε_N contains all residual terms for $n > N$.

In matrix notation, and for each instant of time, (6) becomes

$$\boldsymbol{\Psi} = \mathbf{Y} \boldsymbol{\Psi}_{\mathbf{nm}} + \boldsymbol{\varepsilon}, \quad (7)$$

where

$$\boldsymbol{\Psi} = [\psi(\vec{r}_1, t), \ldots, \psi(\vec{r}_L, t)]^T, \quad (8)$$

$$\boldsymbol{\Psi}_{\mathbf{nm}} = [\psi_{00}, \psi_{1-1}, \ldots, \psi_{NN}]^T, \quad (9)$$

$$\boldsymbol{\varepsilon} = [\varepsilon_N(\vec{r}_1, t), \ldots, \varepsilon_N(\vec{r}_L, t)]^T, \quad (10)$$

and

$$\mathbf{Y} = \begin{bmatrix} Y_{00}(\theta_1,\phi_1) & \ldots & Y_{NN}(\theta_1,\phi_1) \\ \vdots & \vdots & \vdots \\ Y_{00}(\theta_L,\phi_L) & \ldots & Y_{NN}(\theta_L,\phi_L) \end{bmatrix}. \quad (11)$$

The matrix $\mathbf{Y}$ of size $L \times (N+1)^2$ contains the spherical harmonic functions and the ^{T} symbol denotes matrix transpose. The approximate solution to (7) defines the discrete SFT

$$\hat{\boldsymbol{\Psi}}_{\mathbf{nm}} = \mathbf{Y}^{+} \boldsymbol{\Psi}, \quad (12)$$

where the $^+$ symbol denotes pseudo-inverse.

C. Regularized SFT

Tikhonov regularization [15], [16] defines $\mathbf{Y}^+$ in (12) as

$$\mathbf{Y}^{+} = (\mathbf{Y}^T \mathbf{Y} + \lambda^2 \mathbf{I})^{-1} \mathbf{Y}, \quad (13)$$

where λ is known as the regularization parameter. An equivalent expression is obtained by means of the singular value decomposition (SVD),

$$\mathbf{Y} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T, \quad (14)$$

its inversion, and the smoothing of $\boldsymbol{\Sigma}$ as follows:

$$\mathbf{Y}^{+} = \mathbf{V} \boldsymbol{\Sigma}_{\text{reg}}^{-\mathbf{1}} \mathbf{U}^T. \quad (15)$$

The matrices $\mathbf{U}$ and $\mathbf{V}$ are unitary and $\boldsymbol{\Sigma}$ is a rectangular, diagonal matrix containing the singular values σ_ℓ, where $\ell = 1, ..., L$. The regularized inverse of $\boldsymbol{\Sigma}$ is defined as

$$\boldsymbol{\Sigma}_{\text{reg}}^{-1} = \text{diag}\left(\frac{|\sigma_\ell|^2}{|\sigma_\ell|^2 + \lambda^2} \times \frac{1}{\sigma_\ell} \right). \quad (16)$$

In summary, the regularized discrete SFT is defined by (12), (15), and (16), whereas the discrete ISFT is defined by

$$\hat{\boldsymbol{\Psi}} = \mathbf{Y} \hat{\boldsymbol{\Psi}}_{\mathbf{nm}}. \quad (17)$$

III. Evaluation of RIR interpolation

Sparse and dense RIR datasets were calculated using the algorithm in [17] to evaluate the proposed interpolation approach described in Fig 1. This algorithm calculates the sound pressure on a rigid spherical baffle placed inside a rectangular parallelepiped room. The calculated sound pressure corresponds to the total reverberant field, which includes the scattering from the rigid sphere and the high-order reflections from the walls. The origin of the room coordinate system was a bottom corner of the room. The dimensions of the room were 4.62 m wide (along x), 3.84 m long (along y), and 3 m high (along z). In room coordinates, the omnidirectional source position was $(1.5,\ 2,\ 1)$ m, whereas the center of the spherical microphone array, that is, the center of the spherical coordinate system shown in Fig. 2, was $(2.5, 2, 1.5)$ m. The reverberation time of the room was 0.2 s and the reflection order, 36. A sampling frequency of 16 kHz was used to calculate 3200 samples in time, corresponding to a duration of 200 ms.

