ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

Ear centering in the spatial and transform domains for near-field head-related transfer functions

César SALVADOR[(1)], Ayrton URVIOLA[(1)], Shuichi SAKAMOTO[(2)]

[(1)]Perception Research, Peru, salvador@perception3d.com, aurviola@perception3d.com
[(2)]RIEC and GSIS, Tohoku University, Japan, saka@ais.riec.tohoku.ac.jp

ABSTRACT

The head-related transfer function (HRTF) is a major tool in spatial audio. The HRTF for a point source is defined as the ratio between the sound pressure at the ear position and the free-field sound pressure at a reference position. The reference is typically placed at the center of the head. However, when using spherical Fourier transforms (SFT) and distance-varying filters (DVF) to synthesize HRTFs for point sources very close to the head, the synthesized HRTF assumes that the measurement position and the reference position are the same. Ear centering is a technique that overcomes this ambiguity. Ear centering can be performed with translation operators in the spatial domain (the unit sphere) or with DVFs in the transform domain (the SFT domain). This paper presents a review of ear centering and shows that operating in the spatial domain is computationally more efficient than operating in the transform domain. The reason behind this is that transform-domain ear centering requires DVFs that depend on two source distances, from the reference and from the ear, whereas spatial-domain ear centering requires DVFs that depend on the distance from the reference only. Moreover, operating in the spatial domain is more accurate.

Keywords: Head-related transfer functions, ear centering, translation operator, spherical Fourier transform, distance-varying filter.

1 INTRODUCTION

There is a growing interest in accurately synthesizing the head-related transfer function (HRTF) for arbitrary points in the near field, that is, in the region of space within 1 m from the center of the head [1, 2]. Interests include the design of near-field binaural recording systems [3], the development of near-field auditory displays [4], and the consideration of distance in auditory attention experiments [5, 6].

A promising synthesis approach extrapolates near-field HRTFs from far-field ones using the spherical Fourier transform (SFT), distance-varying filters (DVF), and the inverse spherical Fourier transform (ISFT) [7, 8, 9]. When using the SFT to represent spherical HRTF datasets, the default spherical symmetry of the SFT is specified with respect to the reference position. The default spherical symmetry, however, does not allow to distinguish between the reference position (the head center) and the measurement positions (the ears). Such mismatch produces a demand of a high number of basis functions in the SFT representation and, therefore, affects the synthesis accuracy.

Ear centering overcomes the mismatch between the default SFT center and the ear position. Ear centering is performed with acoustic operators that translate the SFT center from the head center to the ears. Table 1 overviews research related to ear centering interpreted in terms of translation operators. In summary, translation operators can be applied in the spatial domain (the unit sphere) [10, 11, 12, 13, 14, 15, 16] or in the transform domain (the SFT domain) [17], can be used in far-field [10, 11, 12, 13, 14] or near-field [15, 16, 17] HRTF synthesis, and can consider acoustic propagation in the free field [13, 14, 15, 16, 17] or include an acoustically rigid scatterer that mimics a simple head [10, 11, 12].

Table 1. Overview of translation operators used for ear centering in HRTF synthesis.

Reference	Domain	Distance	Translation model
Porschmann *et al.*, 2019 [10, 11] and Arend *et al.*, 2021 [12]	Spatial domain	Far field	Plane wave with rigid sphere
Zaunschirm *et al.*, 2018 [13]	Spatial domain	Far field	Plane wave
Ben-Hur *et al.*, 2019 [14]	Spatial domain	Far field	Plane wave
Urviola *et al.*, 2021 [15] and 2022 [16]	Spatial domain	Near field	Spherical wave
Richter *et al.*, 2014 [17]	Transform domain	Near field	Spherical wave

In our previous papers [15, 16], a question remained open on whether translation operators perform better in the spatial domain or in the transform domain when synthesizing near-field HRTFs. In this paper, we address the question by contrasting the performance of the spatial-domain method proposed in [15, 16] and the transform-domain method proposed in [17].

