Graph Neural Field with Spatial-Correlation Augmentation for HRTF Personalization

De Hu[1, *], **Junsheng Hu**[1], **Cuicui Jiang**[1]

[1]College of Computer Science, Inner Mongolia University, China
cshood@imu.edu.cn, hujunsheng@mail.imu.edu.cn, jiangcuicui@mail.imu.edu.cn

Abstract

To achieve immersive spatial audio rendering on VR/AR devices, high-quality Head-Related Transfer Functions (HRTFs) are essential. In general, HRTFs are subject-dependent and position-dependent, and their measurement is time-consuming and tedious. To address this challenge, we propose the Graph Neural Field with Spatial-Correlation Augmentation (GraphNF-SCA) for HRTF personalization, which can be used to generate individual HRTFs for unseen subjects. The GraphNF-SCA consists of three key components: an HRTF personalization (HRTF-P) module, an HRTF upsampling (HRTF-U) module, and a fine-tuning stage. In the HRTF-P module, we predict HRTFs of the target subject via the Graph Neural Network (GNN) with an encoder-decoder architecture, where the encoder extracts universal features and the decoder incorporates the target-relevant features and produces individualized HRTFs. The HRTF-U module employs another GNN to model spatial correlations across HRTFs. This module is fine-tuned using the output of the HRTF-P module, thereby enhancing the spatial consistency of the predicted HRTFs. Unlike existing methods that estimate individual HRTFs position-by-position without spatial correlation modeling, the GraphNF-SCA effectively leverages inherent spatial correlations across HRTFs to enhance the performance of HRTF personalization. Experimental results demonstrate that the GraphNF-SCA achieves state-of-the-art results.

Code — https://github.com/hu-junsheng/GraphNF-SCA
Extended version — https://arxiv.org/abs/2511.10697

Introduction

Spatial audio rendering finds various applications, including virtual reality (VR) (Johansson 2019; Zotter and Frank 2019), augmented reality (AR) (Sundareswaran et al. 2003; Yang and Mattern 2019), teleconferencing (Ramos et al. 2017), and hearing assistive devices (Du et al. 2023; Vickers et al. 2021). To generate high-quality immersive audio over headphones, head-related impulse responses (HRIRs), known as head-related transfer functions (HRTFs) in the frequency domain, are often used to simulate the spatial filtering effects, as sound travels from its source to each ear. When virtual source positions are given, binaural signals can be synthesized by the direct convolution of HRIRs and source signals.

In general, accurate HRTFs can be obtained by practical measurement. For example, the CIPIC database (Algazi et al. 2001) collected HRTF data from subjects in multiple spatial orientations in an anechoic chamber, by placing miniature microphones in the subjects' ear canals and utilizing a robotic arm to control the loudspeaker. However, acquiring HRTFs through physical measurements is time-consuming and labor-intensive, thereby limiting the scale of existing HRTF datasets (Sridhar, Tylka, and Choueiri 2017; Bomhardt, de la Fuente Klein, and Fels 2016; Carpentier et al. 2014). To alleviate this issue, `HRTF upsampling` is explored to generate dense HRTFs from sparsely measured HRTFs (Pörschmann, Arend, and Brinkmann 2019). Furthermore, as depicted in Figure 1, HRTFs are different for each subject due to their close relationship with anatomical traits (e.g., torso, head, and pinnae). As a result, the use of existing HRTF datasets for new users may lead to anatomical mismatches, which in turn degrade the spatial audio experience (Simon, Zacharov, and Katz 2016; Jenny, Reuter et al. 2020). To avoid re-measurement of HRTFs for each new user, `HRTF personalization` is employed to generate HRTFs matching the individual's anatomy via existing datasets.

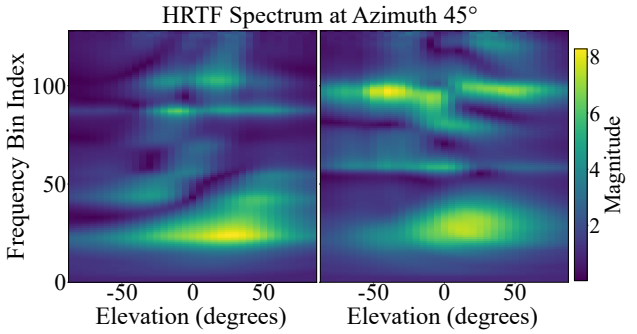

Figure 1: Illustration of position-dependent and subject-dependent nature of HRTFs: The magnitude of HRTFs corresponding to subjects PP1 (left) and PP3 (right) at an azimuth angle of 45°, obtained from the HUTUBS database (Brinkmann et al. 2019).

*Corresponding Author

Related Works

HRTF Personalization (HRTF-P) To generate individual HRTFs for unseen subjects, several HRTF-P methods have been proposed. The most straightforward approaches (Zhou, Jiang, and Ithapu 2021; Teng and Zhong 2023; Ko et al. 2023) were to directly learn the mapping from anatomical features to individual HRTFs. For example, 3D head scans were utilized as anthropometric features in (Zhou, Jiang, and Ithapu 2021), while anthropometric parameters (about head, torso, and pinna) were adopted in (Teng and Zhong 2023; Ko et al. 2023). Nevertheless, learning direct anthropometric-to-HRTF mappings faces fundamental challenges in data-scarce scenarios, due to the fact that the underlying wave-anatomy interactions (e.g., diffraction, scattering, reflection, and resonance) exhibit strong nonlinearity (Takemoto et al. 2012; Algazi et al. 2002; Fantini et al. 2025). An alternative scheme is to estimate the individual HRTF without direct mapping, but by retrieving and synthesizing anatomically similar subjects' HRTFs from a database (Masuyama et al. 2025, 2024; Zotkin et al. 2003; Geronazzo et al. 2019). The simplest retrieval-based approaches (Zotkin et al. 2003; Geronazzo et al. 2019; Katz and Parseihian 2012) output a best-match HRTF by searching the subject with the most similar anthropometric features in a dataset. In such a manner, it is unable to generalize the relationship between the input and the HRTF, limiting the performance in small-scale datasets (Brinkmann et al. 2019; Pelzer et al. 2020). Instead of using individual anatomy, Masuyama et al. (Masuyama et al. 2025) proposed a retrieval-augmented neural field (RANF) to carry out HRTF personalization, in which the interaural time difference (ITD) features were employed to retrieve multiple subjects from a given HRTF dataset. This method requires measuring the target user's HRTFs from only 3-5 directions to calculate ITD features, dispensing with time-consuming anthropometric feature acquisition. Unfortunately, like other HRTF-P approaches mentioned above, it cannot generate personalized HRTFs at non-sampled positions (that are outside the given dataset). Furthermore, the estimation accuracy of personalized HRTFs is still limited due to the complicated wave-anatomy interactions.

