MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

HRTF Personalization via Sim-to-Real Neural Field

Masuyama, Yoshiki; Wichern, Gordon; Boeddeker, Christoph; Richter, Julius; Edo, Takahiro; Bhosale, Swapnil; Le Roux, Jonathan

TR2026-135 September 29, 2026

Abstract

High-fidelity immersive audio experiences demand personalized head-related transfer functions (HRTFs), because generic HRTFs yield inaccurate perceptual localization. Neural fields (NFs) that take the sound source direction together with a compact set of subject-specific parameters to generate personalized HRTFs have achieved accurate HRTF spatial upsampling. These parameters are typically optimized from measured HRTFs, which requires an intricate measurement process in an anechoic chamber. Towards more accessible HRTF personalization, we propose a sim-to-real NF (S2RNF) that converts HRTFs simulated from an individual's 3D head mesh into their realistic counterparts. Specifically, we infer S2RNF's subjectspecific parameters from the simulated HRTFs, and use those parameters to predict the realistic HRTFs. Our experiments confirm that S2RNF outperforms the original simulations and existing sim-to-real methods.

Interspeech 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

HRTF Personalization via Sim-to-Real Neural Field

Yoshiki Masuyama[1], *Gordon Wichern*[1], *Christoph Boeddeker*[1], *Julius Richter*[1], *Takahiro Edo*[1], *Swapnil Bhosale*[1,2], *Jonathan Le Roux*[1]

[1] Mitsubishi Electric Research Laboratories (MERL), Cambridge, USA
[2] University of Surrey, UK

{masuyama,wichern,boeddeker,richter,leroux}@merl.com

Abstract

High-fidelity immersive audio experiences demand personalized head-related transfer functions (HRTFs), because generic HRTFs yield inaccurate perceptual localization. Neural fields (NFs) that take the sound source direction together with a compact set of subject-specific parameters to generate personalized HRTFs have achieved accurate HRTF spatial upsampling. These parameters are typically optimized from measured HRTFs, which requires an intricate measurement process in an anechoic chamber. Towards more accessible HRTF personalization, we propose a sim-to-real NF (S2RNF) that converts HRTFs simulated from an individual's 3D head mesh into their realistic counterparts. Specifically, we infer S2RNF's subject-specific parameters from the simulated HRTFs, and use those parameters to predict the realistic HRTFs. Our experiments confirm that S2RNF outperforms the original simulations and existing sim-to-real methods.

Index Terms: head-related transfer function, spatial audio, immersive audio, neural field, sim-to-real

1. Introduction

Head-related transfer functions (HRTFs) describe direction-dependent acoustic transfer functions imposed by an individual's pinna, head, and torso [1]. By using individual HRTFs, we can synthesize high-fidelity binaural audio and control perceptual localization. Such immersive audio technology is in demand for a wide range of applications such as virtual and augmented reality [2]. HRTFs differ from person to person as they depend on the shape of the pinna, head, and upper torso. Consequently, a generic HRTF leads to suboptimal perceptual localization, e.g., front-back confusion [3,4], and thus it is preferable to use individual HRTFs measured on dense spatial grids. However, the measurement process requires substantial time and effort under controlled laboratory environments.

To reduce the measurement effort, HRTF spatial upsampling predicts individual HRTFs for unmeasured directions from sparse measurements. Early works relied on signal processing, including amplitude panning [5, 6] and spatial basis decomposition [7, 8]. Rapid advances in deep learning have led to substantial improvements in upsampling performance [9–13], particularly with neural fields (NFs) [14–18]. NFs are trained to predict the HRTF from a given sound source direction in a gridless manner. Recent works typically share an NF across multiple subjects and personalize the generic NF by conditioning it using a compact set of subject-specific parameters [15,16]. Alongside community efforts such as the 2024 Listener Acoustic Personalization (LAP) Challenge [19], state-of-the-art methods have demonstrated promising upsampling performance even from only three measurements [18].

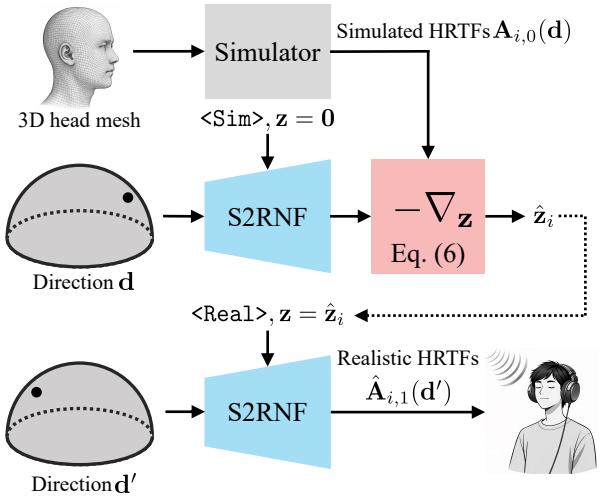

Figure 1: *Overview of S2RNF's inference procedure. S2RNF infers subject-specific parameters* $\mathbf{z}_i$ *so that the reconstruction error on simulated HRTFs* $\mathbf{A}_{i,0}$ *is minimized. Realistic HRTFs* $\hat{\mathbf{A}}_{i,1}$ *are obtained using the estimated parameters* $\hat{\mathbf{z}}_i$ *and changing the condition from simulated to real.*

