From **Neural Fields** to **Perception-Informed Learning**

Scalable and Perceptually Grounded HRTF Personalization

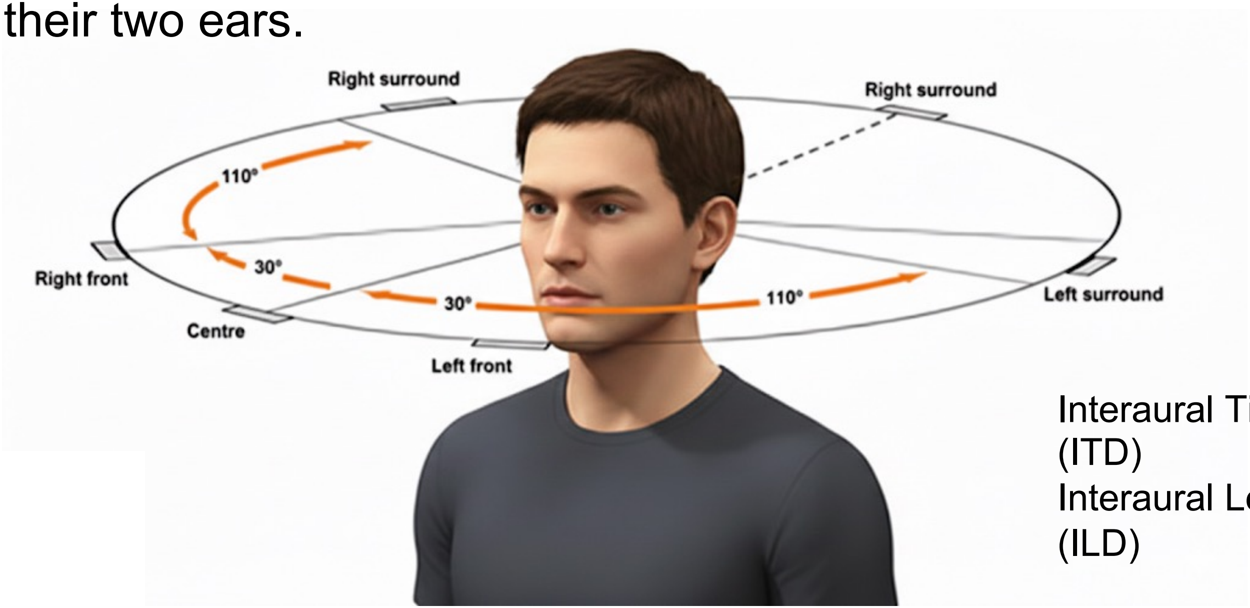
You (Neil) Zhang

University of Rochester

Invited Seminar, Centre for Digital Music
Queen Mary University of London
February 17, 2026

Spatial Effects and Sound Localization

Humans localize sound sources by processing the differences between sounds received by their two ears.

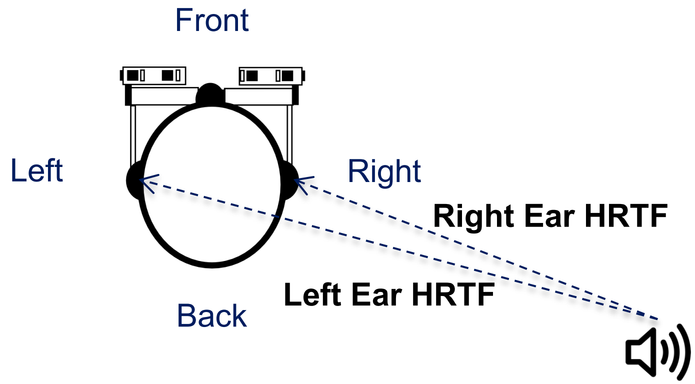


Interaural Time Difference (ITD)
Interaural Level Difference (ILD)

Figure adapted from https://www.soundonsound.com/reviews/mp3-surround, processed by Gemini

Head-Related Transfer Function (HRTF)

HRTF models the **acoustic filtering** effect of a listener's **head, ears, and torso** to enable 3D sound localization.



