Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Thuillier, Etienne; Jin, Craig; Valimaki, Vesa

HRTF Interpolation using a Spherical Neural Process Meta-Learner

Published in:
IEEE/ACM Transactions on Audio Speech and Language Processing

DOI:
10.1109/TASLP.2024.3361377

Published: 01/01/2024

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Thuillier, E., Jin, C., & Valimaki, V. (2024). HRTF Interpolation using a Spherical Neural Process Meta-Learner. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *32*, 1790-1802. https://doi.org/10.1109/TASLP.2024.3361377

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

HRTF Interpolation Using a Spherical Neural Process Meta-Learner

Etienne Thuillier, *Member, IEEE*, Craig T. Jin, *Senior Member, IEEE*, and Vesa Välimäki, *Fellow, IEEE*

***Abstract*—Several individualization methods have recently been proposed to estimate a subject's Head-Related Transfer Function (HRTF) using convenient input modalities such as anthropometric measurements or pinnae photographs. There exists a need for adaptively correcting the estimation error committed by such methods using a few data point samples from the subject's HRTF, acquired using acoustic measurements or perceptual feedback. To facilitate this, we introduce a Convolutional Conditional Neural Process meta-learner specialized in HRTF error interpolation. In particular, the model includes a Spherical Convolutional Neural Network component to accommodate the spherical geometry of HRTF data. It also exploits potential symmetries between the HRTF's left and right channels about the median plane. In this work, we evaluate the proposed model's performance purely on time-aligned spectrum interpolation grounds under a simplified setup where a generic population-mean HRTF forms the initial estimates prior to corrections instead of individualized ones. The trained model achieves up to 3 dB relative error reduction compared to state-of-the-art interpolation methods despite being trained using only 85 subjects. This improvement translates up to nearly a halving of the data point count required to achieve comparable accuracy, in particular from 50 to 28 points to reach an average of -20 dB relative error per interpolated feature. Moreover, we show that the trained model provides well-calibrated uncertainty estimates. Accordingly, such estimates could inform the sequential decision problem of acquiring as few correcting HRTF data points as needed to meet a desired level of HRTF individualization accuracy.**

***Index Terms*—Audio systems, representation learning, spatial audio, uncertainty.**

I. Introduction

RECENT adoption of augmented and virtual reality interfaces has pushed the need for immersive spatial audio rendering solutions that scale to mass market [1], [2], [3], [4]. The Head Related Transfer Function (HRTF) is a key component of current systems: it simulates the effect of the subject's body on the acoustic transmission channels between the subject's ears and sound sources as a function of their locations raround the subject [5]. Crucially, the HRTF is a function of the subject's morphology and is specific to each individual. Studies have shown that spatial audio percepts deteriorate when a generic HRTF is used for all subjects in a population compared to using individualized HRTF estimates [5], [6]. Models of this phenomenon are actively being developed [7], [8]. In this work, we propose the first HRTF interpolation method that provides well-calibrated uncertainty estimates. The method also demonstrates significantly improved interpolation accuracy with regards to the state of the art.

A. Prior Art

A recent review paper classifies HRTF individualization techniques into four categories defined by the source of HRTF information: acoustic measurements, numerical simulation, anthropometric data, and perceptual feedback [6]. As an alternative approach useful to our discussion, we classify below the individualization techniques into two broad classes according to the way in which the subject's individualized HRTF is represented.

A first class of methods represents the subject's individualized HRTF using a sparse set of observed HRTF data points. In such approaches, interpolation methods are applied downstream to provide HRTF filter estimates at specified directions of arrival between the observed locations. Typically, the observations are collected using acoustic measurements [9], but recommender systems have also been proposed for composing the sparse set with HRTF filters derived from a pre-existing database and according to perceptual feedback obtained from the user [10].

Improvements in interpolation methods result in sparser sets of observations becoming sufficient for meeting a required accuracy threshold, thereby accelerating the individualized HRTF acquisition process. Early methods include barycentric interpolation [11], bilinear interpolation [12], natural neighbour interpolation [13], interpolation using spherical harmonics [14], thin-plate spherical spline interpolation [15] and Gaussian process regression [16]. In particular, use of a learned spherical harmonic subspace allows for recovering the full HRTF from a significantly reduced set of data points [17]. More recently, pre-processing has been shown to significantly reduce the required density of HRTF measurements needed to meet a given interpolation accuracy requirement [18], [19], [20], [21]. Neural-network regressor models have also been proposed [22], [23], including a spherical convolutional neural network performing interpolation from a relatively dense equiangular grid counting 120 data points [24]. Related works includes HRTF upsampling approaches using generative models [25]. However, such models currently provide improvements in the sparsest regimes only.

Manuscript received 18 October 2023; revised 23 January 2024; accepted 25 January 2024. Date of publication 1 February 2024; date of current version 11 March 2024. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jan Skoglund. *(Corresponding author: Etienne Thuillier.)*

Etienne Thuillier and Vesa Välimäki are with the Acoustics Lab, Department of Information and Communications Engineering, Aalto University, FI-02150 Espoo, Finland (e-mail: etienne.thuillier@aalto.fi; vesa.valimaki@aalto.fi).

Craig T. Jin is with the School of Electrical and Computer Engineering, University of Sydney, Sydney, NSW 2050, Australia (e-mail: craig.jin@sydney.edu.au).

Digital Object Identifier 10.1109/TASLP.2024.3361377

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

A second class of methods parametrizes individualized HRTFs using low-dimensional latent-space representations embodied by fixed-length vectors of adjustable coefficients. Various approaches have been proposed to predict the coefficients of the representation including the use of anthropometric measurements [26], [27], pinnae photographs [28], HRTF observations [22], perceptual feedback [29] or combinations thereof [30]. This provides a convenient means for promptly estimating a subject's HRTF from one or several input modalities. However, a common design compromise facing techniques in this class lies in providing a representation that is compact enough that prediction is facilitated, while retaining sufficient expressiveness that HRTF variability across the population is faithfully represented. Due to the fixed dimensionality of latent representations in particular, and unlike non-parametric interpolation methods, the expressiveness of the model does not scale with additional data points provided to it. More importantly, any resulting change to the HRTF representation is in this case global, such that any resulting local improvement is susceptible, at least in principle, to adversely affect the representation elsewhere. This contrasts with interpolation methods providing strictly local representations of HRTFs, for example barycentric interpolation.

B. Problem

There is a need for adaptively refining the individualized HRTF estimate provided by a parametric method until a predefined criterion of suitability is achieved, for example a user performance metric threshold under a listening test experiment. To this end, we advocate for a hybrid approach to HRTF individualization in which the parametric estimate is corrected by integrating a few observations of the subject's HRTF using an interpolation method. Under this approach, the HRTF refinement problem can be framed as a sequential decision problem: that of acquiring as few correcting HRTF data points as needed to meet the performance requirement, using measurements or perceptual feedback. Such a problem would benefit from using an accurate interpolation method that also provides well-calibrated uncertainty estimates. When suitably calibrated, uncertainty estimates can indeed be used to inform the choice of the next location to observe. Under a perceptual feedback acquisition scheme, they can additionally inform the selection of proposal HRTF filters to be submitted as queries to the subject. Finally, there also exists a need within augmented reality settings, for matching the rendered sound field with the user's surrounding acoustic environment. Such a problem could also be addressed using the suggested approach by adaptively refining the Binaural Room Transfer Function instead of the HRTF.

C. Solution

To facilitate the hybrid approach mentioned above, we introduce a novel model that we name Spherical Convolutional Conditional Neural Process (SConvCNP). The proposed model is a Convolutional Conditional Neural Process (ConvCNP) meta-learner [31] specialized in HRTF error interpolation. It accommodates the spherical geometry of HRTF data. To this end, it includes a Spherical Convolutional Neural Network component [32], [33] which executes rotation-equivariant feature transforms. It also exploits the approximate symmetry between the HRTF's left and right channels about the median plane. To the authors' best knowledge, this work is the first application of a Neural Process model to spherical data.

The proposed model learns a functional representation of the set of observed HRTF data points that preserves spatial structure and can be addressed at any location on the unit sphere. Furthermore, the representation is learned using rotation equivariant mappings which ensures the same transformation is applied with shared parameters everywhere on the sphere, irrespective of feature location. These aspects allow for learning local interpolations of the HRTF features in a sample-effective fashion. Moreover, the possibility, afforded by the model, to address any location on the unit sphere provides native compatibility for training on any HRTF databases irrespective of its data point grid layout.

This work implements and tests the SConvCNP model in a simplified experimental setup. Firstly, the interpolation is applied on the HRTF spectrum after time-alignment [18], [20], [21], leaving pure delay interpolation as future work for brevity. Secondly, a generic population-mean time-aligned spectrum is used as generic estimate for all subjects before correction, instead of individualized time-aligned spectra. This allows to evaluate the merits of the model purely from an interpolation performance standpoint, leaving the application to individualized HRTF correction as future work. The model is shown to achieve up to 3 dB of relative error reduction compared to state-of-the-art interpolation methods. This translates to nearly a halving of the required data to achieve a comparable level of accuracy. Moreover, our model is shown to provide well-calibrated uncertainty estimates.

This paper is organized as follows. Section II provides background on the ConvCNP model and its meta-training procedure. Section III introduces the SConvCNP model, defines the interpolation tasks on which the model is trained, and proposes baseline and metrics for evaluating the model's performance both in terms of interpolation accuracy and uncertainty calibration. Section IV presents and discusses the experimental results. Section V concludes this paper.

II. Background

In this section, we provide a technical review of the ConvCNP model and its meta-training procedure as background for the introduction of the SConvCNP model in Section III.

