Retrieval-Augmented Neural Field for HRTF Upsampling and Personalization

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

TR2025-029   March 08, 2025

Abstract

Head-related transfer functions (HRTFs) with dense spatial grids are desired for immersive binaural audio generation, but their recording is time-consuming. Although HRTF spatial upsampling has shown remarkable progress with neural fields, spatial upsampling only from a few measured directions, e.g., 3 or 5 measurements, is still challenging. To tackle this problem, we propose a retrieval-augmented neural field (RANF). RANF retrieves a subject whose HRTFs are close to those of the target subject from a dataset. The HRTF of the retrieved subject at the desired direction is fed into the neural field in addition to the sound source direction itself. Furthermore, we present a neural network that can efficiently handle multiple retrieved subjects, inspired by a multi- channel processing technique called transform-average-concatenate. Our experiments confirm the benefits of RANF on the SONICOM dataset, and it is a key component in the winning solution of Task 2 of the listener acoustic personalization challenge 2024.

Retrieval-Augmented Neural Field
for HRTF Upsampling and Personalization

Yoshiki Masuyama*, Gordon Wichern*, François G. Germain*, Christopher Ick*†, and Jonathan Le Roux*

*Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
†Music and Audio Research Laboratory, New York University, Brooklyn, NY, USA

Abstract—Head-related transfer functions (HRTFs) with dense spatial grids are desired for immersive binaural audio generation, but their recording is time-consuming. Although HRTF spatial upsampling has shown remarkable progress with neural fields, spatial upsampling only from a few measured directions, e.g., 3 or 5 measurements, is still challenging. To tackle this problem, we propose a retrieval-augmented neural field (RANF). RANF retrieves a subject whose HRTFs are close to those of the target subject from a dataset. The HRTF of the retrieved subject at the desired direction is fed into the neural field in addition to the sound source direction itself. Furthermore, we present a neural network that can efficiently handle multiple retrieved subjects, inspired by a multi-channel processing technique called transform-average-concatenate. Our experiments confirm the benefits of RANF on the SONICOM dataset, and it is a key component in the winning solution of Task 2 of the listener acoustic personalization challenge 2024.

Index Terms—Head-related transfer function, spatial audio, neural field, retrieval-augmented generation

I. INTRODUCTION

Spatial audio rendering has been extensively studied for various applications, including telepresence systems [1] and mixed reality technologies [2], [3]. To generate high-quality immersive audio over headphones, the spatial filtering effects as sound travels from its source to each ear should be mimicked by head-related transfer functions (HRTFs). These filtering effects are caused by the reflection and scattering of sound on the upper torso, head, and pinnae. Hence, HRTFs are different for each subject, and the use of a generic HRTF may result in suboptimal perceptual localization [4], [5]. Although individual HRTFs should ideally be measured with dense spatial grids, this requires an enormous amount of time and effort [6]. To mitigate this problem, various HRTF spatial upsampling and personalization methods have been developed [7]–[11].

Neural networks have recently been applied to spatial upsampling, motivated by their adaptive and powerful modeling capabilities [12]–[20]. In particular, neural field (NF)-based methods can exploit a dataset with arbitrary spatial grids and have shown promising performance by representing HRTFs as a function of the sound source direction [16]–[19]. NF was first used to represent the HRTFs of a single subject, then extended to represent those of multiple subjects by incorporating a small number of subject-specific parameters [18], [19]. As a result, a pre-trained generic NF can be easily adapted to a new subject by optimizing the subject-specific parameters while freezing most of the model parameters. While this adaptation works well when a large number of measurements are available, spatial upsampling from fewer than ten measurements remains highly challenging.

