IEEE Open Journal of Signal Processing

Received 25 February 2025; revised 20 June 2025; accepted 28 November 2025. Date of publication 4 December 2025; date of current version 1 January 2026. The review of this article was arranged by Associate Editor M. Geronazzo.

Digital Object Identifier 10.1109/OJSP.2025.3640517

RANF: Neural Field-Based HRTF Spatial Upsampling With Retrieval Augmentation and Parameter Efficient Fine-Tuning

YOSHIKI MASUYAMA [1] (Member, IEEE), GORDON WICHERN [1] (Member, IEEE), FRANÇOIS G. GERMAIN [1] (Member, IEEE), CHRISTOPHER ICK[1,2], AND JONATHAN LE ROUX [1] (Fellow, IEEE)

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139 USA
[2]Music and Audio Research Laboratory, New York University, Brooklyn, NY 10012 USA

CORRESPONDING AUTHOR: YOSHIKI MASUYAMA (email: masuyama@merl.com).

ABSTRACT This paper gives an in-depth description of our submission to Task 2 of the Listener Acoustic Personalization (LAP) challenge 2024, which aims to reconstruct head-related transfer functions (HRTFs) with dense spatial grids from sparse measurements. Neural fields (NFs) with parameter-efficient fine-tuning (PEFT) have led to dramatic performance improvements in HRTF spatial upsampling and personalization. Despite these advances, spatial upsampling performance remains limited in scenarios with very sparse measurements. Our proposed system, named retrieval-augmented NF (RANF), incorporates HRTFs retrieved from a dataset as auxiliary inputs. We leverage multiple retrievals via transform-average-concatenate and adopt a PEFT technique tailored for retrieval augmentation. Furthermore, we capitalize on the results of a signal-processing-based spatial upsampling method as optional inputs. By incorporating these auxiliary inputs, our system demonstrated state-of-the-art performance on the SONICOM dataset and placed first in Task 2 of the LAP challenge 2024.

INDEX TERMS Head-related transfer function, spatial audio, neural field, retrieval-augmented generation.

I. INTRODUCTION

Head-related transfer functions (HRTFs) describe the acoustic transfer functions from a sound source to both ears under anechoic conditions. HRTFs contain all essential binaural cues for sound source localization: interaural time differences (ITDs), interaural level differences (ILDs), and spectral coloration [1], [2], [3]. Hence, by convolving HRTFs with a dry source signal, we can make a subject perceive the sound as coming from a specific 3D position. Such immersive audio generation techniques have a wide range of applications such as telepresence systems [4], hearing aids [5], and mixed reality systems [6], [7].

HRTFs are unique to each subject as they result from the reflection and scattering of sound on the upper torso, head, and pinnae. It is thus preferable to use individual HRTFs for accurate perceptual localization [8], [9]. However, recording individual HRTFs is time-consuming [10]. To mitigate this problem, various HRTF spatial upsampling and personalization methods have been developed [11], [12], [13], [14], [15], [16].

A straightforward approach is to select suitable HRTFs for a target subject from existing datasets based on a chosen criterion, including some criteria that do not require subject-specific HRTF measurements [17], [18], [19], [20]. Meanwhile, signal-processing-based spatial upsampling has been widely used when HRTFs for the target subject can be recorded at multiple directions. The panning-based methods [11], [12], [13] predict the HRTF at the desired direction as a weighted sum of the measured HRTFs. In particular, vector base amplitude panning [11] was adopted by the ISO/IEC MPEG-H 3D Audio standard due to its efficiency [21]. Spatial-decomposition-based methods [14], [15], [16] encode the measurements into global spatial basis functions and decode them for a new direction. The performance of both these signal-processing-based methods deteriorates when the measurements are too few [22], [23].

© 2025 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

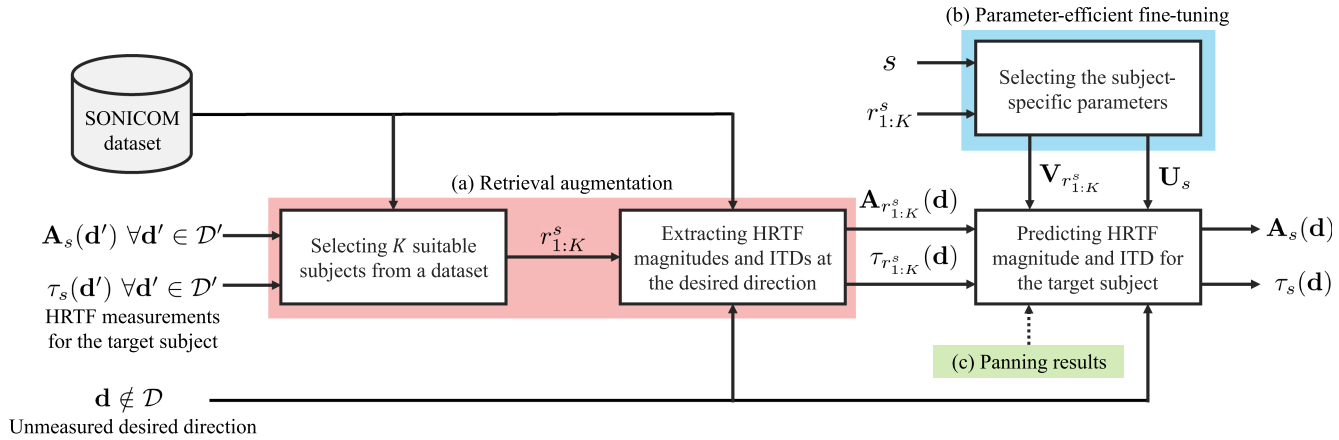

FIGURE 1. **Overall pipeline of RANF. As retrieval augmentation in (a), we select *K* subjects suitable for a target subject *s*, and their HRTF magnitude and ITD at the desired direction d are extracted. In (b), we select the subject-specific parameters based on the target subject *s* and the retrieved subjects $r^s_{1:K}$. Lastly, we predict the HRTF magnitude and ITD for the target subject *s* at the desired direction d by leveraging retrieved HRTFs. Depending on the sparsity level, the results of a panning-based method [12] can be optionally fed into the NF as in (c).**

Deep learning has significantly advanced HRTF spatial upsampling in a data-driven manner by leveraging datasets with dense spatial grids [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. In particular, neural fields (NFs), originally developed in computer vision [32], [33], have recently gained attention due to their grid-agnostic nature [23], [29], [30], [31]. In HRTF spatial upsampling, NFs are trained to model HRTFs as a function of the sound source direction. We can share an NF across multiple subjects except for a small number of subject-specific parameters and adapt the pre-trained NF to a new subject through parameter-efficient fine-tuning (PEFT) [23], [31]. This adaptation technique has made it easy to train an NF for a target subject, but HRTF spatial upsampling from fewer than ten measurements remains a challenging task.

In this paper, we propose a retrieval-augmented NF (RANF) that incorporates the results of HRTF selection into the NF as depicted in Fig. 1. RANF selects multiple subjects whose HRTFs are similar to those of the target subject and takes their HRTFs at the desired direction as input. We design a neural network that efficiently leverages multiple retrievals with a tailored PEFT strategy. Moreover, the results of a panning-based method can be used as an optional input to RANF, which takes the local characteristics of the HRTFs into account. We participated in Task 2 of the Listener Acoustic Personalization (LAP) challenge 2024 and achieved state-of-the-art results on the SONICOM dataset [34].