Spatial upsampling still requires a few individual HRTF measurements, i.e., each subject needs to visit a laboratory with specialized equipment. In contrast, measurement-free personalization aims to predict personalized HRTFs from only anthropometric features and/or 3D meshes that are measurable even in everyday environments. One approach is to numerically simulate personalized HRTFs from a 3D mesh [20, 21], e.g., using Mesh2HRTF [22]. This approach can exploit well-developed prior knowledge of sound propagation but typically causes artifacts at high frequencies. An alternative is a data-driven approach that learns a mapping from anthropometric features to personalized HRTFs [23–29]. A promising direction is to combine both approaches by refining the simulated HRTFs using a neural network to suppress the simulation-based artifacts [30].

In this paper, we propose a sim-to-real NF (S2RNF) that personalizes an NF predicting realistic HRTFs using simulated HRTFs. Building upon NFs with subject-specific parameters [15, 16], S2RNF further incorporates domain-specific parameters that specify the domain (simulated or real) of the output HRTFs as illustrated in Fig. 1. These parameters allow S2RNF to compute the subject-specific parameters from only the simulated HRTFs. Then, S2RNF predicts realistic, personalized HRTFs for arbitrary directions by changing the domain-

This work was done while S. Bhosale was an intern at MERL.

specific parameters to those of the real domain. In our experiments, S2RNF provides substantial performance gains over simulated HRTFs and outperforms an existing method [30].

2. Preliminaries

2.1. Problem setting

Let us define an HRTF at both ears for subject i as $\mathbf{H}_i(\mathbf{d}) \in \mathbb{C}^{2\times F}$, where F is the number of frequency bins, $\mathbf{d} = (\theta, \phi)$ is the sound source direction, and $\theta \in [0, 2\pi)$ and $\phi \in [-\pi/2, \pi/2]$ are the azimuth and elevation, respectively. This paper focuses on modeling the log-magnitude response:

$$\mathbf{A}_i(\mathbf{d}) = 20\log_{10}(|\mathbf{H}_i(\mathbf{d})|) \in \mathbb{R}^{2\times F}. \tag{1}$$

The magnitude response is crucial for the elevation localization [31] and inherently covers the interaural level difference that is important for azimuth localization. HRTFs vary across subjects, i.e., $\mathbf{A}_{i_1}(\mathbf{d}) \neq \mathbf{A}_{i_2}(\mathbf{d})$ for $i_1 \neq i_2$, due to the differences in head and ear shapes. HRTF personalization aims to approximate a function that maps direction $\mathbf{d}$ to the individual HRTF $\mathbf{A}_i(\mathbf{d})$ for each subject i. Once the magnitude response is predicted, the corresponding phase can be obtained via minimum-phase reconstruction[1] [32].

2.2. Neural fields (NFs) for HRTF spatial upsampling

NFs have been widely used to represent HRTFs as a function of $\mathbf{d}$ due to their powerful modeling capability [15–18]. Recent works share an NF across subjects, and condition it by a compact set of subject-specific parameters $\mathbf{z}_i$ [15–17]:

$$\mathbf{A}_i(\mathbf{d}) \approx \texttt{NF}(\mathbf{d}, \mathbf{z}_i), \tag{2}$$

where $\mathbf{z}_i$ allows for the personalization of the output HRTF. The parameters can be a vector concatenated with $\mathbf{d}$ as input [15] or a tuple of low-rank matrices for advanced conditioning [16]. The NF in (2) is typically trained on HRTFs for multiple subjects to minimize the log-spectral distortion (LSD) [19]:

$$\mathcal{L}(\mathbf{A}_i(\mathbf{d}), \hat{\mathbf{A}}_i(\mathbf{d})) = \frac{1}{2\sqrt{F}} \sum_{\text{ch}\in\mathcal{C}} \|\mathbf{A}_i^{\text{ch}}(\mathbf{d}) - \hat{\mathbf{A}}_i^{\text{ch}}(\mathbf{d})\|_2, \tag{3}$$

where $\hat{\mathbf{A}}_i^{\text{ch}}(\mathbf{d}) \in \mathbb{R}^F$ is the predicted HRTF for each ear, and $\mathcal{C} = \{\text{left}, \text{right}\}$.