HRTF Upsampling (HRTF-U) To generate HRTFs at unsampled positions, HRTF-U can be achieved based on traditional spatial interpolation methods, such as barycentric interpolation (David and Katz 2018; Hartung, Braasch, and Sterbing 1999) and spherical harmonic interpolation (Evans, Angus, and Tew 1998; Arend, Brinkmann, and Pörschmann 2021; Engel, Goodman, and Picinali 2022). These methods achieve high interpolation accuracy if the measurement points in the adopted HRTF dataset are distributed densely and evenly. This is because HRTFs tend to be similar for points that are very close to each other, as shown in Figure 1. However, existing databases are typically limited in scale, leading to significant performance degradation in conventional interpolation methods. With the rapid development of deep learning, data-driven HRTF interpolations have gained increasing attention (Hogg et al. 2024; Gebru et al. 2021). Recently, Lee et al. (Lee, Lee, and Lee 2023) interpolated HRTFs by mapping those HRTFs from neighboring points to target points via the neural network that was built upon Feature-wise Linear Modulation (FiLM). Since the input size is required to be fixed in such an architecture, a repeated sampling strategy was used to select a fixed number of neighbors around the target point. Alternatively, the state-of-the-art work (Hu et al. 2025) introduced a dual-graph attention network with variable input sizes, which shows the strong capability of graph neural networks in modeling spatial correlation among HRTFs. It can theoretically predict subject-specific HRTFs at any position, but cannot personalize them for unseen subjects.

Comparison and Analysis As discussed above, the existing HRTF generation approaches can be divided into two classes, i.e., HRTF-P and HRTF-U. For a given HRTF dataset containing position-specific and subject-specific measurements, HRTF-P estimates personalized HRTFs for arbitrary subjects at sampled positions, while HRTF-U predicts HRTFs at arbitrary positions for known subjects. By comparison, HRTF-U often performs well since it focuses more on the spatial correlation among HRTFs. In contrast, HRTF-P demonstrates constrained performance due to the inherent complexity of modeling wave-anatomy interactions. Until now, to the best of our knowledge, HRTF-P has always been modeled independently from HRTF-U, which may fail to capture the geometric relationship among personalized HRTFs. *This naturally raises a critical question: Could developing a collaborative mechanism between HRTF-P and HRTF-U enhance the HRTF-P performance?*

Contribution

Motivated by the above problem, we propose a novel HRTF-P framework (termed GraphNF-SCA) in this work. The contributions are threefold. First, we develop a deep learning model for HRTF-P via graph neural networks (GNNs), which employs an encoder to extract universal HRTF features and then integrates target subject-specific characteristics through a decoder to generate personalized HRTFs. Second, we construct an HRTF-U module based on another GNN, which comprehensively learns the geometric correlations of HRTFs across spatial positions. Third, and most critically, we feed the output of the HRTF-P module into the fine-tuned HRTF-U module, which reinforces the initially position-independent personalized HRTFs by leveraging spatial correlations. Experimental results demonstrate that the estimation accuracy of spatial-correlation augmented HRTFs is significantly higher than that of position-independent baseline methods. These findings validate our initial hypothesis that integrating HRTF-U can effectively enhance HRTF-P performance.

Preliminaries

Problem Statement

Our objective is to generate personalized HRTFs for unseen subjects, using a small-scale dataset with subject-specific and position-specific HRTF measurements. Due to the fact that the phase of HRTFs can be effectively reconstructed using the minimum-phase approximation (Cuevas Rodríguez et al. 2019; Hogg et al. 2024; Masuyama et al. 2024), this work focuses only on the magnitude of HRTFs. Let

$\mathbf{H}_s^d \in \mathbb{R}^{2K}$ be the HRTF magnitude of subject s at direction $d = (\theta, \varphi)$, where $\theta \in [0, 2\pi)$ and $\varphi \in [-\pi/2, \pi/2]$ are the azimuth angle and the elevation angle, respectively, and K is the number of frequency bins per ear. For a given HRTF dataset, we have $s \in \mathcal{S}$ and $d \in \mathcal{D}$ with $\mathcal{S}$ and $\mathcal{D}$ being the subject set and the direction set, respectively.

For an unseen subject $\hat{s} \notin \mathcal{S}$, HRTF-P can be carried out in a retrieval-augmented manner, by learning a mapping

$$\Phi_p \left(\mathbf{H}_{s \in \mathcal{N}_{\hat{s}}}^d \right) \mapsto \mathbf{H}_{\hat{s}}^d, \ \hat{s} \notin \mathcal{S} \tag{1}$$

where $\mathcal{N}_{\hat{s}}$ is the neighboring set of subject $\hat{s}$, which can be acquired by retrieving subjects that have an anatomical feature similar to the target subject, i.e.,

$$\mathcal{N}_{\hat{s}} = \{ s \mid \|a_s - a_{\hat{s}}\|_2 < \delta_s \} \tag{2}$$

where a_s is the subject-relevant characteristic of subject s, and $||\cdot||_2$ represents the l_2 norm. In the state-of-the-art work (Masuyama et al. 2025), a_s is determined by the ITD (if the ITDs of the target subject are measured at a small number of directions in advance) rather than human anatomy.

For an unmeasured spatial position $\hat{d} \notin \mathcal{D}$, recent advances (Hu et al. 2025; Lee, Lee, and Lee 2023) in HRTF-U are also moving toward using the retrieval-augmented strategy to learn the following mapping

$$\Phi_u \left(\mathbf{H}_s^{d \in \mathcal{N}_{\hat{d}}} \right) \mapsto \mathbf{H}_s^{\hat{d}}, \ \hat{d} \notin \mathcal{D} \tag{3}$$

where $\mathcal{N}_{\hat{d}}$ is the neighboring set of point $\hat{d}$, which can be obtained by retrieving the points near the target point, i.e.,

$$\mathcal{N}_{\hat{d}} = \left\{ d \mid \|d - \hat{d}\|_2 < \delta_d \right\} \tag{4}$$

where $\delta_d > 0$ is a proper threshold.