This paper is an extension of our conference paper [35], which introduced RANF with preliminary experiments. The first contribution of this paper is to provide an extended analysis of RANF on the challenge's evaluation set. The hidden evaluation set was released after our conference paper submission [35]. Our second contribution is the novel integration of the results of a panning method into RANF. These contributions will present a full picture of our winning solution for the LAP challenge. Training and inference scripts will be available at https://github.com/merlresearch/ranf-hrtf.

II. PRELIMINARIES

A. PROBLEM SETTING

An HRTF is a frequency-domain acoustic transfer function from a sound source at (θ, ϕ) to both ears, where $\theta \in [0, 2\pi)$ and $\phi \in [-\pi/2, \pi/2]$ are the azimuth and elevation, respectively. The azimuth increases counter-clockwise, $\theta = 0$ corresponding to the front of the subject, while the elevation increases upward, $\phi = 0$ indicating the equatorial plane. An HRTF depends not only on the sound source direction but also on anthropometric features, and thus it differs for each subject s. We model the HRTF for both ears as a function $\mathbf{H}_s(\mathbf{d}) \in \mathbb{C}^{F\times 2}$ of sound source direction $\mathbf{d} = (\theta, \phi)$, where F is the number of frequency bins. It can be characterized by ITD, ILD, and spectral coloration. ITD and ILD support azimuth localization [1], while elevation localization depends on spectral coloration [2]. We focus on modeling the HRTF magnitude $\mathbf{A}_s(\mathbf{d}) \in \mathbb{R}_+^{F\times 2}$, which captures ILD and spectral coloration, and ITD $\tau_s(\mathbf{d})$ in samples. Once the magnitude response and ITD are predicted, we can compute complex spectra using minimum-phase reconstruction [36] and compensate for the ITD in the time domain after applying an inverse discrete Fourier transform (IDFT).

With the above notation, Task 2 of the LAP challenge [37], [38] corresponds to predicting $\mathbf{A}_{\bar{s}}(\mathbf{d})$ and $\tau_{\bar{s}}(\mathbf{d})$ for each target subject $\bar{s}$ at all directions $\mathbf{d} \in \mathcal{D}$. Here, $\mathcal{D}$ is the complete set of the sound source directions for the SONICOM dataset [34], with $|\mathcal{D}| = 793$. For a given target subject $\bar{s}$, a limited number of measurements at $\mathbf{d}' \in \mathcal{D}'$ are given, where $\mathcal{D}' \subset \mathcal{D}$. In the challenge, $|\mathcal{D}'|$ varies from 3 to 100. We are provided with $\mathbf{A}_s(\mathbf{d})$ and $\tau_s(\mathbf{d})$ at all $\mathbf{d} \in \mathcal{D}$ for 200 subjects in the SONICOM dataset, which we can use for pre-training of NFs. The set of these pre-training subjects is denoted as $\mathcal{S}$, and target subjects are such that $\bar{s} \notin \mathcal{S}$.

B. HRTF SELECTION

An efficient method for providing suitable HRTFs is HRTF selection, which retrieves the best-fitting HRTFs from a dataset

with dense spatial grids. Early studies require subjective evaluations to identify the best-fitting subject [17], [18]. More recent studies aim to automatically select the appropriate subject based on the similarity of anthropometric features [19], [20] or on HRTFs at a few directions [26]. These methods have moderately improved sound source localization performance over averaging non-individual HRTFs [20]. Inspired by these findings, our system exploits the results of HRTF selection as auxiliary inputs, where we can leverage the best one or more suitable subjects.

C. HRTF SPATIAL UPSAMPLING VIA PANNING

While HRTF selection retrieves suitable HRTFs from a dataset, panning-based methods aim to upsample sparsely measured HRTFs [11], [12], [13]. These methods perform a weighted sum of the HRTFs at the surrounding positions, relying only on the HRTFs of the target subject. In the LAP challenge, the barycentric interpolation [13] was adopted as one of the baselines.

We use a variant of vector base amplitude panning [12] as our baseline. Let $\mathbf{q} \in \mathbb{R}^3$ be the sound source position in the Cartesian coordinates, and $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{|\mathcal{D}'|}]$ be a matrix of the measured positions. This method determines the weight $\mathbf{g} \in \mathbf{R}^{|\mathcal{D}'|}$ for a desired position $\mathbf{p} \in \mathbb{R}^3$ by solving the following problem:

$$\min_{\mathbf{g}} \ \|\mathbf{g}\|_1 \ \text{ s.t. } \ \mathbf{Q}\mathbf{g} = \mathbf{p}, \ \ \mathbf{g} \geq \mathbf{0}, \tag{1}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm, and $\mathbf{0}$ is the zero vector. After optimizing the weight, we predict the HRTF magnitude and ITD by using the corresponding weighted sum with $\mathbf{g}$. Note that $\mathbf{g}$ depends only on the measured and desired directions regardless of the subjects, and thus we only need to solve (1) for the relevant directions once irrespective of the number of subjects. Although this method shows promising performance when we have a substantial number of measurements (e.g., 100 measurements), it is not applicable to highly sparse scenarios as the optimization problem in (1) could be infeasible [12].

D. HRTF SPATIAL UPSAMPLING VIA NEURAL FIELDS

NFs have originally been developed to synthesize novel views of a 3D scene from multiple 2D observations [32], [33], and they have been extended to spatial audio applications [23], [29], [30], [31], [39], [40], [41]. In HRTF spatial upsampling, HRTF field was first proposed to represent the magnitude response $\mathbf{A}_s(\mathbf{d})$ as follows [31]:

$$\mathbf{A}_s(\mathbf{d}) = \mathtt{NF}(\mathbf{d} \mid \mathbf{\Xi}_s), \tag{2}$$

where $\mathbf{\Xi}_s$ denotes model parameters, and the NF is trained separately for each subject s. Meanwhile, HRTFs are similar across different subjects, and a single generic NF can represent HRTFs of multiple subjects with a small number of subject-specific parameters [23], [31]:

$$\mathbf{A}_s(\mathbf{d}) = \mathtt{NF}(\mathbf{d} \mid \mathbf{\Gamma}, \mathbf{\Xi}_s), \tag{3}$$

where $\mathbf{\Gamma}$ denotes the generic parameters, and the subject-specific parameters $\mathbf{\Xi}_s$ steer the generic model to represent the HRTF of a specific subject s. The NF in (3) is first pretrained on a set of subjects $\mathcal{S}$ by optimizing $\mathbf{\Gamma}$ and $\mathbf{\Xi}_s$ for all $s \in \mathcal{S}$. Then, it is adapted to a target subject $\bar{s}$ by optimizing the subject-specific parameters $\mathbf{\Xi}_{\bar{s}}$ on the available measurements, while freezing the generic parameters $\mathbf{\Gamma}$. The size of $\mathbf{\Xi}_s$ in (3) is typically much smaller than that in (2), and thus the adaptation is more efficient than training a new NF in (2) from scratch.