2 EAR CENTERING IN THE SPATIAL AND TRANSFORM DOMAINS

The HRTF is defined as the sound pressure at the left or right ear position $\mathbf{r}_{\text{ear}}$ due to a point source at the source position $\mathbf{r}$, divided by the free-field sound pressure at the reference position $\mathbf{r}_{\text{ref}}$ [18]. The HRTF, denoted by $\mathcal{H}$, is defined as

$$\mathcal{H}(\mathbf{r}, \mathbf{r}_{\text{ear}}, \mathbf{r}_{\text{ref}}) = \frac{\Psi(\mathbf{r}, \mathbf{r}_{\text{ear}})}{\Psi_{FF}(\mathbf{r}, \mathbf{r}_{\text{ref}})}, \tag{1}$$

where $\Psi(\mathbf{r}_{\text{s}}, \mathbf{r}_{\text{r}})$ denotes the pressure emanated from a source position $\mathbf{r}_{\text{s}}$ measured at a receiver position $\mathbf{r}_{\text{r}}$ and the sub-index FF stands for "free-field", which indicates that the head is not present.

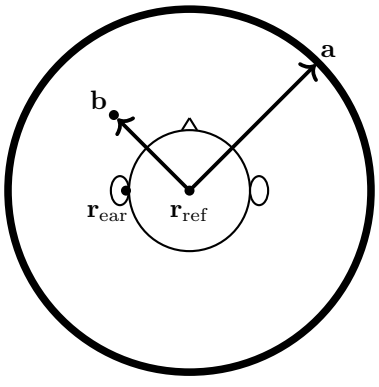


Figure 1. Geometry for near-field HRTF synthesis.

Figure 1 shows the top-view geometry for theoretical near-field HRTF synthesis. The center of the head coincides with the reference position $\mathbf{r}_{\text{ref}}$ and the ear position is denoted by $\mathbf{r}_{\text{ear}}$. The point $\mathbf{a}$ is a point in a continuous, spherical distribution at a far distance $\|\mathbf{a} - \mathbf{r}_{\text{ref}}\|$ from $\mathbf{r}_{\text{ref}}$. The point $\mathbf{b}$ is an

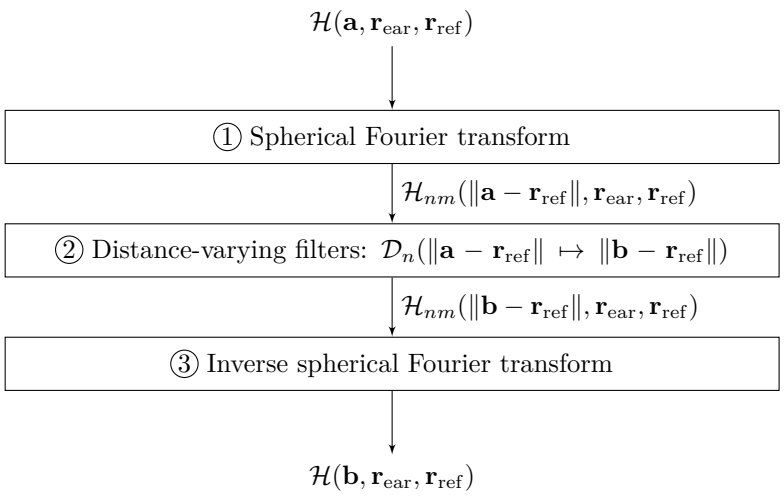


Figure 2. Near-field HRTF synthesis without ear centering [7, 8, 9].