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

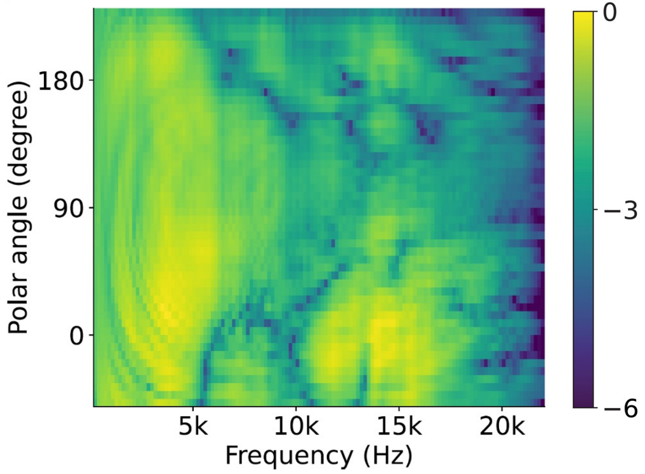


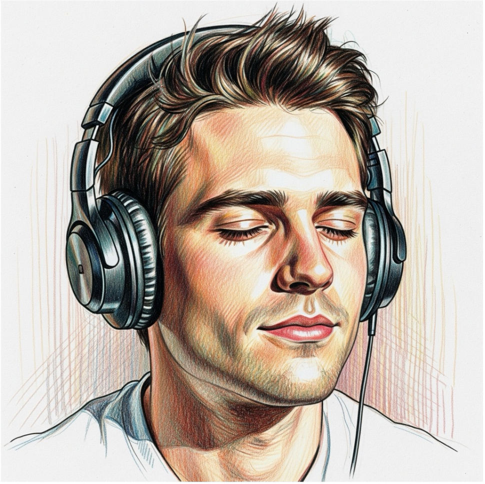
Figure from [Zhang+2023]

HRTF is unique to each person due to differences in ear, head, and torso shape.

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

HRTF Applications: Virtual Spatial Audio Rendering

HRTFs encode human spatial cues to deliver immersive 3D sound.



Headphones

AR smart glasses

VR headsets

Figures generated by Gemini

Measure HRTFs

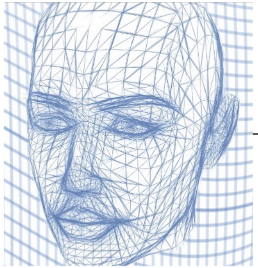
- An anechoic room
- Multiple loudspeakers on motorized arc
- Two microphones
- Head motion control

Time-consuming & Resource-intensive!

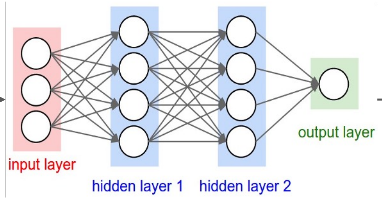
Figure from https://facebookresearch.github.io/SS2_HRTF/

Personalizing HRTF with Machine Learning

Leverage measured data for personalized HRTF prediction

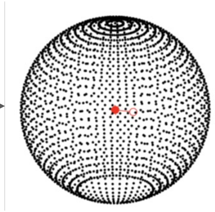


Human physical geometry

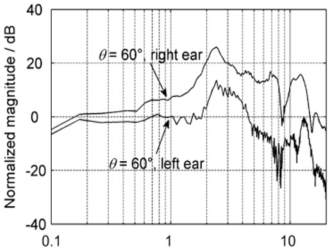


HRTF prediction network

Representation of HRTFs



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



HRTFs at a particular position

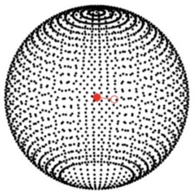
Assumption: Many characteristics are shared across individuals and captured by the **model**, while personalized effects are addressed by adapting the **input**.

Outline

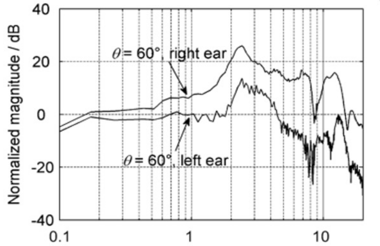
- Background
- Challenges and Research Questions
- **Model**: Neural Fields for HRTF Modeling
- **Data**: Position-Dependent HRTF Normalization
- **Perception**: Perception-Informed Representation Learning
- Conclusion and Outlook

HRTF is High-Dimensional

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!*

Small Datasets and Different Measurement Setups

Existing measured HRTF databases each only contain dozens of subjects.

