Tutorial: Personalizing Spatial Audio with Machine Learning

Audio Engineering Society

2024 AES 6th International Conference on Audio for Games

audio for games

Machine Learning for Personalized Head-Related Transfer Functions (HRTFs) Modeling in Gaming

2024 AES 6TH INTERNATIONAL CONFERENCE ON

Audio for Games

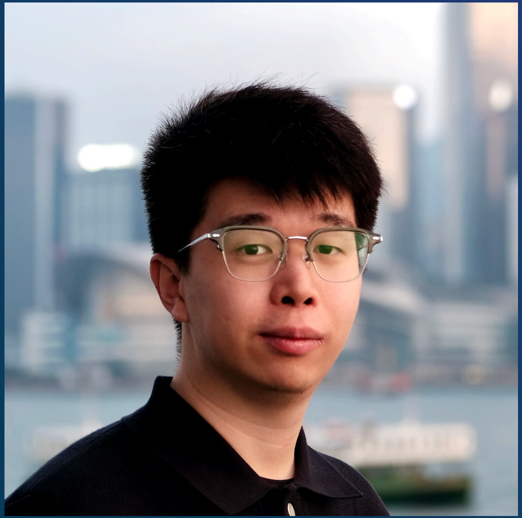
Interactive Audio Innovation, Rediscovering the Heritage.

APRIL 27-29, 2024 | SENJU CAMPUS, TOKYO UNIVERSITY OF THE ARTS

Audio Engineering Society

2024 6TH AES INTERNATIONAL CONFERENCE
ON AUDIO FOR GAMES
APRIL 27-29, 2024, TOKYO

audio for games



You (Neil) Zhang

PhD Candidate
University of Rochester, New York, USA

> **Working on Audio and Speech Machine Learning**

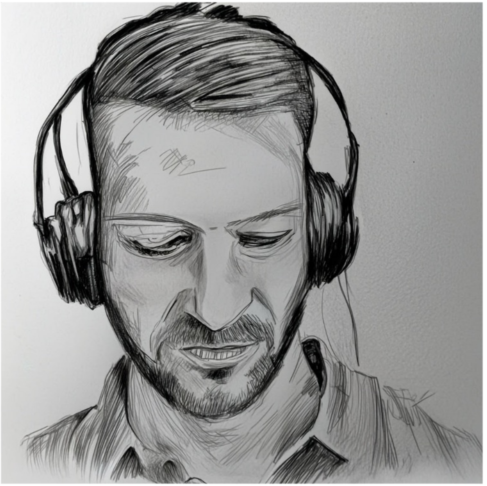
LinkedIn

Personal website

Immersive Audio Environment for Gaming

Figures generated by DALL-E 3

Spatial Audio Rendering



Headphone

Loudspeakers

VR headset

Figures generated by Duet AI

Spatial Effects and Sound Localization

Localize sound sources with differences between sounds received by two ears.

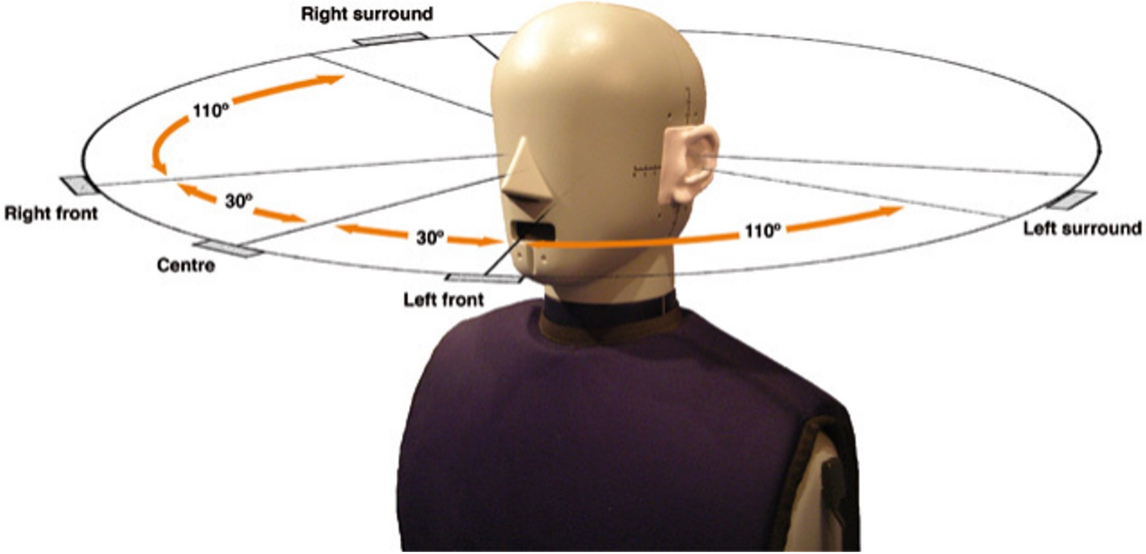


Figure from https://www.soundonsound.com/reviews/mp3-surround

Head-Related Transfer Function (HRTF)

Sound propagation is modeled as a linear **filtering** process from source to ears, including **spectral changes** due to the shape of ear, head, and torso.

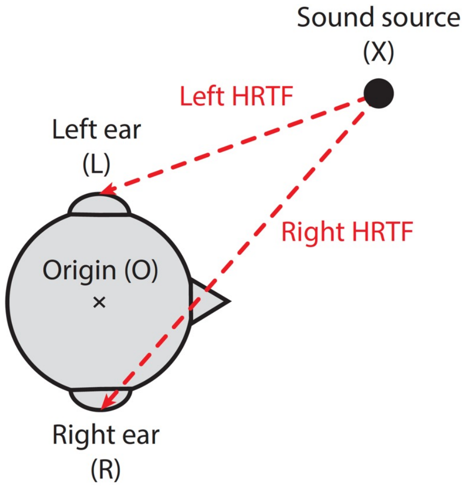


Figure from Isaac Engel's thesis

Left ear HRTF magnitudes (dB) of the midsagittal plane of one subject

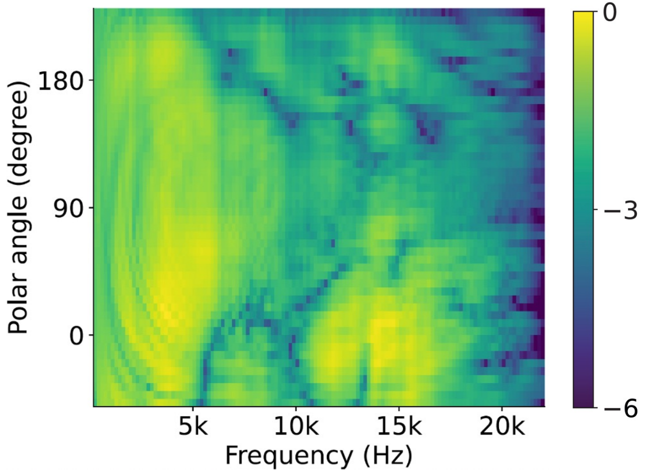


Figure from [Zhang+2023]