When sparse HRTF measurements for the target subject are available, HRTF personalization can be realized by adapting $\mathbf{z}_i$ based on the measurements. In [15], $\mathbf{z}_i$ is extracted from the measured HRTFs using the implicit gradient origin network (IGON) framework [33]. Specifically, $\mathbf{z}_i$ is updated by one gradient-descent step from the origin $\mathbf{z} = \mathbf{0}$ with unit step size:

$$\hat{\mathbf{z}}_i = -\frac{1}{|\mathcal{B}|} \sum_{b\in\mathcal{B}} \nabla_{\mathbf{z}} \mathcal{L}(\mathbf{A}_i(\mathbf{d}_b), \texttt{NF}(\mathbf{d}_b, \mathbf{z}))\big|_{\mathbf{z}=\mathbf{0}}, \tag{4}$$

where $b \in \mathcal{B}$ indexes the sparse measurements, and $|\mathcal{B}|$ denotes the number of measurements. IGON allows encoding the measured HRTFs into $\mathbf{z}_i$ without additional encoders, but requires at least a few HRTF measurements per subject.

2.3. Measurement-free HRTF personalization

HRTF personalization without individual measurements is desirable, as the measurement process requires a specialized environment such as an anechoic chamber. A common approach is to train a neural network to predict the personalized HRTFs from anthropometric features [23–29]. Recent works factorize this task using a pre-trained HRTF autoencoder [27, 28]. The anthropometric features are first mapped to a subject-specific latent variable of the autoencoder, similar to $\mathbf{z}_i$ in (2). The latent variable is then decoded to a magnitude response by using the pre-trained decoder. Another promising approach is to simulate HRTFs from an individual's 3D mesh and refine the simulated HRTFs using a neural network [30]. This approach can exploit a well-developed physics-based simulator and reduce the artifacts from the simulation in a data-driven manner. We will follow the latter approach, building on recent advances in NFs.

[1] The interaural time difference should also be personalized to improve the immersive audio experience. We defer this to future work.

3. Sim-to-Real Neural Field (S2RNF)

3.1. Overview of S2RNF

Building on the success of NFs in HRTF spatial upsampling, we formulate HRTF personalization as the adaptation of a generic NF to a target subject. The problem reduces to inferring the subject-specific parameters $\mathbf{z}_i$ in (2) from simulated HRTFs, thereby avoiding the requirement of individual HRTFs that should be measured in an anechoic chamber. The simulated HRTFs are considerably more accessible because they can be simulated from a 3D head mesh acquired in typical indoor environments using open-source software (e.g., Mesh2HRTF [22]).

To this end, we propose S2RNF, which enables sim-to-real transfer of HRTFs by inferring subject-specific parameters shared across simulated and measured HRTFs. Specifically, S2RNF leverages additional domain-specific parameters $\mathbf{w}_j$, where $j = 0$ and $j = 1$ correspond to the simulated and measured HRTFs, respectively:

$$\mathbf{A}_{i,j}(\mathbf{d}) \approx \texttt{S2RNF}(\mathbf{d}, \mathbf{z}_i, \mathbf{w}_j), \tag{5}$$

where $\mathbf{A}_{i,j}(\mathbf{d})$ is the simulated ($j = 0$) or measured ($j = 1$) HRTF for subject i and direction $\mathbf{d}$. As depicted in Figure 1, we first infer $\mathbf{z}_i$ from the simulated HRTFs $\mathbf{A}_{i,0}(\mathbf{d})$ with $\mathbf{w}_0$. Then, by switching the domain-specific parameters to $\mathbf{w}_1$, we can predict realistic HRTFs $\hat{\mathbf{A}}_{i,1}(\mathbf{d})$. In contrast to an existing sim-to-real method that directly refines the simulated HRTFs [30], S2RNF uses the simulated HRTFs for only computing $\mathbf{z}_i$. S2RNF can thus predict realistic HRTFs in a gridless manner regardless of the spatial grids for the simulation.

The design of S2RNF is inspired by the subject- and dataset-aware NF (SuDaField) [34], which proposed the use of domain-specific parameters to train an NF on multiple HRTF datasets. HRTFs from different datasets exhibit distinctive characteristics due to mismatched measurement conditions [35], and the domain-specific parameters are used to capture these mismatches. While SuDaField may combine simulated and measured datasets for training, it was not explicitly designed for and evaluated in sim-to-real transfer. In contrast, S2RNF aims at sim-to-real transfer, and we accordingly design the training procedure, as described in the following section.

