NIIRF: NEURAL IIR FILTER FIELD FOR HRTF UPSAMPLING AND PERSONALIZATION

Masuyama, Yoshiki; Wichern, Gordon; Germain, François G; Pan, Zexu; Khurana, Sameer; Hori, Chiori; Le Roux, Jonathan

TR2024-026   March 19, 2024

Abstract

Head-related transfer functions (HRTFs) are important for immersive audio, and their spatial interpolation has been studied to up- sample finite measurements. Recently, neural fields (NFs) which map from sound source direction to HRTF have gained attention. Existing NF-based methods focused on estimating the magnitude of the HRTF from a given sound source direction, and the magnitude is converted to a finite impulse response (FIR) filter. We propose the neural infinite impulse response filter field (NIIRF) method that instead estimates the coefficients of cascaded IIR filters. IIR filters mimic the modal nature of HRTFs, thus needing fewer coefficients to approximate them well compared to FIR filters. We find that our method can match the performance of existing NF-based methods on multiple datasets, even outperforming them when measurements are sparse. We also explore approaches to personalize the NF to a subject and experimentally find low-rank adaptation to be effective.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2024

NIIRF: NEURAL IIR FILTER FIELD FOR HRTF UPSAMPLING AND PERSONALIZATION

Yoshiki Masuyama[1,2], Gordon Wichern[1], François G. Germain[1], Zexu Pan[1],
Sameer Khurana[1], Chiori Hori[1], Jonathan Le Roux[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Tokyo Metropolitan University, Tokyo, Japan

ABSTRACT

Head-related transfer functions (HRTFs) are important for immersive audio, and their spatial interpolation has been studied to upsample finite measurements. Recently, neural fields (NFs) which map from sound source direction to HRTF have gained attention. Existing NF-based methods focused on estimating the magnitude of the HRTF from a given sound source direction, and the magnitude is converted to a finite impulse response (FIR) filter. We propose the neural infinite impulse response filter field (NIIRF) method that instead estimates the coefficients of cascaded IIR filters. IIR filters mimic the modal nature of HRTFs, thus needing fewer coefficients to approximate them well compared to FIR filters. We find that our method can match the performance of existing NF-based methods on multiple datasets, even outperforming them when measurements are sparse. We also explore approaches to personalize the NF to a subject and experimentally find low-rank adaptation to be effective.

Index Terms— Head-related transfer function, neural field, implicit neural representations, differentiable digital signal processing, parameter efficient fine-tuning

1. INTRODUCTION

Head-related transfer functions (HRTFs) capture the sound transmission from a sound source to to both ears. They are essential to many applications, including telepresence systems [1] and virtual reality technologies [2, 3]. As HRTFs depend on the shape of the pinnae, head, and upper torso, they vary from person to person, and individual HRTFs may lead to high-quality immersive audio experiences [4, 5]. When the individual HRTFs for all possible source directions are available, we can generate an immersive binaural signal by convolving the HRTFs with the source signals. The ideal approach is to measure a sufficient number of HRTFs for each person, but it requires much time and effort because HRTFs are usually measured sequentially over several hundreds of directions [6]. Hence, both spatial upsampling and personalization of HRTFs have been the focus of intense study [7–9].

To upsample the HRTF measurements, various techniques have been developed such as panning-based methods [10, 11] and the spatial decomposition approach [12–14]. The former approach performs a weighted sum of the measured HRTFs for neighboring directions. For instance, vector-base amplitude panning (VBAP) [10] has been widely used due to its low computational complexity, and it was adopted by the ISO/IEC MPEG-H 3D Audio standard [15]. The latter approach expands the HRTFs to spatial basis functions, e.g., spherical harmonics [13, 14], and uses a global representation to estimate the HRTF for the target direction. This approach can be inter-
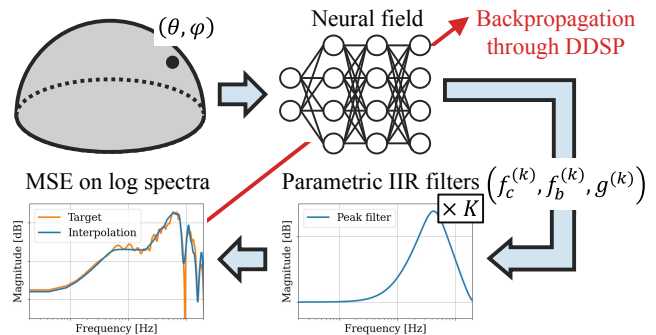
Fig. 1. Illustration of the proposed NF-based HRTF modeling method named NIIRF. The cascade of IIR filters consist of K peaking filters and two shelf filters for both left and right channels.

preted as an optimization problem of the spatial coefficients [12,13], which becomes underdetermined when the number of measurements is limited.