Audio Engineering Society

2024 6TH AES INTERNATIONAL CONFERENCE
ON AUDIO FOR GAMES
APRIL 27-29, 2024, TOKYO

audio for games

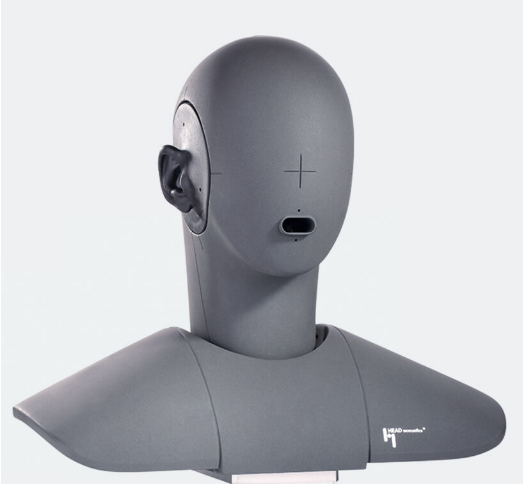
HRTF & Machine Learning Basics

Generic HRTF

Based on worldwide average human head and torso dimensions

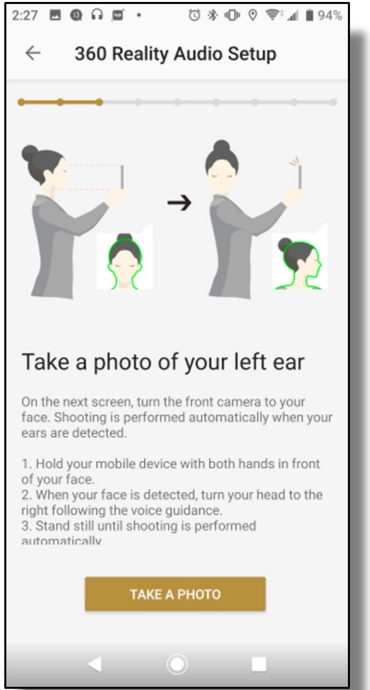
Neumann KU-100

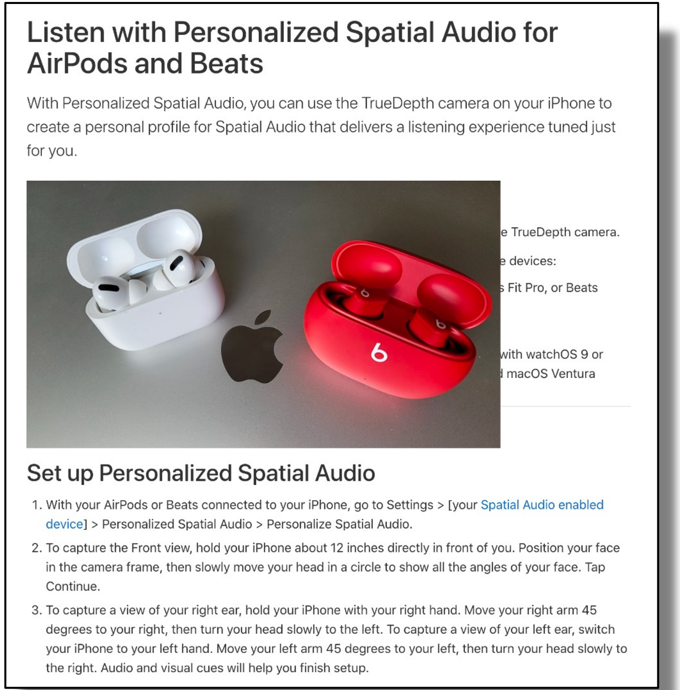
KEMAR 45BB-1



HEAD Acoustics HMS II.5

Personalized HRTF in the Latest Devices





Listen with Personalized Spatial Audio for AirPods and Beats

With Personalized Spatial Audio, you can use the TrueDepth camera on your iPhone to create a personal profile for Spatial Audio that delivers a listening experience tuned just for you.

e TrueDepth camera.
e devices:
s Fit Pro, or Beats
with watchOS 9 or
l macOS Ventura

Set up Personalized Spatial Audio

1. With your AirPods or Beats connected to your iPhone, go to Settings > [your Spatial Audio enabled device] > Personalized Spatial Audio > Personalize Spatial Audio.
2. To capture the Front view, hold your iPhone about 12 inches directly in front of you. Position your face in the camera frame, then slowly move your head in a circle to show all the angles of your face. Tap Continue.
3. To capture a view of your right ear, hold your iPhone with your right hand. Move your right arm 45 degrees to your right, then turn your head slowly to the left. To capture a view of your left ear, switch your iPhone to your left hand. Move your left arm 45 degrees to your left, then turn your head slowly to the right. Audio and visual cues will help you finish setup.

Why Personalized HRTFs?

Benefits:

- Optimal sound source localization perception [Majdak+2013]
- Natural coloration [Brinkmann+2017]
- Easier to localize, easier to externalize, and more natural in timbre [Jenny&Reuter2020]

Important in spatial audio for games!

Majdak, Piotr, Bruno Masiero, and Janina Fels. “Sound localization in individualized and non-individualized crosstalk cancellation systems.” *JASA* 2013.
Brinkmann, Fabian, Alexander Lindau, and Stefan Weinzierl. “On the authenticity of individual dynamic binaural synthesis.” *JASA* 2017.
Jenny, Claudia, and Christoph Reuter. "Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization." *JMIR Serious Games* 2020.

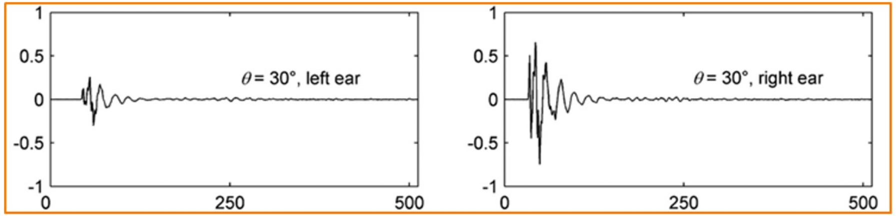
Measure Personalized HRTFs

- Two microphones were inserted in the listeners' ears.
- Multiple loudspeakers are arranged around a vertical arc, which rotates horizontally.
- Drawbacks:
  - Requires an anechoic room
  - Time-consuming
  - Cannot measure arbitrary locations

Figure from https://ieeexplore.ieee.org/document/7099223

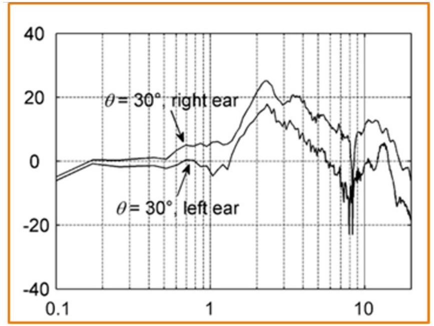
HRIR and HRFR

HRIR - Head-Related Impulse Response



HRFR - Head-Related Frequency Response

Fourier Transform



Personalizing HRTF with Simulation

Finite difference method (FDM) [Tian&Liu2003], Boundary element method (BEM) [Kreuzer+2009], Finite element method (FEM) [Ma+2015]

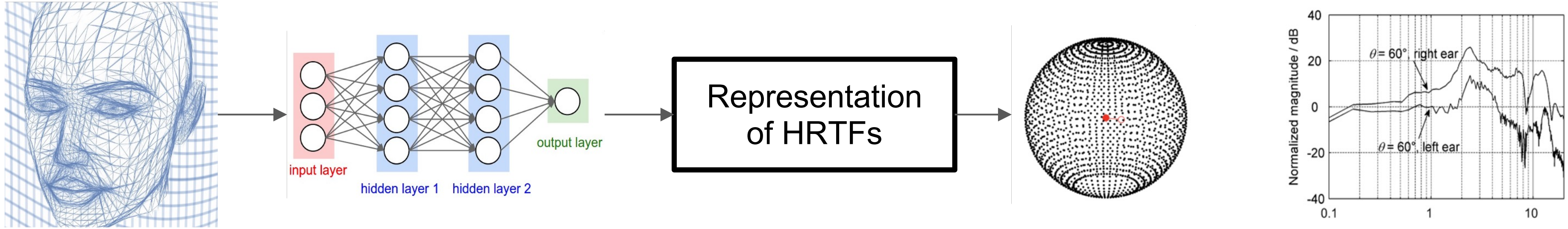
Drawbacks:

- Depend on the availability of precise 3D geometry
- Under unrealistic physics assumptions
- Computationally expensive

Xiao, Tian, and Qing Huo Liu. "Finite difference computation of head-related transfer function for human hearing." *JASA* 2003.
Kreuzer, Wolfgang, Piotr Majdak, and Zhengsheng Chen. "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range." *JASA* 2009.
Ma, Fuyin, et al. "Finite element determination of the head-related transfer function." *JMMB* 2015.

