HEAD-RELATED TRANSFER FUNCTION INTERPOLATION WITH A SPHERICAL CNN

Xingyu Chen, Fei Ma, Yile Zhang, Amy Bastine, Prasanga N. Samarasinghe

Audio and Acoustic Signal Processing Group, Australian National University
E-mail: {xingyu.chen1, fei.ma, yile.zhang, amy.bastine, prasanga.samarasinghe}@anu.edu.au

ABSTRACT

Head-related transfer functions (HRTFs) are crucial for spatial soundfield reproduction in virtual reality applications. However, obtaining personalized, high-resolution HRTFs is a time-consuming and costly task. Recently, deep learning-based methods showed promise in interpolating high-resolution HRTFs from sparse measurements. Some of these methods treat HRTF interpolation as an image super-resolution task, which neglects spatial acoustic features. This paper proposes a spherical convolutional neural network method for HRTF interpolation. The proposed method realizes the convolution process by decomposing and reconstructing HRTF through the Spherical Harmonics (SHs). The SHs, an orthogonal function set defined on a sphere, allow the convolution layers to effectively capture the spatial features of HRTFs which are sampled on a sphere. Simulation results demonstrate the effectiveness of the proposed method in achieving accurate interpolation from sparse measurements, outperforming the SH method and learning-based methods.

Index Terms— Head-related transfer function (HRTF), spherical CNN, interpolation, spatial audio, soundfield reproduction

1. INTRODUCTION

The growing popularity of virtual reality and augmented reality applications has sparked a research interest in spatial audio rendering methods. One of the most popular methods is to reproduce spatial audio binaurally through headphones. Binaural reproduction relies on the head-related transfer function (HRTF) [1] which represents the scattering effect of human anatomy with respect to the direction of sound. The dependence of HRTF on the anatomy of the listener (e.g., pinna, head, and torso shape), makes it highly individual [1]. Consequently, accurate HRTF measurement over a large number of directions is desirable for authentic spatial audio reproduction.

However, HRTF measurement is a time-consuming and expensive task [1]. To alleviate the burden of HRTF measurement, researchers proposed to interpolate dense HRTF from sparsely measured HRTF using methods such as bilinear [2] and barycentric interpolation method [3, 4]. Spatial domain interpolation was proposed to offer the advantage of preserving HRTF spatial features and ensuring the interpolation results were physically valid. The HRTF is decomposed onto the spatial basis functions such as spherical harmonics [5, 6] and principal components [7, 8].

Recently, researchers proposed convolutional neural network (CNN) based methods [9, 10, 11] for HRTF interpolation, and showed promising results. However, CNN was primarily designed for processing signals in a 2D plane. Thus, to process HRTF which is defined on a sphere, CNN-based methods require complicated plane-sphere projection which does not necessarily capture the spatial features of HRTF. There exists the need to address the plane-sphere projection while interpolating sparsely measured HRTF with a CNN.

In this paper, we propose a spherical CNN method [12, 13] for HRTF interpolation. The spherical CNN method exploits the spherical harmonics (SHs), an orthogonal function set defined on a sphere, to resolve the plane-sphere projection problem. We transform the magnitude of sparsely measured HRTF into SH coefficients which are convoluted with kernel functions expressed in the SH domain. The kernel function captures the spatial features of HRTF. Then we convert the convolution result into the magnitude of dense interpolated HRTF. We demonstrate the HRTF interpolation performance of the proposed spherical CNN method by comparing it with the spherical harmonic method and learning-based methods. Our code is available at github.com/xingyuaudio/HRTF-SCNN.

2. PROBLEM FORMULATION

2.1. HRTF interpolation

HRTF depends on the anatomy of a subject and the position of a sound source [1]. Let $H_{\mathrm{id}}^{\mathrm{left(right)}}(\Omega, l)$ denote the HRTF for left(right) ear of subject $\mathrm{id} \in \{1, \ldots, \mathrm{ID}\}$, source position Ω, and frequency bin $l \in \{1, \ldots, L\}$, ID is the number of subjects, and L is the number of frequency bins. Source position $\Omega = (r, \theta, \phi)$ is defined by spherical coordinates, the radius r, the elevation $\theta = [-\frac{\pi}{2}, \frac{\pi}{2}]$ and the azimuth $\phi = [0, 2\pi)$. Hereafter, the superscript for the left(right) ear and the subscript for the subject number are omitted for notation simplicity.