Recently, the machine learning approach has gained increasing attention since this approach can flexibly exploit anthropometric features [16, 17] and HRTFs for multiple subjects [18–21]. Despite recent progress, challenges remain regarding how to exploit a variable number of HRTF measurements and how to estimate HRTFs at arbitrary directions. To tackle these challenges, several works have leveraged the neural field (NF), or implicit neural representation, where the HRTF is represented as a function of the sound source direction [17, 20, 21]. NFs have been developed in computer vision to reconstruct 3D scenes from multiple 2D views [22,23] and applied to spatial audio modeling [24, 25]. In HRTF modeling by NFs, prior works have shown the potential to estimate the magnitude response of the HRTFs [17, 21]. In these methods, the interpolated HRTFs are converted to time-domain finite impulse response (FIR) filters by the inverse discrete Fourier transform (DFT) with minimum phase [26].

Meanwhile, infinite impulse response (IIR) filters are advantageous for approximating HRTFs with fewer coefficients than FIR filters [27, 28]. This is preferable in terms of computational complexity and memory footprint. While various algorithms are applicable to optimize the filter coefficients, back-propagation has recently been used to leverage powerful automatic differentiation libraries [29–31]. This approach is in the family of differentiable digital signal processing (DDSP) [32] and has been applied to HRTF modeling [33]. In addition, a neural network that estimates an IIR filter from a HRTF magnitude response was trained with DDSP [34]. These previous studies focused on approximating HRTFs only for certain directions and were not aimed at spatial upsampling.

In this paper, we propose to integrate NF-based spatial upsam-

pling with differentiable IIR filters as a neural IIR filter field (NI-IRF). As depicted in Fig. 1, our NF estimates the parameters of the cascaded IIR filters instead of the magnitude response or the time-domain FIR filter. Building upon DDSP, we optimize the NF to minimize the error of the generated binaural filter responses by back-propagation. Furthermore, we explore various conditioning approaches for personalizing the NF to new subjects. Our experimental results confirm that the proposed method outperforms the classical panning-based baseline. In addition, the proposed method improves the upsampling accuracy upon existing NF-based methods [21], when the number of measurements is limited.

2. PROPOSED METHOD: NIIRF

2.1. General formulation of HRTF modeling via NF

Defining θ as the azimuth in $[0, 2\pi)$ with $\theta = 0$ in front of the subject, and ϕ as the elevation in $[-\pi/2, \pi/2]$ with $\phi = 0$ on the equatorial plane, we aim to continuously interpolate HRTFs at any direction (θ, ϕ). Each HRTF consists of a left-right pair of discrete-time filters, but our modeling does not include the interaural time difference (ITD) information, as our phase-insensitive loss allows for dealing with it independently similar to previous studies[1] [21, 27].

In practice, the neural field NF has been used to map directions to intermediary parameters $\mathbf{\Psi}_{\theta,\phi}$ that are then mapped onto an HRTF through a DSP algorithm. This can be formulated as:

$$\mathcal{H}_{\theta,\phi} = \text{DSP} \circ \text{NF}(\theta, \phi) = \text{DSP}(\mathbf{\Psi}_{\theta,\phi}), \qquad (1)$$

where $\mathcal{H}_{\theta,\phi}$ is a filter representation of HRTF. In prior works [17, 21], $\mathbf{\Psi}_{\theta,\phi}$ is the magnitude of the DFT coefficients $|H_{\theta,\phi}[m]|$, where $m = 0, \ldots, M-1$ is the frequency bin index. To train NF which estimates the magnitude response, we use the following mean squared error (MSE) [21]:

$$\mathcal{L} = \frac{1}{|\mathcal{D}|M} \sum_{(\theta,\phi)\in\mathcal{D}} \sum_{m=0}^{M-1} \left(20 \log_{10} \left| \frac{H_{\theta,\phi}[m]}{\widetilde{H}_{\theta,\phi}[m]} \right| \right)^2, \qquad (2)$$

where $\widetilde{H}_{\theta,\phi}$ is the DFT of the measured HRTF, and $|\mathcal{D}|$ is the size of the dataset $\mathcal{D}$. In prior works following (1), DSP is a phase estimation followed by an inverse DFT to estimate an FIR filter. The phase is commonly recovered through minimum-phase reconstruction [26].

2.2. Proposed integration of NF and DDSP

Instead of estimating the DFT magnitude of the HRTF, we exploit an NF to estimate parameters of IIR filters as depicted in Fig. 1. This can be formulated as follows:

$$\mathcal{H}_{\theta,\phi} = \text{DDSP} \circ \text{NF}_{\text{IIR}}(\theta, \phi) = \text{DDSP}(\mathbf{\Psi}_{\theta,\phi}). \qquad (3)$$

In the proposed method, $\mathcal{H}_{\theta,\phi}$ is obtained via a DDSP module as a cascade of $K + 2$ IIR filters for each ear, $\mathbf{\Psi}_{\theta,\phi}$ denoting a tuple of parameters for the IIR filters.

More precisely, the DDSP module in (3) converts the IIR parameters $\mathbf{\Psi}_{\theta,\phi}$ into a cascade of a first-order low-shelf (LFS) filter, K second-order peaking filters, and a first-order high-shelf (HFS) filter as illustrated in Fig. 2. Using filter index $k = 0, \ldots, K+1$, the LFS and HFS filters are parameterized by their cut frequency $f_c^{(k)}$ in Hz and their gain $g^{(k)}$ in dB, and the peaking filters are parameterized

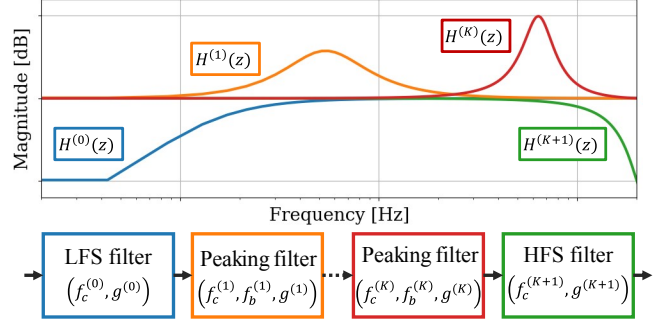[1]The ITD can be estimated using several methods [26, 27].

Fig. 2. Illustration of the cascade of parametric IIR filters.

by their center frequency $f_c^{(k)}$, their bandwidth $f_b^{(k)}$, and their gain $g^{(k)}$ following [33, 35]. From these parameters, we can compute filter coefficients $b_n^{(k)}$ and $a_n^{(k)}$ as shown in the Appendix, where all operations are fully differentiable. The transfer function of the kth filter is given by

$$H^{(k)}(z) = \frac{b_0^{(k)} z^0 + \cdots + b_N^{(k)} z^{-N}}{1 + \cdots + a_N^{(k)} z^{-N}}, \qquad (4)$$

where $N = 1$ for the shelf filters and $N = 2$ for the peaking filters. Instead of directly estimating the filter coefficients, we use parametric IIR filters to guarantee the stability of the estimated filters.

To train NF_{IIR} with the loss function in (2), we use the frequency-sampling method, i.e., uniformly sample the frequency axis into M steps to get DFT coefficients as follows:

$$H_{\theta,\phi}[m] = \prod_{k=0}^{K+1} H_{\theta,\phi}^{(k)}(z_m), \qquad (5)$$

where $z_m = \exp(2\pi j m/M)$, and j is the imaginary unit[2].

Once NF_{IIR} is trained, we can compute the IIR filters for arbitrary directions and apply the filters in the time domain. The proposed method has the potential to reduce the memory footprint because IIR filters can better model acoustic systems with fewer coefficients than FIR filters, such as those obtained through the existing NF-based method [21]. We also expect the parameters of the IIR filters to be easier to interpolate than the magnitude response.