Personalizing HRTF with Machine Learning

Leverage measured data for personalized HRTF prediction



Human physical geometry

HRTF prediction network

HRTFs at various spatial locations (of arbitrary spatial sampling schemes)

HRTFs at a particular position

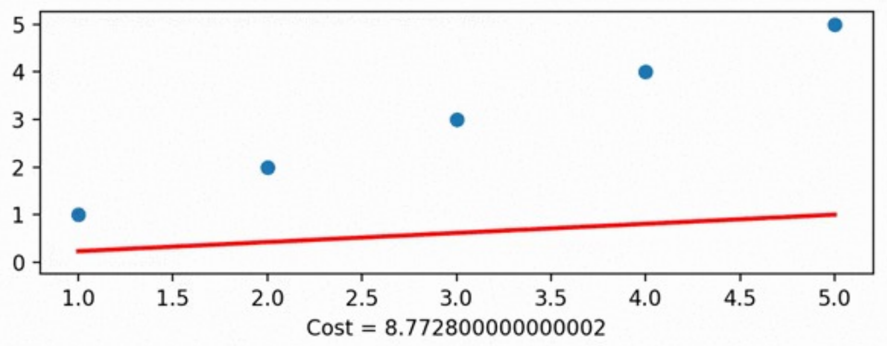
Assumption: Many things are common across people (captured by the model), and other effects are personalized (captured by adapting the input).

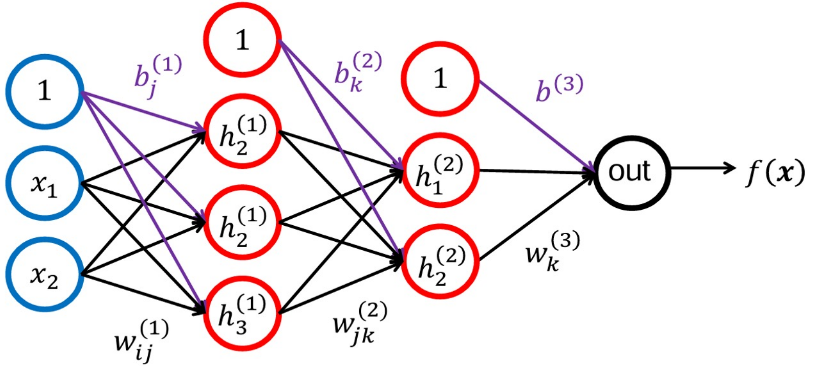
Machine Learning & Deep Learning

Learning from data observations

Gradient descent with loss functions

Neural networks better model non-linearity.





$$f(\boldsymbol{x}) = \sigma\left(\sum_k w_k^{(3)} \sigma\left(\sum_j w_{jk}^{(2)} \sigma\left(\sum_i w_{ij}^{(1)} x_i + b_j^{(1)}\right) + b_k^{(2)}\right) + b^{(3)}\right)$$

$$f(\boldsymbol{x}) = \sigma\left(\boldsymbol{W}_3^T \boldsymbol{\sigma}\left(\boldsymbol{W}_2^T \boldsymbol{\sigma}\left(\boldsymbol{W}_1^T \boldsymbol{x} + \boldsymbol{b}_1\right) + \boldsymbol{b}_2\right) + b_3\right)$$

Data hungry!

Figure from https://youtu.be/rqENjJYWo34

Machine Learning & Deep Learning (Cont'd)

Given examples $(\boldsymbol{X}, \boldsymbol{y})$, learn model $f: \boldsymbol{x} \mapsto y$

Weight (g)	Color	Shape	Taste (1-5)	Calories	Water (%)	**Label**
200	red	round	4	100	85	**apple**
140	orange	round	5	66	86	**orange**
120	yellow	long	5	105	75	**banana**
150	green	round	1	70	90	**apple**
110	green	long	2	100	73	**banana**
200	orange	round	3	85	90	**orange**
118	yellow	long	3	103	72	**banana**
180	red	round	3	81	87	**apple**

Training set: develop the model f; Validation set: tune hyperparameters

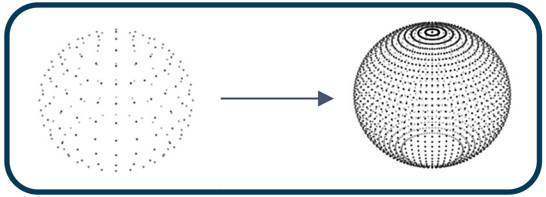
Test set: evaluation the model f

Tasks & Existing Methods

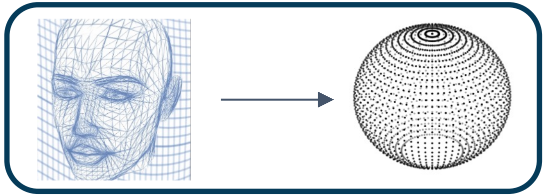
Personalized HRTF Modeling

Two research tasks:

- HRTF Upsampling / Interpolation
  (use known locations to predict unknown)



- HRTF Personalization from Human Input
  (anthropometry, ear shape, head mesh)



Evaluation Metric

Objective evaluation: Log-spectral distortion (LSD)

$$LSD(\mathrm{H}, \hat{\mathrm{H}}) = \sqrt{\frac{1}{LK} \sum_{\theta,\phi} \sum_{k} \left(20 \log_{10} \left| \frac{\mathrm{H}(\theta, \phi, k)}{\hat{\mathrm{H}}(\theta, \phi, k)} \right| \right)^2}$$

ground-truth linear-scale magnitude

spatial locations

predicted linear-scale magnitude

frequency bins

frequency index

Evaluation Metric (Cont’d)

Subjective evaluation

- Auditory models

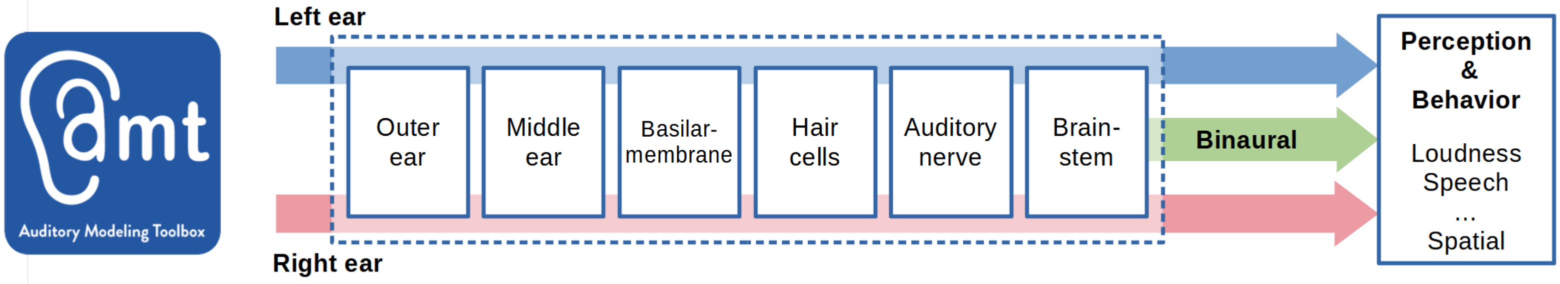


Figure from https://amtoolbox.org/

- Human listening test

Signal Processing-Based Methods for Interpolation

- Vector-based amplitude panning (VBAP) [Pulkki1997]
- 3D bilinear interpolation [Freeland+2004]
- Spherical harmonics [Zotkin+2009]
- Tetrahedral interpolation with barycentric weights [Gamper2013]