In this paper, we focus on the far-field HRTF magnitude

spectra $H_\text{M}(\Omega, l)$, i.e., the sound source is at a distance $r > 1.2$ m from the center of subject head [1],

$$H_\text{M}(\Omega, l) = 20 \log_{10} \left(|H(\Omega, l)| \right), \qquad (1)$$

which is a logarithmic scale similar to human auditory perception [14]. The far-field HRTF magnitude spectra $H_\text{M}(\Omega, l)$ exhibits minimal dependence on r [1], and thus we redefine $\Omega = (\theta, \phi)$. In this regard, we treat $H_\text{M}(\Omega, l)$ as a function of spherical signal $H : \mathbb{S}^2 \to \mathbb{R}^L$, where $\mathbb{S}^2$ is a two-dimensional manifold parameterized by (θ, ϕ).

The problem of interest is to interpolate spatially dense HRTFs $\left\{ H_\text{M}(\Omega, l), \Omega \in \Omega^\text{dense} \right\}$ from spatially sparse HRTFs $\left\{ H_\text{M}(\Omega, l), \Omega \in \Omega^\text{sparse} \right\}$, where $\Omega^\text{sparse} \subset \Omega^\text{dense}$.

2.2. Conventional CNN

HRTF interpolation was considered as an image super-resolution task by some researchers who adopt CNN layers [9, 10, 11]. The conventional CNN layer was originally designed for planar images,

$$(I * k)(x) = \sum_{y \in \mathbb{Z}^2} I(y) k(x - y), \qquad (2)$$

where I is the input image on the 2D plane $\mathbb{Z}^2$ and k is the learnable convolutional kernel [12]. This convolution exhibits shift equivariance [12]

$$[L_t I] * k = L_t [I * k], \qquad (3)$$

where L_t is translation operation and $L_t I$ is a shifted image. The shift equivariance Eq. (3) implies that a kernel k shares the same weight parameters at different regions of the image I. The weight sharing effectively reduces the number of model parameters and enhances model generalization capacity.

However, HRTF is not on a plane but on a sphere where the shift equivariance does not hold. This fact makes a direct application of conventional CNNs unfeasible and necessitates complicated projections of HRTF data [10, 11]. Considering the limitations of existing methods, we propose a spherical CNN method for HRTF interpolation. This method captures spherical features, enabling efficient processing of HRTF data which is distributed on a sphere.

3. METHODOLOGY

In this section, we propose a spherical CNN method for HRTF interpolation. We first introduce spherical convolution, which extends the idea of CNN to spherical signals, and then introduce spectral transformation, which reduces the computational complexity of spherical convolution, and at last describe the model structure.

3.1. Spherical convolution

Spherical CNNs generalize CNN to operate on functions defined on a sphere by using spherical convolution

$$(f * k)(\Omega) = \int_{g \in \mathbf{SO}(3)} f(g\eta) k \left(g^{-1} \Omega \right) \mathrm{d}g, \qquad (4)$$

where f and k are spherical functions, k is treated as a learnable convolutional kernel, g is the rotational operator in $\mathbf{SO}(3)$, η is the north pole [13]. The spherical convolution, Eq. (4), exhibits rotational equivariance

$$[gf] * k = g[f * k], \qquad (5)$$

which enables a spherical CNN to share the same set of weight parameters within a kernel k across different regions on a sphere. When considering H_M, our goal is use $H_\text{M} * k$ to effectively capture spatial features.

3.2. Spectral transformation

In this section, we present spectral transformation [13], a practical method that compute the spherical convolution Eq. (4) which does not have a closed-form solution. Specifically, the spherical convolution Eq. (4) is approximated by three steps: (i) Spherical Harmonic Transform (SHT): decompose spherical domain signals into SH coefficients, (ii) multiplication of SH coefficients and kernel coefficients, and (iii) Inverse Spherical Harmonic Transform (ISHT): converts the multiplication back into the spherical domain.