Various designs of the subject-specific parameters have been explored [23], [31]. In HRTF field [31], a subject-specific latent vector is concatenated to the model input as conditioning-by-concatenation (CbC) [33]. Meanwhile, our previous study explored PEFT including low-rank adaptation (LoRA) [42] and demonstrated its efficacy in HRTF personalization [23]. LoRA adjusts the weight of the lth layer $\mathbf{W}_l \in \mathbb{R}^{N_l \times M_l}$ for each subject s as follows:

$$\mathbf{W}_{l,s} = \mathbf{W}_l + \mathbf{u}_{l,s}\mathbf{v}_{l,s}^{\mathsf{T}}, \tag{4}$$

where $(\cdot)^{\mathsf{T}}$ denotes the transpose operation, and $\mathbf{u}_{l,s} \in \mathbb{R}^{N_l}$ and $\mathbf{v}_{l,s} \in \mathbb{R}^{M_l}$ are subject-specific vectors for constructing a rank-1 matrix.[1] During the adaptation, we freeze $\mathbf{W}_l$ and optimize only $\mathbf{u}_{l,s}$ and $\mathbf{v}_{l,s}$.

Existing NF-based HRTF spatial upsampling has focused on approximating the HRTF magnitude $\mathbf{A}_s(\mathbf{d})$ [23], [31]. However, ITD $\tau_s(\mathbf{d})$ is also essential to obtain the appropriate time-domain filters. We thus modify our NF to predict both HRTF magnitude and ITD in a multi-task fashion:

$$\mathbf{A}_s(\mathbf{d}), \tau_s(\mathbf{d}) = \mathtt{NF}(\mathbf{d} \mid \mathbf{\Gamma}, \mathbf{U}_s, \mathbf{V}_s), \tag{5}$$

where the subject-specific parameters $\mathbf{\Xi}_s$ in (3) are replaced by $\mathbf{U}_s = [\mathbf{u}_{1,s}, \ldots, \mathbf{u}_{L,s}]$ and $\mathbf{V}_s = [\mathbf{v}_{1,s}, \ldots, \mathbf{v}_{L,s}]$ for LoRA, and L is the number of layers. The predicted HRTF magnitudes are converted to complex spectra using minimum-phase reconstruction. Then, we apply IDFT to the complex spectra and shift the time-domain filters to match the predicted ITD.

III. RETRIEVAL-AUGMENTED NEURAL FIELD

We propose to incorporate the results of HRTF selection into the NF-based HRTF spatial upsampling motivated by the recent success of retrieval-augmented generation (RAG). RAG [43] has shown remarkable success in various modalities [43], [44], [45], [46], [47], [48] by selectively exploiting relevant information from external resources as the context for generation. We interpret the selected HRTFs as the context for generating HRTFs of the target subject and feed them into an NF.

[1] We show the rank-1 case of LoRA in (4) for simplicity, but LoRA can use higher ranks in general.

A. OVERVIEW OF RANF

An overview of RANF is depicted in Fig. 1. It comprises three stages: retrieval, selection of subject-specific parameters, and prediction of HRTF magnitudes and ITD. As shown in block (a) of Fig. 1, we first select K subjects $r_k^s \in \mathcal{S}$ based on HRTF similarity, where $k = 1, \ldots, K$ is the index of the retrieved subjects. To avoid clutter, we omit the superscript s in r_k^s hereafter. Then, subject-specific parameters are selected in block (b) of Fig. 1. According to the retrieval, we replace the $\mathbf{V}_s$ part of the subject-specific parameters in (5) with $\mathbf{V}_{r_{1:K}} = \{\mathbf{V}_{r_1}, \ldots, \mathbf{V}_{r_k}\}$ that rely on the retrieved subjects r_k. The details of LoRA for RANF will be explained in Section III-D. Lastly, HRTF magnitude and ITD for the target subject at the desired direction are predicted as follows:

$$\mathbf{A}_s(\mathbf{d}), \tau_s(\mathbf{d}) = \mathtt{NF}(\mathbf{d}, \mathbf{A}_{r_{1:K}}(\mathbf{d}), \tau_{r_{1:K}}(\mathbf{d}) \mid \mathbf{\Gamma}, \mathbf{U}_s, \mathbf{V}_{r_{1:K}}). \quad (6)$$

The main difference from (5) is the utilization of the retrieved HRTF magnitudes and ITDs. Since they are supposed to be close to the target HRTF magnitude and ITD, the retrieval augmentation should ease the HRTF prediction.

The NF is first pre-trained on the SONICOM dataset where both generic parameters $\mathbf{\Gamma}$ and subject-specific parameters $(\mathbf{U}_s, \mathbf{V}_{r_{1:K}})$ are jointly optimized. During this pre-training, r_k is retrieved from $\mathcal{S}$ for each $s \in \mathcal{S}$, and HRTFs at all the sound source directions $\mathbf{d} \in \mathcal{D}$ are available. The parameters are optimized to minimize the sum of the log-spectral distortion (LSD) on the HRTF magnitudes and the ϵ-insensitive mean absolute error (MAE) on the ITDs:

$$\mathcal{L}_{\text{pre-train}} = \sum_{s \in \mathcal{S}} \sum_{\mathbf{d} \in \mathcal{D}} \mathcal{L}(\mathbf{H}_s^\star(\mathbf{d}), \mathbf{H}_s(\mathbf{d})), \quad (7)$$

$$\mathcal{L}(\mathbf{H}^\star(\mathbf{d}), \mathbf{H}(\mathbf{d})) = \mathtt{LSD}(\mathbf{A}^\star(\mathbf{d}), \mathbf{A}(\mathbf{d})) + \lambda \mathtt{MAE}_\epsilon(\tau^\star(\mathbf{d}), \tau(\mathbf{d})), \quad (8)$$

where $\lambda \in \mathbb{R}_+$ balances the two terms, and $(\cdot)^\star$ denotes the oracle value. The LSD is calculated as

$$\mathtt{LSD}(\mathbf{A}^\star(\mathbf{d}), \mathbf{A}(\mathbf{d})) = \frac{1}{2} \sum_{c=1}^{2} \sqrt{\frac{1}{F} \sum\nolimits_{f=1}^{F} \left(20 \log_{10} \frac{A_{c,f}(\mathbf{d})}{A_{c,f}^\star(\mathbf{d})}\right)^2}, \quad (9)$$

where $c \in \{1, 2\}$ and $f = 1, \ldots, F$ are the channel (left and right ears) and frequency indices, respectively. On the other hand, the loss function for ITDs is formulated as

$$\mathtt{MAE}_\epsilon(\tau^\star(\mathbf{d}), \tau(\mathbf{d})) = \mathrm{Max}(|\tau^\star(\mathbf{d}) - \tau(\mathbf{d})|, \epsilon). \quad (10)$$

The oracle ITD $\tau^\star(\mathbf{d})$ is obtained as the integer time sample that maximizes the cross correlation between channels. We set ϵ in (10) to 0.5 to accommodate the quantization in the oracle ITD computation.

The pre-trained NF is then adapted to each target subject $\bar{s} \notin \mathcal{S}$ by optimizing only $\mathbf{U}_{\bar{s}}$ to minimize the loss function over the measured directions $\mathbf{d}' \in \mathcal{D}'$:

$$\mathcal{L}_{\text{adaptation}} = \sum_{\mathbf{d}' \in \mathcal{D}'} \mathcal{L}(\mathbf{H}_{\bar{s}}^\star(\mathbf{d}'), \mathbf{H}_{\bar{s}}(\mathbf{d}')), \quad (11)$$

where we freeze the generic parameters $\mathbf{\Gamma}$ and the parameters relying on the retrieved subjects $\mathbf{V}_{r_{1:K}}$.