3.2. Training of S2RNF

Inspired by HRTF field [15], we encode the simulated HRTFs into $\mathbf{z}_i$ using the IGON framework [33]:

$$\hat{\mathbf{z}}_i = -\frac{1}{|\mathcal{B}_0^{\text{enc}}|} \sum_{b\in\mathcal{B}_0^{\text{enc}}} \nabla_{\mathbf{z}} \mathcal{L}(\mathbf{A}_{i,0}(\mathbf{d}_b), \texttt{S2RNF}(\mathbf{d}_b, \mathbf{z}, \mathbf{w}_0))\big|_{\mathbf{z}=\mathbf{0}}, \tag{6}$$

where $\mathcal{B}_0^{\text{enc}} \subset \mathcal{B}_0$ is a set of direction indices for inferring $\mathbf{z}_i$, and $\mathcal{B}_0$ is a set of direction indices for the simulated HRTFs. Note that the gradient in (6) can be derived using automatic differentiation in a deep learning framework. Once $\hat{\mathbf{z}}_i$ is obtained, S2RNF estimates realistic, personalized HRTFs by swapping the domain-specific parameters to $\mathbf{w}_1$:

$$\hat{\mathbf{A}}_{i,1}(\mathbf{d}) = \texttt{S2RNF}(\mathbf{d}, \hat{\mathbf{z}}_i, \mathbf{w}_1). \tag{7}$$

Thanks to the efficient one-step gradient descent of the IGON framework, we can explicitly transfer the simulated HRTFs to realistic ones at each training step, through (6)–(7), in a manner consistent with the inference procedure. This allows us to optimize the network using a loss function defined on the predicted realistic HRTFs $\hat{\mathbf{A}}_{i,1}(\mathbf{d})$. This approach is much simpler than the iterative optimization of $\mathbf{z}_i$ in SuDaField [34], which would make training through explicit sim-to-real transfer impractical.

We jointly optimize S2RNF's parameters and the domain-specific parameters $\mathbf{w}_j$ for $j \in \{0, 1\}$ by minimizing the sum of the following loss functions $\mathcal{L}_1 + \lambda\mathcal{L}_0$ via backpropagation:

$$\mathcal{L}_1 = \sum_{i\in\mathcal{I}} \sum_{b\in\mathcal{B}_1^{\text{dec}}} \mathcal{L}(\mathbf{A}_{i,1}(\mathbf{d}_b), \texttt{S2RNF}(\mathbf{d}_b, \hat{\mathbf{z}}_i, \mathbf{w}_1)), \tag{8}$$

$$\mathcal{L}_0 = \sum_{i\in\mathcal{I}} \sum_{b\in\mathcal{B}_0^{\text{dec}}} \mathcal{L}(\mathbf{A}_{i,0}(\mathbf{d}_b), \texttt{S2RNF}(\mathbf{d}_b, \hat{\mathbf{z}}_i, \mathbf{w}_0)), \tag{9}$$

where $\mathcal{I}$ is a set of subjects within a training dataset, $\mathcal{B}_0^{\text{dec}} \subset \mathcal{B}_0$, $\mathcal{B}_1^{\text{dec}} \subset \mathcal{B}_1$, and $\mathcal{B}_1$ is a set of direction indices for the measured HRTFs. We set the balancing parameter λ to 1. Our final goal is to minimize $\mathcal{L}_1$ in (8) on unseen test subjects, but we find that including $\mathcal{L}_0$ in (9) helps the training of S2RNF. Note that the sum of two loss functions is back-propagated all the way to $\mathbf{w}_0$ in (6) by keeping the entire computational graph.

During training, we first sample a random half subset of directions from $\mathcal{B}_0$ as $\mathcal{B}_0^{\text{enc}}$ in (6), and assign the other directions to $\mathcal{B}_0^{\text{dec}}$ to calculate the loss function $\mathcal{L}_0$. While that is not required, we share the same directions between $\mathcal{B}_0^{\text{dec}}$ and $\mathcal{B}_1^{\text{dec}}$ to simplify the implementation by assuming the simulated and measured HRTFs share the same spatial grids. At inference, we use all the directions in $\mathcal{B}_0$ to infer $\hat{\mathbf{z}}_i$, i.e., $\mathcal{B}_0^{\text{enc}} = \mathcal{B}_0$.