As an orthogonal direction to spatial upsampling, HRTF selection has also been explored to synthesize plausible binaural audio. From a given dataset with a dense spatial grid, this method aims to select a subject whose HRTFs are close to HRTFs of the target subject. In [21], [22], a subject manually selects the best-fitting subject from a
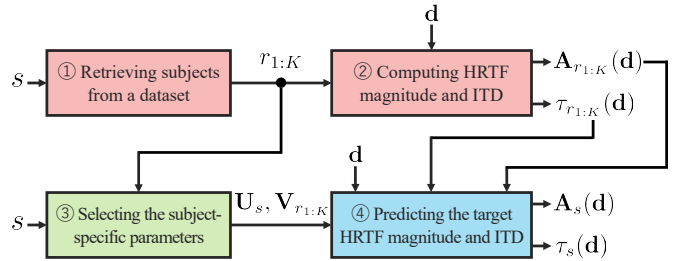
This work was performed while C. Ick was an intern at MERL.

Fig. 1. Overview of RANF. It aims to predict the HRTF magnitude and ITD of a target subject s at a desired direction $\mathbf{d}$ as in ④. We retrieve K subjects in ① and compute their HRTF magnitude and ITD at the desired direction $\mathbf{d}$ in ②. The retrieved HRTF magnitude and ITD are fed into NF with the sound source direction and the subject-specific parameters selected in ③.

dataset through listening tests. To reduce the effort of the target subject, objective selection methods without human evaluation have been developed by using the similarities of HRTFs at a limited number of directions [15] or anthropometric measurements [2], [23], [24]. These methods can achieve moderate performance with a limited number of HRTF recordings or even only with anthropometric measurements.

HRTF selection motivates us to enhance NF for HRTF spatial upsampling by using selected HRTFs. Retrieval-augmented generation (RAG) [25] has shown remarkable success in various modalities including natural language processing [25], [26], computer vision [27], [28], and audio processing [29], [30]. RAG retrieves relevant information from external resources and feeds it into a neural network as a context for data generation. This can improve the generalization capability to out-of-distribution data and the performance without increasing the model parameters. We can interpret the selected HRTFs as the context for generating HRTFs of the target subject.

In this paper, we propose a retrieval-augmented NF (RANF) to address the challenges in HRTF spatial upsampling from sparse measurements. The overview of RANF is depicted in Fig. 1. Inspired by HRTF selection, RANF retrieves multiple subjects whose HRTFs are close to those of the target subject at the measured directions. Then, RANF uses HRTFs of the retrieved subjects in the desired direction as an auxiliary input to NF. Furthermore, we propose a network architecture that can efficiently exploit multiple retrievals to improve the HRTF prediction performance. In our experiments on the SONICOM dataset [31], RANF outperformed the NF without retrieval and the HRTF selection methods.

II. PRELIMINARIES

A. Problem setting

Let $\mathbf{H}_s(\mathbf{d}) \in \mathbb{C}^{F \times 2}$ denote the HRTF for a subject s and direction $\mathbf{d} = (\theta, \phi)$, where F is the number of frequency bins, $\theta \in [0, 2\pi)$ is the azimuth, and $\phi \in [-\pi/2, \pi/2]$ is the elevation. The azimuth

increases counter-clockwise with $\theta = 0$ directly in front of the subject, while the elevation increases upward with $\phi = 0$ corresponding to the equatorial plane. Let $\mathcal{S}$ be the training set of subjects whose HRTFs are measured at all directions $\mathbf{d} \in \mathcal{D}$. Meanwhile, HRTFs of a new subject $\bar{s} \notin \mathcal{S}$ are measured at a limited number of directions $\mathbf{d} \in \mathcal{D}'$, where $\mathcal{D}' \subset \mathcal{D}$. We aim to predict $\mathbf{H}_{\bar{s}}(\mathbf{d})$ for a new subject $\bar{s}$ and all directions $\mathbf{d} \in \mathcal{D}$ from only the measurements for $\mathbf{d} \in \mathcal{D}'$.

Existing studies have shown that azimuth localization mainly relies on the interaural time difference (ITD) and the interaural level difference (ILD) [32], while elevation localization depends on spectral coloration [33]. Hence, HRTF phase is not essential to localization as long as ITD is preserved. This motivates us to approximate the complex HRTF $\mathbf{H}_s(\mathbf{d})$ via its magnitude response $\mathbf{A}_s(\mathbf{d}) \in \mathbb{R}_+^{F \times 2}$ and ITD $\tau_s(\mathbf{d}) \in \mathbb{R}$. The predicted HRTF magnitude can be converted to the time domain with the minimum phase [34], and ITD is compensated by shifting the time-domain impulse responses.