The sparse grids were random distributions on a rigid sphere of radius r; each microphone position is denoted by $\vec{r}_\ell = (r, \theta_\ell, \phi_\ell)$, where $\ell = 1, 2, ..., L$. The sparse RIRs are denoted by $\psi_{\text{sparse}}(\vec{r}_\ell, t)$. A low-pass filter was applied to ψ_{sparse} to account for spatial aliasing, setting the maximum frequency of reliable interpolation to

$$f_{\text{max}} = \frac{cN_{\text{max}}}{2\pi r}, \tag{18}$$

where the speed of sound in air was set to $c = 343$ m/s. The maximum order used to compute the SFT was determined by the limiting case of (5); in effect,

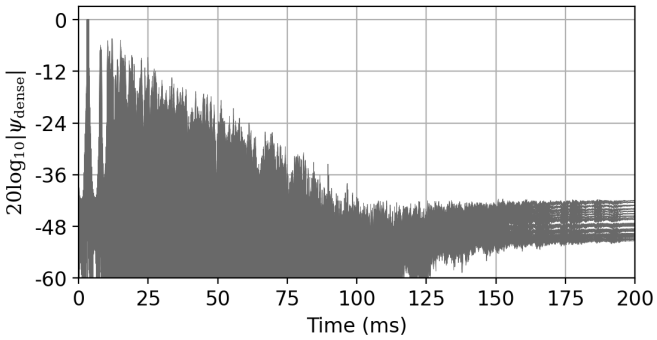
$$N_{\text{max}} = \lfloor \sqrt{L} - 1 \rfloor. \tag{19}$$

The dense grid was a spherical grid based on subdivisions of the icosahedron, denoted by $\vec{p}_i = (r, \theta_i, \phi_i)$, where $i = 1, 2, ..., I$. The dense RIRs used as target data are denoted by $\psi_{\text{dense}}(\vec{p}_i, t)$. This target data was also low-pass filtered according to (18). The dense RIRs, interpolated from ψ_{sparse} with the proposal, are denoted by $\hat{\psi}_{\text{dense}}(\vec{p}_i, t)$. The interpolation error is defined as follows:

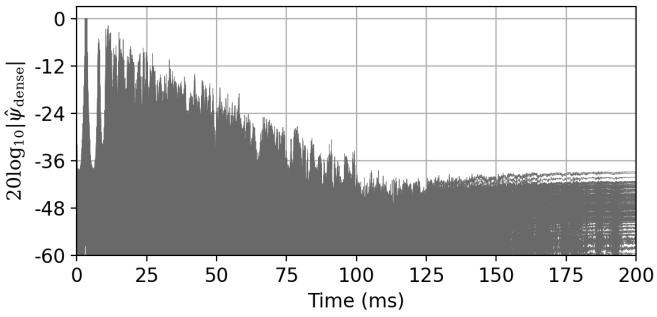
$$E(t) = \frac{\underset{\vec{p}_i}{\text{RMS}} \left(\psi_{\text{dense}}(\vec{p}_i, t) - \hat{\psi}_{\text{dense}}(\vec{p}_i, t) \right)}{\underset{\vec{p}_i}{\text{RMS}} \left(\psi_{\text{dense}}(\vec{p}_i, t) \right)}, \tag{20}$$

where RMS stands for root mean square along $\vec{p}_i$.

Considering a rigid spherical baffle of radius $r = 8$ cm, four sparse grids were evaluated: $L = 49, 36, 25, 16$; correspondingly, $N_{\text{max}} = 6, 5, 3, 2$. In all cases, the dense grid was $I = 162$. Fig. 3 shows the energy of the RIRs in all microphones for target data $\psi_{\text{dense}}(\vec{p}_i, t)$ and interpolated data $\hat{\psi}_{\text{dense}}(\vec{p}_i, t)$. It is observed that, in both cases, the energy remains concentrated within the first 100 ms; after this time, the background noise is only observed below -36 dB. Therefore, the region of interest comprises the first 100 ms. A comparison between Fig. 3(a) and Fig. 3(b) shows that the envelopes are very similar in the region of interest.



