SuDaField: Subject- and Dataset-Aware Neural Field for HRTF Modeling

Masuyama, Yoshiki; Wichern, Gordon; Germain, François G; Ick, Christopher; Le Roux, Jonathan

TR2026-009 January 07, 2026

Abstract

This paper presents SuDaField, a subject- and dataset-aware neural field (NF) that can leverage multiple head-related transfer function (HRTF) datasets. NF-based HRTF modeling has gained much attention because its grid-agnostic formulation accommodates any spatial grids during training and inference. While NFs are grid-agnostic, their training on multiple datasets remains challenging, as HRTFs from different datasets exhibit distinct characteristics due to variations in measurement setups. To mitigate this issue, Task 1 of the Listener Acoustic Personalization (LAP) Challenge 2024 proposed the task of HRTF harmonization, which aims to compensate for dataset-specific effects while preserving spatial cues of the original HRTFs. The harmonization itself is still hindered by the difference in spatial grids and the ill-defined nature of ideal harmonized HRTFs. We thus propose a well-defined framework of HRTF conversion and realize this by concurrently performing NF training and disentanglement of subject- and dataset-specific information. Our NF adopts dataset-specific parameters shared across all subjects within each dataset, with these parameters capturing the influence of the measurement setups. By replacing the dataset-specific parameters with those of another dataset, we can convert HRTFs recorded in one environment to what they would be if recorded in another environment. Our experimental results show that the dataset-specific parameters allow us to effectively perform HRTF conversion, achieving state-of- the-art performance on Task 1 of the LAP Challenge 2024.

IEEE Open Journal of Signal Processing 2025

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

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

SuDaField: Subject- and Dataset-Aware Neural Field for HRTF Modeling

Yoshiki Masuyama[1], Gordon Wichern[1], François G. Germain[1], (Member, IEEE), Christopher Ick[1, 2] (Student Member, IEEE), Jonathan Le Roux[1] (Fellow, IEEE)

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Music and Audio Research Laboratory, New York University, Brooklyn, NY, USA

Corresponding author: Yoshiki Masuyama (email: masuyama@merl.com).

This work was performed while C. Ick was an intern at MERL.

ABSTRACT This paper presents SuDaField, a subject- and dataset-aware neural field (NF) that can leverage multiple head-related transfer function (HRTF) datasets. NF-based HRTF modeling has gained much attention because its grid-agnostic formulation accommodates any spatial grids during training and inference. While NFs are grid-agnostic, their training on multiple datasets remains challenging, as HRTFs from different datasets exhibit distinct characteristics due to variations in measurement setups. To mitigate this issue, Task 1 of the Listener Acoustic Personalization (LAP) Challenge 2024 proposed the task of *HRTF harmonization*, which aims to compensate for dataset-specific effects while preserving spatial cues of the original HRTFs. The harmonization itself is still hindered by the difference in spatial grids and the ill-defined nature of ideal harmonized HRTFs. We thus propose a well-defined framework of HRTF conversion and realize this by concurrently performing NF training and disentanglement of subject- and dataset-specific information. Our NF adopts dataset-specific parameters shared across all subjects within each dataset, with these parameters capturing the influence of the measurement setups. By replacing the dataset-specific parameters with those of another dataset, we can convert HRTFs recorded in one environment to what they would be if recorded in another environment. Our experimental results show that the dataset-specific parameters allow us to effectively perform HRTF conversion, achieving state-of-the-art performance on Task 1 of the LAP Challenge 2024.

INDEX TERMS Head-related transfer function, spatial audio, neural field, disentangled representation learning, HRTF conversion

I. INTRODUCTION

HEAD-related transfer functions (HRTFs) represent direction-dependent filtering effects caused by reflection and scattering around the ears, head, and torso as sound travels to the entrance of the ear canals. We can generate immersive binaural audio by applying HRTFs to dry monaural source signals, with many applications including mixed reality [1], [2]. HRTFs are unique to each subject due to the difference in anthropometric features. Thus, we would ideally want to provide each subject with their own HRTF measurements to maximize perceptual accuracy [3]. For example, the interaural time differences (ITDs) and interaural level differences (ILDs) are affected by the head size and are essential for azimuth localization [4]. Meanwhile, spectral coloration depending on the pinna is important for elevation localization and to avoid front and back confusion [5]. However, measuring HRTFs for each subject with dense spatial grids is time-consuming [6] and not easy to scale. As a more tractable alternative, HRTF spatial upsampling aims to estimate HRTFs on dense spatial grids from sparse measurements for the target subject. Meanwhile, HRTF personalization predicts HRTFs for the target subject by explicitly leveraging dense-grid HRTFs from a separate set of reference subjects. Various techniques have been developed for both approaches [7]–[12].

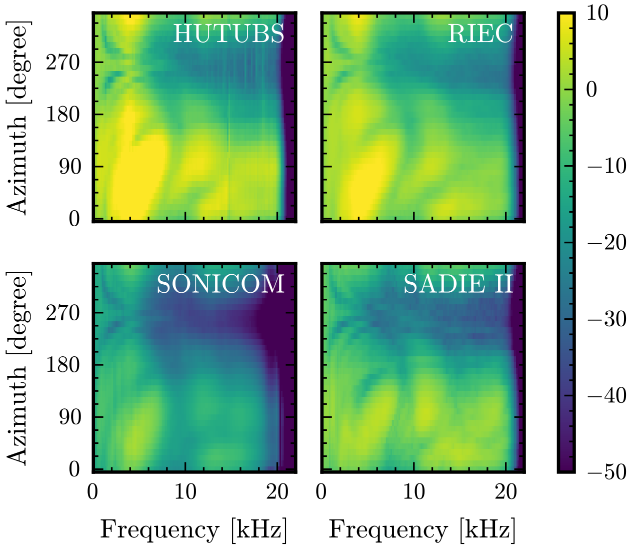

FIGURE 1. Average HRTF magnitudes in dB scale for the left ear on the horizontal plane. For each dataset, averages are computed over all subjects available in the SOFA repository.

Recently, machine-learning-based methods have shown promising performance for these tasks [13]–[26]. In particular, neural field (NF)-based methods for HRTF modeling have gained increasing attention due to their flexibility [21]–[25]. NFs are trained to map a given sound source direction to the corresponding HRTF, realizing grid-agnostic inference. Existing studies have explored various strategies to efficiently adapt an NF pre-trained on HRTFs from multiple subjects to a new subject [21], [23]. To achieve good generalization capability, these methods require a critical mass of data for pre-training, while existing HRTF datasets consist of well under a thousand subjects. On the other hand, naively combining multiple datasets to increase the number of training subjects presents challenges, as different recording setups result in noticeably different HRTF measurements [27], [28]. Figure 1 illustrates this variability. While the averages for HUTUBS [29] and RIEC [6] are indeed similar, the others differ not only in global gain but also in lower-level features like peak and notch locations. Ultimately, without additional care, this variability impedes the training of NFs on multiple datasets and adaptation to HRTFs from new measurement setups.

To develop methods mitigating this issue, Task 1 of the Listener Acoustic Personalization (LAP) Challenge 2024 asked participants to harmonize HRTFs for different subjects measured with different setups [30][1]. Specifically, this task, referred to as *HRTF harmonization*, aims to compensate for the influence of each measurement setup. During the challenge, harmonized HRTFs are evaluated on two aspects: preserving the localization cues [31] of the original HRTFs, and making their original measurement setup undetectable [32].