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

Now we have SONICOM (Increased to 300 subjects in 2025)

Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.
Poole, Katarina C., et al. "The extended SONICOM HRTF dataset and spatial audio metrics toolbox." *Forum Acusticum 2025*.

Research Questions

Low-dimensional modeling: PCA, Spatial PCA, Autoencoder, VAE, SHT, etc.

What is a scalable representation of HRTFs across subjects and datasets?

Most models are trained and evaluated on a single dataset.
Cross-dataset generalization remains unclear.

Can we merge existing HRTF datasets? If so, how do we mitigate measurement biases?

Most learning objectives only minimize spectral distortion.

How do we learn HRTF representations that reflect human perception?

HRTF Field: Unifying Measured HRTF Magnitude Representation with Neural Fields

You Zhang, Yuxiang Wang, Zhiyao Duan

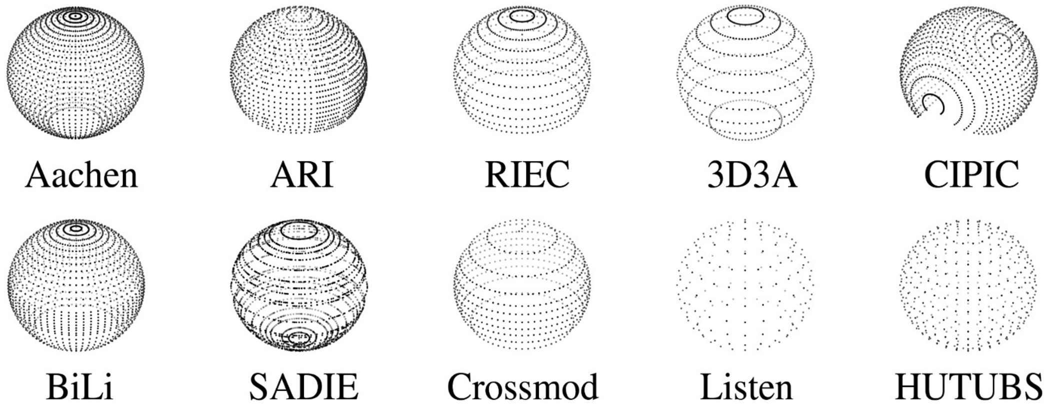
University of Rochester

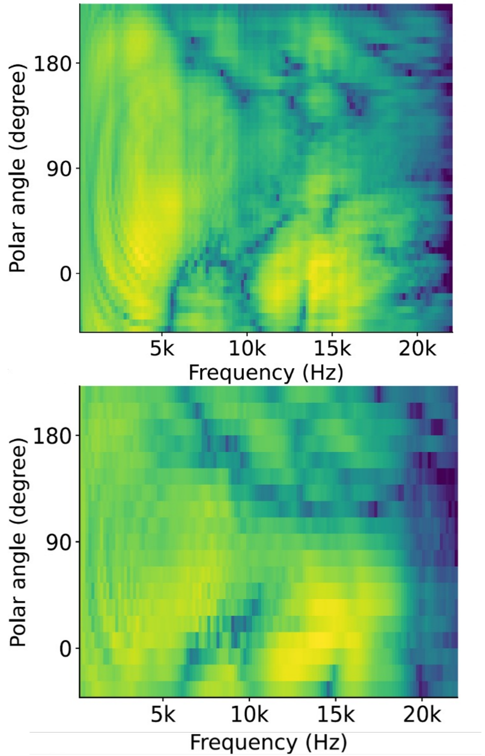
ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

Rhodes Island, Greece

Spatial Sampling Schemes

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





Take A Step Back …

HRTFs of one ear of a subject are a function defined on the **continuous** sphere.

azimuth angle

received pressure

$$\mathrm{HRTF}(\theta, \phi) = \frac{\mathbf{p}(\theta, \phi)}{\mathbf{p}_0}$$

elevation angle

source pressure

Key idea: model the continuous function directly.