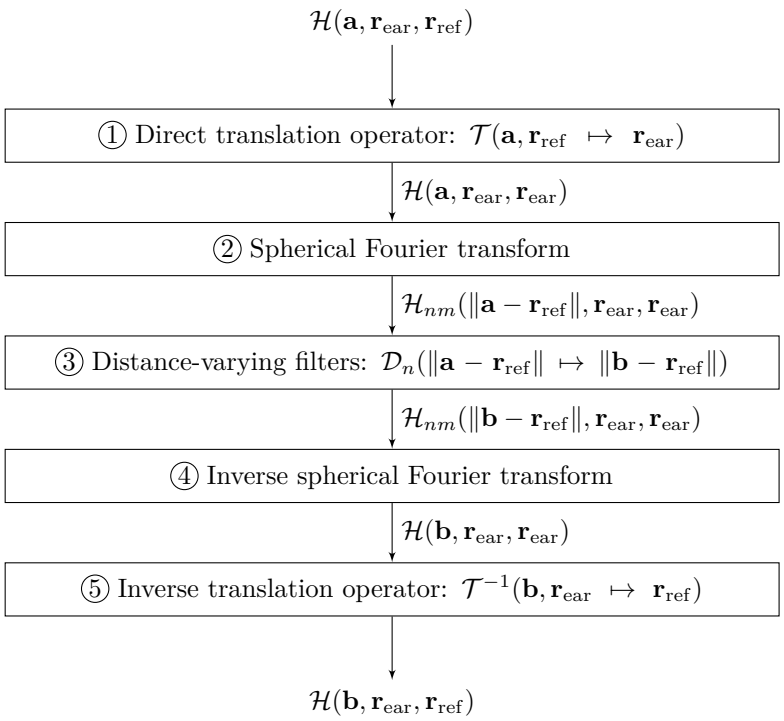


Figure 3. Near-field HRTF synthesis with ear centering in the spatial domain [15, 16].

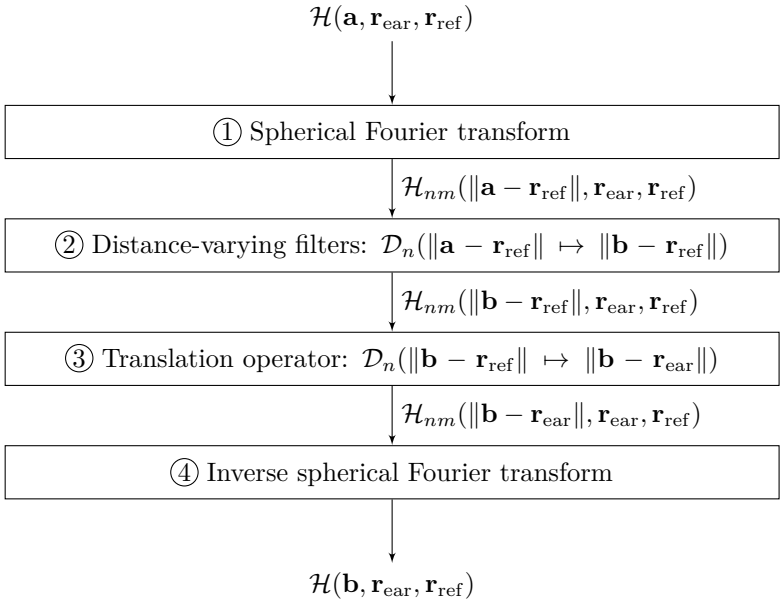


Figure 4. Near-field HRTF synthesis with ear centering in the transform domain [17].

arbitrary point at a near distance $\|\mathbf{b}-\mathbf{r}_{\text{ref}}\|$ from $\mathbf{r}_{\text{ref}}$. The distances to $\mathbf{a}$ and $\mathbf{b}$ from $\mathbf{r}_{\text{ear}}$ are respectively denoted by $\|\mathbf{a}-\mathbf{r}_{\text{ear}}\|$ and $\|\mathbf{b}-\mathbf{r}_{\text{ear}}\|$.

Figure 2 overviews the process for near-field HRTF synthesis without ear centering [7, 8, 9]. The input is a continuous, spherical distribution of HRTFs denoted by $\mathcal{H}(\mathbf{a}, \mathbf{r}_{\text{ear}}, \mathbf{r}_{\text{ref}})$ whereas the output is a synthesized HRTF denoted by $\mathcal{H}(\mathbf{b}, \mathbf{r}_{\text{ear}}, \mathbf{r}_{\text{ref}})$. The encoding stage is performed by the SFT in block ①. Range extrapolation from distance $\|\mathbf{a}-\mathbf{r}_{\text{ref}}\|$ to distance $\|\mathbf{b}-\mathbf{r}_{\text{ref}}\|$ is performed by block ② using DVFs; they are transform-domain acoustic propagators based on a ratio of spherical Hankel functions [7, 8, 9]. The decoding stage is performed by the ISFT in ③.