The ideal results of HRTF harmonization are, however, ill-defined because it is impossible to measure real-world HRTFs without any influence from the measurement setup. We instead propose a proxy task for harmonization named *HRTF conversion* (HC), akin to voice conversion in speech synthesis [33]. A method solving HC is expected to successfully convert HRTFs recorded in one environment to what they would have been if recorded in another one. This would realize the conversion of all HRTFs to a reference measurement setup, eliminating dataset-specific variability, and thereby satisfying the challenge criteria. Directly solving the HC task by optimizing for conversion performance is still generally infeasible due to the lack of HRTFs for identical subjects in different measurement setups, i.e., ground-truth "converted" HRTFs are not available.

In this paper, we instead approach this problem via a disentanglement perspective. Unlike existing methods that perform HRTF harmonization and NF training separately [21], [22], we propose to solve both tasks concurrently using an NF that includes explicit modeling of the dataset-specific effects. Specifically, we combine multiple HRTF datasets without harmonization and train an NF with subject- and dataset-specific parameters (see Fig. 2). This subject- and dataset-aware NF is referred to as SuDaField. While subject-specific parameters, similarly to those proposed in prior work [21]–[24], change from subject to subject, dataset-specific parameters are shared across subjects within each dataset. By decoupling the dataset-specific parameters from the subject-specific ones, we aim to disentangle the influence of subject-specific features, e.g., anthropometric features, from that of the recording setup. Through our experiments, we show how the decoupled parameters allow us to perform HC, resulting in the state-of-the-art performance on Task 1 of the LAP Challenge. Furthermore, we investigate the generalization capability of SuDaField when performing adaptation to HRTFs of new subjects. Our training and inference scripts are available online[2].

The remainder of this paper is organized as follows. Section II explains the problem setup and HRTF modeling by NFs. Then, in Section III, we describe our proposed method that decouples the subject- and dataset-specific parameters. In Section IV, we validate the effectiveness of SuDaField on Task 1 of the LAP Challenge 2024. Furthermore, Section V assesses the adaptation capability of SuDaField to HRTFs of new subjects. We conclude the paper in Section VI.

II. BACKGROUND

A. NOTATION

In principle, an HRTF describes an acoustic transfer function from a sound source in anechoic conditions to both ears, which is primarily affected by the sound source position and anthropometric features of the subject [4]. Let $\mathbf{H}(\theta, \phi, r, s) \in \mathbb{C}^{F\times 2}$ denote an HRTF for subject s and a sound source

[1] https://www.sonicom.eu/lap-challenge/

[2] https://github.com/merlresearch/SuDaField — contents to be added upon acceptance of the paper

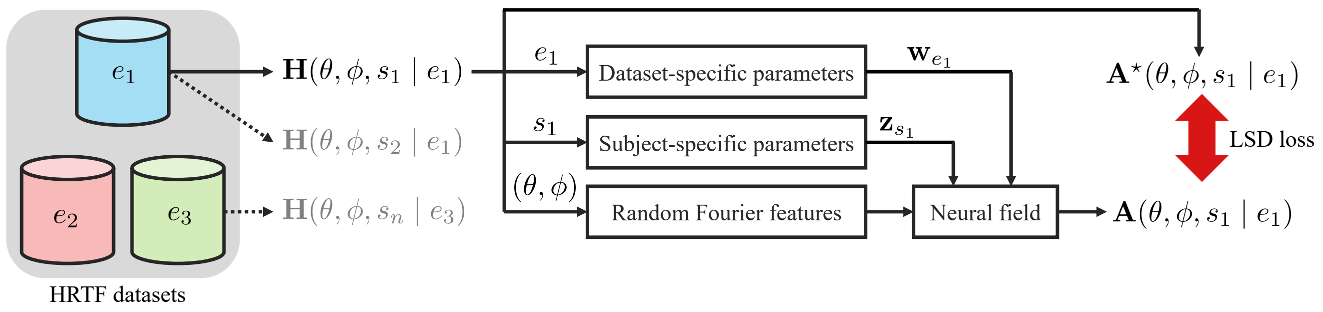
e_1
e_2
e_3
HRTF datasets
$\mathbf{H}(\theta, \phi, s_1 \mid e_1)$
$\mathbf{H}(\theta, \phi, s_2 \mid e_1)$
$\mathbf{H}(\theta, \phi, s_n \mid e_3)$
e_1
s_1
(θ, ϕ)
Dataset-specific parameters
Subject-specific parameters
Random Fourier features
$\mathbf{w}_{e_1}$
$\mathbf{z}_{s_1}$
Neural field
$\mathbf{A}^\star(\theta, \phi, s_1 \mid e_1)$
LSD loss
$\mathbf{A}(\theta, \phi, s_1 \mid e_1)$

FIGURE 2. **Overview of the proposed NF with $\mathbf{z}_{s_1}$ the subject-specific parameters and $\mathbf{w}_{e_1}$ the dataset-specific parameters.**

position (θ, ϕ, r) with respect to the center of the subject's head. Here, $\theta \in [0, 2\pi)$, $\phi \in [-\pi/2, \pi/2]$, and $r \in \mathbb{R}_+$ are the azimuth, elevation, and radius, respectively, and F is the number of frequency bins. The azimuth increases counterclockwise, $\theta = 0$ corresponding to the front of the subject, while the elevation increases upward, $\phi = 0$ indicating the equatorial plane.

In practice, measured HRTFs are distorted by their measurement setup, e.g., microphone and speaker responses. Denoting the dataset-specific measurement setup as e, HRTFs can be modeled as $\mathbf{H}(\theta, \phi, r, s \mid e)$. In this paper, we denote the set of datasets as $\mathcal{E} = \{e_1, e_2, \ldots, e_E\}$, where E is the total number of datasets. Then, we denote the set of subjects from HRTF dataset e as $\mathcal{S}_e$. Typically, the sets of subjects for different datasets are disjoint, i.e., $\mathcal{S}_{e_i} \cap \mathcal{S}_{e_j} = \emptyset$ for all $i \neq j$, although this is not a requirement for our method. Going forward, we omit r. First, r is typically fixed for each dataset and thus can be considered a dataset-specific parameter. Additionally, the influence of r on inter-dataset variability would be negligible in many of the LAP Challenge datasets because the source distances are around $1\,\mathrm{m}$ or beyond. We can thus reasonably introduce a far-field assumption for typical head sizes [34].

Finally, in line with current practices, we equate estimating $\mathbf{H}(\theta, \phi, s \mid e)$ with estimating its magnitude $\mathbf{A}(\theta, \phi, s \mid e) = |\mathbf{H}(\theta, \phi, s \mid e)|$, with $|\cdot|$ the elementwise absolute value, and its ITD $\tau(\theta, \phi, s \mid e)$ between left and right ears. The HRTFs are then recovered by computing the minimum-phase filter [35] corresponding to the magnitude and compensating for ITD by shifting the filters in the time domain [25].