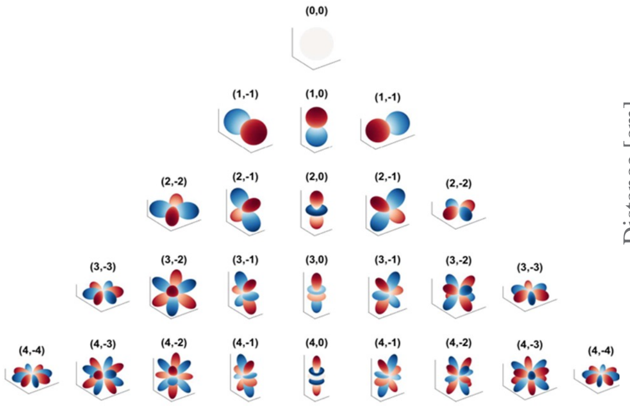


Figure from [Wang+2020]

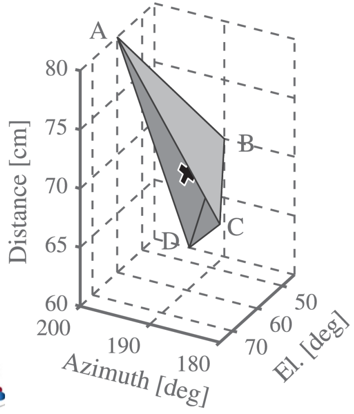


Figure from [Gamper2013]

Pulkki, Ville. “Virtual sound source positioning using vector base amplitude panning.” *JAES* 1997.
Freeland, Fábio P., Luiz WP Biscainho, and Paulo SR Diniz. "Interpolation of head-related transfer functions (HRTFs): A multi-source approach." *ESPC* 2004.
Zotkin, Dmitry N., Ramani Duraiswami, and Nail A. Gumerov. "Regularized HRTF fitting using spherical harmonics." *WASPAA* 2009.
Gamper, Hannes. “Head-related transfer function interpolation in azimuth, elevation, and distance.” *JASA* 2013.

Machine Learning-Based Methods for Interpolation

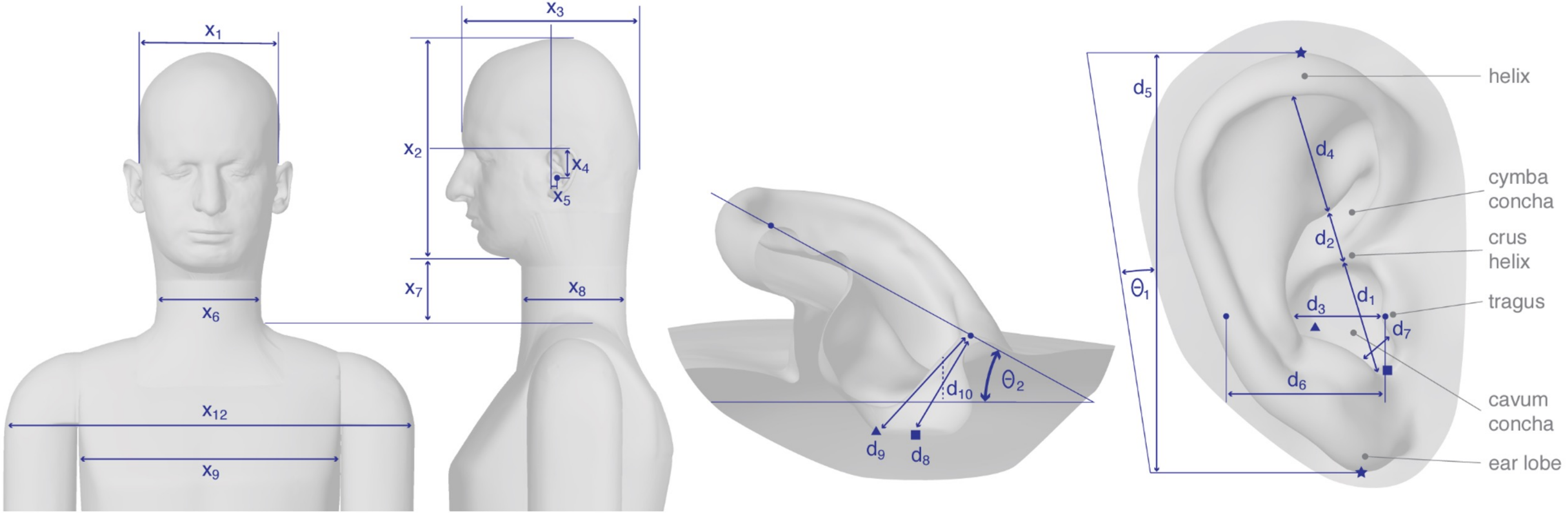
Use datasets to train machine learning models to capture the prior

- Principal component analysis (PCA) [Xie2012]
- Convolutional neural network (CNN) [Jiang+2023]
- Pointwise convolution + FiLM + Hyper-convolution [Lee+2023]
- Neural fields [Zhang+2023]
- Spherical convolutional neural network [Chen+2023]
- Physics-informed neural network [Ma+2023]

Xie, Bo-Sun. “Recovery of individual head-related transfer functions from a small set of measurements.” *JASA* 2012.
Jiang, Ziran, et al. "Modeling individual head-related transfer functions from sparse measurements using a convolutional neural network." *JASA* 2023.
Lee, Jin Woo, Sungho Lee, and Kyogu Lee. "Global HRTF interpolation via learned affine transformation of hyper-conditioned features." *ICASSP* 2023.
Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.
Chen, Xingyu, et al. “Head-Related Transfer Function Interpolation with a Spherical CNN.” *arXiv* 2023.
Ma, Fei, et al. "Physics informed neural network for head-related transfer function upsampling." *arXiv* 2023.

HRTF Personalization from Human Input

Anthropometric measurements



Brinkmann, Fabian, et al. "The HUTUBS HRTF database." 2019.

HRTF Personalization from Human Input (Cont'd)

Ear images or head mesh

Figure from VisiSonics

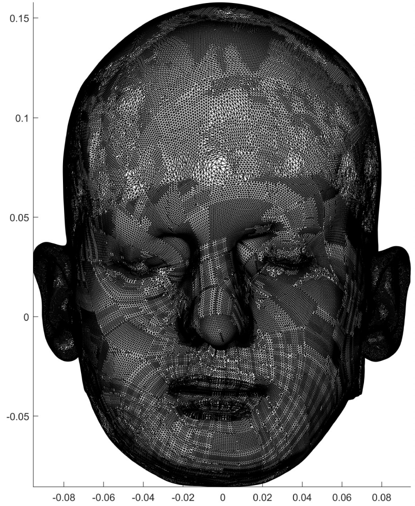


Figure from [Wang+2022]

Wang, Yuxiang, et al. "Predicting global head-related transfer functions from scanned head geometry using deep learning and compact representations." *arXiv* 2022.