Note that HRTF-P and HRTF-U in (1) and (3) are modeled independently, where the former exhibits the effect of human anatomy on HRTFs while the latter reflects the effect of source directions on HRTFs. *Typically, mapping (3) can be learned more easily, whereas mapping (1) presents challenges due to complicated wave-anatomy interactions.* To this end, we construct a novel mapping Ψ_p:

$$\Psi_p : \boxed{\tilde{\Phi}_u \left(\Phi_p(\mathbf{H}_{s \in \mathcal{N}_{\hat{s}}}^d) \mapsto \mathbf{H}_{\hat{s}}^d, d \in \mathcal{N}_d \right) \mapsto \mathbf{H}_{\hat{s}}^d, \ \hat{s} \notin \mathcal{S}} \tag{5}$$

where $\tilde{\Phi}_u$ is the fine-tuning version of Φ_u, which is used to reinforce Φ_p by taking advantage of the spatial relationship among $\mathbf{H}_{\hat{s}}^{d \in \mathcal{N}_d}$.

Method

In this work, the graph neural network (GNN) is adopted, due to the fact that

- GNNs are well suited for modeling the relationship between nodes based on their local similarities, e.g., anthropometric features in (2) and spatial distances in (4).
- GNN supports inputs with non-fixed sizes, which allows for retrieving sets with flexible sizes in (2) and (4).

An overview of the proposed HRTF personalization framework is shown in Figure 2, which consists of three core parts: an HRTF-P module, an HRTF-U module, and a fine-tuning stage. Specifically, the HRTFs of the subjects retrieved in (2) are fed into the GNN-based HRTF-P module that is established using the graph-based network architecture (Figure 3). At the same time, HRTFs corresponding to the positions retrieved in (4) are fed into the GNN-based HRTF-U module that is constructed using another GNN-based architecture (Figure 4). After pretraining the above two modules, the output $\mathbf{H}_{\hat{s}}^d (d \in \mathcal{N}_d)$ of the HRTF-P module is used to fine-tune the HRTF-U module, and the latter can generate high-quality personalized HRTFs reinforced by spatial correlations. The technical details of each part are described in the following subsections.

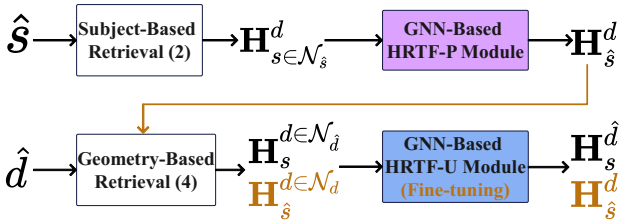

Figure 2: Overview of the proposed HRTF personalization framework (referred to as GraphNF-SCA), where the pretrained HRTF-U module is fine-tuned to reinforce the spatial correlation among the outputs of the HRTF-P module.

GNN-Based HRTF-P Module

As shown in Figure 3, the GNN-based HRTF-P module is composed of a pre-processing unit followed by an encoder-decoder architecture. To be specific, the pre-processing unit prepares inputs for the encoder that extract subject-shared features, while the decoder outputs the personalized HRTFs after embedding subject-specific features.

Pre-processing Unit For a target subject $\hat{s}$, we first construct a graph $\mathcal{G}_{\hat{s}}$ as

$$\mathcal{G}_{\hat{s}} = \{ \mathcal{V}_{\hat{s}}, \mathcal{E}_{\hat{s}}, \mathcal{W}_{\hat{s}} \} \tag{6}$$

where $\mathcal{V}_{\hat{s}}$, $\mathcal{E}_{\hat{s}}$, and $\mathcal{W}_{\hat{s}}$ are the vertex set, the edge set, and the edge-weight set, which are defined by

$$\mathcal{V}_{\hat{s}} = \{ \mathbf{H}_s^d \mid s \in \mathcal{N}_{\hat{s}} \} \tag{7a}$$

$$\mathcal{E}_{\hat{s}} = \{ (s, q) \mid s, q \in \mathcal{N}_{\hat{s}} \} \tag{7b}$$

$$\mathcal{W}_{\hat{s}} = \{ 1 \mid (s, q) \in \mathcal{E}_{\hat{s}} \} \ . \tag{7c}$$

That is, $\mathcal{G}_{\hat{s}}$ is set to a fully-connected graph for mining subject-shared features.

In addition to (6), we also concatenate the source direction d and the subject-relevant characteristic a_s to build the following clue

$$\mathbf{C}_s^d = d \oplus a_s, \ s \in \mathcal{N}_{\hat{s}} \tag{8}$$

with $\oplus$ being the concatenation operator.

Encoder The encoder is dedicated to extract the HRTF-relevant feature $\tilde{\mathbf{H}}_s^d \in \mathbb{R}^{|\mathcal{N}_{\hat{s}}| \times 2K * N}$ from graph $\mathcal{G}_{\hat{s}}$ and the clue-relevant feature $\tilde{\mathbf{C}}_s^d \in \mathbb{R}^{|\mathcal{N}_{\hat{s}}| \times 2K}$ from $\mathbf{C}_s^d$, then fuse them to output the subject-shared feature $\tilde{\mathbf{F}}^d \in \mathbb{R}^{N \times 2K}$, where N represents the number of attention heads used in

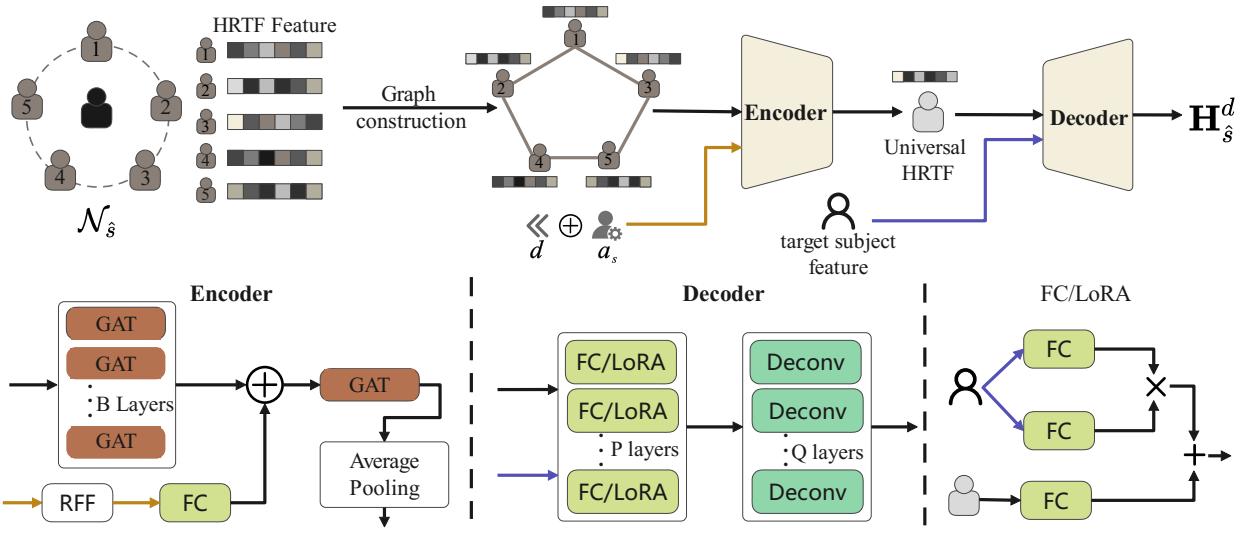

Figure 3: Network architecture for GNN-based HRTF-P module.

the multi-head attention mechanism of the first graph attention (GAT) layer.