B. HRTF MODELING WITH NEURAL FIELDS

NF is a class of neural network that maps a given coordinate to a quantity [36], and it has been successfully applied to spatial audio [19]–[24], [37]–[39]. For HRTF modeling, existing methods have so far been formulated in a dataset-agnostic way, i.e., without an explicit mechanism to absorb the effect of the recording setup e. In [21], an NF was trained to map the sound source direction (θ, ϕ) to the corresponding HRTF magnitude of subject s, i.e.,

$$\mathbf{A}(\theta, \phi, s) = \mathtt{NF}_s(\theta, \phi). \tag{1}$$

NFs are particularly appealing because they can straightforwardly accommodate training and inference with arbitrary spatial grids.

While the NF in (1) could be trained separately for each subject s, recent work [21], [23] also showed that it is typically beneficial to share most of the model parameters, as HRTFs are similar across subjects. Thus, using only a few subject-specific parameters $\mathbf{z}_s$, we can formulate the modeling of multiple subjects using a single shared NF as

$$\mathbf{A}(\theta, \phi, s) = \mathtt{NF}(\theta, \phi \mid \mathbf{z}_s). \tag{2}$$

In [21], $\mathbf{z}_s$ is a vector that is concatenated with the sound source direction as the input to the NF, i.e., conditioning by concatenation (CbC). Another approach is to slightly modify the model parameters based on $\mathbf{z}_s$ similarly to the parameter-efficient fine-tuning (PEFT) [23]. Both approaches have demonstrated promising adaptation capability to a new subject by optimizing the subject-specific parameters.

Note that, if this method were to be applied across multiple datasets, $\mathbf{z}_s$ should also implicitly contain the dataset-specific information [24], as each subject is associated with a dataset.

C. MULTI-DATASET TRAINING

Due to the absence of a single large-scale dataset, combining multiple datasets is highly desirable to train NFs, yet it introduces specific challenges. First, datasets typically use different spatial grids, although this is straightforwardly handled by the grid-agnostic NFs. Second, as already mentioned, each recording setup introduces specific distortion into the measured HRTFs [27], [28]. The distortion is substantial enough that the dataset of origin e of HRTFs $\mathbf{H}(\theta, \phi, s \mid e)$ can easily be determined with a simple support vector machine (SVM) classifier [32].

A simple approach to disentangling the inter-dataset variability of the HRTF magnitudes $\mathbf{A}(\theta, \phi, s \mid e)$ is to simply divide them by a dataset-specific normalization factor. For example, the average magnitude $\mathbf{A}_e^{\mathrm{Avg}} \in \mathbb{R}_+^{F \times 2}$ over all subjects and specific directions was used in a previous study [21]. Then, HRTF modeling methods can be trained on the normalized

$$\mathbf{A}^{\mathrm{Norm}}(\theta, \phi, s) = \mathbf{A}(\theta, \phi, s \mid e) \oslash \mathbf{A}_e^{\mathrm{Avg}}, \tag{3}$$

irrespective of their dataset of origin, where $\oslash$ denotes the element-wise division. At inference, we then recover HRTFs corresponding to a target environment e by denormalizing the NF output, multiplying it by $\mathbf{A}_e^{\text{Avg}}$. This simple normalization and its variants [22] have reasonably alleviated inter-dataset differences in HRTF magnitude. However, the simple normalization in (3) might not be sufficient to compensate for the influence of the measurement setups. In addition, it is uncertain how many subjects are needed to obtain a reliable average magnitude for each dataset. Hence, a more flexible way to handle HRTFs from different datasets is needed.

III. PROPOSED METHOD

A. MOTIVATION FOR DATASET-SPECIFIC PARAMETERS

The subject-specific parameters $\mathbf{z}_s$ in (2) should end up containing the dataset-specific information when training the generic NF in (2) on multiple HRTF datasets. This is because they are the only parameters allowed to be "sample"-dependent. Consequently, the subject- and dataset-specific information are entangled in $\mathbf{z}_s$. A recent work relevant to ours treats this point by domain adversarial learning, optimizing the parameters $\mathbf{z}_s$ at training to obscure their dataset of origin e [24]. This approach aims to eliminate the dataset-specific information $\mathbf{z}_s$, which is inconsistent with the training of the NF to reconstruct $\mathbf{A}(\theta, \phi, s \mid e)$. To disentangle the dataset-specific information from the subject-specific parameters, we need additional parameters to represent the difference in measurement setups.

B. DECOUPLING OF SUBJECT- AND DATASET-SPECIFIC PARAMETERS

As illustrated in Fig. 2, our SuDaField introduces dataset-specific parameters $\mathbf{w}_e$ that are shared across subjects $s \in \mathcal{S}_e$ as follows:

$$\mathbf{A}(\theta, \phi, s \mid e) = \mathtt{NF}(\theta, \phi \mid \mathbf{z}_s, \mathbf{w}_e). \tag{4}$$

By modeling the dataset-specific effect with $\mathbf{w}_e$, we expect that the subject-specific parameters $\mathbf{z}_s$ will focus on representing the subject-specific information, e.g., the anthropometric features.

The NF with the decoupled parameters can be trained similarly to existing NFs in (2). During the training, we select the parameters $(\mathbf{z}_s, \mathbf{w}_e)$ for each target $s \in \mathcal{S}_e$ and compute $\mathbf{A}(\theta, \phi, s \mid e)$. Then, the loss for HRTF magnitude estimation corresponds to the log-spectral distortion (LSD) between ground-truth and estimated magnitudes:

$$\begin{aligned}&\mathtt{LSD}(\mathbf{A}^\star(\theta, \phi, s \mid e), \mathbf{A}(\theta, \phi, s \mid e)) \\ &= \frac{1}{2}\sum_{c=1}^{2}\sqrt{\frac{1}{F}\sum_{f=1}^{F}\left(20\log_{10}\frac{A_{c,f}(\theta, \phi, s \mid e)}{A^\star_{c,f}(\theta, \phi, s \mid e)}\right)^2},\end{aligned} \tag{5}$$

where $(\cdot)^\star$ denotes the oracle value, and $c \in \{1, 2\}$ and $f = 1, \dots, F$ are the channel (left and right ear) and frequency indices, respectively. We can further modify the NF in (4) to jointly predict ITD in samples as $\tau(\theta, \phi, s \mid e)$ [25]. Then, the loss function for ITDs is given by

$$\begin{aligned}&\mathtt{MAE}_\epsilon\big(\tau^\star(\theta, \phi, s \mid e), \tau(\theta, \phi, s \mid e)\big) \\ &\quad = \mathrm{Max}\big(\big|\tau^\star(\theta, \phi, s \mid e) - \tau(\theta, \phi, s \mid e)\big|, \epsilon\big),\end{aligned} \tag{6}$$

where "ground-truth" ITD $\tau^\star(\theta, \phi, s \mid e)$ is the delay in integer samples that maximizes the cross correlation between left-ear and right-ear HRFTs in the time domain[3], and $\epsilon = 0.5$ is for accommodating the quantization error in the ground-truth ITD. Modeling the ITD is essential especially when some simulated datasets are included in the training datasets, as simulated HRTFs sometimes have distinctive ITDs compared with real recordings.