3.3. Network architecture for S2RNF

While various neural network architectures are applicable to S2RNF, this paper focuses on neural networks with the bias-terms fine-tuning (BitFit) [36] following [16, 34]. As depicted in Fig. 2, our network mainly consists of fully-connected (FC) layers that take the random Fourier features (RFF) [37] of the sound source direction as input:

$$\boldsymbol{\mu} = [\sin(\theta - \pi), \cos(\theta - \pi), \sin(\phi), \cos(\phi)]^\mathsf{T}, \tag{10}$$

$$\mathbf{x}_{\text{RFF}} = [\sin(\mathbf{Q}\boldsymbol{\mu})^\mathsf{T}, \cos(\mathbf{Q}\boldsymbol{\mu})^\mathsf{T}]^\mathsf{T}, \tag{11}$$

where $\mathbf{Q} \in \mathbb{R}^{(C/2)\times 4}$ is sampled from an isotropic Gaussian distribution, C is the number of hidden units of the following FC layers, and $(\cdot)^\mathsf{T}$ denotes the transpose. Then, $\mathbf{x}_{\text{RFF}}$ is passed to an FC layer, resulting in $\mathbf{x}_0$, and fed into two core blocks with skip connections (shaded blocks in Fig. 2). In the core blocks, we apply subject-specific or domain-specific biases ($\mathbf{z}_{i,l} \in \mathbb{R}^C$ or $\mathbf{w}_{j,l} \in \mathbb{R}^C$) at each FC layer as follows:

$$\mathbf{x}_l = \begin{cases} \text{Swish}(\mathbf{P}_l\mathbf{x}_{l-1} + \mathbf{b}_l + \mathbf{z}_{i,l}) & \text{Subject-specific,} \\ \text{Swish}(\mathbf{P}_l\mathbf{x}_{l-1} + \mathbf{b}_l + \mathbf{w}_{j,l}) & \text{Domain-specific,} \end{cases} \tag{12}$$

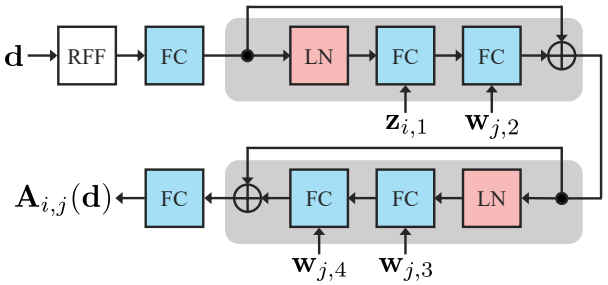

Figure 2: *Network architecture for S2RNF. FC and LN stand for a fully-connected layer and layer normalization, respectively.*

where $l = 1, \ldots, L$ denotes the index of the FC layers inside the core blocks, and $\mathbf{P}_l \in \mathbb{R}^{C\times C}$ and $\mathbf{b}_l \in \mathbb{R}^C$ respectively are the weight matrix and bias at the lth layer. In Fig. 2, L is set to 4, and the first FC layer leverages the subject-specific bias, while the other layers take the domain-specific ones, following the main experiment of SuDaField [34]. The prediction head outputs the two-channel log-magnitude spectra without any activation function.

4. Experiments

4.1. Comparison with baselines on SONICOM dataset

We first validate the personalization performance of S2RNF on an extended version [38] of the SONICOM dataset [39]. The extended version provides 200 pairs of measured and simulated HRTFs. HRTFs were measured at 793 directions for each subject, and we used those with a sampling rate of 48 kHz. The simulated HRTFs were computed for the same source directions from 3D head meshes using Mesh2HRTF [22]. The time-domain impulse responses were converted to the frequency domain using the discrete Fourier transform (DFT) with 256 points, i.e., $F = 129$. The HRTF pairs were randomly split into 160, 15, and 25 subjects for the training, validation, and test sets, respectively.

The architecture of S2RNF is depicted in Fig. 2, where the four hidden FC layers have 512 units with the Swish activation, i.e., $C = 512$. No activation function is applied in the prediction head. We trained S2RNF for 1500 epochs using the RAdam optimizer with a constant learning rate of 0.0005. We used the checkpoint with the lowest validation LSD on the measured HRTFs for evaluation. Our code is available online[2].

For evaluation, we used the root mean squared error (RMSE) and the mean absolute error (MAE) between the true and predicted log-magnitude spectra. The RMSE coincides with LSD in (3). The errors were calculated up to 20 kHz except for the DC component as in previous works [19, 30]. Furthermore, we evaluated a perceptual metric, Polar RMSE (PolRMSE) [40]. Specifically, an auditory model estimates the source direction from each HRTF [41], and RMSE between the estimated directions is used as a metric. As we focus on magnitude modeling, we adopt only polar angles following [42].