The HRTF-relevant feature $\tilde{\mathbf{H}}_s^d$ is derived from the vertex embeddings of the graph $\mathcal{G}_{\hat{s}}$ transformed by B successive GAT layers. In each GAT layer, we concatenate N attention heads activated by the Exponential Linear Unit (ELU) function (Clevert, Unterthiner, and Hochreiter 2020). Taking the first layer as an example, its output $\tilde{\mathbf{H}}_s^d(1)$ can be expressed as

$$\tilde{\mathbf{H}}_s^d(1) = \bigoplus_{n=1}^{N} f_e\Big(\alpha_{s,s}^{(n)} \mathbf{W}_s^{(n)} \tilde{\mathbf{H}}_s^d + \sum_{q \in \mathcal{N}_{\hat{s}}} \alpha_{s,q}^{(n)} \mathbf{W}^{(n)} \tilde{\mathbf{H}}_q^d\Big) \quad (9)$$

where $s \in \mathcal{N}_{\hat{s}}$ is the node index, $f_e(\cdot)$ represents the ELU activation function, $\mathbf{W}_s^{(n)} \in \mathbb{R}^{M \times 2K}$ and $\mathbf{W}^{(n)} \in \mathbb{R}^{M \times 2K}$ are two learnable transformation matrices in the n-th attention head with M being the output feature dimension, and $\alpha_{s,q}^{(n)}$ represents the attention coefficient between nodes s and q, i.e.,

$$\alpha_{s,q}^{(n)} = f\left[f_l(\mathbf{a}_s^{(n)\top} \mathbf{W}_s^{(n)} \tilde{\mathbf{H}}_s^d + \mathbf{a}^{(n)\top} \mathbf{W}^{(n)} \tilde{\mathbf{H}}_q^d)\right] \quad (10)$$

where $\top$ denotes the transpose operator, $\mathbf{a}_s^{(n)} \in \mathbb{R}^M$ and $\mathbf{a}^{(n)} \in \mathbb{R}^M$ are two learnable weight vectors, $f_l(\cdot)$ denotes the LeakyReLU activation function (Maas et al. 2013), and $f[\cdot]$ represents the Softmax function, which normalizes the computed attention coefficients to facilitate comparison between different nodes.

The clue-relevant feature $\tilde{\mathbf{C}}_s^d$ is obtained by feeding the random Fourier feature (RFF) (Tancik et al. 2020) of $\mathbf{C}_s^d$ into a fully connected (FC) layer. Here, the dimension of $\tilde{\mathbf{C}}_s^d$ is fixed to $2K$ by adjusting the dimension of the FC layer.

The feature fusion is carried out as follows: The HRTF-relevant feature $\tilde{\mathbf{H}}_s^d$ and the clue-relevant feature $\tilde{\mathbf{C}}_s^d$ are first concatenated to construct a new node feature $\mathbf{F}_s^d \in \mathbb{R}^{|\mathcal{N}_{\hat{s}}| \times (2K*N+2K)}$. By treating $\mathbf{F}_s^d$ ($s \in \mathcal{N}_{\hat{s}}$) as a vertex set, with edges and weights defined by (7b) and (7c) respectively, we construct a graph processed through a single GAT layer and average pooling, yielding the fused feature $\tilde{\mathbf{F}}^d$. After average pooling, the resulting feature $\tilde{\mathbf{F}}^d$ eliminates individual differences and is transformed into a universal HRTF that is independent of the subject. Such universal representations can then be reconstructed as personalized HRTFs through the decoder.

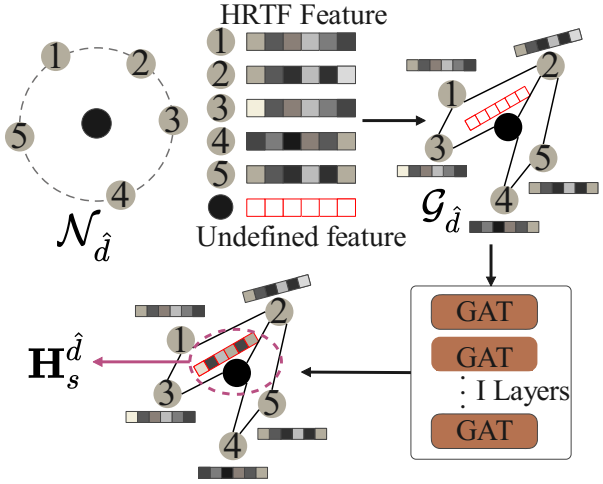

Figure 4: Network architecture for HRTF-U module.

Decoder The decoder is used to embed the target subject's feature into the universal representation $\tilde{\mathbf{F}}^d$, generating the personalized HRTF $\mathbf{H}_{\hat{s}}^d$. Motivated by (Masuyama et al. 2025), we adopt P FC layers with Low-Rank Adaptation (LoRA) (Hu et al. 2022) and Q deconvolution layers to build the decoder. The main difference from (Masuyama et al. 2025) is that we do not split $\tilde{\mathbf{F}}^d$ into HRTF-relevant and clue-relevant parts. In each layer, the target subject's features are extracted by two FC layers (as depicted in Figure 3), generating two low-rank matrices $\mathbf{u} \in \mathbb{R}^{2K \times 1}$ and $\mathbf{v} \in \mathbb{R}^{2K \times 1}$, respectively. By using these two low-rank matrices, we can adaptively adjust the fused feature $\tilde{\mathbf{F}}^d$ as

$$\bar{\mathbf{F}}^d = \tilde{\mathbf{F}}^d + \mathbf{u}\mathbf{v}^\top \quad (11)$$

where $\bar{\mathbf{F}}^d$ is generated by passing $\tilde{\mathbf{F}}^d$ through an FC layer, which has the same dimension with $\tilde{\mathbf{F}}^d$. This adaptive adjustment enables the universal feature $\tilde{\mathbf{F}}^d$ to better match the personalized characteristics of the target subject.