A. ConvCNP Architecture

Neural Processes form a class of deep neural networks operating on sets to model stochastic processes [34], [35]. In neural process models, a set of observed location-feature data point pairs $\{(x_c, y_c)\}_{c=1}^{C}$ at the input informs a predictive distribution provided at the output for unseen values y_t at target locations x_t, much in the same fashion as in Gaussian Processes [36]. In particular, the elements of the input set are subsumed into a representation embedding, which allows for handling sets of

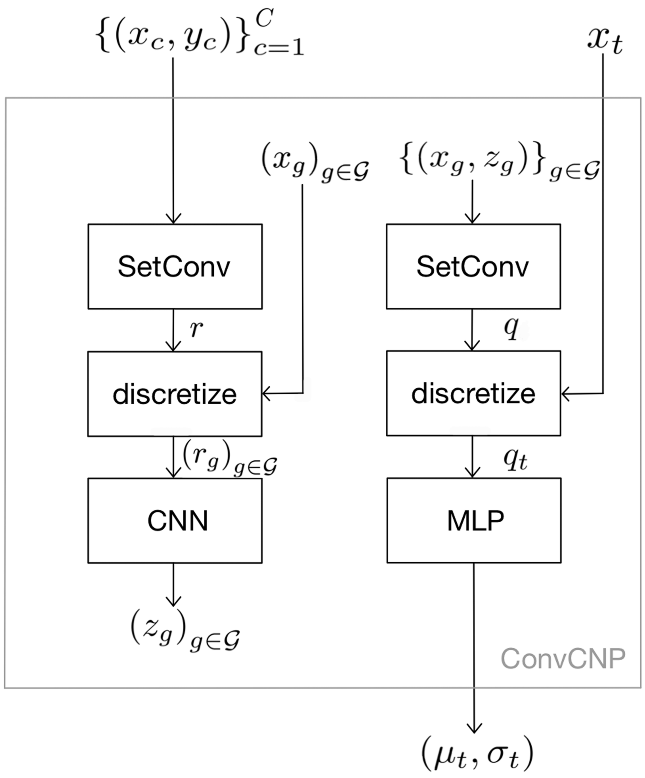
$\{(x_c, y_c)\}_{c=1}^{C}$
x_t
$(x_g)_{g\in\mathcal{G}}$
$\{(x_g, z_g)\}_{g\in\mathcal{G}}$
SetConv
r
discretize
$(r_g)_{g\in\mathcal{G}}$
CNN
$(z_g)_{g\in\mathcal{G}}$
SetConv
q
discretize
q_t
MLP
ConvCNP
(μ_t, σ_t)

Fig. 1. Schematic block diagram of a typical ConvCNP model's architecture.

different sizes and ensures invariance in the ordering of set elements [37]. Recently, functional representation embeddings have been proposed that preserve the spatial structure in the input set and are addressable at any location coordinates x_t. These functional embeddings enable constructing translation equivariant neural process models, as appropriate when modeling stationary data [31], [38], [39]. The ConvCNP is an example of such a model [31]. A description of its architecture is given in the current section.

A typical example of ConvCNP model architecture is given in the block diagram of Fig. 1. The model includes a first set convolution (block *SetConv*) which maps a set of observed data points $\{(x_c, y_c)\}_{c=1}^{C}$ into a functional representation [31], [39]

$$r = \text{SetConv}\left(\{(x_c, y_c)\}_{c=1}^{C}\right), \tag{1}$$

which, assuming a multiplicity of one for data set elements [31], returns a vector-valued point-wise representation

$$r(x) = \left(\sum_{c=1}^{C} K(x_c, x),\ \frac{\sum_{c=1}^{C} y_c K(x_c, x)}{\sum_{c=1}^{C} K(x_c, x)}\right), \tag{2}$$

at any specified location x. In the above expression, $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ denotes a positive definite kernel with learnable parameter(s), for example, a Gaussian kernel in the case of planar data such as images [40]. Accordingly, the first channel of functional embedding r before discretization, forms a kernel density of the observed locations: the result of a convolution between filter $K(x_c, \cdot)$ and a sum of unit-weighted Dirac distributions centered at locations $\{x_c\}_{c=1}^{C}$. The second channel forms an interpolant of the observed data points $\{(x_c, y_c)\}_{c=1}^{C}$ following the Nadaraya-Watson kernel regression method [41].

In ConvCNP models, translation equivariant representation learning is performed downstream of the set convolution using a Convolutional Neural Network (CNN). As pictured in Fig. 1, representation r is first discretized following a grid $(x_g)_{g\in\mathcal{G}}$ of regularly-spaced coordinates. Assuming two-dimensional planar data for example: $\mathcal{G} = \{1, \ldots, G\} \times \{1, \ldots, G\}$, in which G denotes the number of samples of the grid in each dimension. A second set convolution converts—at least implicitly—the learned representation at the output of the CNN back to a functional one [39], denoted q in the diagram. Crucially, q's second channel[1] forms an interpolant of the learned representation $\{(x_g, z_g)\}_{g\in\mathcal{G}}$ following (2).

Given the above, q forms a learned functional representation of the input set which is spatially-structured and addressable at any user-specified target location x_t. In the example of Fig. 1, the resulting point-wise representation q_t is decoded using a fully-connected neural network (MLP)[2] decoder, which maps the target location x_t to mean μ_t and standard deviation σ_t values specifying a predictive distribution for the target features y_t at that location:

$$p\left(y_t \,\middle|\, x_t,\ \{(x_c, y_c)\}_{c=1}^{C}\right) \approx \mathcal{N}\left(y_t;\ \mu_t, \sigma_t^2\right), \tag{3}$$

where we assume uni-variate features for simplicity and $\mathcal{N}$ denotes the normal distribution.

B. Meta-Training

Features and advantages of the model are best understood under the lens of the meta-learning framework [42]. Under this perspective, the observed data-points $\{(x_c, y_c)\}_{c=1}^{C}$ form a task-specific train set and the ConvCNP model forms a meta-learning algorithm mapping the set to a trained discriminator MLP $\circ\, q$ that,[3] given query location x_t, returns a predictive distribution $\mathcal{N}(y_t; \mu_t, \sigma_t^2)$. In particular, the learned functional embedding q of train set $\{(x_c, y_c)\}_{c=1}^{C}$ parametrizes said discriminator MLP $\circ\, q$ such that the set's data point locations $\{x_c\}_{c=1}^{C}$ can be leveraged to provide well-calibrated uncertainty estimates σ_t. This is unlike common learning methods which generally discard such information and, as a consequence, provide trained models that cannot recognize when queried far from the points of the train set. While it remains possible for these models to quantify the uncertainty resulting from noise present in the labels ("data uncertainty", "aleatoric uncertainty"), the uncertainty in the choice of model and the value of its parameters ("model uncertainty", "epistemic uncertainty"), in particular as it relates to the train set used to optimize parameter values, is typically unaccounted for. In contrast, ConvCNP models are shown to provide well-behaved uncertainty estimates for stationary data [31]. This results in part from translation equivariance. Indeed, this

[1] The first channel, representative of density, is less informative at this stage than in the case of the first set convolution and can optionally be discarded.

[2] MLP is the acronym for multilayer perceptron. As is common practice, it stands here as a synonym of fully-connected neural network despite it being a slight misnomer.

[3] Here, MLP $\circ\, q$ denotes the composition of functions q and MLP.

ensures the model's outputs can be computed as function of distance to the context set's data points but not as a function of the data points' coordinate values themselves.

Model optimization within the meta-learning framework is carried out on a set of learning tasks (the meta-training set), each defined by a context (train) set and target (test) set pair. In particular, ConvCNP models can be trained following the maximum-likelihood objective [34]

$$\max_{\theta} \sum_{(\mathcal{C},\mathcal{T})\in\mathcal{M}} \; \sum_{(x,y)\in\mathcal{T}} \log p(y\,|x,\mathcal{C};\;\theta)\,, \tag{4}$$

where θ denotes the coefficient vector of the model's learnable parameters, $\mathcal{M} = \{(\mathcal{C}_m, \mathcal{T}_m)\}_{m=1}^{M}$ denotes the meta-training set, $\mathcal{C}_m = \{(x_c, y_c)\}_{c=1}^{C}$ forms a context (train) set, and $\mathcal{T}_m = \{(x_t, y_t)\}_{t=1}^{T}$ forms a target (test) set. Additional validation and test meta-sets composed of held-out data are used to perform model selection and evaluate generalization performance.

III. Novel Model and Method

In this section, we introduce the SConvCNP model and define the interpolation tasks on which the model is trained. Furthermore, we select the baselines and metrics for the purpose of evaluating the model's performance both in terms of interpolation accuracy and uncertainty calibration.

A. Interpolation Task

When applied to the problem of HRTF interpolation, each task $(\mathcal{C}, \mathcal{T})$ composing the meta-training set $\mathcal{M}$ consists in interpolating a given subject's HRTF to specified unseen (target, test) locations given a set of observed (context, train) HRTF data points acquired from the subject. A detailed description of the specific interpolation task studied in this work follows.