B. Neural field (NF) for HRTF spatial upsampling

An NF for HRTF spatial upsampling is trained to be a map from a sound source direction to a representation of the HRTF. The HRTF field [18] predicts the magnitude response $\mathbf{A}_s(\mathbf{d})$ as follows:

$$\mathbf{A}_s(\mathbf{d}) = \text{NF}(\mathbf{d} \mid \boldsymbol{\Xi}_s), \tag{1}$$

where NF is specialized for each subject s, and $\boldsymbol{\Xi}_s$ denotes its parameters. Since HRTFs have similarities across different subjects, we can pre-train a generic NF for multiple subjects with a small number of subject-specific parameters [18], [19]:

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

where $\boldsymbol{\Gamma}$ denotes the parameters shared across different subjects, and $\boldsymbol{\Xi}_s$ steers the generic model to represent the HRTF of a target subject s. During its pre-training, NF is trained on given pairs of subject $s \in \mathcal{S}$ and direction $\mathbf{d} \in \mathcal{D}$, jointly optimizing $\boldsymbol{\Gamma}$ and $\boldsymbol{\Xi}_s, \forall s \in \mathcal{S}$. Then, we adapt NF to a new target subject $\bar{s}$ by optimizing the subject-specific parameters $\boldsymbol{\Xi}_{\bar{s}}$ on the limited number of measurements, while freezing the shared parameters $\boldsymbol{\Gamma}$. This adaptation is more efficient than training a new NF in (1) from scratch. In [19], the low-rank adaptation (LoRA) [35] shows promising adaptation performance by modifying the weight of the lth layer $\mathbf{W}_l \in \mathbb{R}^{N_l \times M_l}$ as follows:

$$\mathbf{W}_{l,s} = \mathbf{W}_l + \mathbf{u}_{l,s} \mathbf{v}_{l,s}^\top, \tag{3}$$

where $(\cdot)^\top$ denotes the transpose, 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 building a rank-1 matrix.

While existing NF-based methods have focused on approximating the HRTF magnitude, we predict ITD in sample $\tau_s(\mathbf{d})$ as well:

$$\mathbf{A}_s(\mathbf{d}), \tau_s(\mathbf{d}) = \text{NF}(\mathbf{d} \mid \boldsymbol{\Gamma}, \mathbf{U}_s, \mathbf{V}_s), \tag{4}$$

where the subject-specific parameters $\boldsymbol{\Xi}_s$ in (2) are split into $\mathbf{U}_s = [\mathbf{u}_{1,s}, \ldots, \mathbf{u}_{L,s}]$ and $\mathbf{V}_s = [\mathbf{v}_{1,s}, \ldots, \mathbf{v}_{L,s}]$ for LoRA. This allows us to reconstruct the head-related impulse responses except for the distance-dependent time-of-arrival and to evaluate the impulse responses with the objective measures used in the listener acoustic personalization (LAP) challenge[1].

III. Proposed Method: RANF

A. Overview of RANF

As ① in Fig. 1, we first retrieve K subjects $r_k^s \in \mathcal{S}$ for a given target subject s, where $k = 1, \ldots, K$ is the index of the retrieved subjects. The K subjects are selected on the basis of the similarity of

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

the ITD and/or HRTF magnitude to those of a target subject s at the measured directions in $\mathcal{D}'$. To avoid clutter, we omit the superscript s in r_k^s hereafter. In ②, we compute the HRTF magnitudes $\mathbf{A}_{r_{1:k}}(\mathbf{d})$ and ITDs $\tau_{r_{1:k}}(\mathbf{d})$ for the retrieved subjects at the desired direction $\mathbf{d}$. These are fed into NF as additional clues to predict the HRTF of the target subject in ④:

$$\mathbf{A}_s(\mathbf{d}), \tau_s(\mathbf{d}) = \text{RANF}(\mathbf{d}, \mathbf{A}_{r_{1:k}}(\mathbf{d}), \tau_{r_{1:k}}(\mathbf{d}) \mid \boldsymbol{\Gamma}, \mathbf{U}_s, \mathbf{V}_{r_{1:k}}), \tag{5}$$

where subject-specific parameters are selected in ③, combining parameters $\mathbf{U}_s$ for the target subject s and parameters $\mathbf{V}_{r_{1:k}} = \{\mathbf{V}_{r_1}, \ldots, \mathbf{V}_{r_k}\}$ for the retrieved subjects r_k. This allows RANF to perform retrieval-dependent processing.