Firstly, let the HRTF be measured at directions $\{\Omega_p\}_{p=1}^P$ and L frequency bins. we decompose HRTF magnitude spectra $H_\text{M}(\Omega, l)$ onto the SHs:

$$\mathbf{H} = \mathbf{Ya}, \qquad (6)$$

where

$$\mathbf{H} = \begin{bmatrix} H_\text{M}(\Omega_1, 1) & \cdots & H_\text{M}(\Omega_1, L) \\ \vdots & \ddots & \vdots \\ H_\text{M}(\Omega_P, 1) & \cdots & H_\text{M}(\Omega_P, L) \end{bmatrix}$$

is a $P \times (N_H + 1)^2$ matrix containing the real SH basis functions,

$$\mathbf{Y} = \begin{bmatrix} Y_{00}(\Omega_1) & \cdots & Y_{N_H N_H}(\Omega_1) \\ \vdots & \ddots & \vdots \\ Y_{00}(\Omega_P) & \cdots & Y_{N_H N_H}(\Omega_P) \end{bmatrix}$$

is $Y_{nm}(\cdot)$ [15] of order n and mode m evaluated at P directions, and

$$\mathbf{a} = \begin{bmatrix} \alpha_{00}(1) & \cdots & \alpha_{00}(L) \\ \vdots & \ddots & \vdots \\ \alpha_{nn}(1) & \cdots & \alpha_{nn}(L) \end{bmatrix}$$

is a $(N_H + 1)^2 \times L$ matrix containing the SH coefficients. To obtain accurate coefficients, the number of HRTF measurements should greatly exceed the SH decomposition order, $P \gg (N_H + 1)^2$. $\mathbf{a}$ can be calculated through

$$\mathbf{a} = \left(\mathbf{Y}^\intercal \mathbf{Y} \right)^{-1} \mathbf{Y}^\intercal \mathbf{H}. \qquad (7)$$
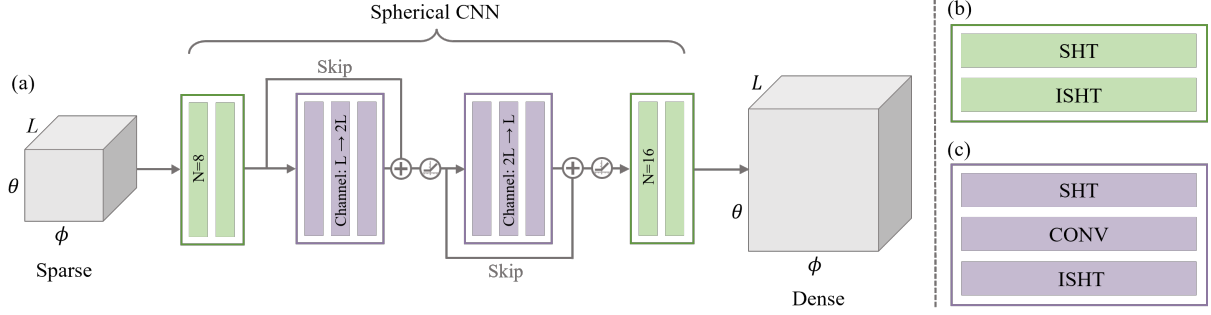
Fig. 1: The network structures of the model. (a): It starts with a Mapping Block for sparse to dense signal transformation, followed by two Convolutional Blocks for spatial feature learning, and ends with a Mapping Block for dense domain mapping. (b): The Mapping block: processes the spherical signal through SHT to convert it into SH coefficients, followed by ISHT maps back to the spherical signal region. (c): The Convolutional block: performs SHT on the spherical signal, followed by coefficients multiplication and ISHT.