2.3. NF for estimating IIR filter parameters

For NF_{IIR} in (3), we use a multi-layer perceptron (MLP) with the GeLU activation functions on the random Fourier features (RFF) [37]. In detail, the sound source direction (θ, ϕ) is converted to the RFF as follows:

$$\mathbf{e} = [\cos([\mathbf{p}_0^{\mathsf{T}} \mathbf{d}), \sin(\mathbf{p}_0^{\mathsf{T}} \mathbf{d}), \ldots, \cos(\mathbf{p}_{C-1}^{\mathsf{T}} \mathbf{d}), \sin(\mathbf{p}_{C-1}^{\mathsf{T}} \mathbf{d})]^{\mathsf{T}}, \qquad (6)$$

where $\mathbf{d} = [\theta - \pi, \phi]^{\mathsf{T}}$, $\mathbf{p}_c \in \mathbb{R}^2$ is sampled from an isotropic Gaussian distribution, $c = 0, \ldots, C-1$ is the channel index, and $(\cdot)^{\mathsf{T}}$ denotes the transpose. In our preliminary experiments, the use of RFF stabilized the learning curve compared with SIREN [38] and slightly improved the performance from the vanilla positional encoding [22].

[2]DFT coefficients in (5) technically correspond to a time-aliased version of the filter response [36]. That error can however be well mitigated by using a large enough M.

At the last layer of the MLP, we employ the sigmoid function to control the frequency range of $f_c^{(k)}$ as follows:

$$f_c^{(k)}(x_c^{(k)}) = (f_{\max}^{(k)} - f_{\min}^{(k)}) \operatorname{sigmoid}(x_c^{(k)}) + f_{\min}^{(k)}, \qquad (7)$$

where $x_c^{(k)}$ is an element of the output of the last affine layer. We design ranges $[f_{\min}^{(k)}, f_{\max}^{(k)}]$ by first finding the frequencies of all peaks and notches in the magnitude response of all HRTFs in a training dataset, sorting these frequencies, and using quantiles to split the interval in equally distributed ranges while ensuring neighboring ranges have 50% overlap. By assigning different frequency ranges to each of the K peak filters, we avoid having a filter cascade with several peaking filters clustered around the same spectral peak. Meanwhile, the range of $f_b^{(k)}$ is fixed as

$$f_b^{(k)}(x_b^{(k)}) = (\tilde{f}_{\max} - \tilde{f}_{\min}) \operatorname{sigmoid}(x_b^{(k)}) + \tilde{f}_{\min}, \qquad (8)$$

where $x_b^{(k)}$ is another element of the last affine layer's output. We do not use any activation functions to obtain $g^{(k)}$ as it is unconstrained.

2.4. Efficient adaption of NF to new subjects

We have so far described an NF for a single subject. As HRTFs are similar across different subjects, we expect that an NF with a small number of subject-specific parameters can efficiently represent the HRTFs of multiple subjects, and define

$$\boldsymbol{\Psi}_{\theta,\phi} = \mathrm{NF}_{\mathrm{IIR}}(\theta, \phi \mid i), \qquad (9)$$

where i is introduced as a subject index to change subject-specific parameters. A popular approach to conditioning NFs is conditioning by concatenation (CbC) [23]. CbC introduces a subject embedding $\mathbf{z}_i$ and concatenates it to the model input; it has been used in HRTF spatial upsampling [21]. Another approach is feature-wise linear modulation (FiLM) [39], which uses another neural network to compute modulation parameters $(\boldsymbol{\mu}_{l,i}, \boldsymbol{\sigma}_{l,i})$ for each layer l from $\mathbf{z}_i$. The output of layer l is then modified from $\mathrm{GeLU}(\mathbf{A}_l \mathbf{x}_l + \mathbf{b}_l)$ to

$$\mathrm{FiLM}(\mathbf{x}_l \mid i) = \boldsymbol{\sigma}_{l,i} \odot \mathrm{GeLU}(\mathbf{A}_l \mathbf{x}_l + \mathbf{b}_l) + \boldsymbol{\mu}_{l,i}, \qquad (10)$$

where $\odot$ denotes the Hadamard product, $\mathbf{x}_l$ is the input of the layer l, and $\mathbf{A}_l$ and $\mathbf{b}_l$ are parameters of the MLP shared across subjects.