Figure 3 overviews the process for near-field HRTF synthesis using ear centering in the spatial domain [15, 16]. The input and output of this process are identical to those shown in Fig. 2. The encoding stage comprises the direct translation operator in block ① and the SFT in block ②. Range extrapolation from distance $\|\mathbf{a} - \mathbf{r}_{\text{ref}}\|$ to distance $\|\mathbf{b} - \mathbf{r}_{\text{ref}}\|$ is performed by block ③ using DVFs; they are transform-domain acoustic propagators based on a ratio of spherical Hankel functions. The decoding stage comprises the ISFT in ④ and the inverse translation operator in ⑤. Note that spatial-domain ear centering is composed of blocks ① and ⑤ which explicitly translate the reference position.

Figure 4 overviews the process for near-field HRTF synthesis using ear centering in the transform domain [17]. The input and output of this process are identical to those shown in Fig. 2. Encoding, range extrapolation and decoding are also identical to Fig. 3. Ear centering is performed by the transform-domain translation operator in block ③ which is implemented with an additional DVF now from distance $\|\mathbf{b} - \mathbf{r}_{\text{ref}}\|$ to distance $\|\mathbf{b} - \mathbf{r}_{\text{ear}}\|$. Although the original proposal in [17] considered an additional optimization stage to modify $\mathbf{r}_{\text{ear}}$ according to frequency, block ③ in our implementation operates with the real position $\mathbf{r}_{\text{ear}}$. Note that transform-domain ear centering does not explicitly translate the reference position.

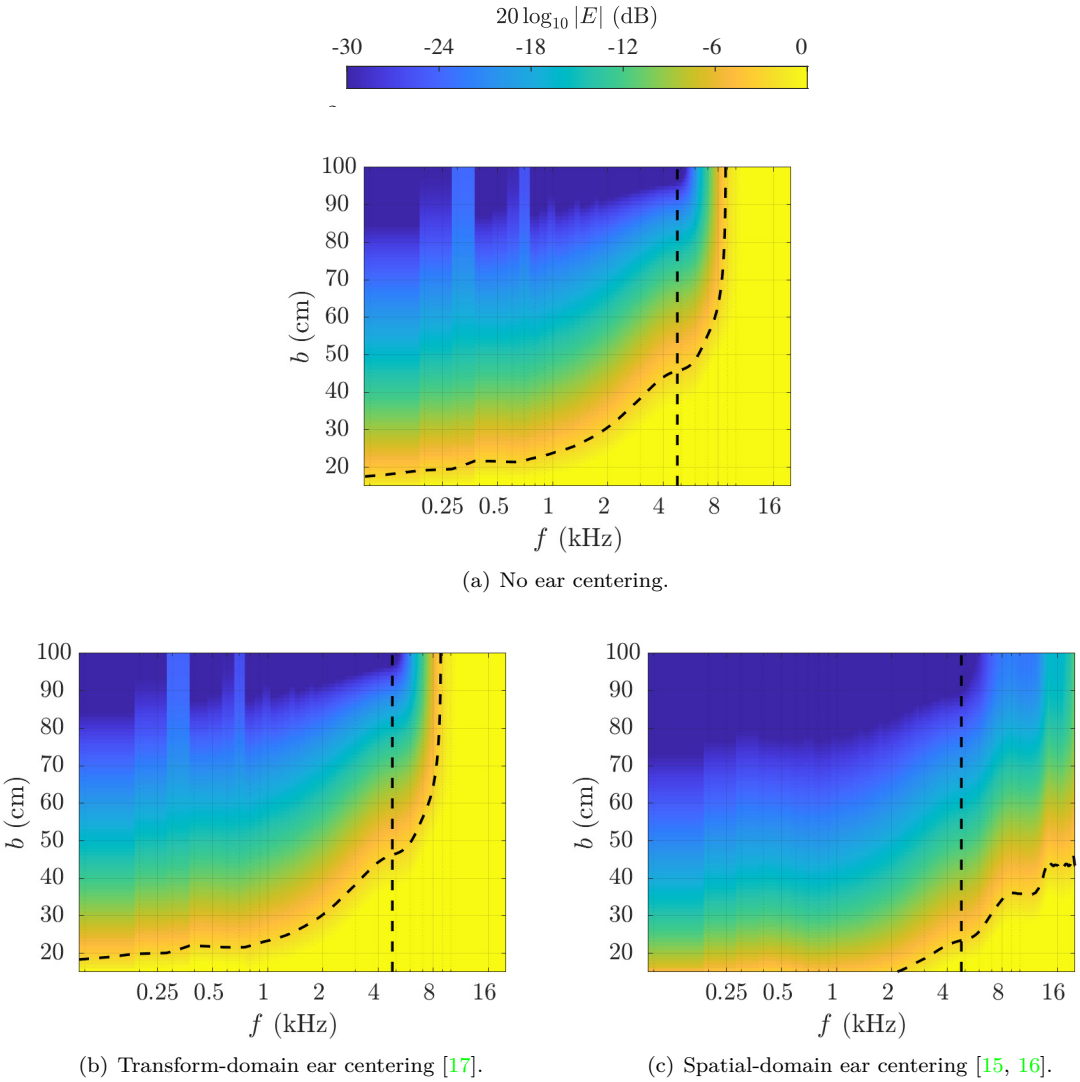
Operating in the spatial domain is computationally more efficient than operating in the transform domain. The reason behind this is that implementing the DVF of block ③ in Fig. 3 only requires distances from $\mathbf{r}_{\text{ref}}$ whereas implementing the transform-domain translation operator of block ③ in Fig. 4 requires DVFs that depend on distances from $\mathbf{r}_{\text{ref}}$ and $\mathbf{r}_{\text{ear}}$.