(a)



(b)

Fig. 3. Energy of RIRs. a) Target. b) Interpolated.

The four panels in Fig. 4 show the interpolation errors calculated with (20) for different values of L and N_{max}. Results are displayed in logarithmic scale. Blue lines correspond to interpolation with the non-regularized SFT ($\lambda = 0$), whereas red lines, to interpolation with the proposed regularized SFT ($\lambda > 0$). Within the high-energy region comprising the first 100 ms, all panels show that the proposal yielded the lower error bounds around -6 dB; these results did not depend on the sparse resolutions. Contrasting all panels, it can be observed that improvements in accuracy were more noticeable at higher sparse resolutions (e.g., $L = 49$). However, the benefits of regularization were also observed at lower sparse resolutions (e.g., $L = 16$). This was of particular interest because it indicates that the proposal can be used with available spherical arrays that have fewer microphones.

IV. Conclusion

This paper has presented an open-source SFT algorithm based on Tikhonov regularization that is suitable for interpolating RIRs captured with microphones that are randomly distributed on a rigid spherical baffle. As a result, the SFT-based interpolation maintained the errors bounded at high-energy values in time and up to a maximum frequency determined by the number of microphones.

Extensions to this work might include physics-based frameworks for the reconstruction of sound pressure fields [14]. This would enable the removal of acoustic scattering from the rigid spherical baffle during interpolation to yield free-field RIRs.