Neural Fields (Implicit Neural Representations)

Representing discrete data as a continuous function

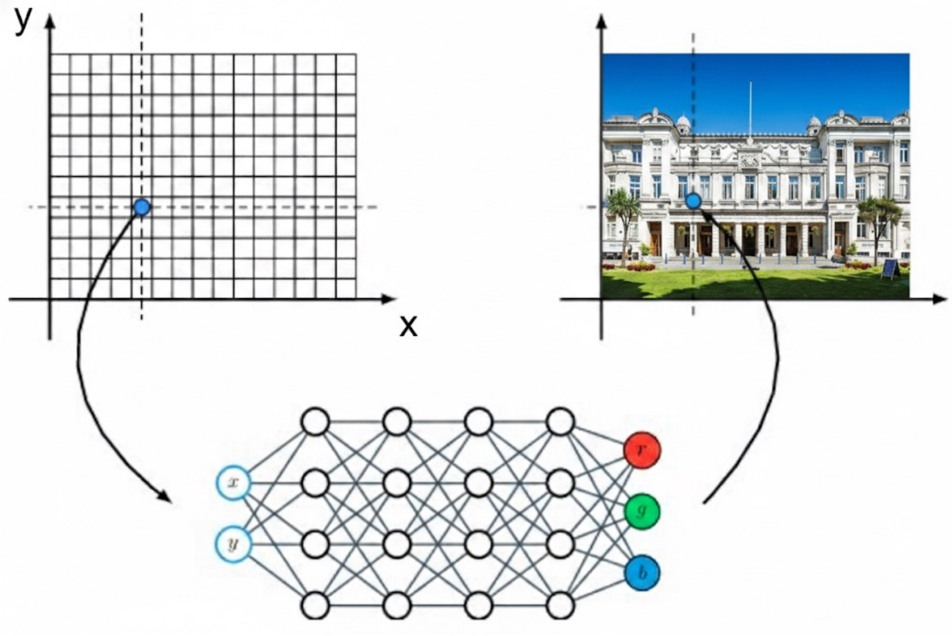
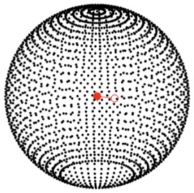


Figure adapted from [Skorokhodov+2021] and QMUL website, processed by Gemini

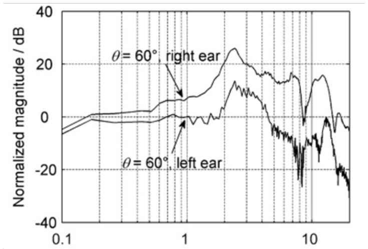
Skorokhodov, Ivan, Savva Ignatyev, and Mohamed Elhoseiny. “Adversarial generation of continuous images.” *CVPR* 2021.

HRTF Field

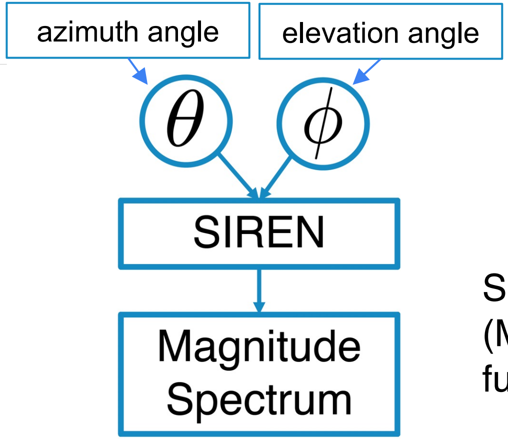
Represent a single subject's HRTFs with a neural field



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



HRTFs at a particular position



frequency bins

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

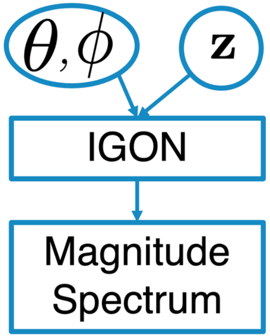
SIREN: a multi-layer perceptron (MLP) with sine activation functions [Sitzmann+2020]

HRTF Field treats HRTFs as continuous functions over direction, decoupling representation from sampling schemes.

Sitzmann, Vincent, et al. "Implicit neural representations with periodic activation functions." *NeurIPS* 2020.