We adopt the log-spectral distortion (LSD) as the loss function to measure the spectral distortion between the predicted and the ground-truth HRTF. The LSD is defined as

$$LSD(\mathbf{H}_{\hat{s}}^d, \bar{\mathbf{H}}_{\hat{s}}^d) = \sqrt{\frac{1}{2K}\sum_{k=1}^{2K}\left(20\log_{10}\frac{\bar{\mathbf{H}}_{\hat{s}}^d(k)}{\mathbf{H}_{\hat{s}}^d(k)}\right)^2} \quad (12)$$

where $\mathbf{H}_{\hat{s}}^d(k)$ represents the k-th element in $\mathbf{H}_{\hat{s}}^d$, and $\bar{\mathbf{H}}_{\hat{s}}^d$ denotes the true HRTF magnitude for the target subject at direction d.

GNN-Based HRTF-U Module

The GNN-based HRTF-U module is built to learn the spatial correlation among HRTFs. The network architecture is depicted in Figure 4, and its input is also a graph, termed $\mathcal{G}_{\hat{d}}$, which is determined by

$$\mathcal{G}_{\hat{d}} = \{\mathcal{V}_{\hat{d}}, \mathcal{E}_{\hat{d}}, \mathcal{W}_{\hat{d}}\} \quad (13)$$

where $\mathcal{V}_{\hat{d}}$, $\mathcal{E}_{\hat{d}}$, and $\mathcal{W}_{\hat{d}}$ are the vertex set, the edge set, and the edge-weight set, which are defined by

$$\mathcal{V}_{\hat{d}} = \{\mathbf{H}_s^d \mid d \in \mathcal{N}_{\hat{d}}\} \cup \mathbf{H}_s^{\hat{d}} \quad (14a)$$

$$\mathcal{E}_{\hat{d}} = \{(d,p) \mid \|d-p\|_2 < a \cdot \delta_d,\ d,p \in \mathcal{N}_{\hat{d}}\} \quad (14b)$$

$$\mathcal{W}_{\hat{d}} = \{\exp(-\frac{\|d-p\|_2^2}{2\sigma^2}) \mid (d,p) \in \mathcal{E}_{\hat{d}}\} \quad (14c)$$

where $0 \le a \le 1$ is the factor that controls the graph topology and σ is the bandwidth parameter of the Gaussian kernel function. Unlike (7), we use a non-fully connected graph, where the edge and the edge weight are determined by the spatial distance between nodes. This emphasizes the spatial correlation between nodes in the subsequent GNN. In addition, we also put the HRTF $\mathbf{H}_s^{\hat{d}}$ at target position $\hat{d}$ into the constructed graph in (14a), which is initialized with all-one vectors. Based on the above, the constructed graph $\mathcal{G}_{\hat{d}}$ is fed into successive I GAT layers and an FC layer. Each GAT layer has the same structure as (9) and (10). In addition, we use the LSD, which is defined similarly to (12), as the loss function. Finally, the output is extracted from the target node at the last GAT layer.

Fine-Tuning

During the pre-training stage, we should train the HRTF-P and HRTF-U models in parallel. The former learns the mapping in (1), while the latter learns the mapping in (3). After that, for an unseen subject $\hat{s} \notin \mathcal{S}$, we run the HRTF-P module at all positions $d \in \mathcal{D}$ to generate $\mathbf{H}_{\hat{s}}^{d\in\mathcal{D}}$. Next, for each position $\hat{d} \in \mathcal{D}$, its neighboring points are retrieved via (14a) to construct the graph (13), which is then fed into the pre-trained HRTF-U module to fine-tune the FC layer. In doing so, personalized HRTFs are no longer position-independent after passing through the fine-tuned HRTF-U module. Consequently, personalized HRTFs are enhanced through spatial correlations, which is expected to improve accuracy.

Experiments

Datasets and Baseline

We conducted evaluations on three public HRTF datasets, including the SONICOM (Engel et al. 2023) dataset, the CIPIC dataset (Algazi et al. 2001), and the HUTUBS dataset (Brinkmann et al. 2019). Details of these datasets can be found in **Appendix A**.

To evaluate performance, we compared the proposed method with several existing approaches. For traditional methods, we selected the nearest-neighbor method and the ITD/LSD-based HRTF selection method. For data-driven methods, we included recent advances, such as NF(CbC) (Zhang, Wang, and Duan 2023), NF(LoRA) (Masuyama et al. 2024), and RANF (Masuyama et al. 2025). Detailed descriptions of these comparison methods can be found in **Appendix B**.

Experimental Setup

For the HRTF-P module, the first graph attention network in the encoder consists of $B = 2$ GAT layers, and the numbers of their attention heads were fixed to 8 and 1 respectively; the second graph attention network in the encoder consisted of a single GAT layer with 6 attention heads; the decoder involved $P = 2$ FC layers and $Q = 4$ deconvolution layers. The HRTF-U module used a two-layer GAT structure (i.e., $I = 2$), which has 8 and 1 attention heads, respectively. For other parameter settings, the Gaussian kernel width σ in (14c) was set to 0.5, a and δ_d in (14b) were set to 0.75 and 20, respectively.

Regarding the training strategy, we adopted the following multiphase optimization strategy. The pre-training phase of the HRTF-P module used the RAdam optimizer (Liu et al. 2019) with an initial learning rate of 0.001, training for 200 epochs. In addition, a dynamic learning rate decay strategy was employed, which multiplies the learning rate by 0.9 whenever the validation loss does not decrease over 10 consecutive epochs. The pre-training phase of the HRTF-U module used the Adam optimizer (Kingma and Ba 2015) with an initial learning rate of 0.002, training for 200 epochs. The learning rate was multiplied by 0.95 if the validation loss did not decrease for 3 consecutive epochs. The final fine-tuning phase also used the Adam optimizer with an initial learning rate of 0.002, applying an exponential decay strategy with a decay rate of 0.95 per epoch, and trained for a total of 20 epochs.