Machine Learning-Based Methods for Personalization

Non-parametric methods: Nearest neighbor

Parameters matching (HRTF selection):

- Anthropometric parameters [Zotkin+2003]
- Frequencies of the two lowest spectral notches [Lida+2014]
- Pinna-related anatomical parameters [Liu&Zhong2016]

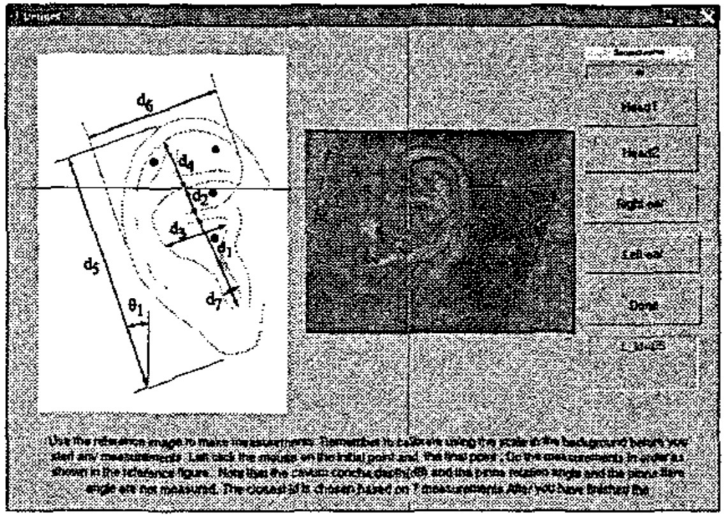


Figure from [Zotkin+2003]

Zotkin, Dmitry N., et al. "HRTF personalization using anthropometric measurements." *WASPAA* 2003.
Iida, Kazuhiro, Yohji Ishii, and Shinsuke Nishioka. “Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener‘s pinnae.” *JASA* 2014.
Liu, Xuejie, and Xiaoli Zhong. “An improved anthropometry-based customization method of individual head-related transfer functions.” *ICASSP* 2016.

Machine Learning-Based Methods for Personalization

Parametric methods: Map the input to learned low-dimensional representation

- Principal component analysis (PCA) [Hu+2008]
- Deep neural network (DNN) [Chun+2017]
- Autoencoder [Chen+2019]
- Variational Autoencoder (VAE) [Miccini&Spagnol2020]
- Spatial principal component analysis (SPCA) [Zhang+2020]
- Spherical harmonics transform (SHT) [Wang+2020]

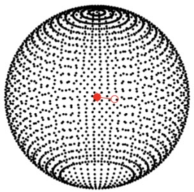
Can handle arbitrary directions!

Hu, Hongmei, et al. "HRTF personalization based on artificial neural network in individual virtual auditory space." *Applied Acoustics* 2008.
Chun, Chan Jun, et al. "Deep neural network based HRTF personalization using anthropometric measurements." *AES Convention* 2017.
Chen, Tzu-Yu, Tzu-Hsuan Kuo, and Tai-Shih Chi. "Autoencoding HRTFs for DNN based HRTF personalization using anthropometric features." *ICASSP* 2019.
Miccini, Riccardo, and Simone Spagnol. "HRTF individualization using deep learning." *VRW* 2020.
Zhang, Mengfan, et al. "Modeling of individual HRTFs based on spatial principal component analysis." *TASLP* 2020.
Wang, Yuxiang, et al. "Global HRTF personalization using anthropometric measures." *AES Convention* 2020.

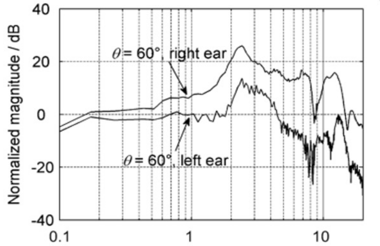
Key Challenges

Challenge1: High-dimensional Data

For each spatial location, and for each ear, HRTF is a function of frequency.



HRTFs at various spatial locations (of arbitrary spatial sampling schemes)



HRTFs at a particular position

$$\boldsymbol{x} \in \mathbb{R}^{L \times F \times 2}$$

L: number of locations (~1000)

F: number of frequency bins (~128)

2: left and right ear

1000 x 128 x 2 = 256,000. *A huge number!*

Challenge1: High-dimensional Data (Cont'd)

Existing measured HRTF databases each only contain dozens of subjects.

Name	# Subjects	# Locations	Elevation Range
3D3A [29]	38	648	$[-57°, 75°]$
Aachen [30]	48	2304	$[-66.24°, 90°]$
ARI	97	1550	$[-30°, 80°]$
BiLi [31]	52	1680	$[-50.5°, 85.5°]$
CIPIC [4]	45	1250	$[-50.62°, 90°]$
Crossmod	24	651	$[-40°, 90°]$
HUTUBS [17]	96	440	$[-90°, 90°]$
Listen	50	187	$[-45°, 90°]$
RIEC [32]	105	865	$[-30°, 90°]$
SADIE II [2]	18	2818	$[-90°, 90°]$

Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.

Challenge1: High-dimensional Data (Cont'd)

Current research status:

Low-dimensional representation: PCA, SPCA, Autoencoder, VAE, SHT, etc.

Open question: What is the intrinsic dimensionality of HRTFs across subjects?

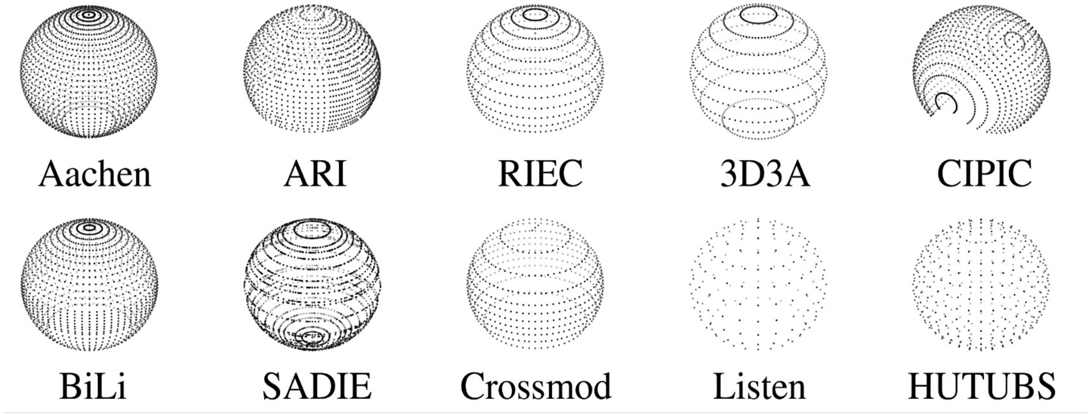
Most of the work trains and evaluates the model on the same database, and it is hard to tell the generalization ability.

- Leave-one-out validation
- Cross-validation

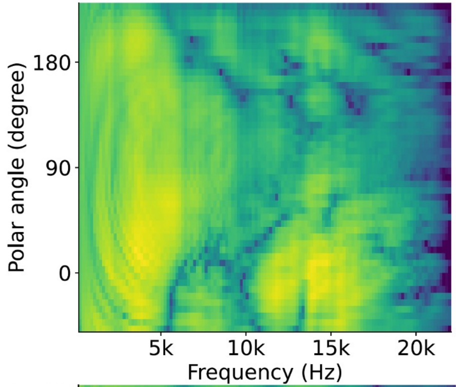
Open question: Can we merge the existing datasets? If so, how?

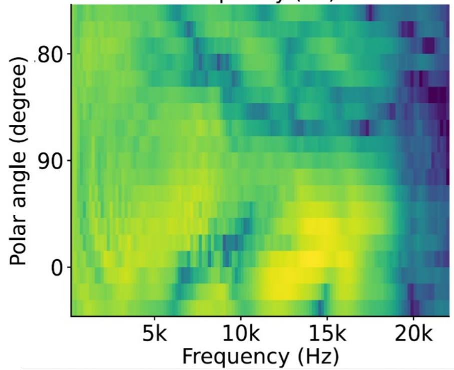
Challenge2: Spatial Sampling Schemes

The source location grids used in HTRF databases differ from one to another, making cross-dataset learning difficult.



Figures from [Zhang+2023]

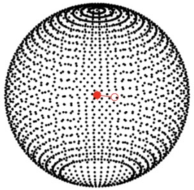




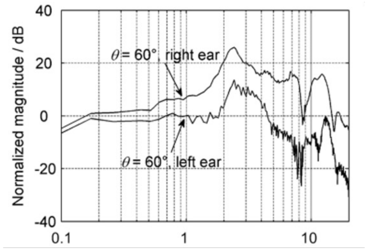
Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.