B. RETRIEVAL AUGMENTATION FOR HRTF SPATIAL UPSAMPLING

Existing retrieval augmented methods typically retrieve the relevant information by k-nearest neighbors in latent domains [45], [46], [47]. Meanwhile, our RANF retrieves the K suitable subjects based on the similarity of HRTFs at the measured directions as follows:

$$r_1, \ldots, r_K \leftarrow \underset{r_{1:K} \in [\mathcal{S}]^K}{\mathrm{argmin}} \sum_{k=1}^{K} \sum_{\mathbf{d}' \in \mathcal{D}'} \mathcal{L}_{\text{retrieval}}(\mathbf{H}_s(\mathbf{d}'), \mathbf{H}_{r_k}(\mathbf{d}')), \quad (12)$$

where $[\mathcal{S}]^K$ denotes the set of all subsets of $\mathcal{S}$ with K distinct elements (if $s \in \mathcal{S}$, which can be the case during training, we restrict these subsets to elements different from s). While any criteria can be used for $\mathcal{L}_{\text{retrieval}}$, we choose the MAE on ITDs based on our preliminary experiment [35]. When multiple subjects result in the same MAE, we select the subjects with lower LSD. Alternatively, we can stochastically sample K subjects from the best-fitting K' chosen as in (12) with $K' > K$. This procedure augments the training data by adding randomness to the pairs of the target and retrieved subjects at each epoch. Then, we extract the HRTF magnitudes and ITDs at the desired direction $\mathbf{d}$ for each retrieved subject r_k. We expect that the HRTFs of the retrieved subjects will reasonably approximate the HRTFs of the target subject even at unmeasured desired directions.

Our retrieval process assumes that the HRTF dataset for retrieval covers all the measured directions in $\mathcal{D}'$ and the desired directions in $\mathcal{D}$. For the LAP challenge, this assumption holds as long as the SONICOM dataset is used for retrieval. However, this may not always be the case, e.g., when retrieving HRTFs from a separate dataset with a different spatial grid. That assumption also compromises the fully grid-agnostic nature of our system. To alleviate these limitations, a straightforward workaround is to interpolate HRTFs for retrieval using an existing method.

C. PROPOSED NETWORK ARCHITECTURE

In RANF, the neural network needs to handle multiple retrieved HRTF magnitudes and ITDs. We thus design a novel network architecture as illustrated in Fig. 2, departing from prior methods which use vanilla fully-connected (FC) layers [23], [31]. Precisely, for a given direction $\mathbf{d} = (\theta, \phi)$, we perform sequence modeling on the corresponding retrieved HRTF magnitudes along the frequency direction by using convolution/deconvolution blocks and bidirectional long short-term memory (BLSTM) networks. We downsample the K retrieved magnitudes separately by using a convolution block, where the input of the initial convolution layer has length F and two channels. The K magnitude encodings have length $F' \leq F$ and C channels. Meanwhile, we calibrate the azimuth value into θ_s and θ_{r_k} for each of the target subject

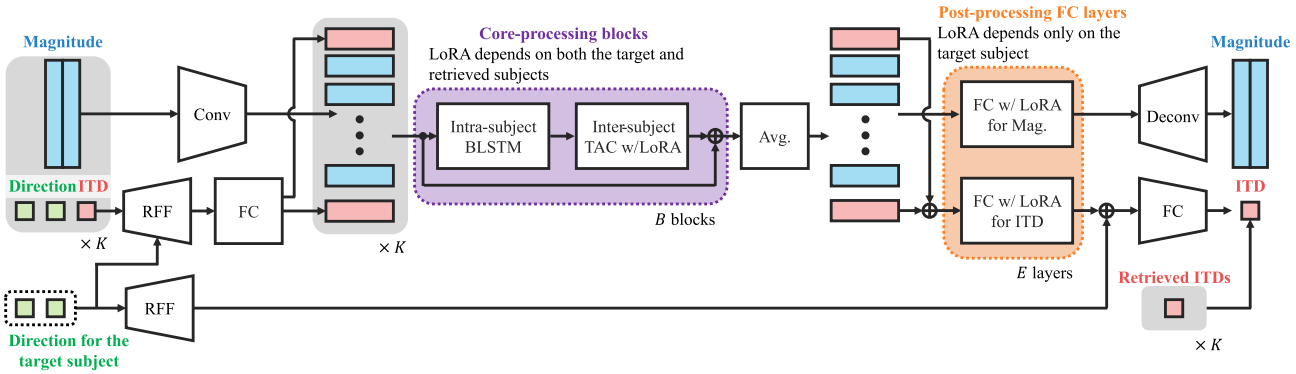

FIGURE 2. **Proposed network architecture. The features in gray boxes are for each of the K retrieved subjects and are processed in parallel except for the inter-subject TAC module. The retrieved ITD indicated by a red square $\tau_{r_k}(\mathbf{d})$ is encoded as RFFs with the sound source directions shown as green squares. Avg. aggregates the K feature sequences, and Conv and Deconv denote convolution and deconvolution blocks, respectively.**

s and retrieved subjects r_k so that their respective ITD at azimuth zero (after calibration) is zero (please refer to the Appendix for details). Then, the corresponding retrieved ITDs for each k are concatenated with these calibrated sound source directions before being transformed into random Fourier features (RFFs) [49]:

$$\tilde{\boldsymbol{\mu}}_k = [\theta_s - \pi, \phi_s, \theta_{r_k} - \pi, \phi_{r_k}, \tau_{r_k}(\mathbf{d})/\tau_{\max}]^\mathsf{T}, \tag{13a}$$

$$\boldsymbol{\mu}_k = [\sin(\tilde{\boldsymbol{\mu}}_k)^\mathsf{T}, \cos(\tilde{\boldsymbol{\mu}}_k)^\mathsf{T}]^\mathsf{T}, \tag{13b}$$

$$\mathbf{e}_k = [\sin(\mathbf{W}_{\mathrm{RFF}}\boldsymbol{\mu}_k)^\mathsf{T}, \cos(\mathbf{W}_{\mathrm{RFF}}\boldsymbol{\mu}_k)^\mathsf{T}]^\mathsf{T}, \tag{13c}$$

where $\tau_{\max}$ is a subject-independent parameter to rescale the retrieved ITDs to be in $[-\pi, \pi)$, the entries of $\mathbf{W}_{\mathrm{RFF}} \in \mathbb{R}^{C\times 10}$ are sampled from a Gaussian distribution, and $\sin(\cdot)$ and $\cos(\cdot)$ are applied entry-wise. The RFF $\mathbf{e}_k$ is converted to two C-dimensional features by an FC layer, and the features are concatenated at both ends of the magnitude encodings. We obtain K feature sequences of length $F'+2$ where each feature has dimension C.