During the pre-training of RANF, we optimize the shared parameters $\boldsymbol{\Gamma}$ and the subject-specific parameters $(\mathbf{U}_s, \mathbf{V}_{r_{1:K}})$ on given pairs of the target subject $s \in \mathcal{S}$ and sound source direction $\mathbf{d} \in \mathcal{D}$. We retrieve the K subjects also from $\mathcal{S}$, where we ensure $r_k \neq s$ for all k. The shared and subject-specific 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[2]:

$$\mathcal{L}_{\text{pre-train}} = \sum_{s \in \mathcal{S}} \sum_{\mathbf{d} \in \mathcal{D}} \mathcal{L}(\mathbf{H}_s^\star(\mathbf{d}), \mathbf{H}_s(\mathbf{d})), \tag{6}$$

$$\mathcal{L}(\mathbf{H}^\star(\mathbf{d}), \mathbf{H}(\mathbf{d})) = \text{LSD}(\mathbf{A}^\star(\mathbf{d}), \mathbf{A}(\mathbf{d})) + \lambda \text{MAE}_\epsilon(\tau^\star(\mathbf{d}), \tau(\mathbf{d})), \tag{7}$$

where $\lambda \in \mathbb{R}_+$ is a hyperparameter, $(\cdot)^\star$ indicates the oracle value, ϵ is set to 0.5 to consider the quantization error in the oracle sample-level ITD, and LSD is defined as

$$\text{LSD}(\mathbf{A}^\star(\mathbf{d}), \mathbf{A}(\mathbf{d})) = \frac{1}{2} \sum_{c=1}^{2} \sqrt{\frac{1}{F} \sum_{f=1}^{F} \left(20 \log_{10} \frac{A_{c,f}(\mathbf{d})}{A_{c,f}^\star(\mathbf{d})} \right)^2}. \tag{8}$$

In (7) and (8), we omit the subscript s for the subject to show a general case, and $c \in \{1, 2\}$ and $f = 1, \ldots, F$ are the channel and frequency indices, respectively. Then, we adapt the pre-trained generic model to a new target subject $\bar{s} \notin \mathcal{S}$ by optimizing only $\mathbf{U}_{\bar{s}}$ to minimize the following loss function:

$$\mathcal{L}_{\text{adaptation}} = \sum_{\mathbf{d} \in \mathcal{D}'} \mathcal{L}(\mathbf{H}_{\bar{s}}^\star(\mathbf{d}), \mathbf{H}_{\bar{s}}(\mathbf{d})). \tag{9}$$

During the adaptation, we freeze the shared parameters $\boldsymbol{\Gamma}$ and the parameters relying on the retrieved subjects $\mathbf{V}_{r_{1:k}}$.

B. HRTF retrieval in ① and ②

To efficiently personalize a generic NF to the target subject s, we retrieve K subjects from a given HRTF dataset whose HRTF will be helpful to predict the HRTF of the target subject. In detail, we retrieve top-K subjects on the basis of a HRTF error measure $\mathcal{L}_{\text{retrieval}}$:

$$\tilde{r}_1, \ldots, \tilde{r}_K \leftarrow \underset{r_{1:K} \in [\mathcal{S}]^K}{\text{argmin}} \sum_{k=1}^{K} \sum_{\mathbf{d} \in \mathcal{D}'} \mathcal{L}_{\text{retrieval}}(\mathbf{H}_s(\mathbf{d}), \mathbf{H}_{r_k}(\mathbf{d})), \tag{10}$$

where $[\mathcal{S}]^K$ denotes the set of all subsets of $\mathcal{S}$ with K distinct elements. The choice of $\mathcal{L}_{\text{retrieval}}$ affects the prediction quality, and we investigate using LSD or MAE on ITDs in Section IV-C.