We can use SuDaField to perform any-to-any HC, i.e., converting HRTFs for a subject in an arbitrary dataset to what they would be if recorded with the setup of another dataset. Specifically, given a source dataset e_i and a target dataset e_j, we can convert the HRTFs of any subject $s \in \mathcal{S}_{e_i}$ to what their HRTFs would be if measured with the setup for dataset e_j by swapping the source dataset-specific parameters $\mathbf{w}_{e_i}$ for the target ones $\mathbf{w}_{e_j}$:

$$\mathbf{A}(\theta, \phi, s \mid e_j), \tau(\theta, \phi, s \mid e_j) = \mathtt{NF}(\theta, \phi \mid \mathbf{z}_s, \mathbf{w}_{e_j}). \tag{7}$$

In particular, we can tackle Task 1 of the LAP Challenge by converting the HRTFs of the subjects of all datasets to what they would be if measured with the setup of a given reference dataset e_{ref}.

C. NETWORK ARCHITECTURE FOR PROPOSED NF

While various conditioning frameworks are applicable to SuDaField, this paper focuses on neural networks with bias-terms fine-tuning (BitFit) [40]. Following our previous study [23], our NF mainly consists of fully-connected (FC) layers that take the random Fourier features (RFF) [41] of the sound source direction as input:

$$\boldsymbol{\mu} = [\sin(\theta - \pi), \cos(\theta - \pi), \sin(\phi), \cos(\phi)]^{\mathsf{T}}, \tag{8}$$

$$\mathbf{x}_0 = [\sin(\mathbf{P_0}\boldsymbol{\mu})^{\mathsf{T}}, \cos(\mathbf{P}_0\boldsymbol{\mu})^{\mathsf{T}}]^{\mathsf{T}}, \tag{9}$$

where $\mathbf{P}_0 \in \mathbb{R}^{(M/2)\times 4}$ is sampled from an isotropic Gaussian distribution, M is the size of the RFFs, and $(\cdot)^{\mathsf{T}}$ denotes the transpose. All hidden FC layers are equipped with GELU activation, and we apply subject-specific or dataset-specific biases inspired by BitFit at select layers:

$$\mathbf{x}_l = \begin{cases}\mathrm{GELU}(\mathbf{P}_l\mathbf{x}_{l-1} + \mathbf{q}_l) & \text{Generic (w/o BitFit)},\\ \mathrm{GELU}(\mathbf{P}_l\mathbf{x}_{l-1} + \mathbf{q}_l + \mathbf{z}_{s,l}) & \text{Subject-specific},\\ \mathrm{GELU}(\mathbf{P}_l\mathbf{x}_{l-1} + \mathbf{q}_l + \mathbf{w}_{e,l}) & \text{Dataset-specific},\end{cases} \tag{10}$$

where $l = 1, \dots, L$ denotes the index of the hidden layers, and $\mathbf{P}_l \in \mathbb{R}^{M\times M}$ and $\mathbf{q}_l \in \mathbb{R}^M$ respectively are the generic weight matrix and bias at the lth layer. The final prediction heads predict HRTF magnitude in the decibel scale and ITD without any activation function. We will show the importance of the selective application of the subject- and dataset-specific biases for disentanglement in Section IV.

[3] We used the `Spatial Audio Metrics` toolbox to compute ITDs: https://github.com/Katarina-Poole/Spatial-Audio-Metrics.

IV. EXPERIMENTS ON HRTF HARMONIZATION

In this experiment, we validate the effectiveness of SuDaField equipping the decoupled subject- and dataset-specific parameters on Task 1 of the LAP Challenge. After training the NF on the combined HRTF datasets, we artificially convert all the HRTFs as if they had been measured with a specific reference setup by swapping the dataset-specific parameters for those of the reference dataset as in (7). In all our experiments, we used HUTUBS as the reference dataset because it has the lowest sampling rate.

A. EXPERIMENTAL SETUP

Following the task description [30], we used eight datasets: 3D3A [42], CHEDAR [43], HUTUBS [29], RIEC [6], SADIE II [44], SONICOM [45], SCUT [34], and WiDESPREaD [46]. Among the eight datasets, CHEDAR and WiDESPREaD are simulated, and WiDESPREaD is designed to only simulate pinna-related filtering effects [46]. Consequently, their time-domain head-related impulse responses have distinctive characteristics, as we will show later. The organizers provided ten subjects for each of the eight datasets. All the HRTFs were downsampled to 44.1 kHz, and a 256-point discrete Fourier transform was performed.

The challenge asked participants to compensate for the influence from different measurement setups. More precisely, each system was evaluated in the following two stages:

Stage 1: This stage extracts a set of localization cues by using an auditory model [31], and the processed HRTFs must preserve the cues of the original HRTFs within a certain threshold. The percentage of the subjects for whom these cues are not preserved should be lower than 20%.

Stage 2: This stage assesses whether the effects of the different measurement setups have been compensated for[4]. For the HRTFs at the common 126 directions, the challenge evaluation scripts consider three kinds of features as the input to the dataset classifiers: HRTF magnitude in the linear scale, HRTF magnitude in the decibel scale, and time-domain impulse responses. Then, multiple classifiers including SVMs are trained to identify which dataset each HRTF originates from. The harmonized HRTFs are evaluated by the accuracy of the best classifier through five-fold cross-validation, and lower accuracy indicates better harmonization.

We trained two kinds of NFs. The first NF was designed to predict only the HRTF magnitude in the decibel scale, i.e., without ITD prediction head. In this scenario, we used oracle ITDs from the original HRTFs to compute the time-domain responses. The second one jointly predicts the HRTF magnitude and ITD. Since the two simulated datasets show atypical ITD behavior, we computed the ITD loss in (6) only on the six other datasets. In both cases, NFs consisted of four hidden FC layers with 512 units each and the prediction heads. We trained the NFs up to 300 epochs with the RAdam optimizer and a learning rate of 1.0×10^{-3}. The training was terminated if the training loss did not decrease for 10 consecutive epochs.

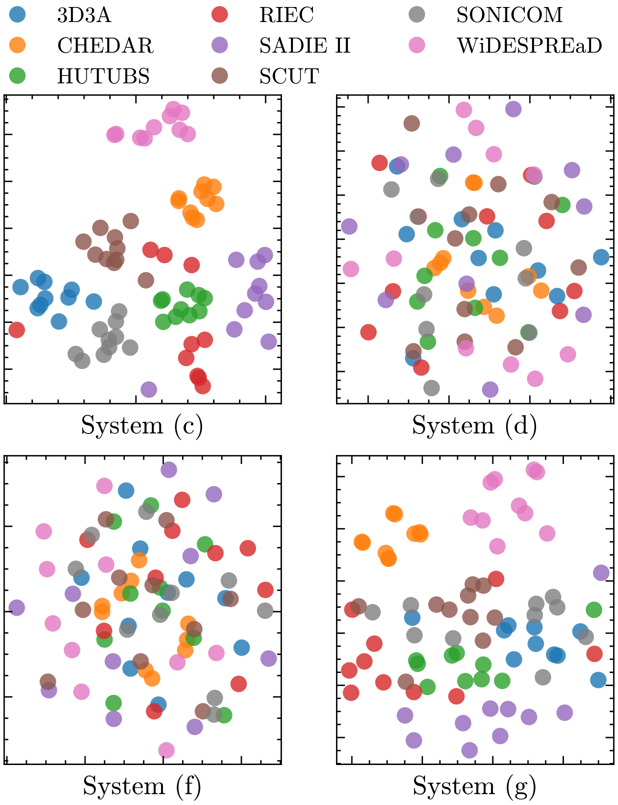

FIGURE 3. 2D visualization of subject-specific biases via t-SNE. Each circle corresponds to a different subject.