The extracted feature sequences are processed by B stacked core-processing blocks, resulting in a single feature sequence for the target subject. Each core-processing block contains an intra-subject BLSTM module and an inter-subject transform-average-concatenate (TAC) module [50] with a residual connection. The intra-subject BLSTM module handles each of the K feature sequences separately and performs sequence modeling along the feature dimension of length $F'+2$. Meanwhile, the inter-subject TAC module mixes the feature vectors across different retrieved subjects, where each of the $F'+2$ feature dimensions is treated separately. Both modules preserve the channel dimension C in their output. The K output sequences of the final core-processing block are averaged into a single feature sequence.

The feature sequence, except for both ends, is passed through E additional FC layers equipped with LoRA that separately process each of the F' entries. This module focuses on the features corresponding to the HRTF magnitude of the target subject, and both components of LoRA depend on the target subject as in (5). Lastly, the output sequence is decoded to the two-channel HRTF magnitude by a deconvolution block. Meanwhile, the features at both ends are passed to a separate set of E FC layers with LoRA. Then, the ITD feature is summed up with an additional RFF computed as in (13) while replacing $\tilde{\boldsymbol{\mu}}_k$ with $[\theta_s - \pi, \phi_s]^\mathsf{T}$. This feature is fed into the subsequent ITD prediction head to lead the final ITD prediction as follows:

$$\tau_s(\mathbf{d}) = \tau_s^{\Delta}(\mathbf{d}) + \frac{1}{K}\sum_{k=1}^{K}\tau_{r_k}(\mathbf{d}), \tag{14}$$

where $\tau_s^{\Delta}(\mathbf{d})$ is the output of the ITD prediction head.

D. INTER-SUBJECT MODELING WITH LORA

For the inter-subject TAC module, we treat each of the $F'+2$ feature vectors in the sequence separately. This module mixes the information across different retrieved subjects and performs subject-specific processing with LoRA. Let $\mathbf{x}_{b,f,r_k} \in \mathbb{R}^C$ be the input feature vector for the kth retrieved subject in the bth block at feature index f. All the features for $k = 1, \ldots, K$ are processed by an FC layer and then averaged over the retrieved subjects:

$$\bar{\mathbf{x}}_{b,f,r_k} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{FC}_b^{\mathrm{Avg}}(\mathbf{x}_{b,f,r_k}), \tag{15}$$

where the output dimension of $\mathrm{FC}_b^{\mathrm{Avg}}(\cdot)$ is $C/2$. The averaged feature is concatenated with the output of another FC layer:

$$\mathbf{z}_{b,f,r_k} = \left[\mathrm{FC}_b^{\mathrm{Pass}}\left(\mathbf{x}_{b,f,r_k}\right)^\mathsf{T}, \bar{\mathbf{x}}_{b,f,r_k}^\mathsf{T}\right]^\mathsf{T}, \tag{16}$$

where the output dimension of $\mathrm{FC}_b^{\mathrm{Pass}}(\cdot)$ is also $C/2$. The TAC operation was originally proposed in array signal processing to handle a variable number of multi-channel features in a permutation invariant manner [50]. We utilize the TAC operation to mix the features for multiple retrieved subjects. By construction, it can scale the model to an arbitrary number of retrieved subjects, which is advantageous for variable computational resources.

The concatenated features $\mathbf{z}_{b,f,r_k}$ are then passed to an FC layer with LoRA:

$$\mathbf{y}_{b,f,r_k} = \mathrm{LN}_b\left(\mathrm{FC}^{\mathrm{LoRA}}_{b,s,r_k}\left(\mathbf{z}_{b,f,r_k}\right)\right), \tag{17}$$

where $\mathrm{LN}(\cdot)$ is the layer normalization, and LoRA in the FC layer differs from (4) as follows:

$$\mathbf{W}_{b,s,r_k} = \mathbf{W}_b + \mathbf{u}_{b,s}\mathbf{v}^{\mathsf{T}}_{b,r_k}. \tag{18}$$

By selecting $\mathbf{v}_{b,r_k}$ based on the retrieved subject, we expect that $\mathbf{u}_{b,s}\mathbf{v}^{\mathsf{T}}_{b,r_k}$ captures the relation between the target and retrieved subjects. Since r_k is always retrieved from $\mathcal{S}$, $\mathbf{v}_{b,r_k}$ can be pre-trained and frozen during the adaptation.

E. INTEGRATION WITH PANNING-BASED METHOD

In the low-sparsity setting with 100 measurements, the panning method in (1) outperformed RANF in terms of ILD and LSD. To build a system capable of leveraging the respective strengths of both RANF and the panning, we propose to additionally feed these results into the NF as shown in block (c) of Fig. 1. Regarding the HRTF magnitudes, we concatenate the predicted magnitudes for the target subject $\hat{\mathbf{A}}^{\mathrm{Pan}}_s(\mathbf{d})$ and the retrieved subject $\hat{\mathbf{A}}^{\mathrm{Pan}}_{r_k}(\mathbf{d})$ to the original retrieved magnitude $\mathbf{A}_{r_k}(\mathbf{d})$. Here, $\hat{\mathbf{A}}^{\mathrm{Pan}}_{r_k}(\mathbf{d})$ is computed from the measurements only at the directions in $\mathcal{D}'$ similar to the panning for the target subject. Accordingly, the input channel dimension of the first convolution layer is now increased to a total of 6 instead of the original 2. Similarly, the predicted ITDs $\hat{\tau}^{\mathrm{Pan}}_s(\mathbf{d})$ and $\hat{\tau}^{\mathrm{Pan}}_{r_k}(\mathbf{d})$ are concatenated with $\tilde{\boldsymbol{\mu}}_k$ in (13) after division by $\tau_{\max}$. The neural network can learn the relation between the target and retrieved subjects more easily by incorporating the panning results for both the target and retrieved subjects as additional guidance. The additional complexity caused by this integration is minimal because it affects only the initial convolution layer and the calculation of the RFFs in (13). We refer to this integration of RANF and panning as RANF+ in our experiment.

Altogether, we expect that the retrieval augmentation and panning improve HRTF spatial upsampling by explicitly providing complementary information. The retrieval delivers HRTFs at the desired direction from similar subjects as approximations. In contrast, the panning provides an additional plausible approximation as a weighted sum of the HRTFs of the target subject at the surrounding positions. Note that our integrated method, RANF+, can accomodate any HRTF spatial interpolation methods, e.g., barycentric interpolation [13] or spatial-decomposition-based methods [14], [15], [16]. More advanced interpolation techniques might improve the overall performance of RANF+, which we defer to future work.

IV. EXPERIMENTAL VALIDATION

A. EXPERIMENTAL SETUP

The LAP Challenge Task 2 asked participants to upsample HRTFs at four sparsity levels:

- Low sparsity level: 100 measurements per subject;
- Mid sparsity level: 19 measurements per subject;
- High sparsity level: 5 measurements per subject clustered around the front;
- Very high sparsity level: 3 measurements (front, left, and above) per subject.

At each sparsity level, evaluation was performed on 3 subjects. For pre-training and retrieval, HRTFs for 200 subjects in the SONICOM dataset [34] were provided, with 793 measurements provided for each subject. We used HRTFs sampled at 48 kHz with a free-field compensation using a minimum-phase filter. The time-domain impulse responses were converted to the frequency domain with the discrete Fourier transform with 256 points. ITDs were extracted by the `Spatial Audio Metrics` toolbox.[2]

For the pre-training of RANF, we split the SONICOM dataset into training and validation sets at each sparsity level. We first computed MAE on ITDs between the 3 evaluation subjects and the 200 SONICOM subjects at the measured directions. Based on the MAEs, we selected the top-5 subjects for each evaluation subject and assigned them to the training set. We wanted to make sure that the generic model is pre-trained on the subjects most similar to the target subjects. The next 3 most similar subjects were assigned to the validation set. The remaining subjects were also used for the training set,[3] excluding subject `P0079` due to its atypical ITD behavior. Finally, the training and validation sets for pre-training consisted of 190 and 9 subjects, respectively.