The subject embedding $\mathbf{z}_i$ can be optimized by back-propagation similar to other parameters, a method referred to as auto-decoding [23]. To personalize the NF to a new subject, we optimize only $\mathbf{z}_i$ to minimize the MSE in (2) while freezing the shared parameters. Auto-decoding can easily handle an arbitrary number of measurements of the target subject because it does not need additional neural networks to extract $\mathbf{z}_i$ from the measurements.

We also explore two parameter-efficient fine-tuning methods: bias-terms fine-tuning (BitFit) [40] and low-rank adaptation (LoRA) [41]. BitFit adapts a neural network to a new domain by fine-tuning the bias term $\mathbf{b}_l$ instead of all parameters. Here, we introduce a different bias term $\mathbf{b}_{l,i}$ for each subject: $\mathrm{BitFit}(\mathbf{x}_l \mid i) = \mathrm{GeLU}(\mathbf{A}_l \mathbf{x}_l + \mathbf{b}_{l,i})$. LoRA introduces an additional rank-1 matrix for each layer as a product of two vectors[3]:

$$\mathrm{LoRA}(\mathbf{x}_l \mid i) = \mathrm{GeLU}(\mathbf{A}_l \mathbf{x}_l + \mathbf{u}_{l,i} \mathbf{v}_{l,i}^\top \mathbf{x}_l + \mathbf{b}_l), \qquad (11)$$

where $\mathbf{u}_{l,i}$ and $\mathbf{v}_{l,i}$ are the subject-specific vectors for building the rank-1 matrix. After optimizing the rank-1 matrix for each target

---

[3]In general, LoRA can also leverage low-rank matrices whose rank is higher than 1. In this paper, we focus on the case with rank-1 for simplicity.
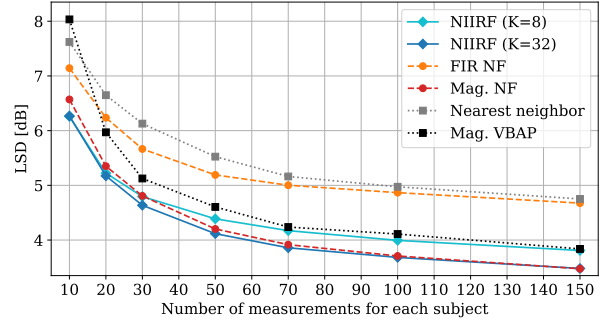

Fig. 3. LSD with different number of measurements for training.

subject, we can add it to the original weight $\mathbf{A}_l$, which reduces the computational cost to obtain the HRTFs for the same subject.

3. EXPERIMENTS

3.1. Comparison with existing single-subject methods

We first validate the interpolation accuracy of the proposed method under a single-subject setting. We used the CIPIC dataset [42], which includes 45 subjects and 1250 measurements per subject. The measurements were randomly split into three sets: 1000 for evaluation, 100 for validation, and 150 for training. The sampling frequency was 44.1 kHz, and we used the DFT with 512 points to compute $\widetilde{H}_{\theta,\phi}[m]$ in (2). The log-spectral distortion (LSD), defined as the square root of $\mathcal{L}$ in (2) measured from 20 Hz to 20 kHz, was used to perform evaluation on the audible frequency range (lower is better).

Our NF ("NIIRF") consists of four hidden layers with 512 units and an output layer whose size depends on the number of peaking filters K. The NF was trained using the RAdam optimizer [43] with a learning rate of 5×10^{-4}. We used the model with the best loss $\mathcal{L}$ on the validation set for evaluation. We compared the proposed method with two baselines: the nearest neighbor algorithm ("Nearest neighbor") and a variant of VBAP [11] ("Mag. VBAP"). In the latter baseline, we performed a weighted sum of the magnitude responses in the training dataset, where the weight is calculated following [11]. NF-based methods estimating the magnitude response [21] ("Mag. NF") and the time-domain FIR filter ("FIR NF") were also evaluated. They differ from the proposed method only by the output layer.