HRTF Field

Learning HRTF representations across subjects



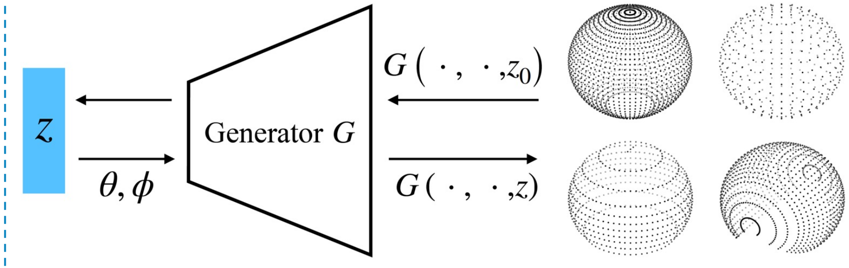
latent code for a subject

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

frequency bins

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

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



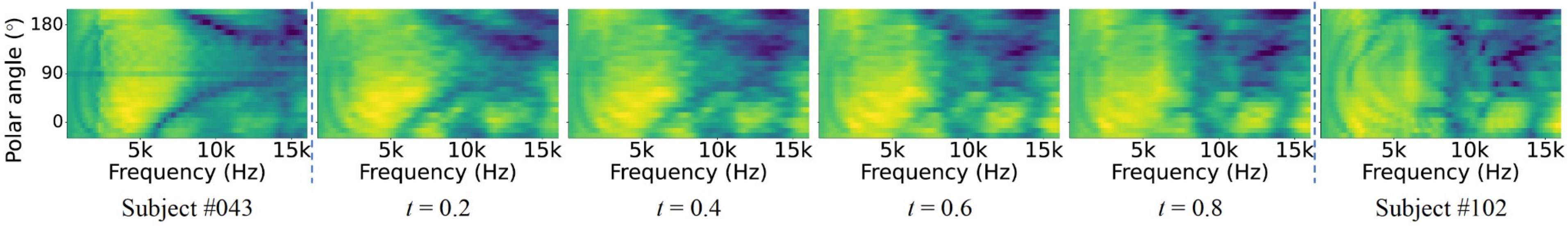
$$\mathcal{L} = \mathcal{L}_{\text{MSE}} \left(\mathbf{x}, G\left(\cdot, \ \cdot, \mathbf{z} \right) \right)$$

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

Bond-Taylor, Sam, and Chris G. Willcocks. "Gradient origin networks." *ICLR* 2021.

HRTF Field Conclusions

- Enables mix-database training and cross-database evaluation
- Supports conditional generation (interpolation / upsampling) from randomly observed locations
- Supports generation from the latent space



Mix-database training solved? Not fully

Mitigating Cross-Database Differences for Learning Unified HRTF Representation

Yutong Wen, You Zhang, Zhiyao Duan

University of Rochester

IEEE WASPAA 2023

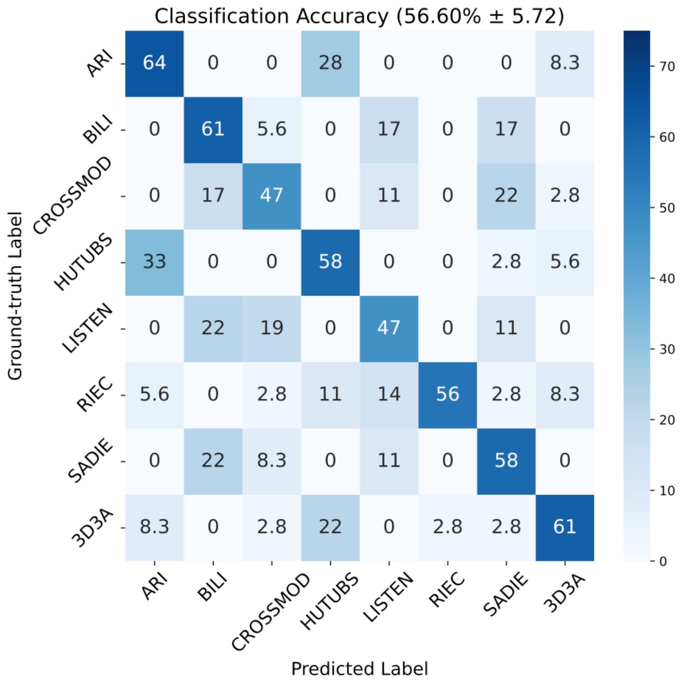
New Paltz, NY, USA

Measurement Setup Differences

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

Reproduced in our work:

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

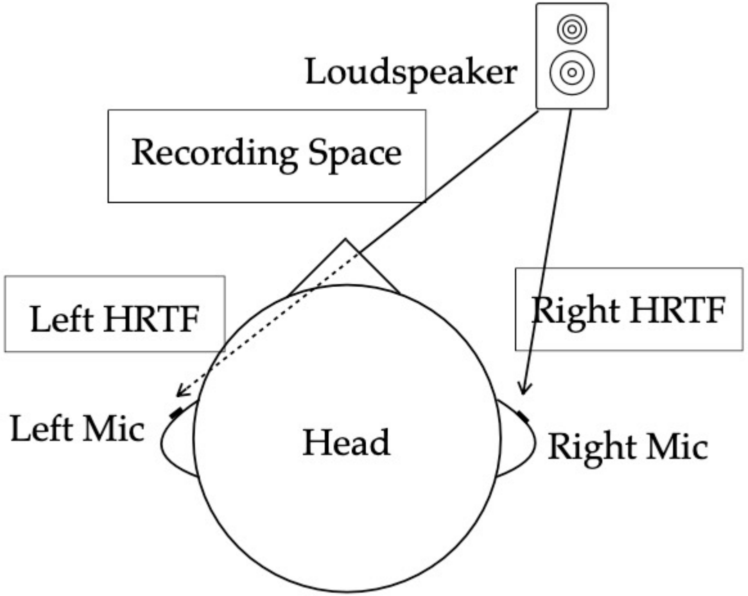


Pauwels, Johan, and Lorenzo Picinali. "On the relevance of the differences between HRTF measurement setups for machine learning." *ICASSP* 2023.

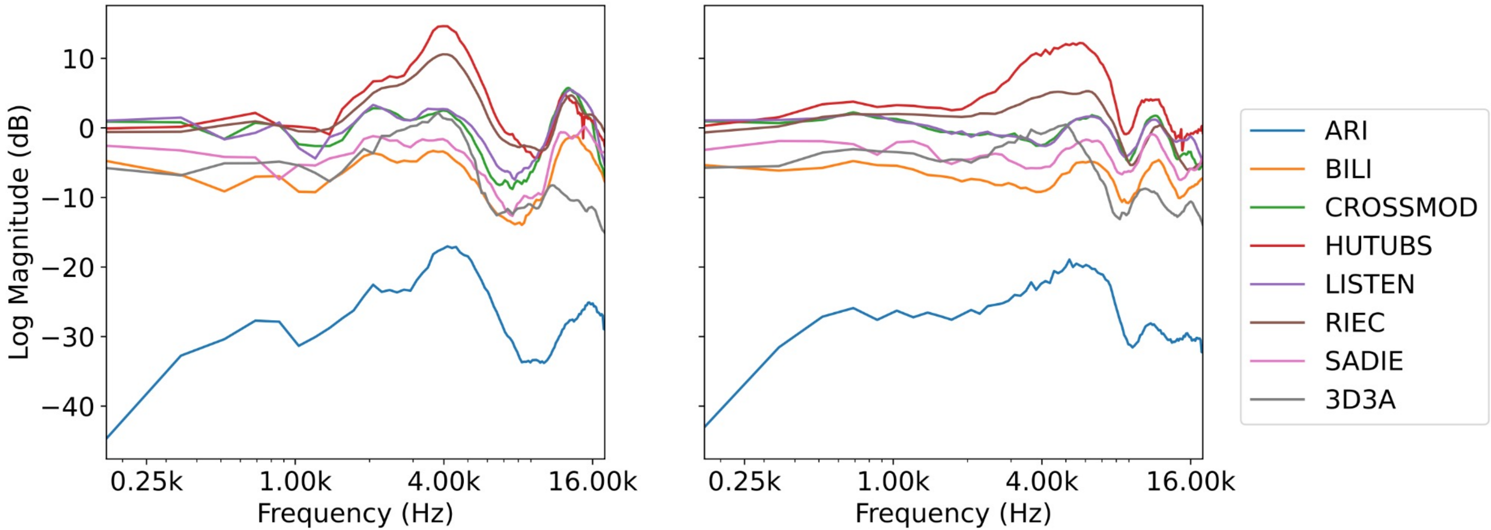
Investigating Cross-Database Differences

Systematic difference could be caused by:

- Different loudspeakers
- Recording space
- Microphones



Average HRTFs Across Subjects in Different Databases

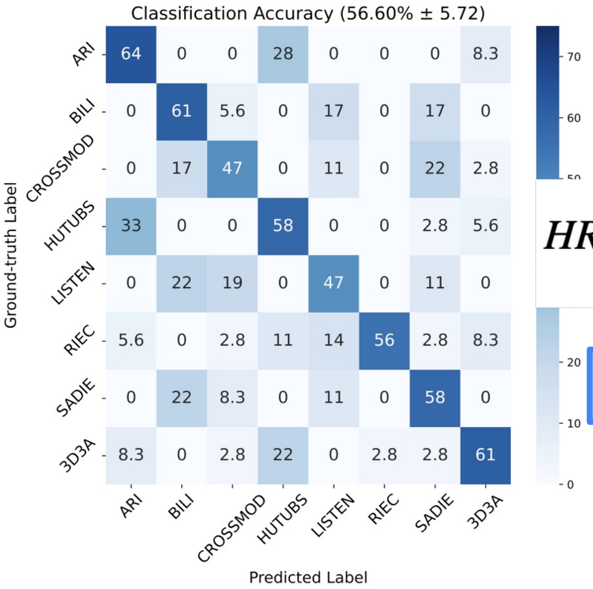


(a) at source position (0, 0) (b) at source position (90, 0)

- There are systematic differences in the measurement system responses at each source position.
- These systematic differences in the measurement system responses are position-dependent.

Our Normalization Successfully Confuses a SVM Classifier

Before normalization

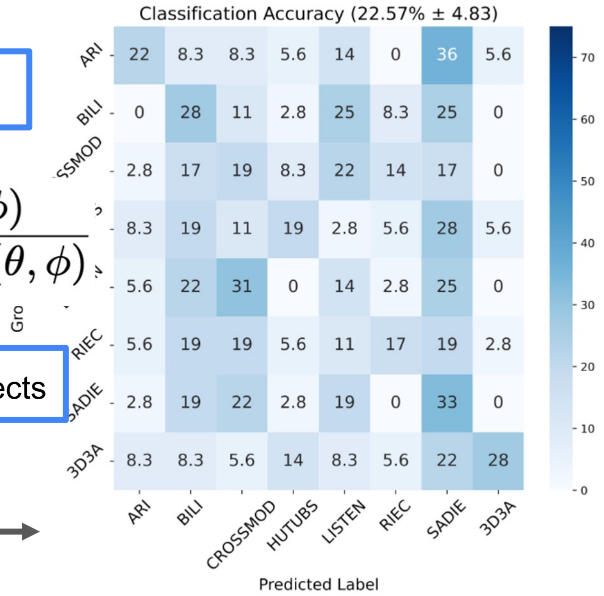


Unnormalized HRTF magnitude

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

Average HRTF magnitude across subjects

After normalization



Our Normalization Improves Cross-Database Reconstruction

The systematic differences across HRTF datasets are position-dependent.

Our proposed normalization methods using average person HRTFs from individual positions are beneficial for cross-database reconstruction.

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

△ Training sets ○ Test sets

From Proposal to Community Adoption

Neural fields are increasingly adopted for HRTF modeling

- Recent extensions: **NIIRF, RANF, SuDaField** (subject- and dataset-aware neural fields)
- Validation at larger scale: **SONICOM** database
- State-of-the-art performance in **upsampling and harmonization** (LAP Challenge)

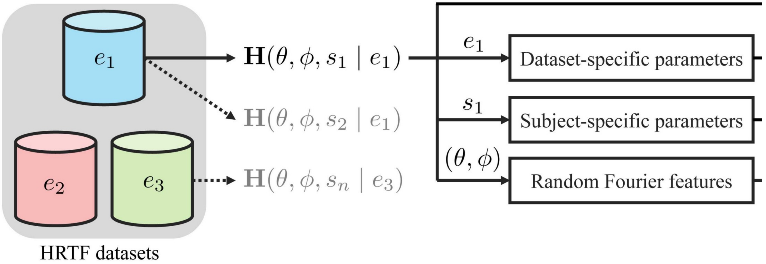


Figure from [Masuyama+2025]

The representation paradigm has evolved beyond the original formulation.

Masuyama, Yoshiki, et al. "SuDaField: Subject-and Dataset-Aware Neural Field for HRTF Modeling." *IEEE OJSP 2025*.

Takeaway

- Personalized HRTFs are important but difficult to measure — motivating data-driven approaches.
- HRTF field, agnostic to spatial sampling schemes, enables unified modeling and cross-dataset learning.
- HRTF databases exhibit position-dependent systematic differences, which hinder generalization.
- Position-wise normalization using average HRTFs effectively mitigates these biases and benefits mix-database training.

Halfway. Questions?

Representation and harmonization enable scalability.