In ②, the HRTF magnitudes and ITDs of the retrieved subjects $r_{1:K}$ at the desired direction $\mathbf{d}$ are computed, assuming that all of the desired directions are in the set of sound source directions $\mathcal{D}$. This assumption is not overly limiting if the HRTF dataset for retrieval has sufficiently dense spatial grids. If the desired direction $\mathbf{d}$ is not in $\mathcal{D}$, we can use another NF to predict the HRTFs at $\mathbf{d}$ for the subjects in the dataset. We shall leave this for future work.

[2]We compute LSD and the oracle ITD based on `Spatial Audio Metrics`: https://github.com/Katarina-Poole/Spatial-Audio-Metrics.
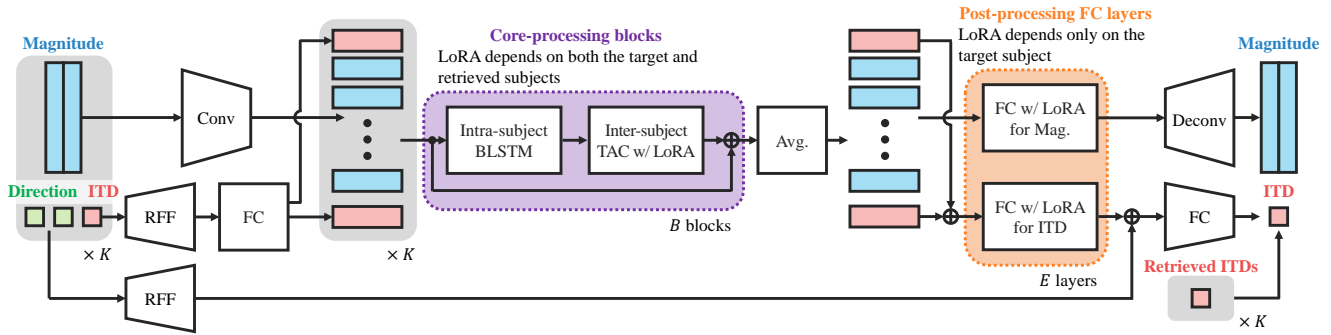
Fig. 2. Network architecture for the proposed RANF. HRTF magnitudes are encoded by a convolution network and decoded by a deconvolution network. The retrieved ITDs are encoded as RFF with the sound source direction and passed to an FC layer. The feature sequence in the gray box is for each of the K retrieved subjects and processed in parallel except for the inter-subject TAC module. The features are aggregated across all retrieved subjects by averaging (Avg.).

C. Network architecture for RANF in ④

Figure 2 illustrates a network architecture for RANF. The HRTF magnitude of each retrieved subject is encoded with a 1D convolution network that downsamples the number of frequency bins F to F', where input and output channels are 2 and C, respectively. The convolution network captures the local frequency information in the retrieved HRTF magnitude. Meanwhile, the retrieved ITD is transformed into random Fourier features (RFF) [36] with the azimuth and elevation of the sound source. The RFF is converted into two C-dimensional feature vectors by a fully connected (FC) layer and concatenated at both ends of the encoded magnitude. Hence, ITD-related features are passed to the beginning of both forward and backward LSTMs in the following intra-subject BLSTM modules. We obtain K feature sequences with shape $(F' + 2) \times C$.

Next, we feed the feature sequences into B core-processing blocks, each consisting of an intra-subject BLSTM module and an inter-subject transform-average-concatenate (TAC) module [37] with a residual connection. In the intra-subject BLSTM module, we handle each of the K feature sequences separately and perform sequence modeling along the $F' + 2$ feature dimensions. On the other hand, the inter-subject TAC module mixes the feature vectors across different retrieved subjects, where each of $F' + 2$ feature dimensions is treated separately. This module includes LoRA to perform retrieval and target-dependent processing.