Consider the following time-alignment factorization of the HRTF spectrum [18], [20], [21]:

$$h(x) = \left(\mathrm{e}^{\mathrm{i}\frac{2\pi n}{N}\tau(x)}\right)_{n=0}^{N/2} \odot m(x), \tag{5}$$

where

- $\odot$ denotes the Hadamard (element-wise) product,
- $x \in \mathcal{S}^2 = \{x \in \mathbb{R}^3 \mid \|x\|_2 = 1\}$ denotes the sound source direction represented in cartesian coordinates on the unit sphere,
- N denotes the filter tap count of the Head-Related Impulse Response (HRIR),
- $\tau : \mathcal{S}^2 \to [0, \infty)^2$ returns the pure delay values for both ears at specified location x,
- $m : \mathcal{S}^2 \to \mathbb{C}^{(N/2+1)\times 2}$ returns the positive frequency side of the time-aligned HRTF spectrum for both ears at specified location x,
- $h : \mathcal{S}^2 \to \mathbb{C}^{(N/2+1)\times 2}$ returns the positive frequency side of the HRTF spectrum for both ears at specified location x.

Under this factorization, the time-aligned spectrum is composed of the minimum-phase and nonlinear phase all-pass components of the HRTF [43].

Inspection of (5) reveals that interpolating the pure delay τ and the time-aligned spectrum m is in principle less challenging than interpolating spectrum h directly. Indeed, the exponential factor in (5) maps pure delay values on the unit sphere to complex values, which real and imaginary parts ripple on the surface of $\mathcal{S}^2$ following the spatial variations of pure delay τ, at a rate proportional to normalized frequency n/N. Consequently, this exponential factor significantly contributes to the irregularity of the HRTF spectrum, especially in the higher portion of the frequency range. In effect, pure delay and aligned spectrum components have been shown to require spherical harmonic representations of greatly reduced order compared to the non-processed spectra for comparable reconstruction accuracy [18], [20], [21], [44].

In this work, we employ simulated HRTFs from the HUTUBS database without changes to its coordinate system, which places the origin at the center of the subject's head [45]. We extract the time-aligned spectrum m by factoring out the pure-delay exponential term out of (5) for each data point of the HRTF set individually. In particular, the pure delay is estimated in a preliminary step as the power-weighted average of excess group delay [43]. More specifically, the weighted-average is computed using frequency bins lying within the 0 to 1.1 kHz frequency range. This avoids sharp group delay jumps occurring around zeros of the HRTF spectrum in the upper frequency range [43]. When applied to the simulated HRTFs of the HUTUBS database, this approach provides pure delay values that are spatially smooth. We apply this time-alignment method to down-sampled versions of the binaural filters from 44.1 to 33.075 kHz. This reduces the HRIR tap count from $N = 256$ to $N = 192$. This is carried-out solely for the purpose of lowering the memory requirements when running the model.

For brevity, we limit the experiments of this work to the interpolation of the time-aligned spectrum m and leave the comparatively less challenging problem of interpolating the pure delay τ as future work. More specifically, we aim to interpolate the time-aligned spectrum centered around the population-mean. Accordingly, the i^{th} data point entering the composition of context or target set $\mathcal{C}, \mathcal{T}$ is given for a particular subject s by

$$\left(x_i,\; y_i^{(s)}\right) = \left(x_i,\; m_i^{(s)} - \bar{m}_i\right), \tag{6}$$

where

$$\bar{m}_i = \frac{1}{S}\sum_{s=1}^{S} m_i^{(s)}, \tag{7}$$

denotes the time-aligned spectrum mean taken across the S subjects of the train set and $m_i^{(s)} = m^{(s)}(x_i)$ denotes the value of time-aligned spectrum specific to subject s at location x_i.

Each task $(\mathcal{C}, \mathcal{T})$ in the train/validate/test meta-set splits is composed using the HRIR filters from a single individual's set in the HUTUBS database [45]. In particular, the context sets $\mathcal{C} = \{(x_c, y_c)\}_{c=1}^{C}$ are of varying size and comprise from zero to a hundred data points sampled on the unit sphere according to an approximately-uniform-grid layout. In practice, one such approximately-uniform grid is prepared beforehand for each possible sample count. For each generated task $(\mathcal{C}, \mathcal{T})$, one

TABLE I
SPLIT OF SUBJECTS FROM THE HUTUBS' SIMULATED HRTF DATABASE [45]

Set	Subjects	Count
Meta-train	All but 1, 4, 18, 27, 28, 30, 53, 65, 67, 88, 96	85
Meta-validate	4, 28, 30, 53, 65	5
Meta-test	1, 18, 27, 67	4

of these grids is randomly drawn, thereby selecting both the number of context data point samples and their relative locations on the unit sphere. Following this, a randomly-determined three-dimensional rotation of the grid is conducted to produce the final set of sampled coordinates on the unit sphere. Finally, the HRTF set data points closest to the coordinates of the rotated grid are elected to form the context set $\mathcal{C}$. The remaining data points of the HRTF set are used to form the target set $\mathcal{T}$.

In order to augment the meta-train set, the uniform grid is replaced by an irregular grid with identical data point count half of the time during training. In particular, the coordinates of the irregular grid are in this case drawn independently following a uniform density across the surface of the sphere. Furthermore, the data points of the task $(\mathcal{C}, \mathcal{T})$ are mirrored about the median plane half of the time. This augments the meta-train set with variants of the original subjects presenting permuted ears.

Given that the simulated HRTF sets from the HUTUBS database comprise 1730 data points per subject, the approach described above provides a great number of interpolation tasks. In practice, each task $(\mathcal{C}, \mathcal{T})$ is generated in real time within the train loop. This results in a meta-training set $\mathcal{M}$ of considerable size from relatively few subjects. A summary of the HUTUBS subjects split among the meta-train, meta-validation and meta-test set is given in Table I. In this split, subjects 88 and 96 are discarded since they form duplicates of subjects 22 and 1 respectively [46].

B. SConvCNP Model

The ConvCNP model was originally introduced with applications on planar data, such as images [31]. Accordingly, we adapt it to the spherical geometry of HRTF data and to the approximate symmetry between the left and right channels of the HRTF about the median plane [17]. A detailed description of the resulting SConvCNP model is provided in this section.

Assuming a subject's morphology is perfectly symmetric about the median plane, the right HRTF channel would be perfectly recoverable from the left, thereby reducing the effective dimensionality of the HRTF feature space by a factor of two. In practice however, subjects are only approximately symmetric. Nevertheless, allowing observed feature values from one channel to inform the values in the opposite channel at the mirrored location should facilitate HRTF interpolation. The SConvCNP ensures this by mirroring the right channel of the data points about the median plane. As shown in Fig. 2, the context set is decomposed (in the "split" block) into two channel-specific context sets at the input of the first discretized set convolution block. In particular, the coordinates perpendicular to the median plane are flipped ("flip" block) in the right channel's context set.

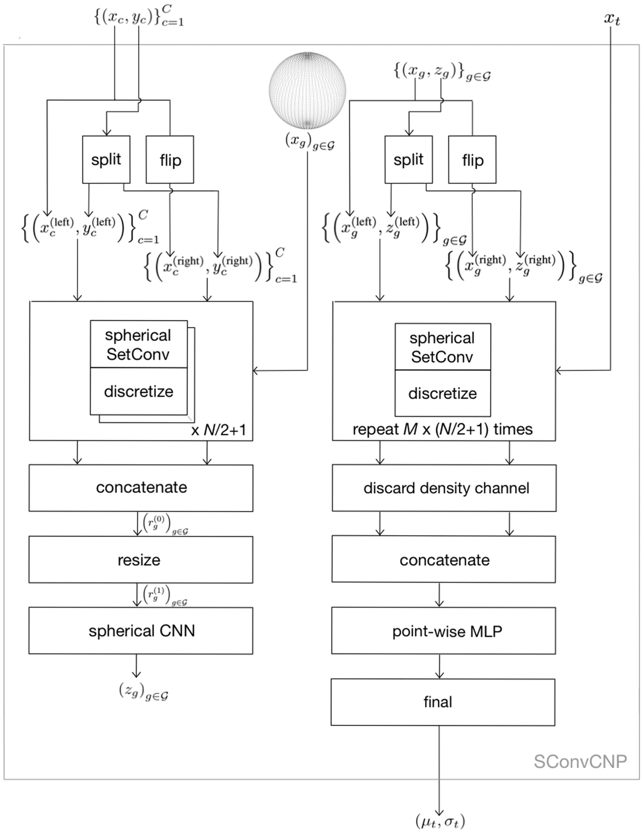
$\{(x_c, y_c)\}_{c=1}^{C}$
x_t
$\{(x_g, z_g)\}_{g\in\mathcal{G}}$
$(x_g)_{g\in\mathcal{G}}$
split
flip
split
flip
$\left\{\left(x_c^{(\text{left})}, y_c^{(\text{left})}\right)\right\}_{c=1}^{C}$
$\left\{\left(x_c^{(\text{right})}, y_c^{(\text{right})}\right)\right\}_{c=1}^{C}$
$\left\{\left(x_g^{(\text{left})}, z_g^{(\text{left})}\right)\right\}_{g\in\mathcal{G}}$
$\left\{\left(x_g^{(\text{right})}, z_g^{(\text{right})}\right)\right\}_{g\in\mathcal{G}}$
spherical SetConv
discretize
x N/2+1
spherical SetConv
discretize
repeat M x (N/2+1) times
concatenate
$\left(r_g^{(0)}\right)_{g\in\mathcal{G}}$
resize
$\left(r_g^{(1)}\right)_{g\in\mathcal{G}}$
spherical CNN
$(z_g)_{g\in\mathcal{G}}$
discard density channel
concatenate
point-wise MLP
final
SConvCNP
(μ_t, σ_t)

Fig. 2. Schematic block diagram of the SConvCNP model. Refer to Table II for tensor dimensions.