Figure 3 illustrates the variations in the LSD averaged over 45 subjects depending on the number of HRTF measurements available for training. The proposed method with 8 filters achieved comparable performance with "Mag. VBAP," even outperforming it when measurements are sparse. The NF-based method for the FIR filter performed poorly compared with that for the magnitude response. This is likely because variation in time domain impulse responses across different directions is much harder to approximate with a fixed model size compared to variations in their spectra. As a result of changing the estimation target to the parameters of the IIR filters, the proposed method with 32 peaking filters slightly improved the performance when the number of HRTF measurements was limited.

Figure 4 shows examples of the responses of the left-channel IIR filter in the frequency and time domains, where we used the model trained for subject 3 with 150 measurements[4]. The time-domain fil-

---

[4]In Fig. 4, we shift the estimated filter response so that its correlation with the target impulse response is maximized, which corresponds to compensating for the transmission delay from a sound source to the ear [27]. Several methods have been developed to interpolate the delay [44].
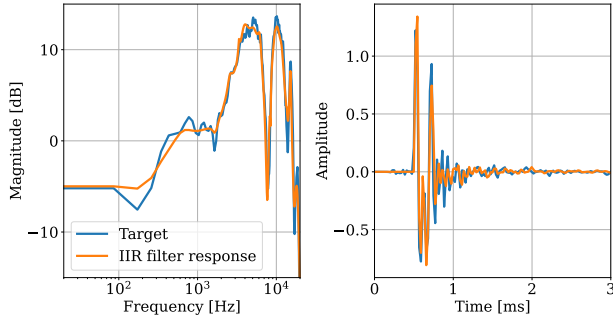
Fig. 4. Estimated filter response in the frequency and time domains.

Table 1. LSD [dB] for different numbers of measurements for adaptation. The number of the subject-specific parameters adapted to the target subject is indicated by Q.

Method	Adaptation	Q	10	20	30	50	100
			Number of measurements				
Directions seen in pre-training (Test1)							
	CoC [21]	32	4.8	4.7	4.6	4.6	4.5
Mag. NF	FiLM	32	**4.3**	4.2	4.2	4.1	4.1
	BitFit	2562	**4.3**	**4.0**	3.9	3.7	**3.5**
	LoRA	5122	**4.3**	**4.0**	**3.8**	**3.6**	**3.5**
	CoC	32	4.8	4.7	4.7	4.6	4.6
NIIRF ($K = 32$)	FiLM	32	4.3	4.2	4.2	4.2	4.2
Proposed	BitFit	2248	**4.3**	**4.0**	3.9	3.7	**3.5**
	LoRA	4808	**4.3**	**4.0**	3.9	3.7	**3.5**
Directions unseen completely (Test2)							
	CoC [21]	32	4.9	4.8	4.8	4.8	4.7
Mag. NF	FiLM	32	**4.5**	**4.4**	4.4	4.4	4.3
	BitFit	2562	5.0	4.8	4.6	4.4	4.4
	LoRA	5122	5.2	5.0	4.9	4.8	4.6
	CoC	32	5.0	4.9	4.9	4.8	4.7
NIIRF (K=32)	FiLM	32	**4.5**	4.5	4.4	4.4	4.4
Proposed	BitFit	2248	4.8	4.5	4.4	4.2	4.1
	LoRA	4808	4.7	**4.4**	**4.2**	**4.1**	**4.0**

ter response captured the shape of the head-related impulse response even though the NF was trained on the phase insensitive loss in (2). This is a favorable feature of the cascaded IIR filters [27, 28].

3.2. Adaptation of multi-subject NF to new subjects

We use the HUTUBS dataset [45] for evaluating adaptation of an NF pre-trained on multiple subjects to new subjects. The dataset contains 94 subjects excluding two repeated subjects [18], and 440 measurements per subject. The first 87 subjects were used for pre-training, and the remaining 7 subjects were used for evaluating the adaptation capability. 100 randomly selected directions are held out from all 94 subjects, and used to form an unseen direction test set ("Test2"). An additional 100 directions are held out from the 7 adaptation subjects to form a seen direction test set ("Test 1"), and also held out from 10 of the training subjects for use as a validation set. All remaining directions are used for either pre-training or adaptation. The NFs were pre-trained on the HRTFs of multiple subjects using the RAdam optimizer, and then the subject-specific parameters were optimized for each target subject with the AdamW optimizer.