Challenge2: Spatial Sampling Schemes (Cont’d)

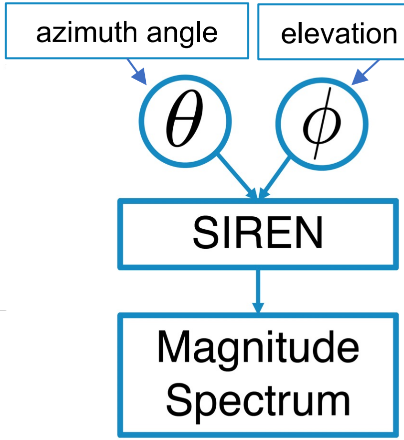
HRTF field [Zhang+2023]: Represent a single subject’s HRTFs with a neural field



HRTFs at various spatial locations (of arbitrary spatial sampling schemes)



HRTFs at a particular position



frequency bins

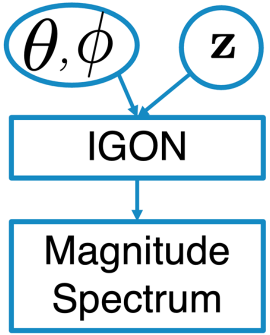
$$f : \mathbb{R}^2 \rightarrow \mathbb{R}^K$$

SIREN: a multi-layer perceptron (MLP) with sine activation functions [Sitzmann+2020]

Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.
Sitzmann, Vincent, et al. “Implicit neural representations with periodic activation functions.” *NeurIPS* 2020.

Challenge2: Spatial Sampling Schemes (Cont'd)

HRTF field [Zhang+2023]: Learning HRTF representations across subjects



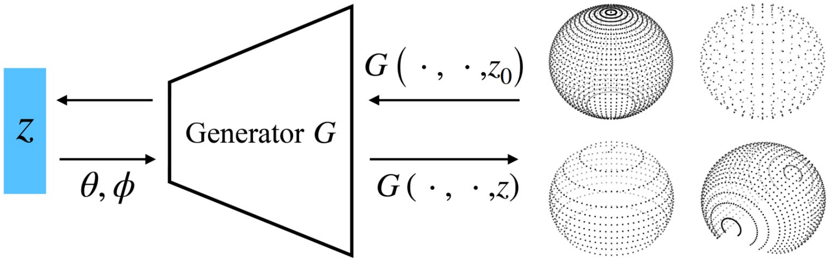
latent code for a subject

$G(\theta, \phi, \mathbf{z})$ $\quad \mathbf{z} \in \mathbb{R}^D$

frequency bins

$G : \mathbb{R}^{2+D} \mapsto \mathbb{R}^K$

$$\mathbf{z} = \mathbf{z}_0 - \nabla_{\mathbf{z}_0} \mathcal{L}_{\text{MSE}}\left(\mathbf{x}, G\left(\cdot, \cdot, \mathbf{z}_0\right)\right)$$



$$\mathcal{L} = \mathcal{L}_{\text{MSE}}\left(\mathbf{x}, G\left(\cdot, \cdot, \mathbf{z}\right)\right)$$

IGON: implicit gradient origin network that uses SIREN architecture [Bond-Taylor&Willcocks2021]

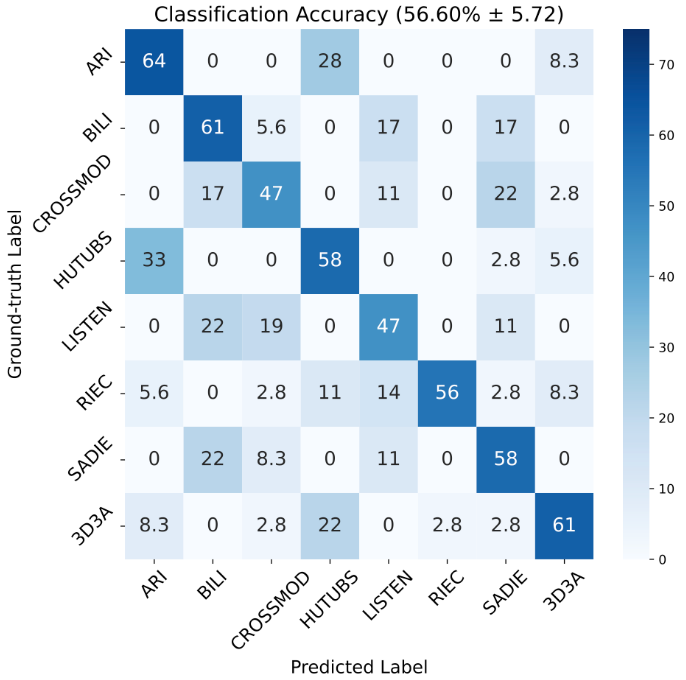
Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.
Bond-Taylor, Sam, and Chris G. Willcocks. "Gradient origin networks." *ICLR* 2021.

Challenge3: Measurement Setup Differences

Another study [Pauwels&Picinali2023] shows that there are other significant differences across HRTF databases, which would hinder the training process.

Reproduced in [Wen+2023]:

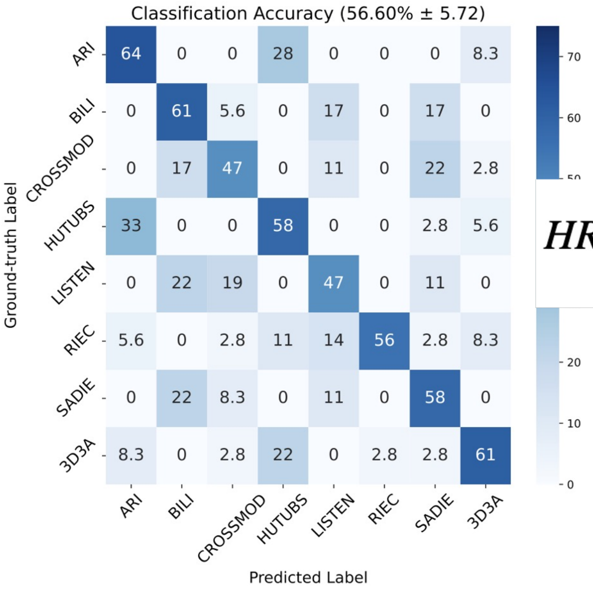
- Total 144 subjects
  - 18 (the smallest size dataset) x 8
  - 432 HRTFs = 18 (subjects) x 12 (common positions) x 2 (ears)
- Model: kernel SVM



Pauwels, Johan, and Lorenzo Picinali. "On the relevance of the differences between HRTF measurement setups for machine learning." *ICASSP* 2023.
Wen, Yutong, You Zhang, and Zhiyao Duan. "Mitigating Cross-Database Differences for Learning Unified HRTF Representation." *WASPAA* 2023.

Challenge3: Measurement Setup Differences (Cont'd)

HRTF normalization / harmonization [Wen+2023]



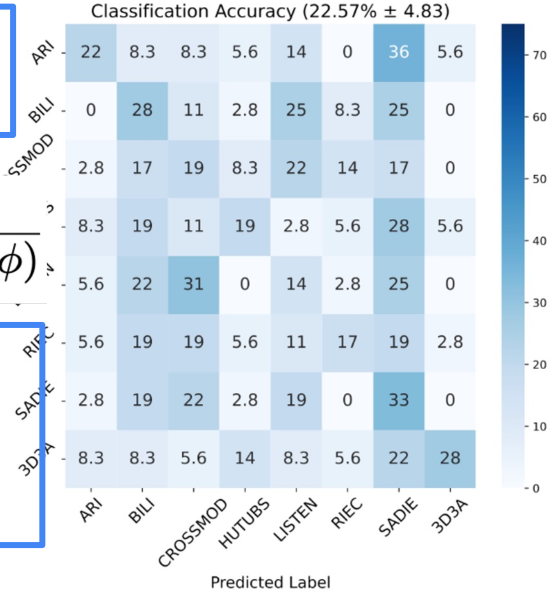
$$HRTF_{\text{normalized}}(\theta, \phi) = \frac{Y(\theta, \phi)}{HRTF_{\text{avg}}(\theta, \phi)}$$

elevation

azimuth

Unnormalized HRTF magnitude

Average HRTF magnitude across subjects



Wen, Yutong, You Zhang, and Zhiyao Duan. "Mitigating Cross-Database Differences for Learning Unified HRTF Representation." *WASPAA* 2023.