The mathematical formulation of the blocks used in Figs. 3 and 4 is available in detail in [15, 16].

3 EVALUATION WITH AN INDIVIDUAL CALCULATED HRTF

This section compares the performance of near-field ear centering in the spatial [15, 16] and transform [17] domains. The DVF and SFT algorithms respectively described in [9] and [19] were adapted to numerically evaluate four scenarios:

- No ear centering
- Transform-domain ear centering [17]
- Spatial-domain ear centering [15, 16].



(a) No ear centering.

(b) Transform-domain ear centering [17].

(c) Spatial-domain ear centering [15, 16].

Figure 5. Synthesis error calculated with (3). Black-dashed curves indicate -3 dB values. Black-dashed lines indicate $f_{\max}$ in (2).

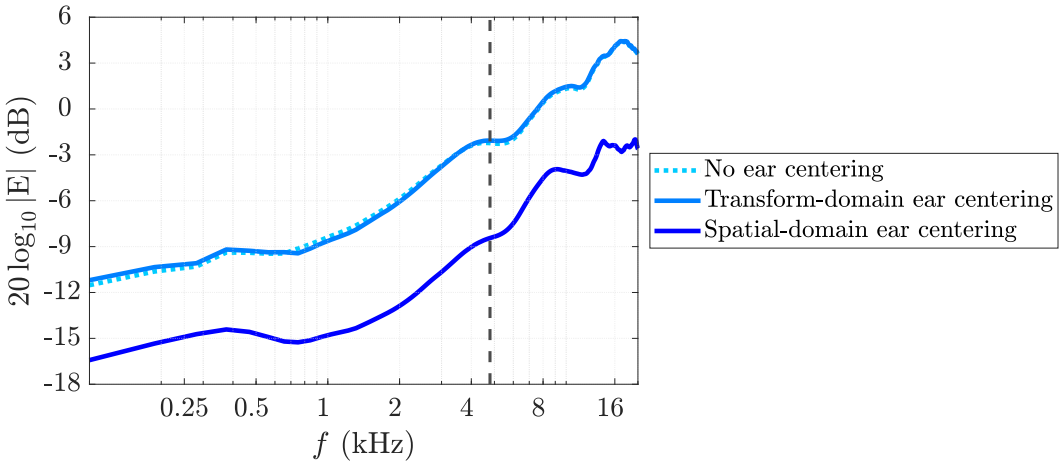


Figure 6. Synthesis error calculated with (4). The black-dashed line indicates $f_{\max}$ in (2).

3.1 Conditions

The conditions for evaluation are similar to the ones used in [15, 16]. Among the existing calculated near-field HRTF collections [20, 21], we chose the one in [21] because its data is open and its resolution across distance is dense. Left-ear HRTFs for one individual head model without torso are used in evaluations. The spatial features due to the torso are prominent at the lower frequencies and can extend up to 3 kHz, whereas the features due to the head and pinna span the middle and high frequencies [22]. Our assessment focuses on head and pinna features, for this reason the lower frequencies are also covered; hence, the absence of torso does not limit our results. The HRIRs have 512 samples along time, were sampled at 48 kHz, and $c = 344$ m/s. The left-ear positions were extracted from the head model.