Evaluation in LAP Challenge 2024

In this part, we evaluated the HRTF-P performance in the listener acoustic personalization (LAP) challenge 2024 using the SONICOM dataset. Due to the fact that the relative relationship between binaural HRTFs fundamentally determines the quality of 3D audio rendering, our evaluation under the LAP Challenge [1] incorporates both the LSD of HRTF magnitude and the interaural level difference (ILD). For the subject-relevant characteristic a_s in (2), inspired by the work

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

Methods	3 measurements		5 measurements		19 measurements		100 measurements	
	ILD[dB]	LSD[dB]	ILD[dB]	LSD[dB]	ILD[dB]	LSD[dB]	ILD[dB]	LSD[dB]
Nearest neighbor	7.64	8.69	4.78	8.30	2.99	5.42	1.35	3.42
HRTF selection(LSD)	1.46	5.78	1.40	5.64	1.54	5.47	1.34	5.41
HRTF selection(ITD)	1.42	6.38	1.46	6.43	1.55	6.71	1.38	6.33
NF(CbC)	1.54	4.87	2.04	5.22	1.79	5.01	1.67	5.12
NF(LoRA)	1.28	4.73	1.37	4.62	1.09	4.07	1.06	3.83
RANF	1.21	4.41	1.31	4.56	0.95	3.58	0.76	3.04
GraphNF	1.22	4.33	1.30	4.53	0.93	3.51	0.76	3.03
GraphNF-SCA	**0.96**	**3.60**	**0.91**	**3.55**	**0.78**	**3.12**	**0.70**	**2.72**

Table 1. LSD, ILD errors for different numbers of measurement directions in LAP challenge 2024.

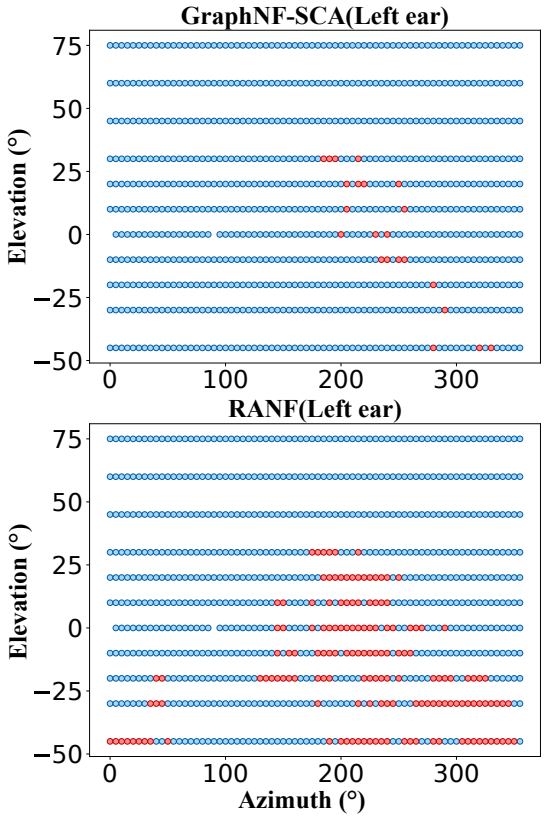

Figure 5: Spatial distribution of LSD errors for unmeasured HRTFs at the left ear. Blue and red dots indicate LSD $\leq \zeta$ and LSD $> \zeta$, respectively.

(Masuyama et al. 2025), we use the ILD feature (assuming that the HRTFs of the target subject are measured at a small number of directions in advance) rather than human anatomy.

In Table 1, we referred to the HRTF-P module (Figure 3) as GraphNF, and the augmented version using spatial correlation (Figure 2) as GraphNF-SCA. The ILD and LSD errors were tested under different numbers of HRTF measurements from different directions (3, 5, 19, and 100). Overall, the performance of all methods improves as the number of measured HRTF directions increases. Traditional methods such as Nearest Neighbor and HRTF selection perform poorly under sparse measurement conditions, with LSD errors reaching approximately 6 dB when only 3–5 directions are available. Data-driven methods outperform traditional approaches, with GraphNF-SCA achieving the best performance. Specifically, GraphNF slightly outperforms the strongest existing baseline, RANF, indicating that incorporating retrieved subject information via graph-based modeling effectively captures inter-subject relationships and enhances HRTF personalization. More importantly, GraphNF-SCA consistently achieves the lowest ILD and LSD errors across all configurations. In an extremely sparse setting with only 3 measurement directions, GraphNF-SCA significantly improves over GraphNF, reducing the LSD from 4.33 dB to 3.60 dB (16.86%) and the ILD from 1.22 dB to 0.96 dB (21.31%). This confirms that the collaborative mechanism between HRTF-P and HRTF-U effectively enhances the performance of HRTF-P and further demonstrates the superiority of GraphNF-SCA in scenarios with severely limited data.

To provide a more intuitive visualization of HRTF estimation results, Figure 5 illustrates the accuracy of HRTF prediction, using the left ear as an example, at all azimuth and elevation angles under the extremely sparse condition of using only 3 measurement directions. The blue and red dots denote HRTFs with errors below and above the threshold $\zeta = 6$ dB, respectively. It can be observed that LSD errors that exceed the threshold ζ (at the left ear) are mainly concentrated in the azimuth range of 180° to 360°, corresponding to sound sources located on the right side of the head. This result aligns with the "acoustic head-shadowing effect", wherein predicting HRTFs for sources located on the contralateral side results in higher errors due to sound waves being attenuated as they propagate through the head. Compared to RANF, GraphNF-SCA exhibits a greater number of low-error LSD points (i.e., below the threshold) in unmeasured directions, with high-error regions mainly concentrated on the contralateral side. To save space, we only present the HRTF estimation performance of RANF and GraphNF-SCA at the left ear. The corresponding results for the right ear, as well as the binaural LSD error distributions of another method, i.e., HRTF Selection (LSD), can be found in **Appendix C**.

Evaluation on Other Databases

To evaluate the generalization ability of the proposed method on other datasets, we carried out experiments on the following databases: $\mathrm{HUTUBS}_{\mathrm{mea}}$, $\mathrm{HUTUBS}_{\mathrm{sim}}$, and CIPIC. Among them, $\mathrm{HUTUBS}_{\mathrm{mea}}$ and $\mathrm{HUTUBS}_{\mathrm{sim}}$ correspond to the measured and simulated data from the