Table 1 presents the LSD with different adaptation methods. As a point of comparison, we trained a single-subject model as used in Section 3.1 for each of the 7 target subjects. For the directions included in the multi-subject pre-training (Test1), the average performance over all subjects was 7.5 dB when training on 10 measurements, and 5.1 dB when training on 100 measurements. We see from Table 1 that all the adaptation methods are effective for HRTF personalization and dramatically improve upon these results. NF-based methods estimating the magnitude response and the parametric IIR filters resulted in a similar performance on Test1, with BitFit and LoRA performing best. However, the proposed NIIRF with LoRA outperformed the existing NF-based method in most cases on Test2. We hypothesize that generalization is better for NIIRF because values specifying parametric filters such as filter center frequencies and gains are more predictable across directions than spectra.

4. CONCLUSION

We proposed to integrate an NF approach and a cascade of differentiable IIR filters for HRTF modeling. The NF is trained to minimize the error of the magnitude response based on DDSP. We demonstrate that the proposed method improves the upsampling accuracy when the number of HRTF measurements is limited. We also investigate various adaptation methods to personalize the NF to a new subject. Our source code is available online[5].

5. APPENDIX: IIR COEFFICIENT COMPUTATION

Following [33, 35], the transfer functions of the first-order LFS and HFS filters are defined as in Table 2, where $\eta^{(k)} = (\rho^{(k)} - 1)/2$, $\rho^{(k)} = 10^{g^{(k)}/20}$, and f_s is the sampling rate. The transfer function of the second-order peaking filter is summarized in Table 3.

Table 2. Coefficients of the first-order shelf filters.

	LFS	HFS
$b_0^{(k)}$	$1 + \eta^{(k)}(1 + \alpha^{(k)})$	$1 + \eta^{(k)}(1 - \alpha^{(k)})$
$b_1^{(k)}$	$\alpha^{(k)} + \eta^{(k)}(\alpha^{(k)} + 1)$	$\alpha^{(k)} + \eta^{(k)}(\alpha^{(k)} - 1)$
$a_1^{(k)}$	$\alpha^{(k)}$	$\alpha^{(k)}$
$\alpha^{(k)}$ if $g^{(k)} \geq 0$	$\frac{\tan(\pi f_c^{(k)}/f_s) - 1}{\tan(\pi f_c^{(k)}/f_s) + 1}$	$\frac{\tan(\pi f_c^{(k)}/f_s) - 1}{\tan(\pi f_c^{(k)}/f_s) + 1}$
$\alpha^{(k)}$ if $g^{(k)} < 0$	$\frac{\tan(\pi f_c^{(k)}/f_s) - \rho^{(k)}}{\tan(\pi f_c^{(k)}/f_s) + \rho^{(k)}}$	$\frac{\rho^{(k)} \tan(\pi f_c^{(k)}/f_s) - 1}{\rho^{(k)} \tan(\pi f_c^{(k)}/f_s) + 1}$

Table 3. Coefficients of the second-order peaking filters.

	Peaking filter
$b_0^{(k)}$	$1 + \eta^{(k)}(1 + \beta^{(k)})$
$b_1^{(k)}$	$\gamma^{(k)}(1 - \beta^{(k)})$
$b_2^{(k)}$	$-\beta^{(k)} - \eta^{(k)}(1 + \beta^{(k)})$
$a_1^{(k)}$	$\gamma^{(k)}(1 - \beta^{(k)})$
$a_2^{(k)}$	$-\beta^{(k)}$
$\beta^{(k)}$ if $g^{(k)} \geq 0$	$\frac{\tan(\pi f_b^{(k)}/f_s) - 1}{\tan(\pi f_b^{(k)}/f_s) + 1}$
$\beta^{(k)}$ if $g^{(k)} < 0$	$\frac{\tan(\pi f_b^{(k)}/f_s) - \rho^{(k)}}{\tan(\pi f_b^{(k)}/f_s) + \rho^{(k)}}$
$\gamma^{(k)}$	$-\cos\left(2\pi \frac{f_c^{(k)}}{f_s}\right)$

6. 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, 2007.