Since we focus on measurement-free HRTF personalization, we cannot fairly compare S2RNF with recent NF-based methods that require at least a few HRTF measurements per subject [17,18]. We thus used the simulated HRTFs and the log-magnitude responses averaged over the training subjects, i.e., non-personalized HRTFs, as baselines. Note that the average

[2] https://github.com/merlresearch/s2rnf

Table 1: *Results on the SONICOM dataset.*

	MAE [dB]	RMSE [dB]	PolRMSE [degree]
Mesh2HRTF	7.50	10.53	42.27
Avg. HRTF	3.78	5.05	41.63
S2RNF	**3.60**	**4.90**	**40.89**
w/o Sim. Loss	3.66	4.98	41.07

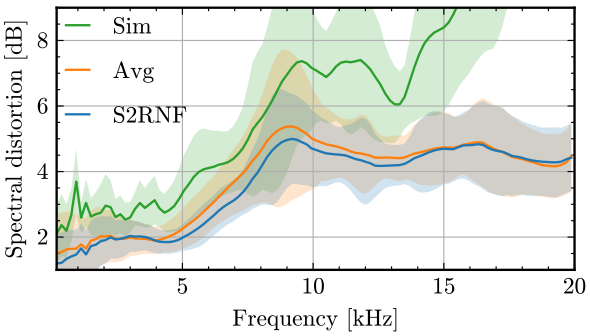

Figure 3: *Spectral distortion averaged over both ears and all directions. The solid lines and shaded areas correspond to the mean and standard deviation over subjects.*

baseline achieves reasonable performance even compared with recent spatial upsampling methods [19] despite its simplicity.

Table 1 presents the results averaged over the 25 test subjects. The HRTFs simulated by Mesh2HRTF result in large errors, which is consistent with findings from [30]. The average baseline, i.e., non-personalized HRTFs, performs much better than the simulated HRTFs. Our S2RNF outperforms the average baseline in all metrics, where the standard deviation of MAE across subjects is 0.48, while the standard deviation for the baseline is 0.59. Although the performance improvement may not seem huge, the effectiveness of S2RNF was validated by a paired one-sided t-test on MAE across test subjects ($p = 1.5\%$). As an ablation study, we compare the performance with S2RNF trained without the loss function on simulated HRTFs $\mathcal{L}_0$ in (9). It consistently degrades the performance as shown in the bottom row of Table 1. It is likely beneficial to fit the model parameters to simulated HRTFs when extracting $\mathbf{z}_i$ from the simulated HRTFs with the IGON framework.

Figure 3 shows the frequency-wise spectral distortion averaged over both channels and all directions. First, we can observe that the spectral distortion for the simulated HRTFs becomes significantly large at high frequency, which is again consistent with [30]. S2RNF reduces spectral distortion, especially around 5 kHz to 15 kHz. Since spectral notches and peaks in this frequency range affect perceptual elevation localization [43], this behavior of S2RNF is preferable.

4.2. Comparison with prior methods on HUTUBS dataset

We compare S2RNF with existing measurement-free HRTF personalization methods on the HUTUBS dataset [44], as it provides manually curated anthropometric features in addition to measured and simulated HRTFs. The HUTUBS dataset contains HRTFs from 96 subjects with 440 source directions per subject. To ensure a fair comparison with prior work relying on the anthropometric features [28], we excluded three subjects whose anthropometric features are unavailable (IDs 18, 79, and 92), as well as two repeated measurements (IDs 88 and 96). We used the first 85 subjects as a fixed training set and performed leave-one-out evaluation over the remaining 6 subjects, using 5 subjects for validation.

Table 2: *Results on the HUTUBS dataset.*

	MAE [dB]	RMSE [dB]	PolRMSE [degree]
Mesh2HRTF†	7.23	-	-
SPCA-DNN† [24]	6.89	-	-
BEM-DNN† [30]	4.80	-	-
Mesh2HRTF‡	7.14	9.38	36.60
Proto. DNN [28]	3.69	5.10	**32.90**
S2RNF	**3.66**	**5.02**	33.22

†Scores are taken from [30], where the test subjects were not reported.
‡The official simulated HRTFs are spatially downsampled to align the grid with the measured HRTFs.

The simulated HRTFs in the official dataset were computed using Mesh2HRTF [22] on a dense spatial grid that does not match the grid for measured HRTFs. We aligned the grid by interpolating the simulated HRTFs using a panning-based method [6]. The time-domain responses were converted to the frequency domain by DFT with 442 points, resulting in a frequency resolution of around 100 Hz [30].

Since the HUTUBS dataset provides additional anthropometric features, various measurement-free methods have been developed and applied to it. SPCA-DNN decomposes HRTFs using spatial principal component analysis (SPCA) and estimates individual principal components from the anthropometric features [24]. BEM-DNN applies SPCA to the mismatch between the simulated and measured HRTFs, refining the simulated HRTFs based on the anthropometric features [30]. Proto. DNN [28] leverages a pre-trained HRTF autoencoder [45] and estimates a subject-specific latent variable of the autoencoder (close to our $\mathbf{z}_i$) from the anthropometric features. We emphasize that S2RNF does not rely on the anthropometric features.