The set convolution processes each context set in sequence with shared parameters and the two resulting tensors are concatenated along the channel dimension downstream ("concatenate" block). Fig. 2 also shows that the mirroring operation is executed a second time for the right channel, upstream of the second discretized set convolution. This recovers proper left-right filter channel pairings at the output.

A significant aspect of the interpolation task described in Section III-A lies in the spherical geometry of the time-aligned spectrum features to be interpolated. Specifically, each data point location takes value on the unit sphere. Accordingly, specialized set convolutions adapted to this spherical geometry are implemented in the SConvCNP model, as pictured in Fig. 2. In this work, we use a spherical Gaussian kernel [40]:

$$K(x_1, x_2) = \mathrm{e}^{-2\beta(1-x_1\cdot x_2)}, \tag{8}$$

where $x_1,\ x_2 \in \{x \in \mathbb{R}^3 \mid \|x\|_2 = 1\}$, $\cdot$ represents the dot product and the precision parameter $\beta \in (0, \infty)$ is learned. As pictured in Fig. 2, the first set convolution block carries out a dedicated spherical set convolution for each frequency bin, ensuring a specific precision parameter β is learned at each frequency. In contrast, the second set convolution block performs a single discretized set convolution operation repeatedly with a single learned precision parameter β shared across all channel-frequency pairs. Furthermore, the density channel at

TABLE II
DIMENSION OF TENSORS IN THE SCONVCNP MODEL

Tensor	Type	Shape	Channels
y_c	$\mathbb{C}$	$(N/2+1)$	2
$y_c^{(\text{left})}$	$\mathbb{C}$	$(N/2+1)$	1
$y_c^{(\text{right})}$	$\mathbb{C}$	$(N/2+1)$	1
$\left(r_g^{(0)}\right)_{g\in\mathcal{G}}$	$\mathbb{C}$	$G\times G\times(N/2+1)$	4
$\left(r_g^{(\ell)}\right)_{g\in\mathcal{G}},\ \ell>0$	$\mathbb{R}$	$G\times G\times(N/2+1)$	M
$(z_g)_{g\in\mathcal{G}}$	$\mathbb{R}$	$G\times G\times(N/2+1)$	M
$\left(z_g^{(\text{left})}\right)_{g\in\mathcal{G}}$	$\mathbb{R}$	$G\times G\times(N/2+1)$	$M/2$
$\left(z_g^{(\text{right})}\right)_{g\in\mathcal{G}}$	$\mathbb{R}$	$G\times G\times(N/2+1)$	$M/2$
$q_t^{(\ell)}$	$\mathbb{R}$	$(N/2+1)$	M
μ_t	$\mathbb{C}$	$(N/2+1)$	2
σ_t	$\mathbb{C}$	$(N/2+1)$	2

the output of the second set convolution is discarded ("discard density channel" block).

To further accommodate the spherical geometry of HRTF data, we substitute planar convolutional layers in the CNN component with recently proposed spherical ones [32], [33]. Correspondingly, rotation equivariance is achieved in place of translation equivariance. We based our implementation on publicly-available code provided for spin-weighted spherical convolution [47].[4] In particular, we recover Esteves' simple zonal filter convolution [32] as a special case discarding all spin directions but the null-valued one. In the resulting layer, the convolution operation is carried out in Spherical Harmonic (SH) space by matrix-multiplication of the input features with the layer's filter coefficients. In practice, the SH representation of the filter is interpolated directly from a few number of learnable SH coefficients. This provides localized zonal filters while simultaneously avoiding the cost of the forward SH transform for that part of the operation [32].

In principle, the frequency dimension could be treated as an additional channel dimension. In this work, we propose to implement (single dimension) planar convolution in this dimension in order to promote the meta-learner's sample efficiency. This results in a three-dimensional hybrid planar-spherical convolution, with one axis for the frequency bins and two for the sound source direction. As reported in Table II, the receptive fields has dimensions $G\times G\times(N/2+1)$ throughout all planar-spherical convolutional layers, where G denotes the number of equiangular samples in each azimuth and elevation directions.

In classical fashion, the spherical CNN component of the model is composed of residual blocks arranged in a sequence. As pictured in Fig. 3, each block follows a single-layer pre-activation architecture [48]. As common in residual architectures, a "resize" layer is positioned at the input of the spherical CNN. Firstly, this block converts the complex-valued input tensor into an equivalent float-valued tensor, by concatenating real and imaginary parts along the channel dimension. Moreover, it scales the number of channels to the specified count M used in the residual blocks of the spherical CNN.

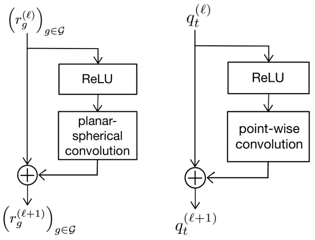
$\left(r_g^{(\ell)}\right)_{g\in\mathcal{G}}$
ReLU
planar-spherical convolution
$\left(r_g^{(\ell+1)}\right)_{g\in\mathcal{G}}$
$q_t^{(\ell)}$
ReLU
point-wise convolution
$q_t^{(\ell+1)}$

Fig. 3. Single-layer pre-activation residual blocks used to compose the spherical CNN (left) and MLP (right) components of the SConvCNP model of Fig. 2. Refer to Table II for dimension of tensors.

The point-wise MLP component of the SConvCNP model is also composed of single-layer pre-activation residual blocks as represented in Fig. 3. Each block is implemented using a point-wise convolution layer for sharing parameters across frequency bins. The model comprises a final layer that resizes and splits the channel dimension to provide a complex-valued predictive mean tensor μ_t and an unconstrained complex-valued standard deviation tensor σ'_t. Furthoremore, this layer provides the predictive standard deviation tensor σ_t using a risen softplus non-linearity forcing the real and imaginary parts of the unconstrained standard deviation coefficients to positive values [31], [49]:

$$\sigma_t = \text{risen\_softplus}\left(\operatorname{Re}\left(\sigma'_t\right)\right) + \ldots \mathrm{i}\,\text{risen\_softplus}\left(\operatorname{Im}\left(\sigma'_t\right)\right), \quad (9)$$

where

$$\text{risen\_softplus}\left(\nu\right) = \sigma_{\text{floor}} + \left(1-\sigma_{\text{floor}}\right)\log\left(1+\mathrm{e}^{\nu}\right), \quad (10)$$

and $\sigma_{\text{floor}}\in(0,\infty)$ is small. This yields the following conditional probability density estimate for the target features y_t:

$$p\left(y_t \middle| x_t, \{(x_c, y_c)\}_{c=1}^{C}\right) \approx \ldots \mathcal{N}\left(y_t^{\text{Re}}; \mu_t^{\text{Re}}, \Sigma^{\text{Re}}\right)\mathcal{N}\left(y_t^{\text{Im}}; \mu_t^{\text{Im}}, \Sigma^{\text{Im}}\right), \quad (11)$$

where $\mathcal{N}$ denotes the multivariate normal distribution, $y_t^{\text{Re}} = \text{flatten}(\operatorname{Re}(y_t))$, $\mu_t^{\text{Re}} = \text{flatten}(\operatorname{Re}(\mu_t))$, $\Sigma^{\text{Re}} = \text{diag}(\text{flatten}(\operatorname{Re}(\sigma_t)))^2$, $y_t^{\text{Im}} = \text{flatten}(\operatorname{Im}(y_t))$, $\mu_t^{\text{Im}} = \text{flatten}(\operatorname{Im}(\mu_t))$, $\Sigma^{\text{Im}} = \text{diag}(\text{flatten}(\operatorname{Im}(\sigma_t)))^2$, and flatten reshapes the tensor provided as argument into a vector.

C. Interpolation Accuracy

In this work, we compare the performance of the SConvCNP model to Gaussian process regressor, thin-plate spherical spline and barycentric interpolation baselines. Interpolation is carried out for all three methods on the SConvCNP model's input features. Similarly to [15], the thin-plate spherical spline method is implemented following Whaba [50], using second-order splines and without smoothing. Gaussian process regression is conducted on a per-frequency basis similarly to Luo et al. [16].

[4] [Online]. Available: https://github.com/google-research/google-research/tree/master/spin_spherical_cnns

At each frequency, we define a covariance function for the real part and one for the imaginary part of the bin using the spherical Gaussian kernel from (8). The observational noise of the model is fixed with a value of 1e-4. The remaining meta-parameters, in particular the precision parameters from the spherical Gaussian kernels, are fitted on 340 tasks from the meta-train set under the log marginal likelihood objective [36]. Meta-parameter values are maintained fixed upon evaluation on the meta-test set. In classical fashion, the barycentric interpolation baseline provides each interpolated feature $\hat{y}$ as a convex combination of the values $\{y_i\}_{i=1}^3$ found at the observed data points defining the smallest spherical triangle enclosing the target point location x, i.e.:

$$\hat{y} = \sum_{i=1}^{3} b_i y_i, \tag{12}$$

where b_i denotes the barycentric coordinate of the target location x associated with the i^{th} vertex of said spherical triangle. The barycentric coordinates b_i are computed as ratios of spherical triangle areas, each computed as the sum of the spherical angles.

We also compare the SConvCNP model's HRTF magnitude interpolation performance specifically, to that of a publicly-available implementation[5] of the natural-neighbors interpolation method [13], [51]. In particular, we apply this implementation directly on the HRTF spectrum after downsampling to 33.075 kHz but without any time-alignment pre-processing. More specifically, we run the implementation provided for the NAT-PH variant, which carries out interpolation on the magnitude and phase of the HRTF as described in [13].