After the B core-processing blocks, we average the K feature sequences into a single target feature sequence. A portion of this sequence corresponding to the HRTF magnitude is processed by E post-processing FC layers with LoRA, where both LoRA vectors are specified by the target subject as in (3). The output sequence is decoded to the two-channel HRTF magnitude by a 1D deconvolution block. Meanwhile, the features at both ends are summed up for ITD prediction and passed to E post-processing FC layers. We then incorporate an additional RFF computed only from the desired direction $\mathbf{d}$ and predict the residual between the oracle ITD and the average ITD of the retrieved subjects via FC layers without LoRA. That is, the network is trained to predict $\tau_s(\mathbf{d}) - (1/K) \sum_{k=1}^{K} \tau_{r_k}(\mathbf{d})$.

D. Inter-subject TAC module

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 by using a TAC module [37] and performs subject-specific processing with LoRA.

Figure 3 shows the inter-subject TAC module, where LN is layer normalization. The input feature vectors are processed by two FC layers, where the outputs of one FC layer are averaged over the retrieved subjects. The averaged vector is concatenated with the outputs
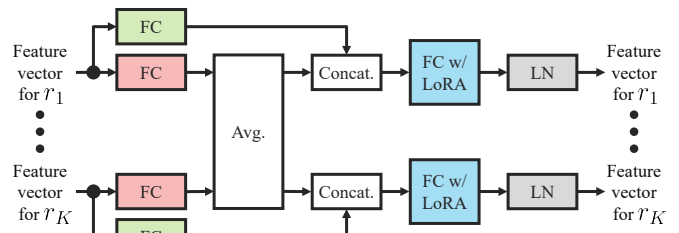

Fig. 3. Inter-subject TAC module. Trainable layers are colored, where the layers with the same color share their parameters.

of the other FC layer. This TAC operation was originally proposed for multi-channel speech separation to handle multi-channel features in a permutation invariant manner [37]. We use the TAC operation to efficiently mix the features for multiple retrieved subjects. It can scale the model to an arbitrary number of retrieved subjects, which is advantageous for environments with variable computational resources.

The concatenated features are then passed to an FC layer with LoRA. Here, LoRA differs from (3) as follows:

$$\mathbf{W}_{b,s,k} = \mathbf{W}_b + \mathbf{u}_{b,s} \mathbf{v}_{b,r_k}^{\mathsf{T}}, \tag{11}$$

where $b = 1, \ldots, B$ indexes the core-processing blocks. With the modification in (11), we aim to take the relation between the target and retrieved subjects into account. Since we retrieve r_k from the set of pre-training subjects $\mathcal{S}$ even in inference, $\mathbf{v}_{b,r_k}$ can be pre-trained and fixed during the adaptation.

IV. EXPERIMENTS

A. Dataset and experimental setup

We evaluate RANF on the SONICOM HRTF dataset [31], which includes 200 subjects and 793 measurements per subject. HRTFs were sampled at 48 kHz, and a free-field compensation was performed with a minimum-phase filter as in the LAP challenge. The first 179 subjects were used for pre-training, and the remaining 20 subjects were used to evaluate the adaptation capability. We removed subject P0079 from the dataset due to its atypical ITD behavior. Within the 179 subjects, the last 19 subjects were used for validation. The upsampling performance was investigated on four different sets $\mathcal{D}'$ of measured directions for the target subjects, where the number of measured directions varied in $(3, 5, 19, 100)$ as in the LAP challenge.

The convolution and deconvolution networks consisted of 4 layers with the PReLU activation, where the encoded dimension C was 128. The number of core-processing blocks B was 4, and the BLSTM had 64 units for each direction in each block. The FC layers in the inter-subject TAC module used the GELU activation. The number of