Secondly, we multiply the HRTF SH coefficients $\mathbf{a}$ with a convolution kernel k by Hadamard product, $\mathbf{a} \odot \mathbf{b}$, where $\mathbf{b}$ is

$$\mathbf{b} = \begin{bmatrix} \beta_{00}(1) & \cdots & \beta_{00}(L) \\ \vdots & \ddots & \vdots \\ \beta_{n0}(1) & \cdots & \beta_{n0}(L) \end{bmatrix}$$

where β_{n0} is the $m = 0$ coefficient of kernel k. We realize Eq. (4) by a linear combination of spherical harmonics and associated coefficients α_{nm} and β_{n0} [16],

$$(f * k)(\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 2\pi \sqrt{\frac{4\pi}{2n+1}} \alpha_{nm} \beta_{n0} Y_{nm}(\Omega). \quad (8)$$

Eq. (8) represents the convolution of a single kernel whose output is summed across the rows $\sum_{i=1}^{L} (\mathbf{a} \odot \mathbf{b})_i$ to a $(N_H + 1)^2 \times 1$ column vector. In a convolutional layer, there can be u kernels and the layer output is u column vectors concatenated to form u channels

$$\mathbf{c} = [(\sum_{i=1}^{L} (\mathbf{a} \odot \mathbf{b})_i)_1, (\sum_{i=1}^{L} (\mathbf{a} \odot \mathbf{b})_i)_2, \ldots, (\sum_{i=1}^{L} (\mathbf{a} \odot \mathbf{b})_i)_u]. \quad (9)$$

Thirdly, we use ISHT to convert the multiplication back into the spherical domain

$$\mathbf{Yc}. \quad (10)$$

3.3. Model

Network structure: The model structure is presented in Fig. 1 (a). Overall this model generates the output block containing the densely interpolated HRTF from an input block containing the sparsely measured HRTF through a learnable kernel.

The input and output are connected by the mapping block Fig. 1 (b) and the convolution block Fig. 1 (c). The mapping block, Fig. 1(c), comprises of the SHT and the ISHT,

which aggregates the learned feature and maps back to the $\{H_{\mathrm{M}}(\Omega, l)\}_{\Omega \in \Omega^{\text{dense}}}$ domain. The convolutional blocks, Fig. 1(c), consist of SHT (Eq. 7), a convolutional layer (Eq. 9), and ISHT (Eq. 10). The sparse HRTFs undergo two convolutional blocks, which contain skip connections [17] and rectified linear unit (ReLU) activation function [18].

Loss function: The logarithmic spectral distortion (LSD) between the true H_{M} and predicted $\hat{H}_{\mathrm{M}}$ is used as the loss function of the model and performance measure across methods in Sec. 4,

$$\text{LSD} = \sqrt{\frac{1}{PL} \sum_{p=1}^{P} \sum_{l=1}^{L} \left| H_{\mathrm{M}}(\Omega_p, l) - \hat{H}_{\mathrm{M}}(\Omega_p, l) \right|^2}. \quad (11)$$

4. SIMULATION

In this section, we conducted experiments on HRTF interpolation to assess the effectiveness of our proposed method, and compare it with the SH method [6], the SH+DNN method [19], HRTF field method [20], and the CNN+GAN method [11].

4.1. Data pre-processing

Considering that multiple data sets need to be consistent [21] and sub-models may appear [22], we chose to conduct experiments exclusively on the HUTUBS dataset due to its extensive subject coverage [23]. We employed 94 subjects with duplicates excluded [23]. We used 77, 10, and 7 subjects to generate training, validation, and test data, respectively. We used ground truth 16 order SH coefficients to generate HRTFs from 480 directions and across $L = 93$ frequency bins, spanning from 172 Hz to 16 kHz.
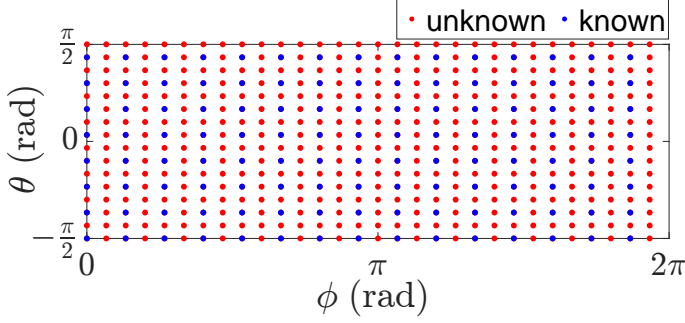
Fig. 2: Directions of 360 unknown and 120 known HRTFs.