In our neural network shown in Fig. 2, the convolution and deconvolution blocks consisted of 4 layers with PReLU activation, where F' and C were 16 and 128, respectively. For the core-processing blocks, we stacked $B = 4$ blocks and used the GELU activation for the FC layers in the inter-subject TAC module. The number of post-processing FC layers E was set to 2, and the final ITD prediction head also consisted of 2 layers. The rank for LoRA was set to 1.

For training, we set λ to $10^6/48000$ in (8) to convert the loss scale from sample to microsecond. The neural network was pre-trained up to 200 epochs with early stopping if the best validation loss did not improve for 20 successive epochs. We used the RAdam optimizer with initial learning rate of 0.001, multiplied by 0.9 when the validation loss did not improve for 10 epochs. Then, the checkpoint with the best validation loss was adapted to the evaluation subjects for 1000 epochs. We did not perform validation during adaptation to exploit all measurements.

We compared RANF to four learning-free baselines and two NF-based methods.

HRTF Selection (ITD): We selected the best-fitting subject from the SONICOM dataset, in terms of the MAE on ITDs at the measured directions.

[2][Online]. Available: https://github.com/Katarina-Poole/Spatial-Audio-Metrics

[3]Note that this setup is not fully identical to the challenge setup since we performed additional cross-validation and ensemble during the challenge.

TABLE 1. Average ITD Error, ILD Error, and LSD on the LAP Challenge Evaluation Set

	3 measurements			5 measurements			19 measurements			100 measurements		
Methods	ITD [μs]	ILD [dB]	LSD [dB]	ITD [μs]	ILD [dB]	LSD [dB]	ITD [μs]	ILD [dB]	LSD [dB]	ITD [μs]	ILD [dB]	LSD [dB]
HRTF selection (ITD)	26.5	1.7	6.5	20.6	1.1	5.6	16.4	1.2	6.5	12.6	0.9	5.4
HRTF selection (LSD)	49.9	1.5	6.3	26.8	1.0	6.1	35.2	1.2	5.2	27.6	0.9	4.9
Nearest neighbor	298.5	7.7	8.6	150.3	4.6	7.8	102.8	2.8	5.2	38.7	1.2	3.0
Panning [12]	×	×	×	×	×	×	29.0	1.7	3.9	12.5	0.4	2.3
NF (CbC) [31]	21.1	1.3	5.3	15.6	1.5	4.7	19.8	1.9	5.0	7.2	1.0	4.2
NF (LoRA) [23]	25.7	1.0	4.8	18.6	**1.0**	4.6	18.1	1.2	3.8	6.1	0.7	3.3
RANF (proposed)	**18.6**	**0.9**	**4.6**	**15.3**	1.1	**4.4**	15.2	0.9	3.3	6.8	0.6	2.5
RANF+ (proposed)	×	×	×	×	×	×	**14.4**	**0.9**	**3.3**	**5.8**	**0.4**	**2.2**

HRTF Selection (LSD): We selected the subject with the minimum LSD at the measured directions, instead of the MAE on ITDs.

Nearest neighbor: We selected as interpolated HRTF the HRTF of the target subject at the nearest position in terms of the Euclidean distance.

Panning: We implemented the panning method in (1) with `CVXPY` [51]. The optimized nonnegative weight $\mathbf{g}$ was directly used for ITD interpolation. Meanwhile, we further normalized the weight by dividing it by its ℓ_1 norm when interpolating HRTF magnitudes [12]. This normalization slightly improved the performance of the magnitude interpolation in our preliminary experiment.

NF (CbC): This is an NF without retrieval, similar to (5) except for the conditioning by a single subject-specific latent vector instead of LoRA [31]. The dimension of the latent vector was 32 following the original paper. The NF consists of 4 hidden FC layers and the heads for predicting HRTF magnitudes and ITD.

NF (LoRA): This is the method formulated in (5). In contrast to RANF, both components of LoRA, $\mathbf{U}_s$ and $\mathbf{V}_s$, depend on the target subject s [23].

RANF: For the proposed RANF, 10 subjects were retrieved from the SONICOM dataset, i.e., $K = 10$. During the training, we first selected 30 subjects, i.e., $K' = 30$, and then randomly sampled 10 subjects as described in Section III-B. At evaluation time, we retrieved the 10 best-fitting subjects.

RANF+: As elaborated in Section III-E, we fed the results of the panning method [12] to RANF. This method was evaluated at only low and mid sparsity levels due to the limited applicability of the panning method.

For evaluation, we used the following three metrics as in the challenge: LSD, MAE on ILDs, and MAE on ITDs without ϵ in (10). ILDs were calculated as the difference between the root mean squares of HRTFs for the left and right channels. The LSD was calculated between 20 Hz and 20 kHz by limiting the range of f in (9). All metrics were calculated using the `Spatial Audio Metrics` toolbox.

B. MAIN RESULTS

Table 1 shows the scores averaged over the three evaluation subjects at each sparsity level. Comparing the learning-free baseline methods, HRTF selection methods resulted in moderate performance even under highly-sparse conditions. Panning is feasible only for the low and mid sparsity levels, but it demonstrated promising performance on the low sparsity level, especially in terms of ILD and LSD.

Both existing NF-based methods improved the performance from the HRTF selections and the nearest neighbor algorithm. NF (LoRA), which includes more subject-specific parameters, achieved a better LSD but increased the ITD error in highly sparse scenarios, likely because LoRA overfits to the limited number of measurements. By incorporating the retrievals, RANF outperformed both existing methods except for the ILD with 5 measurements. Although RANF resulted in a worse LSD than the panning at the low sparsity level (100 measurements), RANF+ performed the best in all metrics by incorporating the panning results as additional inputs. Regarding ILD and LSD, RANF+ performed equally to or slightly better than the best of RANF and the panning. The improvement was more notable for the ITD under both conditions, while the panning itself performed worse than RANF. This result indicates that the complementary combination of retrieval and panning is more beneficial for ITD prediction.

C. QUALITATIVE ANALYSIS

Figure 3 shows examples of the upsampled HRTF magnitude and ITD under the sparsest condition. While the ITD was accurately predicted regardless of the retrieval, RANF demonstrated better magnitude prediction, especially under 15 kHz. Figure 4 illustrates the HRTF magnitudes over the median plane, since the elevation localization mainly relies on spectral coloration. Here, the elevation from -45° to 90° corresponds to the front, and from 90° to 225° indicates the back. Compared with NF (CbC) and NF (LoRA), RANF shows more distinctive spectral notches. In addition, RANF more precisely predicts the smooth spectral peak around 4 kHz compared with HRTF selection (ITD). This result indicates that RANF successfully aggregates the information from multiple retrievals and refines the retrieved HRTF magnitudes. Although the notch frequencies are not completely aligned with the original ones, RANF shows the potential for more accurate modeling.