| Methods	3 measurements			5 measurements			19 measurements			100 measurements		
	ITD [μs]	ILD [dB]	LSD [dB]	ITD [μs]	ILD [dB]	LSD [dB]	ITD [μs]	ILD [dB]	LSD [dB]	ITD [μs]	ILD [dB]	LSD [dB]
Nearest neighbor	274.2	7.6	8.7	154.2	4.8	8.3	108.3	3.0	5.4	44.0	1.4	3.4
HRTF selection (ITD)	26.3	1.4	6.4	24.5	1.5	6.4	23.2	1.6	6.7	20.0	1.4	6.3
HRTF selection (LSD)	37.5	1.5	5.8	35.7	1.4	5.6	37.4	1.5	5.5	31.4	1.3	5.4
NF (CbC) [18]	22.1	1.5	4.9	20.7	2.0	5.2	14.8	1.8	5.0	11.8	1.7	5.1
NF (LoRA) [19]	28.6	1.3	4.7	24.7	1.4	**4.6**	14.7	1.1	4.1	**9.1**	1.1	3.8
RANF (proposed)	**20.5**	**1.2**	**4.4**	**18.7**	**1.2**	4.6	**14.2**	**1.0**	**3.6**	10.0	**0.8**	**3.0**
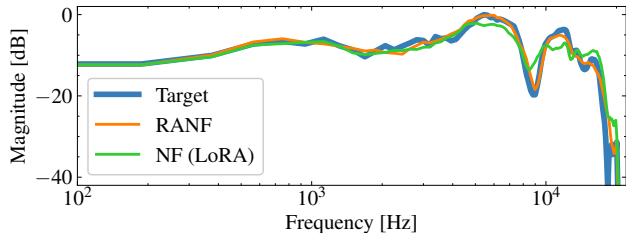
Fig. 4. Ipsilateral HRTF magnitudes for a subject with $(\theta, \phi) = (\pi/2, 0)$.

post-processing FC layers E was set to 2, and the last ITD prediction network consisted of 2 layers. In (7), we set λ to 20.8 to change the loss scale from sample to microsecond. The neural network was first pre-trained for 200 epochs and then adapted for 500 epochs by using the RAdam optimizer with an initial learning rate of 0.001. The learning rate was multiplied by 0.9 when the validation loss did not decrease for 10 successive epochs. The model with the best validation loss was used for adaptation. During adaptation, we did not perform validation to exploit all the measurements for adaptation. Our training and inference scripts are available online[3].

We adopted two baselines: the nearest neighbor direction and HRTF selection. The best-fitting subject was selected from the 179 subjects in terms of LSD or MAE of ITDs. In addition, we compared the proposed method to NFs with two kinds of subject-specific parameters: the conditioning by concatenation (CbC) [18] and LoRA [19]. CbC introduces a subject-specific latent vector and concatenates it to the input. Its dimension was set to 32 following the original paper.

B. Comparison with existing methods

Table I shows the average scores used in the LAP challenge, where we evaluate the performance only on the unmeasured directions $\mathbf{d} \in \mathcal{D}\backslash\mathcal{D}'$. In RANF, we used MAE on ITDs to retrieve 5 subjects based on a preliminary study. The nearest neighbor direction performed poorly with only 3 or 5 measurements, which highlights the difficulty of HRTF spatial upsampling from such a limited number of measurements. Meanwhile, HRTF selection retained reasonable performance even in the sparsest condition.

Regarding the existing NF-based methods, both methods substantially improved LSD from HRTF selection. NF (LoRA), which contains more subject-specific parameters, achieved better LSD but increased the ITD error for 3 and 5 measurements, likely because LoRA overfits for such a limited number of measurements. By incorporating the top-5 retrievals, RANF consistently outperformed the baselines and improved upsampling performance from NF (LoRA) except for the ITD with 100 measurements. Figure 4 shows examples of the predicted HRTF magnitude for subject `P0181` under the 100 measurements condition. We can observe that RANF captured the peak and notch around 10 kHz more accurately than NF (LoRA).

[3]https://github.com/merlresearch/ranf-hrtf

TABLE II
UPSAMPLING RESULTS WITH DIFFERENT RETRIEVAL STRATEGIES.