We also evaluated the normalization technique in (3) as the direction-independent normalization. Since the normalization might eliminate perceptual characteristics of HRTFs in $\mathbf{A}^{\text{Norm}}(\theta, \phi, s)$, we denormalized the normalized HRTFs from different datasets by multiplying by the average from HUTUBS. A variant proposed in [22] calculated the average at each direction and performed direction-dependent normalization. Its applicability is then limited to the directions shared across the source dataset and HUTUBS, i.e., it is no longer grid-agnostic. This is due to the dependence of the denormalization on the direction-wise average of HUTUBS.

B. EXPERIMENTAL RESULTS

Table 1 shows the performance for different output and decoupling conditions. Systems (a) and (b) correspond to the two HRTF normalization baselines. Although the direction-independent normalization has been used as pre-processing of NF training [21], [24], we find it to be insufficient to fully capture the inter-dataset variability, resulting in poor Stage 2 performance. Conversely, direction-dependent normalization [22] worked much better in Stage 2 but substantially degraded localization cues as shown by a Stage 1 failure

[4] We assessed the processed HRTFs with the official script: https://github.com/jpauwels/lap-task1/tree/main.

TABLE 1. **Results on Task 1 of the LAP Challenge 2024. The second and third columns indicate the indices of the layers with subject- and dataset-specific biases in (10), respectively. Loc. Fail. denotes the percentage of subjects whose processed HRTFs did not pass the Stage 1 criteria. The symbol † denotes the systems submitted to the challenge, and their results are taken from the challenge technical report [30].**

	System characteristics			Stage 1 (%) ↓	Stage 2 (%) ↓			
System ID	Subject-specific	Dataset-specific	ITD	Loc. Fail.	Magnitude	Magnitude dB	Time	Max
(a)	Direction-independent normalization			3.8	88.1	96.9	88.1	96.9
(b)	Direction-dependent normalization			63.8	45.0	41.3	53.8	53.8
(c)	1, 2	∅	Original	0.0	98.1	100.0	98.8	100.0
(d)	1	2	Original	17.5	18.8	19.4	41.3	41.3
(e)	1, 2	∅	Predicted	10.0	98.1	100.0	98.8	100.0
(f)	1	2	Predicted	12.5	**15.6**	**16.9**	**18.1**	**18.1**
(g)	1	4	Predicted	5.0	55.6	65.6	65.6	65.6
(h)	1, 2	3, 4	Predicted	7.5	32.5	40.0	33.8	40.0
(i)	EqPCA-UoA†			13.8	-	-	-	95.0
(j)	CoWiDeq†			10.0	-	-	-	92.3
(k)	IOA3D†			5.0	-	-	-	26.9

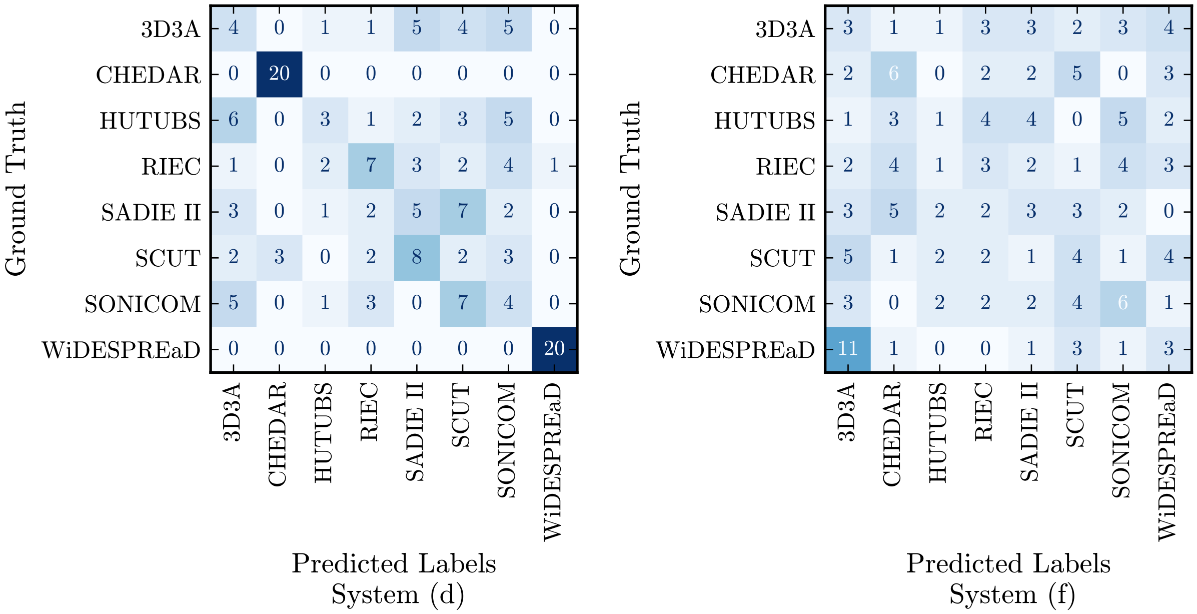

FIGURE 4. **Confusion matrices for dataset classification. The classifiers take harmonized time-domain impulse responses as input, where left and right channels are handled separately.**

rate of 63.8%. Altogether, these results confirm the need for more complex approaches to achieve reasonable performance for the task.

Systems (c) and (d) correspond to our first kind of NFs where only HRTF magnitudes are predicted while oracle ITDs are used. System (c) equips the subject-specific biases at the 1st and 2nd hidden FC layers and no dataset-specific ones. The dataset of origin for the reconstructed HRTFs was easily detected as shown by the perfect 100% maximum Stage 2 classification accuracy. In contrast, system (d), one of our SuDaField variants, shared the bias at the 2nd layer across subjects in the same dataset as the dataset-specific bias and achieved a better 41.3% maximum Stage 2 classification accuracy. This result indicates that the decoupled biases disentangle the effects of the measurement setup and the subject-specific characteristics. To visually support this point, we show the subject-specific biases using the t-SNE method [47] in Fig. 3, where t-SNE nonlinearly projects the concatenation of all the subject-specific biases $\mathbf{z}_{s,l}$ within each system to a 2-dimensional representation. For system (c), we observe clear clusters for the different datasets. This empirically confirms our earlier claim that, in the absence of decoupled dataset-specific biases, the measurement setup information necessarily has to be absorbed by the subject-specific biases. On the other hand, the biases from system (d) overlap well across different datasets, showing that our decoupled dataset-specific biases successfully absorbed most of the inter-dataset variability for magnitudes.