Dataset	Methods	3 measurements		5 measurements		19 measurements		100 measurements	
		ILD[dB]	LSD[dB]	ILD[dB]	LSD[dB]	ILD[dB]	LSD[dB]	ILD[dB]	LSD[dB]
$\text{HUTUBS}_{\text{mea}}$	HRTF selection(LSD)	1.44	5.61	1.32	5.49	1.33	5.42	1.44	5.40
	RANF	1.42	4.69	1.24	4.43	0.93	4.20	0.93	3.82
	GraphNF	1.41	4.63	1.29	4.43	0.91	3.94	0.90	3.52
	GraphNF-SCA	**0.97**	**3.74**	**0.93**	**3.72**	**0.83**	**3.44**	**0.72**	**3.23**
$\text{HUTUBS}_{\text{sim}}$	HRTF selection(LSD)	1.34	4.93	1.23	5.21	1.04	5.12	1.03	5.20
	RANF	1.24	4.13	1.11	4.10	0.93	3.14	0.83	2.77
	GraphNF	1.22	4.12	1.20	3.94	0.96	3.15	0.83	2.77
	GraphNF-SCA	**0.96**	**3.21**	**0.92**	**3.10**	**0.81**	**2.74**	**0.68**	**2.41**
CIPIC	HRTF selection(LSD)	2.33	6.61	2.40	6.42	1.81	6.34	1.70	6.12
	RANF	1.63	4.92	1.80	5.21	1.59	4.32	1.21	3.84
	GraphNF	1.63	4.91	1.52	4.85	1.44	4.21	1.04	3.82
	GraphNF-SCA	**1.14**	**3.85**	**1.13**	**3.83**	**1.07**	**3.52**	**0.91**	**3.32**

Table 2. LSD, ILD errors for different numbers of measurement directions on other datasets.

HUTUBS dataset, respectively. To save space, we only included the best-performing traditional method, namely HRTF selection (LSD). From Table 2, we observe a trend similar to that in Table 1: As the amount of measurement data increases, both the ILD and LSD errors decrease accordingly. Under sparse measurement conditions (3–5 directions), GraphNF achieves an ILD error comparable to the strongest baseline, RANF, while reducing the LSD error by 1.8%, showing a slight advantage. Under higher-density measurement conditions (100 directions), GraphNF reduces the ILD error by 4.3% and the LSD error by 2.1%, indicating a limited improvement. Furthermore, by incorporating spatial correlation modeling, GraphNF-SCA demonstrates clear advantages under sparse measurement conditions. Compared to RANF, GraphNF-SCA achieves a 26.9% reduction in ILD error and a 23.7% reduction in LSD error. Even under high-density measurement conditions, it retains a clear advantage, with average reductions of 18.3% in ILD error and 13.1% in LSD error. The proposed method consistently outperforms all baselines across all evaluated datasets, with particularly strong performance under extremely data-sparse conditions.

Ablation Study

In the ablation study, we first investigated the impact of selecting different subject-relevant characteristics a_s and varying the number $M = |\mathcal{N}_{\hat{s}}|$ of subjects in retrieval (2), under the extremely sparse condition of using only three measurement directions on the SONICOM dataset. As shown in Table 3, using the ITD and ILD features for retrieval results in comparable HRTF personalization performance, both slightly outperforming LSD-based retrieval. In addition, the impact of M is also minor. Notably, GraphNF-SCA shows a clear and consistent improvement over GraphNF across all settings.

The previous experiment settings were kept, and we used the ILD feature to conduct another ablation experiment on the SONICOM dataset. We compared the following variants of GraphNF: the GraphNF without inputting source direction d and subject-related feature a_s (denoted as GraphNF (w/o d, a_s)); the GraphNF incorporating d and a_s but without employing the GAT layer for further feature integration (denoted as GraphNF (w/ d, a_s, w/o g)). As shown in Table 4, the HRTF reconstruction error decreases significantly as more modules are integrated into the model. These results validate the effectiveness of the collaborative mechanism proposed in the GraphNF-SCA network.

Retrieval	M	GraphNF		GraphNF-SCA	
		ILD[dB]	LSD[dB]	ILD[dB]	LSD[dB]
LSD	1	1.23	**4.30**	0.94	3.59
	5	1.26	4.38	0.94	3.63
	10	1.24	4.39	0.93	3.61
ITD	1	1.24	**4.30**	0.92	3.57
	5	**1.22**	4.33	0.93	3.61
	10	1.23	4.39	0.93	3.64
ILD	1	1.23	**4.30**	0.93	**3.56**
	5	**1.22**	4.33	0.96	3.60
	10	1.27	4.37	**0.90**	3.59

Table 3. Impact of different retrieval strategies.

	Model	ILD[dB]	LSD[dB]
(a)	GraphNF (w/o d, a_s, g)	1.69	4.65
(b)	GraphNF (w/ d, a_s, w/o g)	1.34	4.41
(c)	GraphNF	1.22	4.33
(d)	GraphNF-SCA	0.96	3.60

Table 4. Impact of module integration.

For more ablation studies and comparison experiments, please refer to **Appendix D**.

Conclusion

In this work, we propose GraphNF-SCA, a novel framework for personalized HRTF prediction that effectively incorporates spatial correlation modeling into graph-based learning. Our method consists of three key components: the HRTF-P module, which predicts individual HRTFs via a graph-based encoder-decoder; the HRTF-U module, which captures the spatial structure of HRTFs through a second GNN; and a fine-tuning stage that refines the output of the HRTF-P module by using the fine-tuned HRTF-U module. By explicitly modeling the spatial relationship between HRTFs, GraphNF-SCA significantly outperforms existing methods that perform HRTF personalization independently for each position. Extensive experiments demonstrate that our framework achieves state-of-the-art performance. We believe GraphNF-SCA offers a promising direction for accurate and scalable HRTF personalization in immersive 3D audio applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62361045 and 62201297.

References

Algazi, V. R.; Duda, R. O.; Duraiswami, R.; Gumerov, N. A.; and Tang, Z. 2002. Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112(5): 2053–2064.

Algazi, V. R.; Duda, R. O.; Thompson, D. M.; and Avendano, C. 2001. The CIPIC HRTF database. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 99–102.

Arend, J. M.; Brinkmann, F.; and Pörschmann, C. 2021. Assessing spherical harmonics interpolation of time-aligned head-related transfer functions. *Journal of the Audio Engineering Society*, 69: 104–117.

Bomhardt, R.; de la Fuente Klein, M.; and Fels, J. 2016. A high-resolution head-related transfer function and three-dimensional ear model database. In *Proceedings of Meetings on Acoustics*, 050002.

Brinkmann, F.; Dinakaran, M.; Pelzer, R.; Grosche, P.; Voss, D.; and Weinzierl, S. 2019. A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses. *Journal of the Audio Engineering Society*, 67(9): 705–718.

Carpentier, T.; Bahu, H.; Noisternig, M.; and Warusfel, O. 2014. Measurement of a head-related transfer function database with high spatial resolution. In *Forum Acusticum*, 1–6.

Clevert, D. A.; Unterthiner, T.; and Hochreiter, S. 2020. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 10.

Cuevas Rodríguez, M.; Picinali, L.; González Toledo, D.; Garre, C.; de la Rubia Cuestas, E.; Molina Tanco, L.; and Reyes Lecuona, A. 2019. 3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation. *PLoS ONE*, 14(3): 103–104.