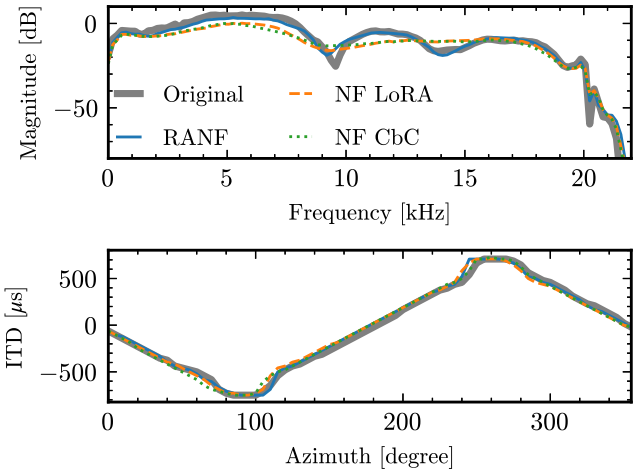

FIGURE 3. Examples of the HRTF magnitude and ITD for subject P0208 upsampled from 3 measurements. The top panel shows the ipsilateral HRTF magnitudes at the right direction. The bottom panel depicts the ITD on the horizontal plane.

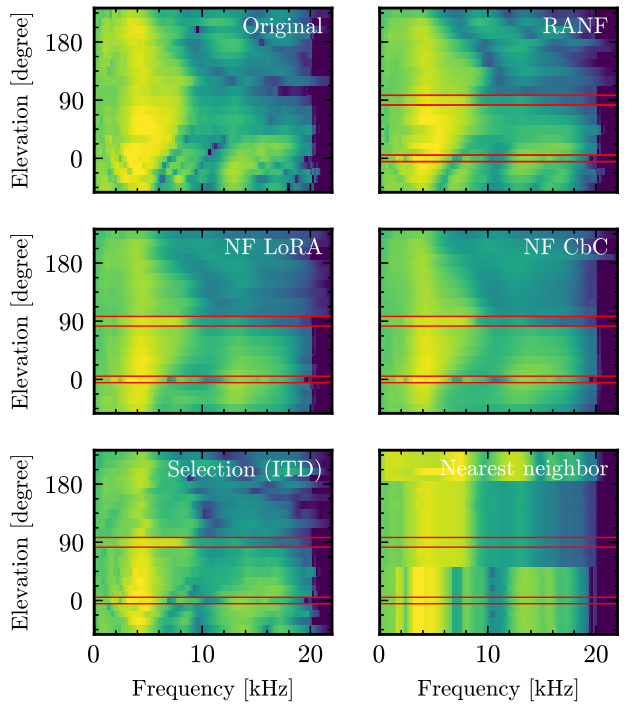

FIGURE 4. Examples of HRTF magnitude over the median plane for subject P0208 upsampled from 3 measurements. The measured directions are surrounded by red frames.

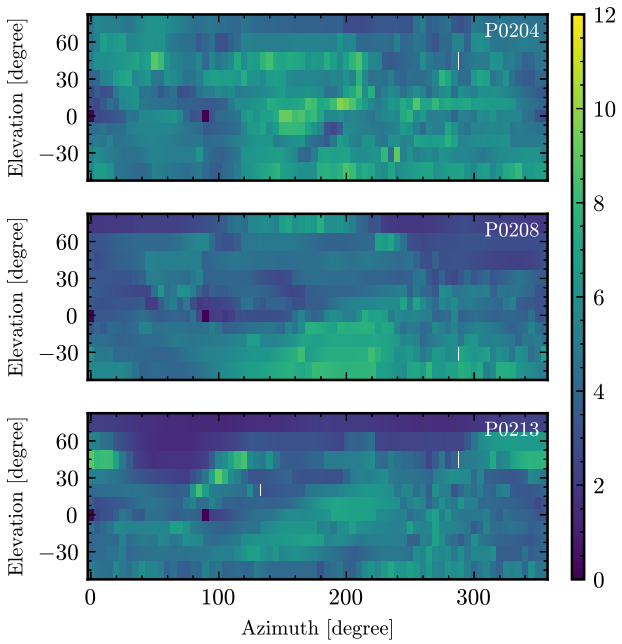

FIGURE 5. Subject-wise LSD across directions for the left channel, where the HRTFs are upsampled from 3 measurements by RANF.

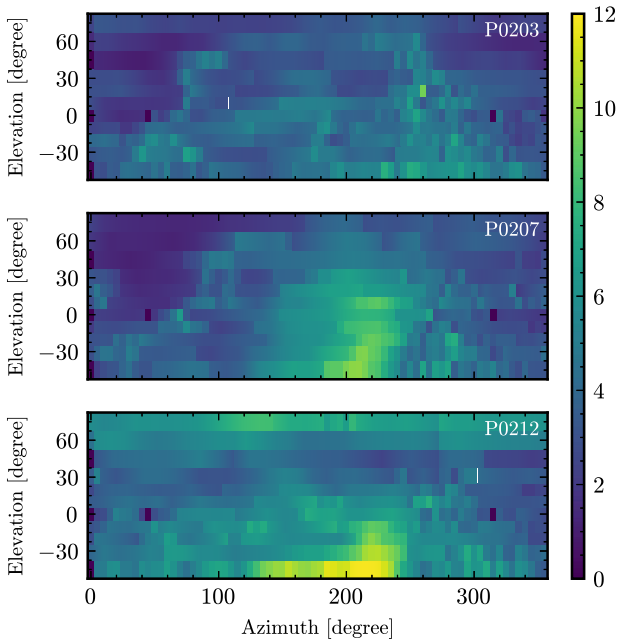

FIGURE 6. Subject-wise LSD across directions for the left channel, where the HRTFs are upsampled from 5 measurements by RANF.

Figures 5 and 6 illustrate the LSD for RANF across directions when upsampling from 3 and 5 measurements, respectively. The measured directions are plotted with zero error, and the zenith is omitted for a simpler visualization. In the sparsest case with 3 measurements, the LSD pattern differs across subjects. With 5 measurements, higher LSD tends to concentrate in the lower rear direction, where subject P0212 resulted in significantly higher error than others. This should be because the lower rear direction is far from the five measured directions. These results indicate that more measurements reduce the variance in the LSD patterns within subjects, but RANF might result in high LSD at the directions far from the measured ones.

V. CONCLUSION

In this paper, we described our submission to Task 2 of the LAP Challenge 2024. We enhanced NF-based HRTF spatial upsampling by incorporating retrieval augmentation and the results of the panning method. In our experiments, our

proposed method RANF outperformed NF-based methods without retrieval in terms of MAE of ITDs and LSDs. Moreover, the integration with the results of the panning method further improved all the metrics, especially under the low-sparsity condition.

The experiments in this paper used only the SONICOM dataset as in the challenge. NFs pre-trained on a single dataset might not generalize well to other datasets due to the domain mismatch caused by the difference in measurement setup. Developing RANF on multiple HRTF datasets might improve the generalization capability of RANF, but it is non-trivial. This is because our current retrieval strategy assumes that the HRTFs at both the desired and measured directions are available for retrieval subjects. Moving forward, we aim to explore grid-agnostic retrieval and investigate the impact of the direction difference between the target and the retrieved HRTFs. In addition, we plan to explore retrieval strategies based on perceptual criteria.