Next, systems (e)–(h) correspond to proposed SuDaField variants predicting both HRTF magnitudes and ITDs. Again, the HRTFs reconstructed by system (e), i.e., an NF without dataset-specific parameters, result in high classification accuracy. System (f) achieves the best, i.e., lowest, Stage 2 classification accuracies, which was the ranking score for the challenge. It is notably lower than the 26.9% maximum accuracy obtained by the challenge winner's submission, system (k). However, when we moved the dataset-specific biases to the 4th layer as system (g), the classification accuracy significantly increased, that is, performance got worse. We suspect this is because having the dataset-specific bias close to the prediction heads limits their modeling capacity, resulting in leakage of the dataset-specific effects into the subject-specific bias. This also can be seen in Fig. 3 when comparing the t-SNE plots: for the subject-specific biases of system (f), we observe no dataset-specific clusters, while for those of system (g), we observe some level of dataset-specific clusters. Finally, system (h) increased the number of layers equipped with subject- and dataset-specific biases, but the small gains in Stage 1 were combined with a much worse performance in Stage 2 compared to system (f). Through our investigations, we find it necessary to manually search for the most appropriate layers for subject- and dataset-specific biases.

C. EXPERIMENTAL DISCUSSION AND LIMITATION

We dig deeper into the best-performing systems by showing the dataset classification confusion matrix of systems (d) with the original ITD and (f) with the converted ITD in Fig. 4. When we only converted HRTF magnitudes (System (d)), the best classifier can still easily distinguish the two simulated datasets, CHEDAR and WiDESPREaD, with 100% accuracy. Meanwhile, by converting both HRTF magnitude and ITD, system (f) successfully decreased the classification accuracy across the board, including these two datasets. This result suggests that ITD conversion is crucial for any successful LAP Challenge Task 1 (i.e., HRTF harmonization) methods. This is further illustrated in Fig. 5, where we overlay the converted ITDs of subjects from the WiDESPREaD dataset when using system (f), alongside the original ITDs of those same subjects and the original ITDs of subjects in HUTUBS. It shows that the original ITDs for WiDESPREaD and HUTUBS have completely different distributions, likely because WiDESPREaD only simulates the pinna-related filtering effect. Hence, we necessarily require ITD conversion to have any chance to confuse the dataset classifier. Conversely, the distribution of the converted ITDs overlaps well with that for the reference dataset HUTUBS, confirming the system has successfully converted the ITDs. The switch to estimated ITDs might degrade the Stage 1 performance on the two simulated datasets. Indeed, the failure examples for system (f) consisted of five samples from CHEDAR, two samples from WiDESPREaD, and only three samples from the six real HRTF datasets.

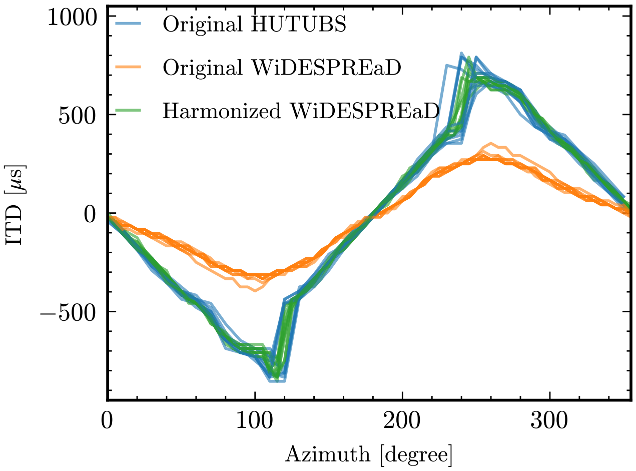

FIGURE 5. Original and harmonized ITDs on the horizontal plane, where system (f) is used for harmonization. Each line corresponds to different subjects.

Figure 6 depicts the HRTF magnitudes over directions in the median plane for the left (ear) channel before and after conversion by system (f). Here, elevations up to 90° correspond to the front of the subject, while angles greater than 90° correspond to the back. Regardless of the original datasets, the harmonized HRTF magnitudes clearly exhibit similar overall gain. Then, comparing two subjects from SONICOM before and after harmonization, we find that several subject-specific notches are preserved in the process, particularly at elevations below zero degrees. This result indicates that our method disentangles well subject- and dataset-specific effects into their corresponding biases.

HC via our best system, system (f), demonstrates promising performance in terms of the LAP Challenge Task 1 evaluation criteria. The converted HRTFs, however, inherit the influence of the reference setup, i.e., HUTUBS, which is misaligned with the ultimate goal of HRTF harmonization: compensating for the influence of any measurement setup. To achieve this goal, HRTF conversion requires a reference dataset without any measurement-related distortion. High-quality simulated HRTFs could serve as the reference, but the current simulated HRTF datasets, CHEDAR and WiDESPREaD, have significantly distinctive characteristics from real recordings, especially in terms of ITD, as shown in Fig. 4. These datasets may suffer from simulation artifacts instead of measurement-related distortion.

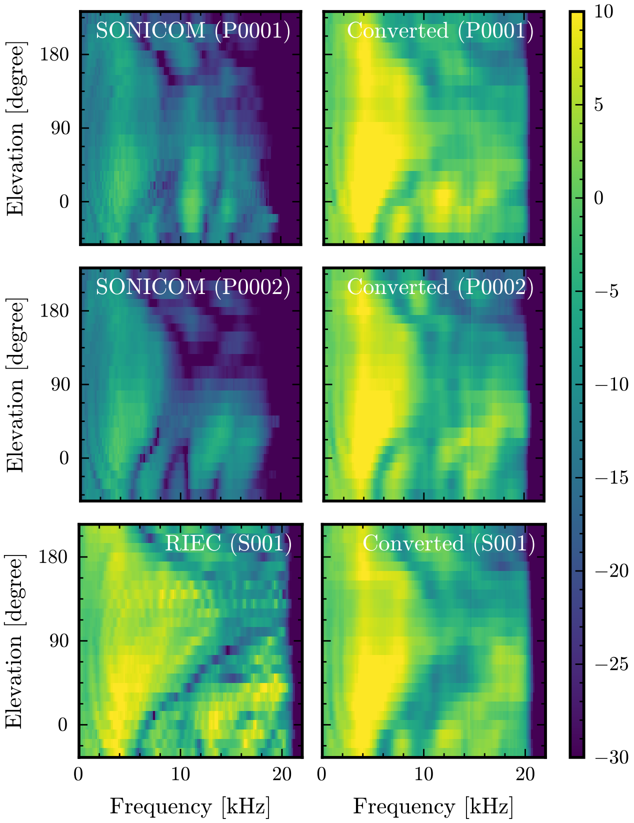

FIGURE 6. HRTF magnitude in dB scale over the median plane before and after conversion by system (f).

V. EXPERIMENTS ON ADAPTATION

In this experiment, we investigate the adaptation capability of SuDaField to new subjects from the SONICOM [45] and CIPIC [48] datasets. Here, we adapted NFs trained on HRTFs from multiple subjects to each target subject by fine-tuning specific parameters. The adaptation was performed with the sparse measurements from the target subjects and evaluated LSD with respect to ground-truth magnitudes in all other directions in the context of HRTF spatial upsampling [30].

A. EXPERIMENTAL SETUP

The adaptation performance of magnitude-prediction-only NFs is validated in two experimental settings, namely the adaptation to new subjects of a known dataset and the adaptation to a new dataset, exploring the influence of various NF and pre-training conditions.