Retrieval criterion	# of retrievals K	ITD [μs]	ILD [dB]	LSD [dB]
Random	1	26.6	1.3	**4.4**
	5	23.8	1.3	4.5
	10	23.5	1.3	**4.4**
LSD	1	29.9	1.3	4.6
	5	22.4	**1.2**	**4.4**
	10	22.7	**1.2**	**4.4**
ITD error	1	23.8	**1.2**	**4.4**
	5	**20.5**	**1.2**	**4.4**
	10	22.7	1.3	**4.4**

Post-processing FC layers	B	Params (M)	ITD [μs]	ILD [dB]	LSD [dB]
✗	4	0.75	24.6	1.4	4.7
✗	5	0.89	24.3	1.3	4.7
✓	4	0.82	**20.5**	**1.2**	**4.4**

C. Ablation study

To show the efficacy of our retrieval strategy and network architecture, we perform an ablation study on the 3 measurement condition. **Retrieval strategy**: In RANF, we need to choose the HRTF error measure $\mathcal{L}_{\text{retrieval}}$ in (10). In addition to top-K retrieval in terms of LSD and the ITD error, we evaluated the random retrieval of subjects for pre-training. Table II presents the HRTF spatial upsampling performance with different retrieval criteria and different numbers of retrievals. The results show comparable LSD and ILD error regardless of the number of retrievals or their criterion. Conversely, ITD upsampling performance was substantially improved when performing retrieval based on the ITD error. It also helped to retrieve more than one subject, and we ultimately found that retrieving 5 subjects with the ITD-based criterion leads to the best results.

Efficacy of post-processing FC layers: Our NF shown in Fig. 2 leverages LoRA in two formulations. LoRA in the core-processing blocks depends on both target and retrieved subjects as in (11), while LoRA in the post-processing FC layers relies solely on the target subject as in (3). Table III presents the upsampling performance with and without the post-processing FC layers, where $K = 5$. It can be observed that the post-processing FC layers bring better performance even with fewer shared trainable parameters.

V. CONCLUSION

We present RANF, an HRTF spatial upsampling method that enhances NF by leveraging HRTF selection. RANF retrieves multiple subjects from a given HRTF dataset, and our experiments demonstrated the benefit of multiple retrievals in terms of ITD error. The winning system of the LAP challenge 2024 is based on RANF, and our future work will include a detailed analysis of RANF after the organizer releases the baselines and full datasets.

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] 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, Jul. 2004.

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

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

[5] 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," *BioRxiv*, 2020.

[6] 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.

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

[8] 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.

[9] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs [head related transfer functions]," in *Proc. ICASSP*, May 2004.

[10] J. Ahrens, M. R. P. Thomas, and I. Tashev, "HRTF magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data," in *Proc. APSIPA*, Dec. 2012, pp. 1–5.

[11] 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.

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

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

[14] 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. IWAENC*, Sep. 2022.

[15] 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.

[16] 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. ICASSP*, Jun. 2021, pp. 3385–3389.

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

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

[19] 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. ICASSP*, Apr. 2024, pp. 1016–1020.

[20] 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.

[21] 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–E43, Nov. 2006.

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

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

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

[25] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. NeurIPS*, vol. 33, Dec. 2020, pp. 9459–9474.

[26] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *Proc. ICML*, Jul. 2020, pp. 3929–3938.

[27] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer, "Retrieval-augmented diffusion models," in *Proc. NeurIPS*, vol. 35, Nov. 2022, pp. 15 309–15 324.

[28] D. Horita, N. Inoue, K. Kikuchi, K. Yamaguchi, and K. Aizawa, "Retrieval-augmented layout transformer for content-aware layout generation," in *Proc. CVPR*, Jun. 2024, pp. 67–76.

[29] Y. Yuan, H. Liu, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Retrieval-augmented text-to-audio generation," in *Proc. ICASSP*, Apr. 2024, pp. 581–585.

[30] M. Wang, I. Shafran, H. Soltau, W. Han, Y. Cao, D. Yu, and L. El Shafey, "Retrieval augmented end-to-end spoken dialog models," in *Proc. ICASSP*, Apr. 2024, pp. 12 056–12 060.

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

[32] A. Carlini, C. Bordeau, and M. Ambard, "Auditory localization: A comprehensive practical review," *Front. Psychol.*, vol. 15, p. 1408073, 2024.

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

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

[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, Apr. 2022.

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

[37] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. ICASSP*, May 2020, pp. 6394–6398.