Candidate methods are compared using common metrics computed on a per-feature basis, including the relative error (LRE)

$$\text{LRE}\left(m_{f,e}, \hat{m}_{f,e}\right) = 20\log_{10}\left|\frac{\hat{m}_{f,e} - m_{f,e}}{m_{f,e}}\right|, \tag{13}$$

and the log-magnitude distance (LMD)

$$\text{LMD}\left(m_{f,e}, \hat{m}_{f,e}\right) = \left|20\log_{10}\left|\frac{\hat{m}_{f,e}}{m_{f,e}}\right|\right|, \tag{14}$$

where in a slight departure of notation, $\hat{m}$ and m denote here the predicted point-wise time-aligned HRTF spectrum value and the ground truth value respectively, f indexes over the frequency bin, and e indexes over the left and right ears. For completeness, we also report the log-spectral distortion (LSD), which is given in prior work as follows for a whole binaural filter [22]:

$$\text{LSD}\left(m, \hat{m}\right) = \frac{1}{2}\sum_{e=1}^{2}\sqrt{\frac{1}{(N/2+1)}\sum\nolimits_{f=1}^{N/2+1}\left(20\log_{10}\left|\frac{\hat{m}_{f,e}}{m_{f,e}}\right|\right)^2}. \tag{15}$$

D. Uncertainty Calibration

Several methods have been proposed for assessing a regressor's ability to gauge the uncertainty it provides alongside its point-wise predictions [52], [53], [54]. In particular, Levi et al. introduce a specific definition of uncertainty calibration according to which the model is calibrated if, in expectation over the data-generating distribution, the predicted variance it provides matches the squared error it commits upon carrying out the point-wise prediction [54]. In principle, this condition must hold across all possible values for the predicted variance.

In practice, an approximate but tractable verification of this condition can be conducted for a limited number of variance values using a data set of finite size [54]. In such an approach, the resulting set of predicted variance and squared error pairs are divided into equally-sized groups forming non-overlapping contiguous interval divisions of the predicted variance axis. The expectation over the data-generating distribution is approximated within each group as the sample mean of squared error values in the group. The resulting mean squared error (MSE) values obtained for all groups are plotted as a function of the groups' respective mean predicted variance values (MPV). This allows for assessing the degree of miss-calibration. In particular, overconfident models produce an MPV versus MSE curve exceeding the identity line. Under-confident ones produce a curve lying under it. Miss-calibration can be summarized by a single-scalar mean-aggregate of the calibration error [54]. In this work we propose to use the following mean calibration distance (MCD) metric:

$$\frac{1}{D}\sum_{i=1}^{D}\left|10\log_{10}\frac{\text{MSE}_i}{\text{MPV}_i}\right|, \tag{16}$$

where D denotes the number of divisions of the predicted variance axis.

IV. Results

This section summarizes the meta-test set performance of a selected SConvCNP model configuration, which, among other candidates, achieved, after early stopping, near-best meta-validation set performance in both mean relative error level and mean calibration distance metrics according to (13) and (16) respectively. All candidate configurations were trained with a batch size of 8. Both meta-test and meta-validation sets comprised 340 tasks. The selected configuration's spherical CNN and point-wise MLP components both comprise five residual blocks with $M = 128$ channels each. The spherical convolution is implemented using a 64×64 equiangular grid ($G = 64$). Each planar-spherical filter is composed of 7 taps of SH representations interpolated from 16 learnable SH coefficients each [32]. The standard deviation floor σ_{floor} value is 1e-4 in the selected model. A version of the model and code is publicly available online.[6]

Fig. 4 provides an example of HRTF interpolation task using the SConvCNP model. In particular, this example is given for the FABIAN head and torso simulator (subject 1 of the HUTUBS dataset) and a specific draw of 20 context point locations represented in the top diagram of the figure (black markers). The diagram further marks the location of three target locations

[5] [Online]. Available: https://github.com/AudioGroupCologne/SUpDEq

[6] [Online]. Available: https://github.com/etienne-thuillier/np_4_hrtf_interpolation

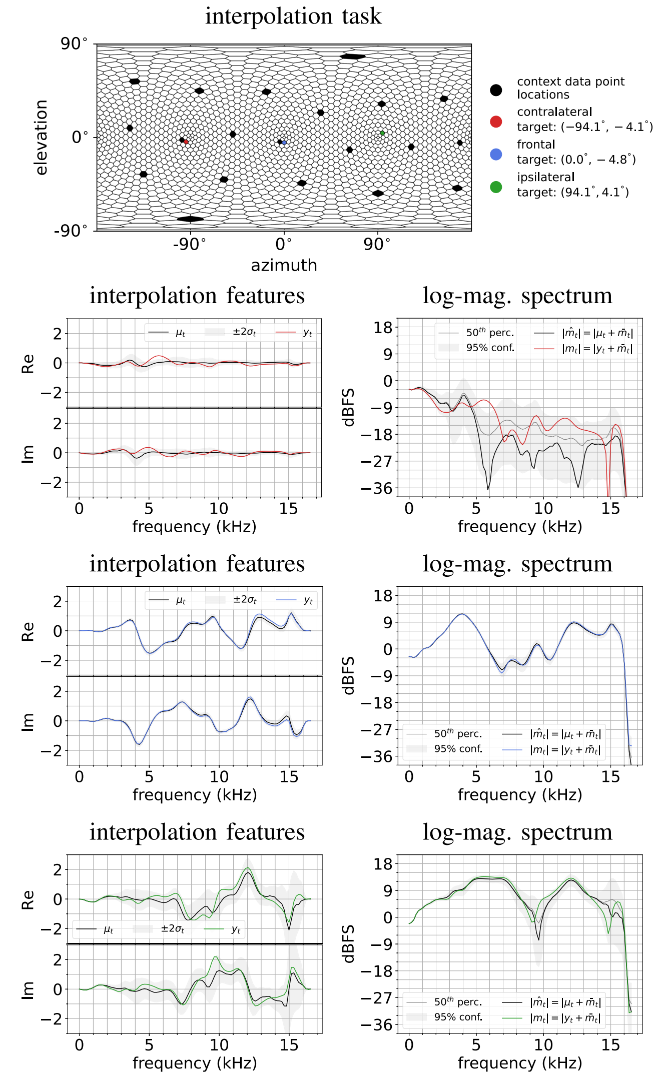

Fig. 4. Time-aligned HRTF spectrum interpolation task example presenting twenty context data points sampled from subject 1 of the HUTUBS database. First row: diagram marking the locations of the context data points (black) as well as three target locations (colored). Left plots: ground truth residual time-aligned HRTF spectrum (colored) and corresponding predictive distribution (black, grey) provided by the SConvCNP for the left channel at the target locations. Right plots: corresponding log-magnitude HRTF spectrum.

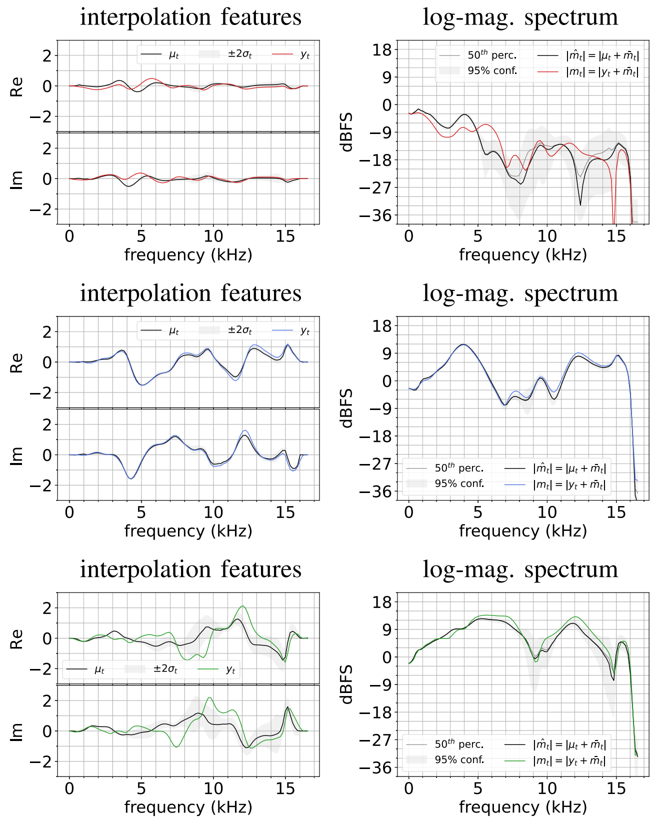

Fig. 5. Left ear channel Gaussian Process solution at the contralateral (top), frontal (middle) and ipsilateral (bottom) target directions of the interpolation task of Fig. 4.

each with a distinct color code. The prediction provided by the SConvCNP model at each target location is reported in a corresponding row of the figure. The left plot of each row compares the predictive distribution to the ground truth. The right plot represents the associated log-magnitude spectra, namely point-wise estimate $\hat{m} = \mu_t + \bar{m}_t$ and ground truth $m = y_t + \bar{m}_t$ where $\bar{m}_t$ denotes the population mean as defined in (7). The 50th percentile (median) and the 95% confidence intervals appearing in the log-magnitude plots are simulated estimates, computed from a population of samples randomly drawn according to the predictive distribution provided by the SConvCNP model.

As pictured, the model's predictive distribution lies in good agreement with the ground truth features y_t. In particular, the ground truth generally falls within the 95% confidence interval ($\pm 2\sigma_t$ range, grey region) around the predictive mean in all three target location cases. Moreover, the predictive mean μ_t of the model (full black line) shows significant correlation with the ground truth y_t in both the ipsilateral direction case (green marker) and frontal direction case (blue marker). Furthermore, the model's prediction is more uncertain when the target point (ipsilateral direction, green marker) lies further away from context data points than in close vicinity (frontal direction, blue marker). This suggests the model's predictive distribution effectively captures model uncertainty.