Table 1: Comparison of model Parameters and average LSD for Unknown HRTFs across Subjects ID 88-94.

HRTF model	Parameters	Unknown
SH $N = 8$ [6]	-	7.24
SH+DNN [19]	803241	5.11
HRTF field [20]	4458589	4.19
CNN+GAN [11]	100827332	4.36
Proposed	**484530**	**2.25**

4.2. Training

Fig. 2 shows the direction of 120 known HRTFs and 360 unknown HRTFs. The LSD (Eq. (11)) serves as the loss function for the training process. The training procedure used the Adam optimizer [24] with a batch size of 14. Our model was trained on a V100 GPU for about 700 epochs using a learning rate of 0.001, with early stopping based on validation data.

4.3. Result and discussion

We conducted a comparative analysis by adopting interpolation methods, including SH, SH+DNN, HRTF field, and CNN+GAN. Table 1 provides insights into the number of parameters and average LSD of learning-based methods for unknown HRTFs across Subjects ID 88-94. Notably, the proposed method exhibits the lowest parameter count due to its streamlined architecture consisting of only two convolution blocks. In contrast, both SH+DNN and HRTF field methods rely on fully connected networks. SH+DNN requires sub-networks training for each frequency bin while CNN+GAN employs a generator with 11 conventional CNN layers. This observation highlights the efficiency of our approach. Furthermore, our method achieves the lowest LSD. The SH method with 8th-order coefficients performed worse than SH+DNN with 4th-order coefficients, due to its inability to accurately compute 8th-order coefficients with 120 known HRTFs. It indicates that the proposed method learns from convolution blocks, rather than solely relying on direct inferences through the mapping blocks.

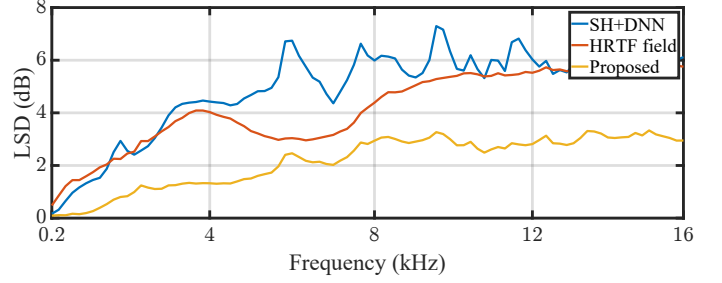We compare the HRTF interpolation performance of the

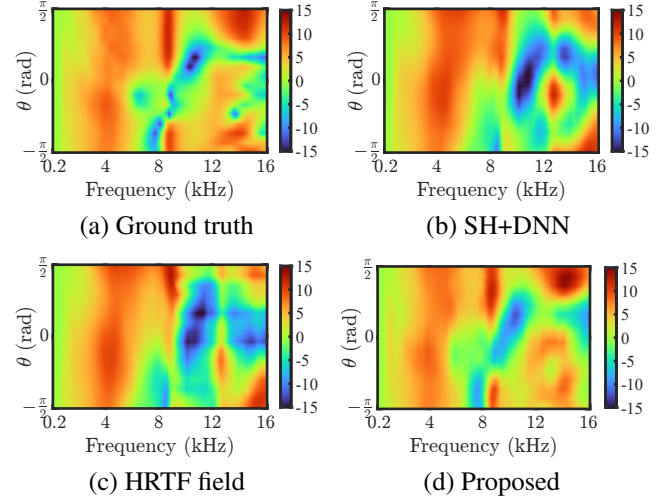Fig. 3: Interpolation LSD over frequencies.

(a) Ground truth

(b) SH+DNN

(c) HRTF field

(d) Proposed

Fig. 4: The left ear HRTF at $\phi = \pi$ of Subject ID 89.

proposed model with the SH+DNN and HRTF field model which has fewer parameters. Interpolating the high-frequency part of HRTF (above 8kHz) is challenging. Fig. 3 confirms that our method performs well in high frequencies. Fig. 4 showcases HRTF interpolation at $\phi = \pi$. It provides a visual assessment of how well the methods fit the *notch*, which represents a notable decrease at high-frequency, crucial for sound source direction detection [1]. The ground truth exhibits a peak in the 8 to 12 kHz frequency range. The other two methods show a different spectral shape compared to the ground truth. In contrast, our proposed method appropriately captures the peaks and notches of the ground truth.