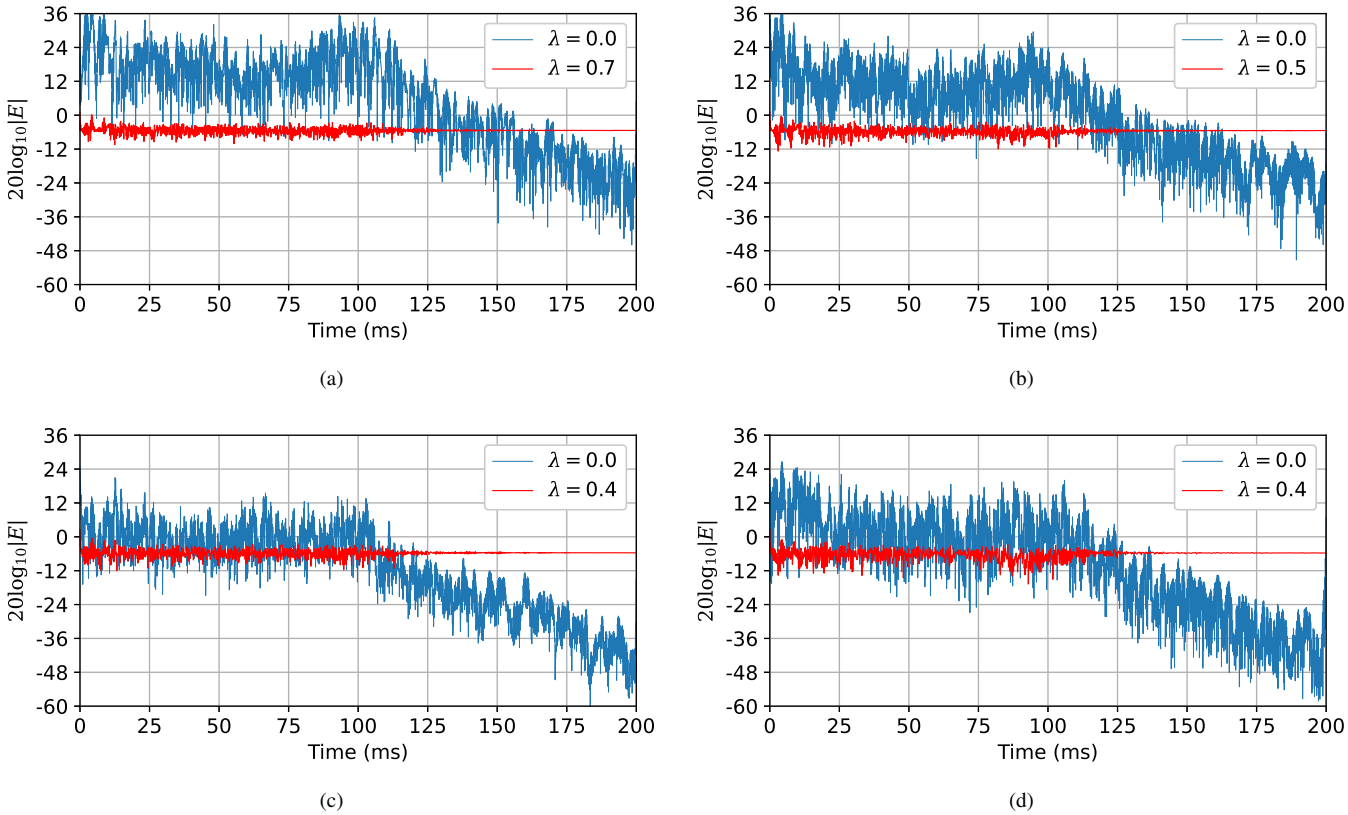


Fig. 4. Interpolation errors. a) $L = 49$, $N_{\max} = 6$. b) $L = 36$, $N_{\max} = 5$. c) $L = 25$, $N_{\max} = 4$. d) $L = 16$, $N_{\max} = 3$.

Alternative regularization techniques that minimizes errors in the ℓ^1-norm instead of the ℓ^2-norm, such as compressive sensing [18], might also be considered when implementing SFT algorithms. Finally, a perceptual evaluation of the interpolated RIRs by means of detectability of differences, and localization tests along angles, could provide more insight into the validity of the suggested approach.

REFERENCES

[1] N. Xiang, *Architectural Acoustics Handbook*. Acoustic Series, J. Ross Publishing, 2017.

[2] M. Vorländer, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer International Publishing, 2nd ed., 2020.

[3] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *J. Acoust. Soc. Am.*, vol. 132, no. 1, pp. 261–270, 2012.

[4] N. Raghuvanshi, R. Narain, and M. C. Lin, "Efficient and accurate sound propagation using adaptive rectangular decomposition," *IEEE Trans. Vis. Comput. Graphics*, vol. 15, pp. 789–801, Sept. 2009.

[5] J. Shi, C. D. Salvador, J. Treviño, S. Sakamoto, and Y. Suzuki, "Spherical harmonic representation of rectangular domain sound fields," *Acoust. Sci. Technol.*, vol. 41, no. 1, pp. 451–453, 2020.

[6] J. Villegas, "Locating virtual sound sources at arbitrary distances in real-time binaural reproduction," *Virtual Reality*, vol. 19, pp. 201–212, Nov. 2015.

[7] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3D tune-in toolkit: An open-source library for real-time binaural spatialisation," *PLOS ONE*, vol. 14, no. 3, pp. 1–37, 2019.

[8] R. Mignot, G. Chardon, and L. Daudet, "Low frequency interpolation of room impulse responses using compressed sensing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 205–216, 2014.

[9] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1929–1941, 2017.

[10] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Enhancement of spatial sound recordings by adding virtual microphones to spherical microphone arrays," *J. Inf. Hiding and Multimedia. Signal Process.*, vol. 8, pp. 1392–1404, Nov. 2017.

[11] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. London, UK: Academic Press, 1999.

[12] D. Healy, Jr., D. Rockmore, P. Kostelec, and S. Moore, "FFTs for the 2-sphere-Improvements and variations," *J. Fourier Anal. Appl.*, vol. 9, pp. 341–385, July 2003.

[13] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech, Audio Process.*, vol. 13, pp. 135–143, Jan. 2005.

[14] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Boundary matching filters for spherical microphone and loudspeaker arrays," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, pp. 461–474, Mar. 2018.

[15] A. Neumaier, "Solving ill-conditioned and singular linear systems: A tutorial on regularization," *SIAM Review*, vol. 40, no. 3, pp. 636–666, 1998.

[16] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Design theory for binaural synthesis: Combining microphone array recordings and head-related transfer function datasets," *Acoust. Sci. Technol.*, vol. 38, pp. 51–62, Mar. 2017.

[17] D. P. Jarret, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, pp. 1462–1472, Sept. 2012.

[18] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, pp. 21–30, Mar. 2008.