In contrast, the predictive mean is much less correlated with the ground truth in the contralateral direction case (red marker) despite the target direction being close to a context data point as indicated by the diagram at the top of Fig. 4. In particular, the predictive mean is practically agnostic above the 5-kHz mark, with a near-zero value throughout, and the standard deviation extends significantly outwards from the abscissa to capture variations in ground truth value. This is not unexpected as interpolation is a harder problem in the contra-lateral region, where the HRTF is spatially more intricate such that correlations between data points would occur within small distances only. Given this, the magnitude spectrum estimate $|\hat{m}_t| = |\mu_t + \bar{m}_t|$ significantly undershoots the ground truth (right plot), while the transformed distribution's median (50% percentile) better predicts the power spectrum of the filter in this case (red curve).

Fig. 5 depicts the solution provided by the Gaussian process regressor baseline for the target directions of Fig. 4's task.

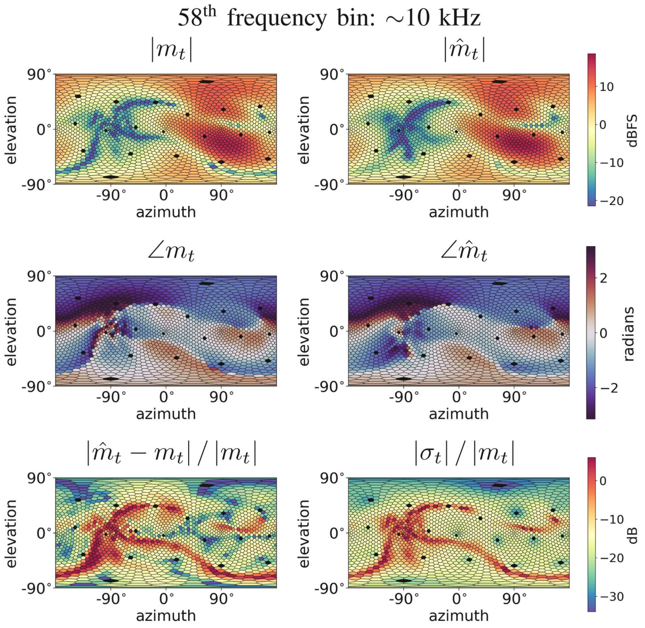

Fig. 6. Left ear channel time-aligned HRTF spectrum as a function of sound source direction x_t for the interpolation task of Fig. 4. Plots are provided for the 58th frequency bin of the spectrum. Top and middle: log-magnitude and phase for ground truth $m_t = y_t + \bar{m}_t$ (left) and SConvCNP model's predictive mean $\hat{m}_t = \mu_t + \bar{m}_t$ (right) of the time-aligned HRTF spectrum. Bottom left: relative error committed by the SConvCNP model. Bottom right: SConvCNP predictive uncertainty relative to ground truth magnitude in decibels.

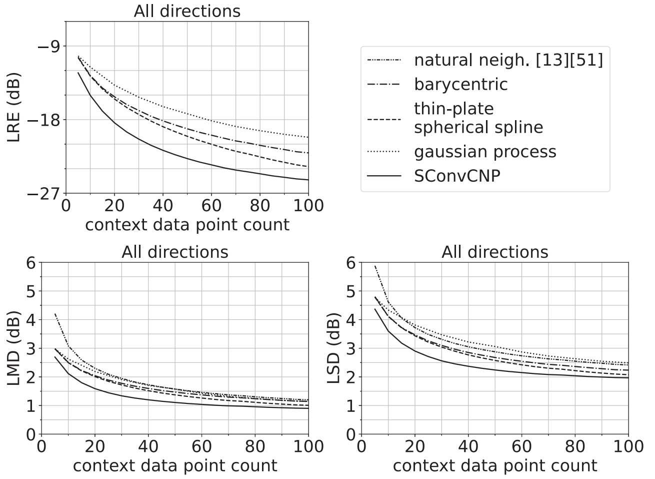

Fig. 7. Average time-aligned HRTF spectrum interpolation error across output features in the 0-15.5 kHz range and meta-test set's interpolation tasks as a function of context data point count. The proposed method (SConvCNP) improves upon all baselines on all three evaluation metrics. Upper-left: relative error level according to (13). Lower-left: log-magnitude distance according to (14). Lower-right: log-spectral distortion according to (15).

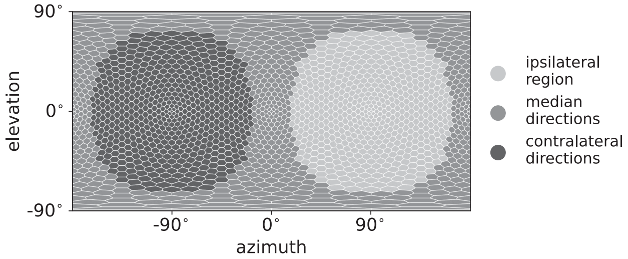

Fig. 8. Definition of HRTF regions used in generating the plots of Figs. 9 and 10. The boundaries separating the regions lies at ±18.1° lateral angle from the median plane, which distributes the HRTF directions of the HUTUBS grid in approximately equal proportions amongst the three specified regions.

Contrary to SConvCNP, the Gaussian process' predictive distribution does not capture the variability of the ground truth features along the frequency axis for any of the targets. Crucially, the uncertainty estimates are of similar value at any given frequency in the contralateral and frontal target cases (top-left and middle-left plots). This is expected since the Gaussian process' predictive uncertainty is solely a function of distance to context data points [36], which is similar for both these target locations in the example. In this baseline specifically, the modeled degree of correlation between feature values at distinct locations on the unit sphere is tuned by the precision meta-parameter of the covariance function, and is hence equal everywhere on the sphere, including in the contralateral and ipsilateral regions. This contrasts starkly with the SConvCNP model's ability to provide well-behaved uncertainty estimates in both these regions as seen in Fig. 4. Contrary to the Gaussian Process model, the SConvCNP is allowed to exploit information from the feature values in order to predict uncertainty, which could explain this ability.

The magnitude and phase responses of time-aligned HRTF spectrum is represented on HUTUBS' data point grid for the 58th frequency bin in Fig. 6. As pictured, the SConvCNP model's mean estimate $\hat{m}_t = \mu_t + \bar{m}_t$ closely matches the ground truth $m_t = y_t + \bar{m}_t$ both in terms of magnitude and phase (top and middle plots). More precisely, the predictive mean solution's error generally lies under the -15 dB threshold relative to ground truth outside low-magnitude areas on the unit sphere (lower left plot). Furthermore, the predictive uncertainty provided by the model seems generally consistent with the observed error (lower right plot).

A sample efficiency comparison of candidate methods is provided in Figs. 7 and 9. In particular, the graphs of these figures report error scores as a function of the number of context data points provided to the interpolation method candidates. Fig. 7 includes plots for the LRE, LMD and LSD metrics as defined in (13), (14) and (15) respectively. Fig. 9 provides further detail for the LRE metric specifically. In this figure, three additional LRE plots are provided for the ipsilateral, median, and contralateral HRTF regions defined in Fig. 8. Error levels are provided in each plot of Fig. 9 for three distinct frequency bands: 0–5 kHz, 5–10 kHz, and 10–15 kHz. The natural neighbor method is intentionally omitted from the LRE plots of Figs. 7 and 9 as this candidate can only be meaningfully compared on magnitude-error-metric grounds since it interpolates the HRTF spectrum without the time-alignement pre-processing. In both Figs. 7 and 9, each curve represents an average error score value taken across tasks, directions, left/right ear channels and, the case being, frequency bins. In particular, the average was

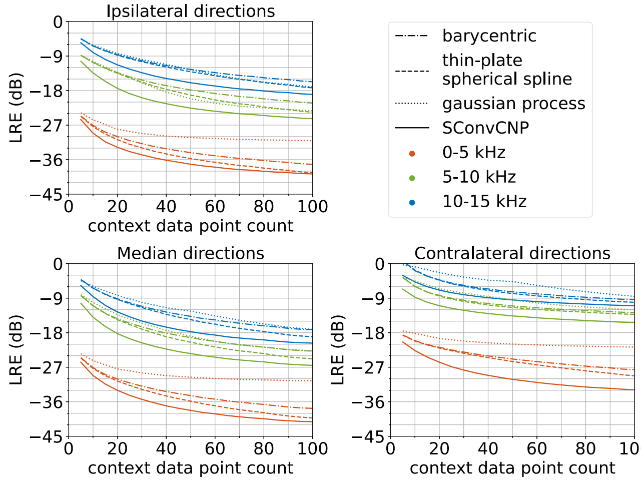

Fig. 9. Per frequency-band average time-aligned HRTF spectrum interpolation error across output features and the meta-test set's interpolation tasks as a function of context data point count. The proposed method (SConvCNP) improves upon all baselines in all specified regions and in each frequency-band. Upper-left, lower-left and lower-right: relative error level according to (13) for HRTF direction sub-regions defined in Fig. 8.

TABLE III
Per-Subject Mean Relative Error Level

thin-plate spherical spline	SConvCNP	difference
-18.31 dB	-20.86 dB	2.56 dB
-17.13 dB	-19.81 dB	2.68 dB
-18.92 dB	-20.90 dB	1.98 dB
-17.21 dB	-19.53 dB	2.32 dB

taken for each reported count over a set of 340 randomly drawn meta-test tasks.