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

[3] S. Serafin, M. Geronazzo, C. Erkut, N. C. Nilsson, and Nordahl R., "Sonic interactions in virtual reality: State of the art, current challenges, and future directions," *IEEE Comput. Graph., Appl.*, vol. 38, no. 2, pp. 31–43, 2018.

[4] 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.

[5] J. S. Andersen, R. Miccini, S. Serafin, and S. Spagnol, "Evaluation of individualized hrtfs in a 3D shooter game," in *Proc. I3DA*, 2021, pp. 1–10.

[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, 2014.

[7] H. Gamper, "Head-related transfer function interpolation in azimuth, elevation, and distance," *J. Acoust. Soc. Am.*, vol. 134, no. 6, pp. EL547–EL553, 2013.

[8] D. Y. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *Proc. WASPAA*, 2003, pp. 157–160.

[9] H. Hu, L. Zhou, H. Ma, and Z. Wu, "HRTF personalization based on artificial neural network in individual virtual auditory space," *Appl. Acout.*, vol. 69, no. 2, pp. 163–172, 2008.

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

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

[12] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs [head related transfer functions]," in *Proc. ICASSP*, 2004, vol. IV, pp. 45–48.

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

[14] 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, 2021.

[15] 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, 2015.

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

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

[18] 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*, 2022.

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

[20] 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*, 2021, pp. 3385–3389.

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

[22] 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, 2022.

[23] 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," *Comput. Graph. Forum*, vol. 41, no. 2, pp. 641–676, 2022.

[24] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," in *Proc. NeurIPS*, 2022, pp. 3165–3177.

[25] 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," *arXiv:2305.04447*, 2023.

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

[27] G. Ramos and M. Cobos, "Parametric head-related transfer function modeling and interpolation for cost-efficient binaural sound applications," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 1735–1738, 2013.

[28] P. Nowak and U. Zölzer, "Spatial interpolation of HRTFs approximated by parametric IIR filters," in *Proc. DAGA*, 2022.

[29] B. Kuznetsov, J. D. Parker, and F. Esqueda, "Differentiable IIR filters for machine learning applications," in *Proc. DAFX*, 2020, pp. 297–303.

[30] S. Nercessian, "Neural parametric equalizer matching using differentiable biquads," in *Proc. DAFX*, 2020, pp. 265–272.

[31] S. Nercessian, A. Sarroff, and K. J. Werner, "Lightweight and interpretable neural modeling of an audio distortion effect using hyperconditioned differentiable biquads," in *Proc. ICASSP*, 2021, pp. 890–894.

[32] J. Engel, C. Gu, A. Roberts, et al., "DDSP: Differentiable digital signal processing," in *Proc. ICLR*, 2019.

[33] P. Bhattacharya, P. Nowak, and U. Zölzer, "Optimization of cascaded parametric peak and shelving filters with backpropagation algorithm," *Proc. DAFX*, pp. 101–108, 2020.

[34] J. T. Colonel, C. J. Steinmetz, M. Michelen, and J. D. Reiss, "Direct design of biquad filter cascades with deep learning by sampling random polynomials," in *Proc. ICASSP*, 2022, pp. 3104–3108.

[35] U. Zölzer, *Digital audio signal processing*, John Wiley & Sons, 2022.

[36] L. Rabiner, B. Gold, and C. McGonegal, "An approach to the approximation problem for nonrecursive digital filters," *IEEE Trans. Audio Electroacust.*, vol. 18, no. 2, pp. 83–106, 1970.

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

[38] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proc. NeurIPS*, 2020, pp. 7462–7473.

[39] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018, pp. 3942–3951.

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

[41] 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*, 2022.

[42] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. WASPAA*, 2001, pp. 99–102.

[43] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. ICLR*, 2020.

[44] F. Brinkmann, R. Roden, A. Lindau, and S. Weinzierl, "Audibility and interpolation of head-above-torso orientation in binaural technology," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 931–942, 2015.

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