The sound sources are distributed in spherical grids based on subdivisions of the edges of the icosahedron. Icosahedral samplings are chosen because they achieve bounded spherical integration errors distributed across all orders, whereas other samplings aiming at perfect quadrature at low SFT orders yield large errors concentrated in the high SFT orders [19].

The maximum frequency up to which reliable synthesis is ensured is calculated as

$$f_{\max} = \frac{cN_{\mathrm{g}}}{2\pi r_{\mathrm{bound}}}, \tag{2}$$

where c is the speed of sound in air, N_{g} is the maximum SFT order of the icosahedral grid, and r_{bound} is the radius of a sphere fully containing the head.

Datasets at distances b, ranging from 20 to 100 cm with 1 cm spacing, are used as target data. For each distance, 642 directions on an icosahedral grid are considered. Datasets are also synthesized for these distributions of points.

3.2 Error Metric

Target and synthesized HRTF datasets are respectively organized as $\boldsymbol{H}(b_i, \Omega_j, f_\kappa)$ and $\hat{\boldsymbol{H}}(b_i, \Omega_j, f_\kappa)$. Index $i = 1, 2, ..., 81$ indicates radial distances; index $j = 1, 2, ..., 642$, directions on the sphere; and index $\kappa = 1, 2, ..., 257$, frequency bins. The overall synthesis error across angles is defined as

$$E(b_i, f_\kappa) = \frac{\underset{\Omega_j}{\mathrm{RMS}}\{\boldsymbol{H} - \hat{\boldsymbol{H}}\}}{\underset{\Omega_j}{\mathrm{RMS}}\{\boldsymbol{H}\}}, \tag{3}$$

and the overall synthesis error across angles and distances as

$$\mathrm{E}(f_\kappa) = \underset{b_i}{\mathrm{RMS}}\left\{ \frac{\underset{\Omega_j}{\mathrm{RMS}}\{\boldsymbol{H} - \hat{\boldsymbol{H}}\}}{\underset{\Omega_j}{\mathrm{RMS}}\{\boldsymbol{H}\}} \right\}, \tag{4}$$

where RMS stands for root mean square along either directions Ω_j or distances b_i.

3.3 Results

Panels in Fig. 5 show synthesis errors calculated with (3) and displayed in a logarithmic scale, from -30 dB to 0 dB, to contrast with the perceivable HRTF dynamic range of around 30 dB, as reported in [23]. The black-dashed curves highlight the -3 dB values and are used as an indicator to compare among panels. We use the -3 dB indicator because this value is commonly considered as a perceivable difference. The black-dashed lines indicate $f_{\max}$ as formulated in (2).

Contrasting panels (a) and (b) in Fig. 5, it is observed that synthesis without ear centering and with ear centering in the transform domain yield similar results. Panel (c) in the same figure, on the other hand, clearly shows that the most accurate synthesis is obtained when applying ear centering in the spatial domain.

Figure 6 shows the overall synthesis errors across angles and distances calculated with (4). These results also indicate that spatial-domain ear centering yields a clear enhancement in near-field HRTF synthesis.

4 CONCLUSIONS

We presented a review of ear-centering methods used in HRTF synthesis. Ear centering can be applied in the spatial or transform domains, can be formulated for sources in the far field or near field, and can consider acoustic propagation in the free field or include an acoustically rigid scatterer that mimics a simple head.

When aiming at synthesizing near-field HRTFs, we showed that operating in the spatial domain is computationally more efficient than operating in the transform domain. Moreover, operating in the spatial domain is more accurate than operating in the transform domain when a frequency-dependent optimization of the ear position is not considered.

Extensions to this work might consider detailed mathematical analyses to contrast the equivalences of the existing methods for ear centering in the spatial and transform domain. Assessments with top-down auditory models and subjective tests could also provide more insight into the validity of ear centering.

ACKNOWLEDGEMENTS

This research was partially supported by JSPS KAKENHI Grant Numbers 19H04145 and 22H00523. The authors wish to thank the SI audio algorithm team for fruitful discussions. The authors also wish to thank Jorge Treviño for his insights and constructive debates.