Table 2 summarizes the results on the HUTUBS dataset. Since the network and training details of [30] were not reported, we borrow their reported scores. The MAE gap between the two simulated HRTFs (Mesh2HRTF) may come from the mismatch of the test subjects and/or the grid interpolation. S2RNF improves the MAE compared to BEM-DNN by over 1 dB, which is much larger than the observed Mesh2HRTF mismatch (0.09 dB). Compared with Proto. DNN, whose training setup is fully aligned with ours, we achieved slightly better spectral distortions. Note that Proto. DNN requires an additional encoder to map the anthropometric features to the latent. Meanwhile, S2RNF can infer $\mathbf{z}_i$ without any encoder thanks to the IGON framework.

5. Conclusion

In this paper, we present S2RNF that transfers individual simulated HRTFs to realistic ones, which enables measurement-free HRTF personalization. By incorporating the domain-specific parameters, we can infer the subject-specific parameters, shared across the simulated and measured HRTFs, from only the simulated HRTFs. S2RNF outperforms the non-personalized average baseline on the SONICOM dataset, and achieves better spectral distortion compared with existing methods on the HUTUBS dataset. While this work relies on the simulated HRTFs in the official datasets, which are computed from high-fidelity 3D head meshes, our future work will explore S2RNF with meshes reconstructed from photogrammetry data [46].

6. Generative AI Use Disclosure

The authors used Generative AI to assist with language editing and polishing of the manuscript. In Fig. 1, the 3D head mesh and listening setup are AI-generated illustrations.

7. References

[1] H. Møller, "Fundamentals of binaural technology," *Appl. Acoust.*, vol. 36, no. 3-4, pp. 171–218, 1992.

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

[3] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123, 1993.

[4] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–464, 1996.

[5] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[6] A. Franck, W. Wang, and F. M. Fazi, "Sparse ℓ_1-optimal multiloudspeaker panning and its relation to vector base amplitude panning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 996–1010, 2017.

[7] 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. APSIPA*, 2012, pp. 1–5.

[8] 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/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2249–2262, 2019.

[9] I. D. Gebru, D. Marković, A. Richard, S. Krenn, G. A. Butler, F. De la Torre, and Y. Sheikh, "Implicit HRTF modeling using temporal convolutional networks," in *Proc. IEEE ICASSP*, 2021, pp. 3385–3389.

[10] X. Chen, F. Ma, and P. N. Samarasinghe, "Head-related transfer functions upsampling with physics-informed spherical convolutional neural network," *J. Acoust. Soc. Am.*, vol. 154, no. 4, pp. A183–A183, 2023.

[11] J. W. Lee, S. Lee, and K. Lee, "Global HRTF interpolation via learned affine transformation of hyper-conditioned features," in *Proc. IEEE ICASSP*, 2023.

[12] E. Thuillier, C. T. Jin, and V. Välimäki, "HRTF interpolation using a spherical neural process meta-learner," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1790–1802, 2024.

[13] X. Hu, J. Li, S. Zhang, S. Goetz, L. Picinali, O. B. Akan, and A. O. Hogg, "HRTFformer: A spatially-aware transformer for personalized HRTF upsampling in immersive audio rendering," *arXiv*, 2025.

[14] X. Lu, Y. Wang, J. Sang, and C. Zheng, "BiCG: Binaural cue generation from unified HRTF datasets," in *Proc. IEEE ICASSP*, 2025.

[15] Y. Zhang, Y. Wang, and Z. Duan, "HRTF field: Unifying measured HRTF magnitude representation with neural fields," in *Proc. IEEE ICASSP*, 2023.

[16] Y. Masuyama, G. Wichern, F. G. Germain, Z. Pan, S. Khurana, C. Hori, and J. Le Roux, "NIIRF: Neural IIR filter field for HRTF upsampling and personalization," in *Proc. IEEE ICASSP*, 2024, pp. 1016–1020.

[17] Y. Masuyama, G. Wichern, F. G. Germain, C. Ick, and J. Le Roux, "Retrieval-augmented neural field for HRTF upsampling and personalization," in *Proc. IEEE ICASSP*, 2025.

[18] D. Hu, J. Hu, and C. Jiang, "Graph neural field with spatial-correlation augmentation for HRTF personalization," *arXiv*, 2025.

[19] A. O. T. Hogg, R. Barumerli, R. Daugintis, K. C. Poole, F. Brinkmann, L. Picinali, and M. Geronazzo, "Listener acoustic personalization challenge - LAP24: Head-related transfer function upsampling," *IEEE Open J. Signal Process.*, vol. 6, pp. 926–941, 2025.