Challenge3: Measurement Setup Differences (Cont'd)

HRTF field, agnostic to spatial sampling schemes, enables cross-dataset learning.

The systematic differences across HRTF datasets are position-dependent.

Our proposed normalization methods using average person HRTFs from individual positions are beneficial.

LSD of cross-dataset HRTF reconstruction

Experiments	1	2	3	4	5
ARI	○	△		△	△
ITA				△	△
Listen	△		○	△	△
Crossmod	△	△	△	△	△
SADIE II	△		△	△	△
BiLi	△	△	△	△	△
HUTUBS		△		△	○
CIPIC				△	△
3D3A				△	△
RIEC		○		○	△
HRTF field [15]	7.47	5.54	4.31	4.43	5.01
Our proposed	**4.69**	**4.82**	**3.89**	**3.73**	**4.04**
w/o position dependency	5.61	5.32	4.32	4.00	4.89
w/o ear dependency	5.11	5.11	3.98	3.94	4.67

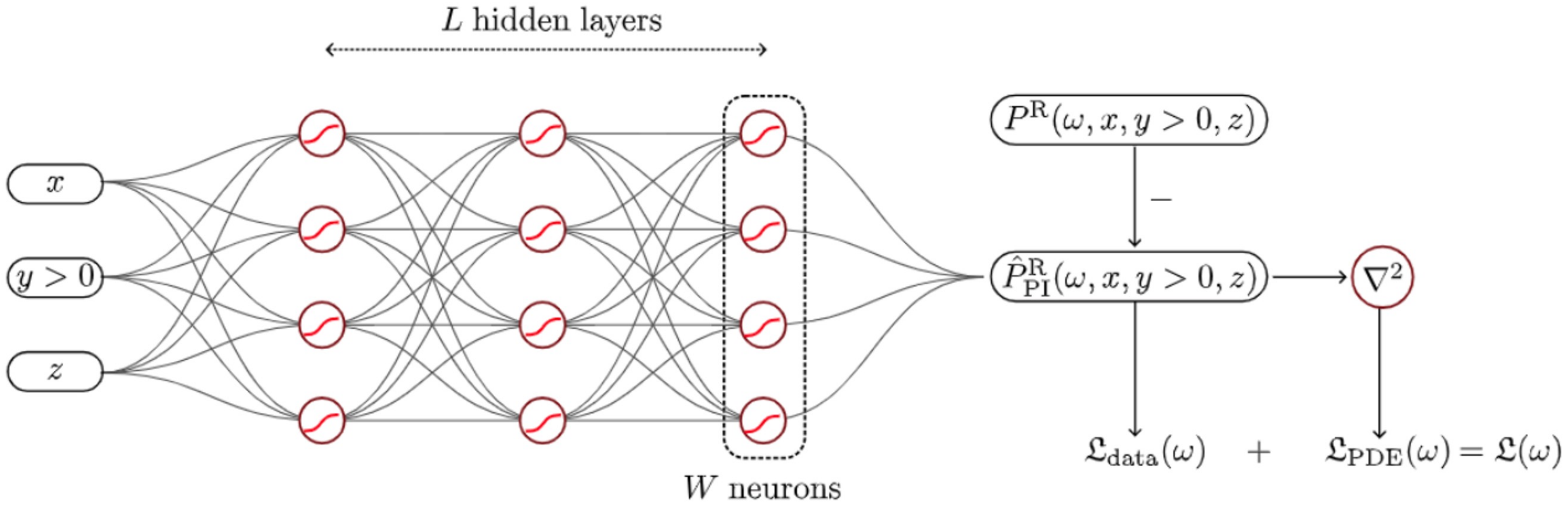
Table from [Wen+2023] △ Training sets ○ Test sets

Wen, Yutong, You Zhang, and Zhiyao Duan. "Mitigating Cross-Database Differences for Learning Unified HRTF Representation." *WASPAA* 2023.
Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.

Emerging & Future Directions

Direction1: Regularize the Model with Priors

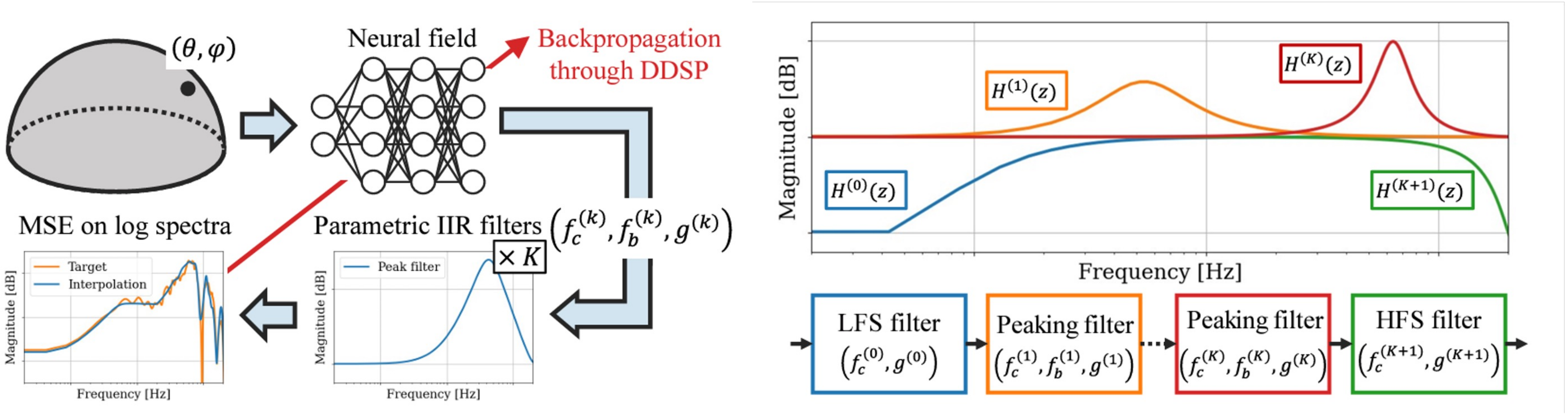
Physics prior: Physics-informed neural network for spatial upsampling [Ma+2023]



Ma, Fei, et al. "Physics informed neural network for head-related transfer function upsampling." *arXiv* 2023.

Direction1: Regularize the Model with Priors (Cont'd)

DSP prior: Model HRTF as IIR filters -- Neural IIR filter field (NIIRF) [Yoshiki+2024]



Masuyama, Yoshiki, et al. "NIIRF: Neural IIR Filter Field for HRTF Upsampling and Personalization." *ICASSP* 2024.

Direction2: Perceptual Loss / Evaluation Metric

Existing works:

The models are trained with MSE loss and evaluated with LSD, which do not reflect perceptual evaluation.