We consider a total of five configurations between pre-training data, subject-specific parameters, and data-specific parameters as summarized in Table 2. Two NFs with only subject-specific biases, $S_{1;\emptyset}$ and $S_{1,2;\emptyset}$, were pre-trained on the ten SONICOM subjects included in the challenge dataset. Here, the subscript before the semicolon corresponds to the layer indices for subject-specific biases in (10), while $\emptyset$ after the semicolon indicates that there were no dataset-specific parameters. Three more NFs were pre-trained on all subjects from all 8 challenge datasets. Here, $A_{1,2;\emptyset}$ is equipped with only subject-specific biases at layers 1 and 2. Meanwhile, $A_{1;2}$ and $A_{1,2;3,4}$ contained both layers with subject-specific and dataset-specific biases. The network architecture and training configuration were the same as in the previous experiment. In addition, we jointly optimized both generic, subject-specific, and dataset-specific parameters during pre-training.

TABLE 2. System configurations and LSD scores when adapting NFs to ten new subjects from SONICOM and CIPIC. Here, the system IDs with S (resp. A) indicate that the pre-training is done only on SONICOM (resp. on all of the LAP Challenge datasets). The second and third columns show the layer indices for the corresponding biases.

	Layers with biases		LSD [dB]	
ID	Subject-specific	Dataset-specific	SONICOM	CIPIC
$S_{1;\emptyset}$	1	$\emptyset$	4.9	17.6
$S_{1,2;\emptyset}$	1, 2	$\emptyset$	4.8	15.8
$A_{1,2;\emptyset}$	1, 2	$\emptyset$	4.7	7.1
$A_{1;2}$	1	2	4.8	7.1
$A_{1,2;3,4}$	1, 2	3, 4	**4.6**	**5.7**

Each pre-trained NF was adapted to 1) ten new subjects from the original SONICOM dataset, and 2) ten subjects from the CIPIC dataset that is not included in the challenge datasets. When adapting the pre-trained NFs to the 10 unseen SONICOM subjects, we optimized the subject-specific biases [21], [23] while freezing all other NF parameters, including the dataset-specific biases found at pre-training for the SONICOM pre-training subjects. For adaptation to the 10 CIPIC subjects, we optimized both subject- and dataset-specific biases since the pre-training datasets do not include CIPIC. By definition, the dataset-specific biases are shared across the 10 subjects during adaptation. For all adaptations, we provide HRTFs at three randomly selected directions for each new (unseen) subject to evaluate the adaptation capability with sparse measurements. We then adapted each NF for 3000 epochs.

B. EXPERIMENTAL RESULTS

The LSD averaged over the 10 unseen subjects from the SONICOM dataset is illustrated in the top panel of Fig. 7. All the NFs ultimately converged to similar LSD values as also summarized in Table 2. Meanwhile, the system pre-trained on all datasets without the dataset-specific biases $A_{1,2;\emptyset}$ resulted in significantly higher LSD than others in early epochs. This could be because the system is required to acquire both subject- and dataset-specific information from scratch. On the other hand, the system with the dataset-specific parameters $A_{1,2;3,4}$ achieved substantially lower LSD even in the early epochs, presumably thanks to the benefits of pre-trained dataset-specific biases.

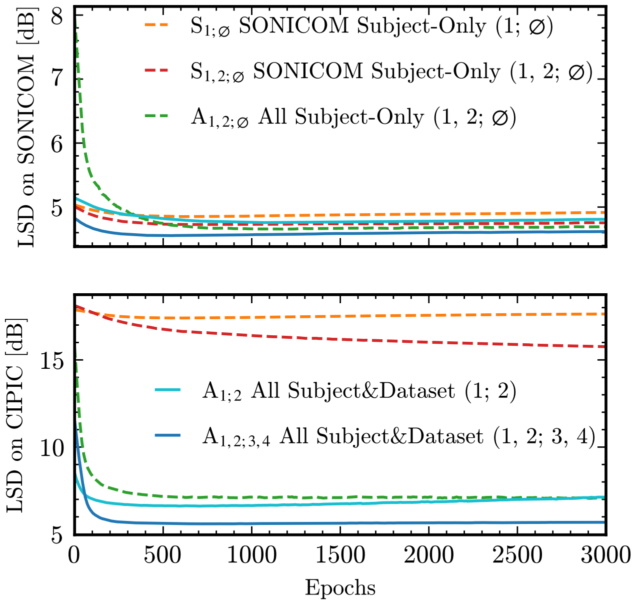

FIGURE 7. Average LSD on the SONICOM and CIPIC datasets. The dotted lines indicate the systems with only subject-specific parameters, while the solid lines show the systems with both subject- and dataset-specific parameters. The numbers before and after the semicolon denote the layer indices for the subject- and dataset-specific biases, respectively.

The bottom panel of Fig. 7 shows the results on the CIPIC dataset. As we can see, systems pre-trained on SONICOM only converge to very high LSD, while systems pre-trained on all the datasets perform much better. This result confirms two crucial points. First, the dataset-specific differences between mainstream real-world high-quality datasets is sufficient to necessitate multi-dataset pre-training to build systems with any kind of generalization capability to new measurement setups. Second, systems with dataset-specific biases (i.e., $A_{1;2}$ and $A_{1,2;3,4}$) unlock faster adaptation speed, and potentially further generalization performance, as $A_{1,2;3,4}$ clearly achieves the best overall adaptation performance.

C. EXPERIMENTAL DISCUSSION AND LIMITATION

According to Table 2, $A_{1,2;3,4}$ outperformed $A_{1;2}$ in adaptation performance, especially for the subjects from the unseen CIPIC dataset. This is inconsistent with the conversion performance in Table 1, where $A_{1;2}$ and $A_{1,2;3,4}$ correspond to Systems (f) and (h), respectively. We suppose that effective adaptation to new subjects from an unseen measurement setup demands a larger number of subject- and dataset-specific parameters, even as an increase in parameters impairs the disentanglement of the information.

In addition, we observed that the subject-specific biases for the CIPIC subjects were outside the distribution of the biases for the pre-training subjects as shown by the 2D t-SNE projections in Fig. 8. This result, unfortunately, indicates that the current model pre-trained on only 80 subjects from Task 1 of the LAP Challenge might not be enough to successfully adapt to a new dataset. From these observations, although SuDaField achieves promising HRTF harmonization performance in terms of Task 1 of the challenge, further investigation remains to improve its ability to generalize to unseen datasets.

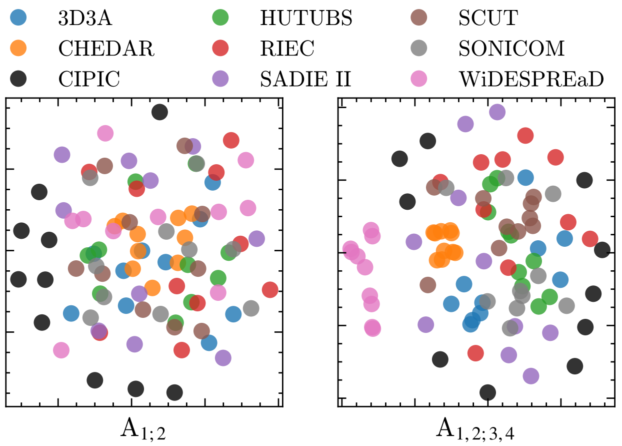

FIGURE 8. 2D visualization of subject-specific biases for the subjects from the pre-training datasets and the adaptation targets from CIPIC.