As expected, all candidates exhibit monotonically decreasing error scores with increased sample count. This observation holds for all error metrics considered in Fig. 7 and for all regions and frequency intervals considered in Fig. 9. Moreover, the SConvCNP model presents significantly lower relative error level compared to the thin-plate spherical spline method, which forms the best baseline: up to 3 dB globally (top left plot of Fig. 7) and up to 4.5 dB in the 0–5 kHz range (contralateral region, bottom right plot of Fig. 9). This improvement translates to nearly a halving of required measurement count to meet an error specification level. For example, meeting a -20 dB average relative error requires approximately 50 measurements using the thin-plate spherical spline method while approximately 28 is sufficient on average using the SConvCNP model. Similar observations can be made in the case of both the LMD and LSD metrics pictured in Fig. 7.

Table III contrasts the relative error level between the proposed method and the best performing baseline (thin plate spherical spline) for all four subjects of the test set taken individually. Each reported mean value is taken over a set of 85 interpolation tasks from the meta-test set described in Section III-A. Each individual mean error difference of the table forms a sample independent from the others. Under the reasonable assumption that the underlying distribution of the mean error difference values is normal, a (paired, one-tailed) Student's t-test analysis can be carried-out to determine if the proposed method provides a statistically significant improvement over the best performing baseline [55]. We conduct this test with a significance level value of 0.05, which translates to a critical value of 2.35. The test's t-value is given as:

$$t = \frac{\mu_{\text{diff}}}{\sigma_{\text{diff}}/\sqrt{n_{\text{test}}}} \tag{17}$$

where μ_{diff} denotes the sample average of the individual mean error difference values reported in Table III, σ_{diff} denotes the associated sample deviation and n_{test} denotes the number of subjects in the test set. The resulting t-value greatly surpasses the critical value of the test: $t = 15.4 > 2.35$. This corresponds to a p-value of $2.9e-4$, which lies two orders of magnitude under the significance level. Hence we conclude that the SConvCNP model does provide a statistically-meaningful reduction of relative error level over the best performing baseline.

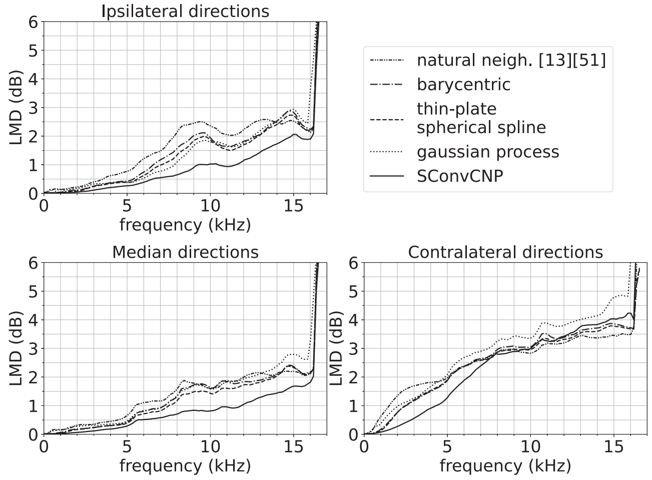

Fig. 10. Average LMD (log-magnitude distance) as a function of frequency for meta-test tasks with numbering 40 context data points. Each plot corresponds to one of the three HRTF direction regions defined in Fig. 8.

Fig. 10 provides a summary of log-magnitude distance level as a function of frequency in the specific case of context sets numbering 40 context data points. The three plots of the figure detail the error levels specific to each region defined in Fig. 8. The SConvCNP model significantly outperforms all baselines in the 0–14 kHz across all regions, except in the contralateral region (top-right plot) where the natural neighbour matches and then outperforms the proposed model from the 7.5 kHz mark onwards. In agreement with the results of Fig. 9, the improvements brought by the SConvCNP model are most significant beyond the 6 kHz mark in the frontal and ipsilateral region, while it is most significant under 6 kHz in the contralateral region. In particular, the SConvCNP model provides an improvement of up to 0.8 dB compared to the best baseline at any frequency, as found in the ipsilateral region around the 9.2 kHz mark.

Miscalibration of the trained SConvCNP model is summarized in Fig. 11. In this figure, the calibration of the trained SConvCNP model is evaluated over a meta-test set of 340 randomly generated tasks using $D = 16$ divisions. In particular, the

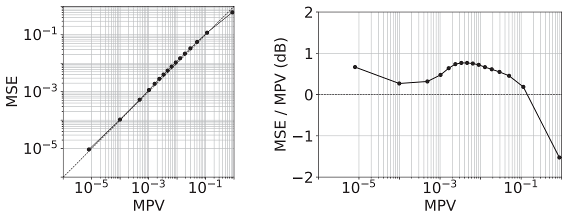

Fig. 11. Miscalibration of the trained SConvCNP model's uncertainty estimates. Left: mean predicted variance (horizontal axis) versus mean square error (vertical axis) plot. Perfect calibration corresponds to the identity line (dashed). Right: miscalibration level in decibels.

predicted variance and squared error pairs observed at the output of the model are pooled across interpolation tasks, data point locations, frequency bins, left/right channels, and real/imaginary parts to form the MSE versus MPV curve shown in Fig. 11.

As pictured in the left plot of Fig. 11, the mean predicted variance closely matches the mean square error. In particular, the resulting curve lies neither significantly above or below the identity line. Hence, the trained SConvCNP model is neither markedly over-confident or under-confident in its predictions. More precisely, the rightmost plot reveals that the effective squared error lies, in expectation and for all but most uncertain predictions, within 1.0 dB of the predicted variance. This level of miscalibration is moderate when put in contrast to the range of relative error reduction that is achievable when using interpolation. In particular, the mean relative error reaches -9.1 dB before acquisition of the first observations (not represented in the figures) and drops under -21 dB at the 40 points mark as pictured in the top-left plot of Fig. 7. Accordingly, we conclude that the model's uncertainty estimates σ_t could usefully inform the problem of acquiring additional HRTF data points to improve HRTF individualization upon a pre-existing HRTF estimate.

V. Conclusion

In this work we introduced a model for HRTF interpolation which, for the first time, outputs well-calibrated uncertainty estimates. We showed the method proved sample efficient on the time-aligned HRTF spectrum interpolation task. In particular, meta-training was carried-out successfully using a modest data set of 85 subjects. Furthermore, the interpolators returned by the proposed meta-learning model were shown to require up to nearly half the number of context data point count compared to state-of-the-art interpolation methods at comparable accuracy level. In particular, as few as 28 points are required to reach an average of -20 dB relative error per interpolated feature, compared to 50 points in the case of the best performing baseline. Contrary to the Gaussian process regression baseline, the interpolators also showed well-calibrated uncertainty estimates.

The proposed model's time and space complexity severely limits its applicability towards real-time interpolation and audio rendering setups. However it can readily be used for offline up-sampling of sparse HRTF sets. Furthermore, a promising application lies in facilitating the sequential decision problem of acquiring as few correcting HRTF data-points as needed to achieve a required degree of HRTF individualization accuracy. In particular, the provided uncertainty estimates could be used to identify the location at which obtaining a new measurement would, in expectation, maximally reduce the model's uncertainty. Furthermore, the predictive distribution could be used to compare the probability of HRTF query candidates conditioned on the data points already acquired for the subject so as to select the most relevant ones to be submitted for perceptual feedback evaluation from the subject. Treatment of such sequential decision problem is left as future research work. Other future development avenues include evaluation of the model's ability to correct HRTF estimates provided by state-of-the-art parametric individualization methods instead of the train set population mean used under the limited scope of this work.

References

[1] P. Crum, "Here come the hearables: Technology tucked inside your ears will augment your daily life," *IEEE Spectr.*, vol. 56, no. 5, pp. 38–43, May 2019.

[2] R. Gupta et al., "Augmented/mixed reality audio for hearables: Sensing, control, and rendering," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 63–89, May 2022.

[3] J. Yang, A. Barde, and M. Billinghurst, "Audio augmented reality: A systematic review of technologies, applications, and future research directions," *J. Audio Eng. Soc.*, vol. 70, no. 10, pp. 788–809, Oct. 2022.

[4] J. Herre and S. Disch, "MPEG-I immersive audio—Reference model for the virtual/augmented reality audio standard," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 229–240, May 2023.

[5] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, 1996.

[6] C. Guezenoc and R. Seguier, "HRTF individualization: A survey," in *Proc. 145th Conv. Audio Eng. Soc.*, 2018. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=19855

[7] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *J. Acoustical Soc. Amer.*, vol. 136, no. 2, pp. 791–802, 2014.

[8] P. Majdak, C. Hollomey, and R. Baumgartner, "AMT 1.x: A toolbox for reproducible research in auditory modeling," *Acta Acustica*, vol. 6, 2022, Art. no. 19, doi: 10.1051/aacus/2022011.

[9] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *J. Audio Eng. Soc.*, vol. 55, no. 7/8, pp. 623–637, 2007.

[10] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Gaussian process models for HRTF based sound-source localization and active-learning," 2015, *arXiv:1502.03163*.

[11] R. Sundareswara and P. Schrater, "Extensible point location algorithm," in *Proc. Int. Conf. Geometric Model. Graph.*, 2003, pp. 84–89.

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

[13] C. Pörschmann, J. M. Arend, D. Bau, and T. Lübeck, "Comparison of spherical harmonics and nearest-neighbor based interpolation of head-related transfer functions," in *Proc. AES Int. Conf. Audio Virtual Augmented Reality*, 2020. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20874

[14] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized HRTF fitting using spherical harmonics," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 257–260.