APPENDIX

ITDs for the front direction should be zero under ideal recording setups, assuming ears are symmetrically equipped on the head. We find however that ITDs computed from the SONICOM dataset contain small offsets for each subject. We compensate this subject-wise offset with a linear model. In detail, we choose two measured directions on the horizontal plane, $\mathbf{d}_1 = (\theta_1, 0)$ and $\mathbf{d}_2 = (\theta_2, 0)$, where $-\pi/2 \leq \theta_1 < \theta_2 \leq \pi/2$. Then, the offset of the azimuth δ_s is calculated as follows

$$\alpha = \frac{\tau_s(\mathbf{d}_2) - \tau_s(\mathbf{d}_1)}{\theta_2 - \theta_1}, \tag{19}$$

$$\delta_s = \frac{\tau_s(\mathbf{0})}{\alpha}, \tag{20}$$

where the zero vector $\mathbf{0}$ indicates the front direction, and δ_s is added to the original azimuth θ. We only tested this simple solution as proof-of-concept of the offset compensation.

ACKNOWLEDGMENT

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

REFERENCES

[1] A. Carlini, C. Bordeau, and M. Ambard, "Auditory localization: A comprehensive practical review," *Front. Psychol.*, vol. 15, 2024, Art. no. 1408073.

[2] 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, Jul. 2018.

[3] A. Andreopoulou and B. F. G. Katz, "Identification of perceptually relevant methods of inter-aural time difference estimation," *J. Acoust. Soc. Amer.*, vol. 142, no. 2, pp. 588–598, 2017.

[4] 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.

[5] G. R. Dabike et al., "The ICASSP SP cadenza challenge: Music demixing/remixing for hearing aids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Workshops*, 2024, pp. 93–94.

[6] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, Aug. 2004.

[7] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*. Plantation, FL, USA: J. Ross Publishing, 2013.

[8] E. M. Wenzel and S. H. Foster, "Perceptual consequences of interpolating head-related transfer functions during spatial synthesis," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, Oct. 1993, pp. 102–105.

[9] J. Oberem, J. G. Richter, D. Setzer, J. Seibold, I. Koch, and J. Fels, "Experiments on localization accuracy with non-individual and individual HRTFs comparing static and dynamic reproduction methods," 2020, *BioRxiv*.

[10] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acout. Sci., Tech*, vol. 35, no. 3, pp. 159–165, Mar. 2014.

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

[12] 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, May 2017.

[13] M. Cuevas-Rodrıǵuez et al., "3D tune-in toolkit: An open-source library for real-time binaural spatialisation," *PLoS One*, vol. 14, no. 3, 2019, Art. no. e0211899.

[14] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs [head related transfer functions]," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, pp. iv–iv.

[15] 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. 2012 Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Dec. 2012, pp. 1–5.

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

[17] Y. Iwaya, "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 340–343, Nov. 2006.

[18] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *J. Acoust. Soc. Amer.*, vol. 131, no. 2, pp. EL99–EL105, Jan. 2012.

[19] D. Schönstein and B. F. G. Katz, "HRTF selection for binaural synthesis from a database using morphological parameters," in *Proc. ICA*, 2010.

[20] S. Spagnol, "HRTF selection by anthropometric regression for improving horizontal localization accuracy," *IEEE Signal Process. Lett.*, vol. 27, pp. 590–594, 2020.

[21] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D Audio—The new standard for coding of immersive spatial audio," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 770–779, Aug. 2015.

[22] 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. Int. Workshop Acoustic Signal Enhancement*, 2022, pp. 1–5.

[23] Y. Masuyama et al., "NIIRF: Neural IIR filter field for HRTF upsampling and personalization," in *Proc. ICASSP 2024-2024 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 1016–1020.

[24] M. Zhu, M. Shahnawaz, S. Tubaro, and A. Sarti, "HRTF personalization based on weighted sparse representation of anthropometric features," in *Proc. Int. Conf. 3D Immersion*, Dec. 2017.

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

[26] 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 Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2085–2099, 2024.

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

[28] 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, Feb. 2024.

[29] I. D. Gebru et al., "Implicit HRTF modeling using temporal convolutional networks," in *Proc. ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech Signal Process.*, Jun. 2021, pp. 3385–3389.
[30] J. W. Lee, S. Lee, and K. Lee, "Global HRTF interpolation via learned affine transformation of hyper-conditioned features," in *Proc. ICASSP 2023-2023 IEEE Int. Conf. Acoust., Speech Signal Process.*, Jun. 2023, pp. 1–5.
[31] Y. Zhang, Y. Wang, and Z. Duan, "HRTF field: Unifying measured HRTF magnitude representation with neural fields," in *Proc. ICASSP 2023-2023 IEEE Int. Conf. Acoust., Speech Signal Process.*, Jun. 2023, pp. 1–5.
[32] B. Mildenhall, P.P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, Jan. 2022.
[33] Y. Xie et al., "Neural fields in visual computing and beyond," *Comput. Graph. Forum*, vol. 41, no. 2, pp. 641–676, May 2022.
[34] I. Engel et al., "The SONICOM HRTF dataset," *J. Audio Eng. Soc.*, vol. 71, pp. 241–253, May 2023.
[35] Y. Masuyama, G. Wichern, F. G. Germain, C. Ick, and J. Le Roux, "Retrieval-augmented neural field for HRTF upsampling and personalization," in *Proc. ICASSP 2025-2025 IEEE Int. Conf. Acous., Speech Signal Process.*, 2025, pp. 1–5.
[36] 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. Amer.*, vol. 91, no. 3, pp. 1637–1647, Mar. 1992.
[37] M. Geronazzo et al., "Technical report: SONICOM}/IEEE listener acoustic personalisation (LAP) challenge - 2024," *TechRxiv*, Nov. 2024.
[38] A. O. T. Hogg et al., "Listener acoustic personalization challenge - LAP24: Head-related transfer function upsampling," *IEEE Open J. Signal Process.*, vol. 6, pp. 926–941, 2025.
[39] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," in *Proc. NeurIPS*, Dec. 2022, pp. 3165–3177.
[40] 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 Int. Conf. Acoust., Speech, Signal Process. Workshops*, 2024, pp. 740–744.
[41] 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 Process. Mag.*, vol. 41, no. 6, pp. 60–71, Nov. 2024.
[42] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022.
[43] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. NeurIPS*, 2020, vol. 33, pp. 9459–9474.
[44] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3929–3938.
[45] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer, "Retrieval-augmented diffusion models," in *Proc. NeurIPS*, 2022, vol. 35, pp. 15309–15324.
[46] D. Horita, N. Inoue, K. Kikuchi, K. Yamaguchi, and K. Aizawa, "Retrieval-augmented layout transformer for content-aware layout generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 67–76.
[47] Y. Yuan, H. Liu, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Retrieval-augmented text-to-audio generation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 581–585.
[48] M. Wang et al., "Retrieval augmented end-to-end spoken dialog models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 12056–120 60.
[49] M. Tancik et al., "Fourier features let networks learn high frequency functions in low dimensional domains," in *Proc. NeurIPS*, Dec. 2020, pp. 7537–7547.
[50] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6394–6398.
[51] S. Diamond and S. Boyd, "CVXPY: A python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 1–5, 2016.