5. CONCLUSION

This paper proposed a spherical CNN method to interpolate HRTFs based on sparse measurements. By leveraging the SH transformation, the proposed method directly exploits spherical information, eliminating the need for plane-sphere projection. Simulation results demonstrated that our method outperforms both conventional SH methods and learning-based methods. This method not only enhances HRTF interpolation but also offers a compelling alternative to the widely employed SH method in soundfield reproduction tasks.

6. REFERENCES

[1] S. Li and J. Peissig, "Measurement of head-related transfer functions: A review," *Appl. Sci.*, vol. 10, no. 14, pp. 5014, 2020.

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

[3] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *Proc. 16th Int. Audio Eng. Soc. Conf. Spatial Sound Reproduction,*, 1999, pp. 319–329.

[4] D. Poirier-Quinot and B. F. G. Katz, "The anaglyph binaural audio engine," in *Proc. Audio Eng. Soc. Conv. 144*, 2018.

[5] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized hrtf fitting using spherical harmonics," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.,*. IEEE, 2009, pp. 257–260.

[6] J. Ahrens, M. R. P. Thomas, and I. Tashev, "Hrtf magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data," in *Proc. Asia Pacific Signal Inf. Process. Association Conf.* IEEE, 2012, pp. 1–5.

[7] B.-S. Xie, "Recovery of individual head-related transfer functions from a small set of measurements," *J. Acoust. Soc. Amer.*, vol. 132, no. 1, pp. 282–294, 2012.

[8] M. Zhang, Z. Ge, T. Liu, X. Wu, and T. Qu, "Modeling of individual hrtfs based on spatial principal component analysis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 785–797, 2020.

[9] Y. Wang, Y. Zhang, Z. Duan, and M. Bocko, "Global hrtf personalization using anthropometric measures," in *Proc. Audio Eng. Soc. Conv. 150*, 2021.

[10] Z. Jiang, J. Sang, C. Zheng, A. Li, and X. Li, "Modeling individual head-related transfer functions from sparse measurements using a convolutional neural network," *J. Acoust. Soc. Amer.*, vol. 153, no. 1, pp. 248–259, 2023.

[11] A. O. T. Hogg, M. Jenkins, H. Liu, I. Squires, S. J. Cooper, and L. Picinali, "Hrtf upsampling with a generative adversarial network using a gnomonic equiangular projection," *arXiv preprint arXiv:2306.05812*, 2023.

[12] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical cnns," in *Int. Conf. Learning Representations*, 2018.

[13] C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis, "Learning so (3) equivariant representations with spherical cnns," in *Proc. Eur. Conf. Computer Vision*, 2018, pp. 52–68.

[14] Julius O. Smith, "Techniques for digital filter design and system identification with application to the violin," M.S. thesis, Stanford University, 06/1983 1983.

[15] E. G. Williams and J. A. Mann III, "Fourier acoustics: sound radiation and nearfield acoustical holography," 2000.

[16] J. R. Driscoll and D. M. Healy, "Computing fourier transforms and convolutions on the 2-sphere," *Adv. Appl. Math.*, vol. 15, no. 2, pp. 202–250, 1994.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[18] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learning*, 2010, pp. 807–814.

[19] J. Xi, W. Zhang, and T. D. Abhayapala, "Magnitude modelling of individualized hrtfs using dnn based spherical harmonic analysis," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.* IEEE, 2021, pp. 266–270.

[20] Y. Zhang, Y. Wang, and Z. Duan, "Hrtf field: Unifying measured hrtf magnitude representation with neural fields," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2023, pp. 1–5.

[21] Y. Wen, Y. Zhang, and Z. Duan, "Mitigating cross-database differences for learning unified hrtf representation," *arXiv preprint arXiv:2307.14547*, 2023.

[22] J. Pauwels and L. Picinali, "On the relevance of the differences between hrtf measurement setups for machine learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2023, pp. 1–5.

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

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.