VI. CONCLUSION

This paper presented SuDaField, a subject- and dataset-aware NF that disentangles subject- and dataset-specific effects into corresponding decoupled parameters. SuDaField allows us to convert HRTFs as if they had been measured with a specific reference setup by swapping the dataset-specific parameters with those of the reference dataset. Furthermore, we explored the generalization capability of SuDaField to new subjects not only from known datasets but also from unknown datasets. Future work includes exploring further the HRTF conversion capabilities of our model, from its applicability as a data augmentation technique for training to its performance for adaptation in terms of perceptual metrics.

REFERENCES

[1] F. Keyrouz and K. Diepold, “Binaural source localization and spatial audio reproduction for telepresence applications,” *Presence: Teleoperators, Virtual Environ.*, vol. 16, no. 5, pp. 509–522, Oct. 2007.

[2] K. Iida, *Head-related transfer function and acoustic virtual reality*. Springer, 2019.

[3] E. M. Wenzel and S. H. Foster, “Perceptual consequences of interpolating head-related transfer functions during spatial synthesis,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 1993, pp. 102–105.

[4] A. Carlini, C. Bordeau, and M. Ambard, “Auditory localization: A comprehensive practical review,” *Frontiers in Psychology*, vol. 15, p. 1408073, 2024.

[5] 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 Transactions on Audio, Speech and Language Processing*, vol. 26, no. 7, pp. 1247–1260, 2018.

[6] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, “Dataset of head-related transfer functions measured with a circular loudspeaker

array," *Acoustical Science and Technology*, vol. 35, no. 3, pp. 159–165, Mar. 2014.

[7] Y. Iwaya, "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 340–E43, Nov. 2006.

[8] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. EL99–EL105, Jan. 2012.

[9] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, Jun. 1997.

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

[11] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs [head related transfer functions]," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2004.

[12] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *Journal of the Audio Engineering Society*, vol. 69, no. 1, pp. 104–117, Jan. 2021.

[13] Y. Ito, T. Nakamura, S. Koyama, and H. Saruwatari, "Head-related transfer function interpolation from spatially sparse measurements using autoencoder with source position conditioning," in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022.

[14] M. Zhu, M. Shahnawaz, S. Tubaro, and A. Sarti, "HRTF personalization based on weighted sparse representation of anthropometric features," in *Proc. International Conference on 3D Imaging (IC3D)*, Dec. 2017.

[15] Y. Zhou, H. Jiang, and V. K. Ithapu, "On the predictability of HRTFs from ear shapes using deep networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 441–445.

[16] 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," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 32, pp. 2085–2099, 2024.

[17] X. Chen, F. Ma, and P. N. Samarasinghe, "Head-related transfer functions upsampling with physics-informed spherical convolutional neural network," *Journal of the Acoustical Society of America*, vol. 154, no. 4, pp. A183–A183, 2023.

[18] E. Thuillier, C. T. Jin, and V. Välimäki, "HRTF interpolation using a spherical neural process meta-learner," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 32, pp. 1790–1802, Feb. 2024.

[19] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 3385–3389.

[20] J. W. Lee, S. Lee, and K. Lee, "Global HRTF interpolation via learned affine transformation of hyper-conditioned features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023.

[21] Y. Zhang, Y. Wang, and Z. Duan, "HRTF field: Unifying measured HRTF magnitude representation with neural fields," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023.

[22] Y. Wen, Y. Zhang, and Z. Duan, "Mitigating cross-database differences for learning unified HRTF representation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023.

[23] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1016–1020.

[24] T. Lobato and R. Sottek, "A process for calibrating HRTFs based on differentiable implicit representations and domain adversarial learning," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2024, pp. 271–275.

[25] Y. Masuyama, G. Wichern, F. G. Germain, C. Ick, and J. Le Roux, "Retrieval-augmented neural field for HRTF upsampling and personalization," *arXiv preprint arXiv:2501.13017*, 2025.

[26] X. Lu, Y. Wang, J. Sang, and C. Zheng, "BiCG: Binaural cue generation from unified HRTF datasets," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[27] S. Li and J. Peissig, "Measurement of head-related transfer functions: A review," *Applied Sciences*, vol. 10, no. 14, 2020.

[28] A. Andreopoulou, D. R. Begault, and B. F. G. Katz, "Inter-laboratory round robin HRTF measurement comparison," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 895–906, 2015.

[29] 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," *Journal of the Audio Engineering Society*, vol. 67, no. 9, pp. 705–718, Sep. 2019.

[30] M. Geronazzo, L. Picinali, A. Hogg, R. Barumerli, K. Poole, R. Daugintis, J. Pauwels, S. Ntalampiras, G. Mclachlan, and F. Brinkmann, "Technical report: SONICOM/IEEE listener acoustic personalisation (LAP) challenge-2024," *TechRxiv*, 2024.

[31] 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 Acustica*, vol. 7, 2023.

[32] J. Pauwels and L. Picinali, "On the relevance of the differences between HRTF measurement setups for machine learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[33] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp. 132–157, 2021.

[34] G. Yu, R. Wu, Y. Liu, and B. Xie, "Near-field head-related transfer-function measurement and database of human subjects," *Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. EL194–EL198, 2018.

[35] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637–1647, Mar. 1992.

[36] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," *Computer Graphics Forum*, vol. 41, no. 2, pp. 641–676, May 2022.

[37] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2022, pp. 3165–3177.

[38] D. Di Carlo, A. A. Nugraha, M. Fontaine, and K. Yoshii, "Neural Steerer: Novel steering vector synthesis with a causal neural field over frequency and source positions," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing Workshops (ICASSPW)*, 2024, pp. 740–744.

[39] S. Koyama, J. G. C. Ribeiro, T. Nakamura, N. Ueno, and M. Pezzoli, "Physics-informed machine learning for sound field estimation: Fundamentals, state of the art, and challenges," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 60–71, 2024.

[40] E. B. Zaken, S. Ravfogel, and Y. Goldberg, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proc. Annual Meeting of the Association for Computational Linguistics*, vol. 2, 2022, pp. 1–9.

[41] 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. Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2020, pp. 7537–7547.

[42] R. Sridhar, J. G. Tylka, and E. Y. Choueiri, "A database of head-related transfer function and morphological measurements," in *Audio Engineering Society Convention*, 2017.

[43] S. Ghorbal, X. Bonjour, and R. Séguier, "Computed HRIRs and ears database for acoustic research," *Journal of the Audio Engineering Society*, no. 10361, 2020.

[44] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, “A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database,” *Applied Sciences*, vol. 8, no. 11, 2018.

[45] I. Engel, R. Daugintis, T. Vicente, A. O. Hogg, J. Pauwels, A. J. Tournier, and L. Picinali, “The SONICOM HRTF dataset,” *Journal of the Audio Engineering Society*, vol. 71, pp. 241–253, May 2023.

[46] C. Guezenoc and R. Seguier, “A wide dataset of ear shapes and pinna-related transfer functions generated by random ear drawings,” *Journal of the Acoustical Society of America*, vol. 147, no. 6, pp. 4087–4096, 2020.

[47] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” *JMLR*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[48] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The CIPIC HRTF database,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2001, pp. 99–102.