Research direction:

Perceptual loss functions and evaluation metrics

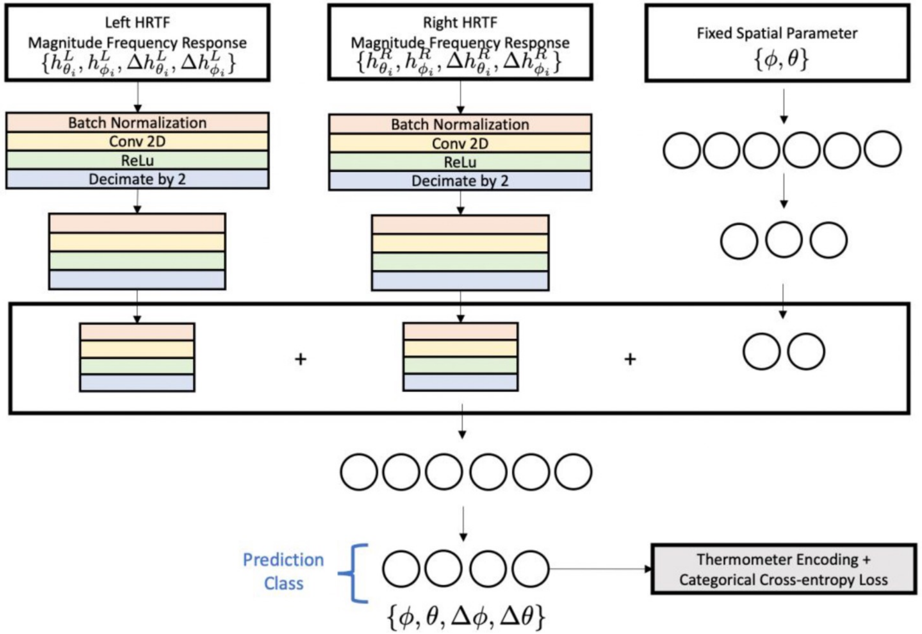
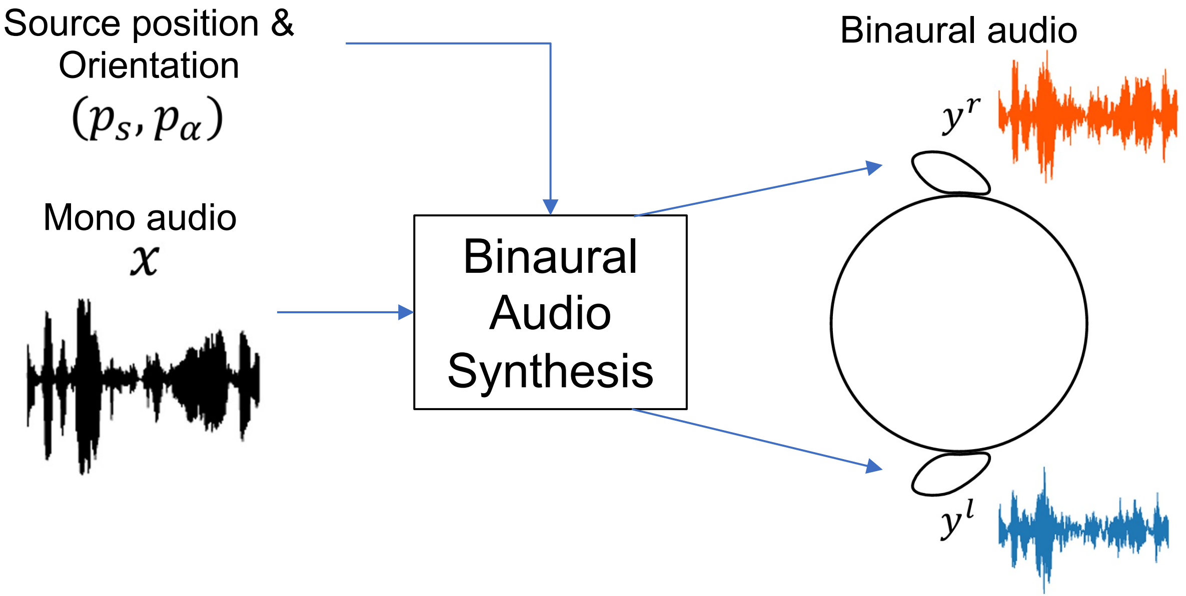


Figure from [Ananthabhotla +2021]

Ananthabhotla, Ishwarya, Vamsi Krishna Ithapu, and W. Owen Brimijoin. "A framework for designing head-related transfer function distance metrics that capture localization perception." *JASA Express Letters* 2021.

Direction3: Binaural Audio Synthesis



Existing methods:

WarpNet
[Richard+2020]

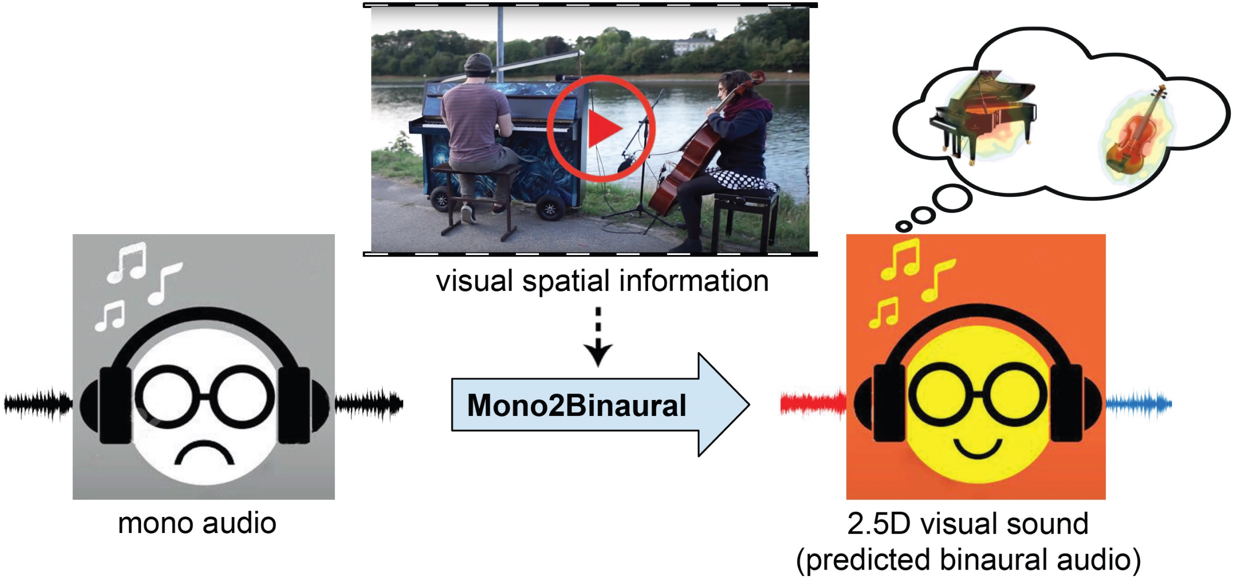
BinauralGrad
[Leng+2022]

Neural Fourier Shift
[Lee & Lee2023]

Richard, Alexander, et al. "Neural synthesis of binaural speech from mono audio." *ICLR* 2020.
Leng, Yichong, et al. “Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis.” *NeurIPS* 2022.
Lee, Jin Woo, and Kyogu Lee. “Neural fourier shift for binaural speech rendering.” *ICASSP* 2023.

Direction3: Binaural Audio Synthesis (Cont'd)

Injecting the spatial information contained in the video frames



Gao, Ruohan, and Kristen Grauman. "2.5 D visual sound." *CVPR* 2019.

Conclusions

Takeaway messages

- Personalized HRTF is important for spatial audio rendering in games.
- Machine learning methods have been evolving quite a lot. Most methods seek a low-dimensional representation of HRTFs.
- The bottleneck of personalized HRTF modeling with machine learning lies in the following:
  - Datasets;
  - HRTF data representation;
  - Evaluation metric.

Summary

- HRTF & Machine Learning Basics
- Tasks & Existing Methods
  - HRTF upsampling/interpolation
  - HRTF personalization from human input
- Key Challenges
  - High-dimensional data vs. small datasets
  - Spatial sampling schemes
  - Measurement setup differences
- Emerging & Future Directions
  - Regularize the model with priors
  - Perceptual loss / evaluation metric
  - Binaural audio synthesis

Thank you! Questions?