[20] T. Xiao and Q. Huo Liu, "Finite difference computation of head-related transfer function for human hearing," *J. Acoust. Soc. Am.*, vol. 113, no. 5, pp. 2434–2441, 2003.

[21] W. Kreuzer, P. Majdak, and Z. Chen, "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1280–1290, 2009.

[22] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "Mesh2HRTF: Open-source software package for the numerical calculation of head-related transfer functions," in *Proc. Int. Congr. Sound, Vib.*, 2015.

[23] M. Zhu, M. Shahnawaz, S. Tubaro, and A. Sarti, "HRTF personalization based on weighted sparse representation of anthropometric features," in *Proc. IC3D*, 2017.

[24] 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, Lang. Process.*, vol. 28, pp. 785–797, 2020.

[25] Y. Zhou, H. Jiang, and V. K. Ithapu, "On the predictability of HRTFs from ear shapes using deep networks," in *Proc. IEEE ICASSP*, 2021, pp. 441–445.

[26] Y. Wang, Y. Zhang, Z. Duan, and M. Bocko, "Global HRTF personalization using anthropometric measures," in *Proc. AES Conv.*, 2021.

[27] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, "Autoencoding HRTFs for DNN based HRTF personalization using anthropometric features," in *Proc. IEEE ICASSP*, 2019, pp. 271–275.

[28] R. Niu, S. Koyama, and T. Nakamura, "Head-related transfer function individualization using anthropometric features and spatially independent latent representation," in *Proc. IEEE WASPAA*, 2025.

[29] J. C. A. Sánchez, L. Comanducci, M. Pezzoli, and F. Antonacci, "Towards HRTF personalization using denoising diffusion models," in *Proc. IEEE ICASSP*, 2025.

[30] M. Zhang, J.-H. Wang, and D. L. James, "Personalized HRTF modeling using DNN-augmented BEM," in *Proc. IEEE ICASSP*, 2021, pp. 451–455.

[31] M. Geronazzo, S. Spagnol, and F. Avanzini, "Do we need individual head-related transfer functions for vertical localization? The case study of a spectral notch distance metric," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1247–1260, 2018.

[32] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1637–1647, 1992.

[33] S. Bond-Taylor and C. G. Willcocks, "Gradient origin networks," in *Proc. ICLR*, 2021.

[34] Y. Masuyama, G. Wichern, F. G. Germain, C. Ick, and J. Le Roux, "SuDaField: Subject- and dataset-aware neural field for HRTF modeling," *IEEE Open J. Signal Process.*, vol. 6, pp. 1169–1178, 2025.

[35] J. Pauwels and L. Picinali, "On the relevance of the differences between HRTF measurement setups for machine learning," in *Proc. IEEE ICASSP*, 2023.

[36] E. B. Zaken, S. Ravfogel, and Y. Goldberg, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proc. ACL*, vol. 2, 2022, pp. 1–9.

[37] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Proc. NeurIPS*, 2020, pp. 7537–7547.

[38] K. C. Poole, J. Meyer, V. Martin, R. Daugintis, N. Marggraf-Turley, J. Webb, L. Pirard, N. La Magna, O. Turvey, and L. Picinali, "The extended SONICOM HRTF dataset and spatial audio metrics toolbox," in *Proc. Conv. Eur. Acoust. Assoc.*, 2025.

[39] I. Engel, R. Daugintis, T. Vicente, A. O. Hogg, J. Pauwels, A. J. Tournier, and L. Picinali, "The SONICOM HRTF dataset," *J. Audio Eng. Soc.*, vol. 71, pp. 241–253, 2023.

[40] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1493–1510, 1999.

[41] R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, and R. Baumgartner, "A Bayesian model for human directional localization of broadband static sound sources," *Acta Acust.*, vol. 7, p. 12, 2023.

[42] Y. Zhang, A. Francl, R. Gao, P. Calamia, Z. Duan, and I. Anantabhotla, "Towards perception-informed latent HRTF representations," in *Proc. IEEE WASPAA*, 2025.

[43] K. Iida, Y. Ishii, and S. Nishioka, "Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 317–333, 2014.

[44] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718, 2019.

[45] Y. Ito, T. Nakamura, S. Koyama, S. Sakamoto, and H. Saruwatari, "Spatial upsampling of head-related transfer function using neural network conditioned on source position and frequency," *IEEE Open J. Signal Process.*, 2025.

[46] L. Pirard, K. C. Poole, and L. Picinali, "Enhancing photogrammetry reconstruction for hrtf synthesis via a graph neural network," *arXiv preprint arXiv:2510.02813*, 2025.