[15] S. Carlile, C. Jin, and J. Leung, "Performance measures of the spatial fidelity of virtual auditory space: Effects of filter compression and spatial sampling," in *Proc. Int. Conf. Auditory Display*, 2002.

[16] Y. Luo, D. N. Zotkin, H. Daumé, and R. Duraiswami, "Kernel regression for head-related transfer function interpolation and spectral extrema extraction," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 256–260.

[17] B.-S. Xie, "Recovery of individual head-related transfer functions from a small set of measurements," *J. Acoust. Soc. Amer.*, vol. 132, no. 1, pp. 282–294, 2012.

[18] I. Ben Hagai, M. Pollow, M. Vorländer, and B. Rafaely, "Acoustic centering of sources measured by surrounding spherical microphone arrays," *J. Acoust. Soc. Amer.*, vol. 130, no. 4, pp. 2003–2015, Oct. 2011, doi: 10.1121/1.3624825.

[19] M. Aussal, "Méthodes numériques pour la spatialisation sonore, de la simulation á la synthése binaurale," Ph.D. dissertation, Ecole Polytechnique X, Palaiseau, France, Oct. 2014. [Online]. Available: https://pastel.hal.science/tel-01095801

[20] J.-G. Richter, M. Pollow, F. Wefers, and J. Fels, "Spherical harmonics based HRTF datasets: Implementation and evaluation for real-time auralization," *Acta Acustica United Acustica*, vol. 100, no. 4, pp. 667–675, 2014.

[21] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Efficient representation and sparse sampling of head-related transfer functions using phase-correction based on ear alignment," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2249–2262, Dec. 2019.

[22] Y. Ito, T. Nakamura, S. Koyama, and H. Saruwatari, "Head-related transfer function interpolation from spatially sparse measurements using autoencoder with source position conditioning," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2022, pp. 1–5.

[23] J. W. Lee, S. Lee, and K. Lee, "Global HRTF interpolation via learned affine transformation of hyper-conditioned features," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

[24] X. Chen, F. Ma, Y. Zhang, A. Bastine, and P. N. Samarasinghe, "Head-related transfer function interpolation with a spherical CNN," 2023, *arXiv:2309.08290*.

[25] A. O. 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," 2023, *arXiv:2306.05812*.

[26] B. Zhi, D. N. Zotkin, and R. Duraiswami, "Towards fast and convenient end-to-end HRTF personalization," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 441–445.

[27] D. Yao et al., "An individualization approach for head-related transfer function in arbitrary directions based on deep learning," *JASA Exp. Lett.*, vol. 2, no. 6, 2022, Art. no. 064401.

[28] G. W. Lee and H. K. Kim, "Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear," *Appl. Sci.*, vol. 8, no. 11, 2018, Art. no. 2180.

[29] K. Yamamoto and T. Igarashi, "Fully perceptual-based 3D spatial sound individualization with an adaptive variational autoencoder," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–13, 2017.

[30] R. Miccini and S. Spagnol, "A hybrid approach to structural modeling of individualized HRTFs," in *Proc. Conf. Virtual Reality 3D User Interfaces Abstr. Workshops*, 2021, pp. 80–85.

[31] J. Gordon, W. P. Bruinsma, A. Y. K. Foong, J. Requeima, Y. Dubois, and R. E. Turner, "Convolutional conditional neural processes," in *Proc. 8th Int. Conf. Learn. Representations*, 2020, pp. 26–30. [Online]. Available: https://openreview.net/forum?id=Skey4eBYPS

[32] C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis, "Learning SO(3) equivariant representations with spherical CNNs," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 52–68.

[33] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical CNNs," in *Proc. 6th Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Hkbd5xZRb

[34] M. Garnelo et al., "Conditional neural processes," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1704–1713. [Online]. Available: https://proceedings.mlr.press/v80/garnelo18a.html

[35] M. Garnelo et al., "Neural processes," 2018, *arXiv:1807.01622*.

[36] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning.*, Cambridge, MA, USA: MIT Press, Nov. 2005, doi: 10.7551/mitpress/3206.001.0001.

[37] M. Zaheer, S. S. K. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 3394–3404. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf

[38] A. Y. K. Foong, W. P. Bruinsma, J. Gordon, Y. Dubois, J. Requeima, and R. E. Turner, "Meta-learning stationary stochastic process prediction with convolutional neural processes," in *Proc. Conf. Neural Inf. Process. Syst.*, 2020, pp. 8284–8295. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/5df0385cba256a135be596dbe28fa7aa-Abstract.html

[39] Y. Dubois, J. Gordon, and A. Y. Foong, "Neural process family," Accessed: Oct. 11, 2023. [Online]. Available: http://yanndubs.github.io/Neural-Process-Family/

[40] G. E. Fasshauer, "Positive definite kernels: Past, present and future," *Dolomites Res. Notes Approximation*, vol. 4, pp. 21–63, 2011.

[41] E. A. Nadaraya, "On estimating regression," *Theory Probability Appl.*, vol. 9, no. 1, pp. 141–142, 1964.

[42] J. Schmidhuber, "Evolutionary principles in self-referential learning, or on learning how to learn: The meta-meta-...hook," Ph.D. dissertation, Technische Universität München, Munich, Germany, 1987.

[43] J. Nam, J. S. Abel, and J. O. Smith III, "A method for estimating interaural time difference for binaural synthesis," in *Proc. 125th Audio Eng. Soc.*, 2008. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14763

[44] C. Pörschmann and J. M. Arend, "A method for spatial upsampling of voice directivity by directional equalization," *J. Audio Eng. Soc.*, vol. 68, no. 9, pp. 649–663, 2020.

[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, Sep. 2019.

[46] B. Fabian et al., "The HUTUBS head-related transfer function (HRTF) database," 2019, doi: 10.14279/depositonce-8487.

[47] C. Esteves, A. Makadia, and K. Daniilidis, "Spin-weighted spherical CNNs," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 8614–8625. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6217b2f7e4634fa665d31d3b4df81b56-Abstract.html

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Comput. Vis. Conf.*, 2016, pp. 630–645.

[49] T. A. Le, H. Kim, M. Garnelo, D. Rosenbaum, J. Schwarz, and Y. W. Teh, "Empirical evaluation of neural process objectives," in *Proc. NeurIPS Workshop Bayesian Deep Learn.*, 2018.

[50] G. Wahba, "Spline interpolation and smoothing on the sphere," *SIAM J. Sci. Stat. Comput.*, vol. 2, no. 1, pp. 5–16, 1981.

[51] J. M. Arend, C. Pörschmann, S. Weinzierl, and F. Brinkmann, "Magnitude-corrected and time-aligned interpolation of head-related transfer functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3783–3799, Sep. 2023, doi: 10.1109/TASLP.2023.3313908.

[52] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 69, no. 2, pp. 243–268, 2007.

[53] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2796–2804. [Online]. Available: https://proceedings.mlr.press/v80/kuleshov18a.html

[54] D. Levi, L. Gispan, N. Giladi, and E. Fetaya, "Evaluating and calibrating uncertainty prediction in regression tasks," *Sensors*, vol. 22, no. 15, 2022, Art. no. 5540.

[55] J. Brownlee, *Statistical Methods for Machine Learning: Discover How to Transform Data into Knowledge With Python.* Vermont, VIC, Australia: Mach. Learn. Mastery, 2018.

Etienne Thuillier (Member, IEEE) is currently working toward the Ph.D. degree from Aalto University, Espoo, Finland. His research focuses on the application of machine learning for HRTF individualization. Originally a telecommunications engineer from the Ecole Polytechnique de Montréal, Etienne Thuillier started his career as a patent consultant before reconverting to technical practice through a startup creation experience and master's studies in audio signal processing and acoustics at Aalto University. He subsequently worked on research and development projects at Apple, Microsoft Research and Meta Reality Labs.

Craig T. Jin (Senior Member, IEEE) received the M.S. degree in applied physics from Caltech, Pasadena, CA, USA, in 1991, and the Ph.D. degree in electrical engineering from the University of Sydney, Sydney, NSW, Australia, in 2001. He is currently an Associate Professor with the School of Electrical and Computer Engineering, University of Sydney, and the Director of the Computing and Audio Research Laboratory. His research interests include experimental and theoretical aspects of acoustic and biomedical signal processing. He has authored more than 200 papers in these research areas, holds eight patents, and founded three start-up companies. He received national recognition in Australia (April 2005, Science in Public Fresh Innovators Program) for his invention of a spatial hearing aid.

Vesa Välimäki received the M.Sc. degree in technology and the Doctor of Science in Technology degree in electrical engineering from the Helsinki University of Technology, Espoo, Finland, in 1992 and 1995, respectively. He is currently a Full Professor of audio signal processing and the Vice Dean of Research with Aalto University, Espoo, Finland. His doctoral dissertation dealt with fractional delay filters and physical modeling of musical instruments. In 1996, he was a Postdoctoral Research Fellow with the University of Westminster, London, U.K. During 2001–2002, he was a Professor of signal processing with the Pori School of Technology and Economics, Tampere University of Technology, Pori, Finland. In 2002, he was appointed Professor of audio signal processing with TKK. During 2008–2009, he was on sabbatical as a Visiting Scholar with the Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA, USA. His research interests include signal processing and machine learning applied to audio and music technology. Prof. Välimäki is a Fellow of the AES, and a Life Member of the Acoustical Society of Finland. During 2015–2020, he was the Senior Area Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He was the General Chair of the 14th Sound and Music Computing Conference SMC-17 in 2017. He is the Editor-in-Chief of the *Journal of the Audio Engineering Society*.