David, P. Q.; and Katz, B. F. 2018. The Anaglyph binaural audio engine. In *Audio Engineering Society Convention 144*, 1–17.

Du, Y. C.; Yu, H.; Ciou, W.; and Li, Y. 2023. A Wearable Assistive Listening Device with Immersive Function Using Sensors Fusion Method for the 3D Space Perception. *IEEE Sensors Journal*, 24(2): 2108–2117.

Engel, I.; Daugintis, R.; Vicente, T.; Hogg, A. O.; Pauwels, J.; Tournier, A. J.; and Picinali, L. 2023. The sonicom HRTF dataset. *The Journal of the Audio Engineering Society*, 71(5): 241–253.

Engel, I.; Goodman, D. F.; and Picinali, L. 2022. Assessing HRTF preprocessing methods for Ambisonics rendering through perceptual models. *Acta Acustica*, 6: 4.

Evans, M. J.; Angus, J. A.; and Tew, A. I. 1998. Analyzing head-related transfer function measurements using surface spherical harmonics. *The Journal of the Acoustical Society of America*, 104(4): 2400–2411.

Fantini, D.; Geronazzo, M.; Avanzini, F.; and Ntalampiras, S. 2025. A Survey on Machine Learning Techniques for Head-Related Transfer Function Individualization. *IEEE Open Journal of Signal Processing*, 6: 30–56.

Gebru, I. D.; Marković, D.; Richard, A.; Krenn, S.; Butler, G. A.; De la Torre, F.; and Sheikh, Y. 2021. Implicit HRTF modeling using temporal convolutional networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3385–3389.

Geronazzo, M.; Peruch, E.; Prandoni, F.; and Avanzini, F. 2019. Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization. *Journal of the Audio Engineering Society*, 67(6): 414–428.

Hartung, K.; Braasch, J.; and Sterbing, S. J. 1999. Comparison of different methods for the interpolation of head-related transfer functions. In *AES International Conference*, 319–329.

Hogg, A. O.; Jenkins, M.; Liu, H.; Squires, I.; Cooper, S. J.; and Picinali, L. 2024. HRTF upsampling with a generative adversarial network using a gnomonic equiangular projection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 2085–2099.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hu, J.; Li, S.; Si, Q.; and Hu, D. 2025. D-GAT: Dual Graph Attention Network for Global HRTF Interpolation. In *Interspeech*, 2510–2514.

Jenny, C.; Reuter, C.; et al. 2020. Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization. *JMIR serious games*, 8(3): e17576.

Johansson, M. 2019. VR For Your Ears: Dynamic 3D audio is key to the immersive experience by mathias johansson · illustration by eddie guy. *IEEE Spectrum*, 56(2): 24–29.

Katz, B. F.; and Parseihian, G. 2012. Perceptually based head-related transfer function database optimization. *The Journal of the Acoustical Society of America*, 131(2): EL99–EL105.

Kingma, D.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Ko, B. Y.; Lee, G. T.; Nam, H.; and Park, Y. H. 2023. PRTFNet: HRTF individualization for accurate spectral cues Using a compact PRTF. *IEEE Access*, 11: 96119–96130.

Lee, J. W.; Lee, S.; and Lee, K. 2023. Global HRTF interpolation via learned affine transformation of hyper-conditioned features. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; and Han, J. 2019. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

Maas, A. L.; Hannun, A. Y.; Ng, A. Y.; et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*.

Masuyama, Y.; Wichern, G.; Germain, F. G.; Ick, C.; and Roux, J. L. 2025. Retrieval-Augmented Neural Field for HRTF Upsampling and Personalization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Masuyama, Y.; Wichern, G.; Germain, F. G.; Pan, Z.; Khurana, S.; Hori, C.; and Le Roux, J. 2024. NIIRF: Neural IIR Filter Field for HRTF Upsampling and Personalization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1016–1020.

Pelzer, R.; Dinakaran, M.; Brinkmann, F.; Lepa, S.; Grosche, P.; and Weinzierl, S. 2020. Head-related transfer function recommendation based on perceptual similarities and anthropometric features. *The Journal of the Acoustical Society of America*, 148(6): 3809–3817.

Pörschmann, C.; Arend, J. M.; and Brinkmann, F. 2019. Directional equalization of sparse head-related transfer function sets for spatial upsampling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(6): 1060–1071.

Ramos, G.; Cobos, M.; Bank, B.; and Belloch, J. A. 2017. A Parallel Approach to HRTF Approximation and Interpolation Based on a Parametric Filter Model. *IEEE Signal Processing Letters*, 24(10): 1507–1511.

Simon, L. S.; Zacharov, N.; and Katz, B. F. 2016. Perceptual attributes for the comparison of head-related transfer functions. *The Journal of the Acoustical Society of America*, 140(5): 3623–3632.

Sridhar, R.; Tylka, J. G.; and Choueiri, E. Y. 2017. A database of head-related transfer function and morphological measurements. In *Audio Engineering Society Convention*.

Sundareswaran, V.; Wang, K.; Chen, S.; Behringer, R.; McGee, J.; Tam, C.; and Zahorik, P. 2003. 3D audio augmented reality: implementation and experiments. In *IEEE/ACM International Symposium on Mixed and Augmented Reality*, 296–297.

Takemoto, H.; Mokhtari, P.; Kato, H.; Nishimura, R.; and Iida, K. 2012. Mechanism for generating peaks and notches of head-related transfer functions in the median plane. *The Journal of the Acoustical Society of America*, 132(6): 3832–3841.

Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.; and Ng, R. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33: 7537–7547.

Teng, Y.; and Zhong, X. 2023. An individualized HRTF model based on random forest and anthropometric parameters. In *International Conference on Intelligent Human-Machine Systems and Cybernetics*, 143–146.

Vickers, D.; et al. 2021. Involving children and teenagers with bilateral cochlear implants in the design of the BEARS (both EARS) virtual reality training suite improves personalization. *Frontiers in Digital Health*, 12(3): 759723.

Yang, J.; and Mattern, F. 2019. Audio augmented reality for human-object interactions. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 408–412.

Zhang, Y.; Wang, Y.; and Duan, Z. 2023. HRTF Field: Unifying Measured HRTF Magnitude Representation with Neural Fields. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

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

Zotkin, D.; Hwang, J.; Duraiswaini, R.; and Davis, L. 2003. HRTF personalization using anthropometric measurements. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 157–160.

Zotter, F.; and Frank, M. 2019. *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Berlin, Germany: Springer Nature.