REFERENCES

[1] S. T. Prepeliţă, J. G. Bolaños, V. Pulkki, L. Savioja, and R. Mehra, "Numerical simulations of near-field head-related transfer functions: Magnitude verification and validation with laser spark sources," *J. Acoust. Soc. Am.*, vol. 148, no. 1, pp. 153–166, 2020.

[2] J. M. Arend, H. R. Liesefeld, and C. Pörschmann, "On the influence of non-individual binaural cues and the impact of level normalization on auditory distance estimation of nearby sound sources," *Acta Acust. United Ac.*, vol. 5, p. 10, 2021.

[3] 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.

[4] D. S. Brungart, "Near-Field Virtual Audio Displays," *Presence: Teleop. Virt. Env.*, vol. 11, pp. 93–106, Feb. 2002.

[5] F. Monasterolo, S. Sakamoto, C. D. Salvador, Z. Cui, and Y. Suzuki, "The effect of target speech distance on reaction time under multi-talker environment," *IEICE Tech. Rep.*, vol. 118, pp. 83–88, Nov. 2018.

[6] S. Sakamoto, F. Monasterolo, C. D. Salvador, Z. Cui, and Y. Suzuki, "Effects of target speech distance on auditory spatial attention in noisy environments," in *Proc. ICA 2019 and EAA Euroregio*, (Aachen, Germany), pp. 2177–2181, Sept. 2019.

[7] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs," in *Proc. IEEE ICASSP*, vol. 4, pp. 45–48, May 2004.

[8] M. Pollow, K.-V. Nguyen, O. Warusfel, T. Carpentier, M. Müller-Trapet, M. Vorländer, and M. Noisternig, "Calculation of head-related transfer functions for arbitrary field points using spherical harmonics," *Acta Acust. United Ac.*, vol. 98, pp. 72–82, Jan. 2012.

[9] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Distance-varying filters to synthesize head-related transfer functions in the horizontal plane from circular boundary values," *Acoust. Sci. Technol.*, vol. 38, pp. 1–13, Jan. 2017.

[10] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 6, pp. 1060–1071, 2019.

[11] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Correction to "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling" [Jun 19 1060-1071]," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2194–2194, 2020.

[12] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *J. Audio Eng. Soc.*, vol. 69, pp. 104–117, Jan. 2021.

[13] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627, 2018.

[14] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Efficient representation and sparse sampling of head-related transfer functions using phase-correction based on ear alignment," *IEEE Trans. Audio, Speech, Language Process.*, pp. 2249–2262, 2019.

[15] A. Urviola, S. Sakamoto, and C. D. Salvador, "Ear centering for near-distance head-related transfer functions," in *Proc. Int. Conf. Immersive and 3D Audio (I3DA): from Architecture to Automotive*, (Bologna, Italy), IEEE, Sept. 2021.

[16] A. Urviola, S. Sakamoto, and C. D. Salvador, "Ear centering for accurate synthesis of near-field head-related transfer functions," *Appl. Sci.*, vol. 12, no. 16, 2022.

[17] J.-G. Richter, M. Pollow, F. Wefers, and J. Fels, "Spherical harmonics based HRTF datasets: Implementation and evaluation for real-time auralization," *Acta Acust. United Ac.*, vol. 100, pp. 667–675, July 2014.

[18] J. Blauert, *Spatial hearing: The psychophysics of human sound localization.* Cambridge, MA, USA; London, England.: MIT Press, revised ed., 1997.

[19] 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.

[20] Y. Rui, G. Yu, B. Xie, and Y. Liu, "Calculation of individualized near-field head-related transfer function database using boundary element method," in *Proc. 134th Convention Audio Eng. Soc.*, May 2013.

[21] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Dataset of near-distance head-related transfer functions calculated using the boundary element method," in *Proc. Audio Eng. Soc. Int. Conf. Spatial Reproduction —Aesthetics and Science—*, (Tokyo, Japan), Audio Engineering Society, Aug. 2018.

[22] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2053–2064, 2002.

[23] E. Rasumow, M. Blau, M. Hansen, S. van de Par, S. Doclo, V. Mellert, and D. Püschel, "Smoothing individual head-related transfer functions in the frequency and spatial domains," *J. Acoust. Soc. Am.*, vol. 135, no. 4, pp. 2012–2025, 2014.