But do our models sound right?

Towards Perception-Informed Latent HRTF Representations

You Zhang [1,2], *Andrew Francl* [2], *Ruohan Gao* [3], *Paul Calamia* [2], *Zhiyao Duan* [1], *Ishwarya Ananthabhotla* [2]

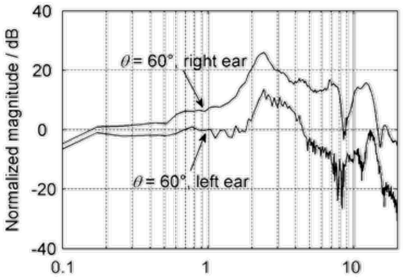
[1] University of Rochester
[2] Meta Reality Labs Research
[3] University of Maryland

IEEE WASPAA 2025

Tahoe City, CA, USA

Motivation

Most **existing** models are trained and evaluated with **spectral reconstruction**.



Spectral reconstruction

Perceptually plausible HRTF

If two HRTFs are close in spectral distance, are they close perceptually?

Our contributions

Goal: Learn HRTF representations that more accurately reflect perceptual correlation, to enable better HRTF personalization for unseen users

- We study how well existing latent HRTF representations preserve perceptual relations, and introduce the benchmark for evaluating this.
- We propose a method for improving on this benchmark.
- We demonstrate practical utility for HRTF personalization.

1. How well do **existing** learned HRTF representations **preserve perceptual relations**?

HRTF Perception

Perceptual benefits of your *personal HRTF*:

- Reduced **Coloration** (less unwanted spectral distortion)
- Improved **Externalization** (sound appears outside the head)
- Enhanced **Localization** (accurately placing sounds in 3D space)

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.

Coloration

Externalization

Localization

How do we mathematically model these?

Computational Auditory Modeling

Coloration: Predicted Binaural Coloration [McKenzie+2022]

Externalization: Auditory Externalization Perception [Baumgartner&Majdak2021]

Localization: Difference of Root Mean Square Error in Polar Angles [Barumerli+2023]

Auditory Modeling Toolbox

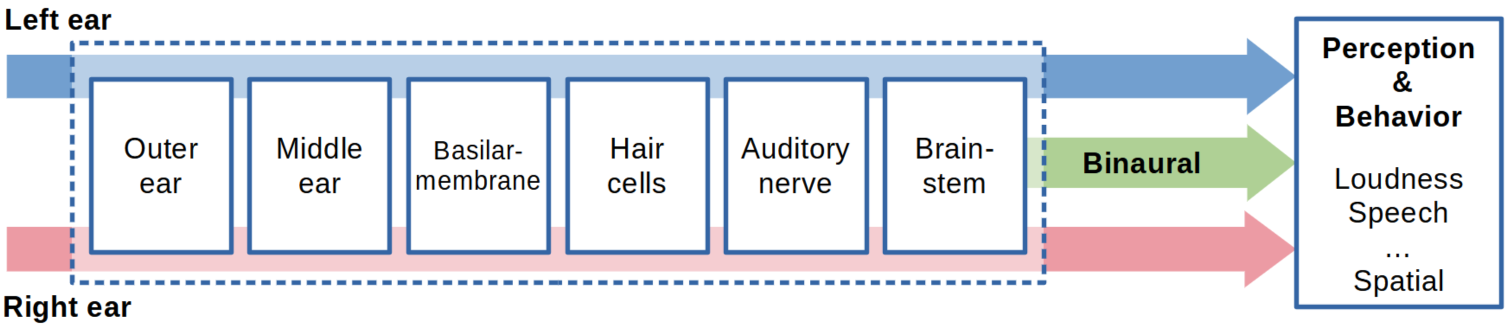


Figure from amtoolbox.org

McKenzie, Thomas, et al. "Predicting the colouration between binaural signals." *Applied Sciences* 2022.
Baumgartner, Robert, and Piotr Majdak. "Decision making in auditory externalization perception:model predictions for static conditions." *Acta Acustica* 2021.
Barumerli, Roberto, et al. "A Bayesian model for human directional localization of broadband static sound sources." *Acta Acustica* 2023.

Computational Auditory Modeling

Coloration: **PBC** [McKenzie+2022]

Externalization: **AEP** [Baumgartner&Majdak2021]

Localization: **DRMSP** [Barumerli+2023]

Objective Perceptual Metrics

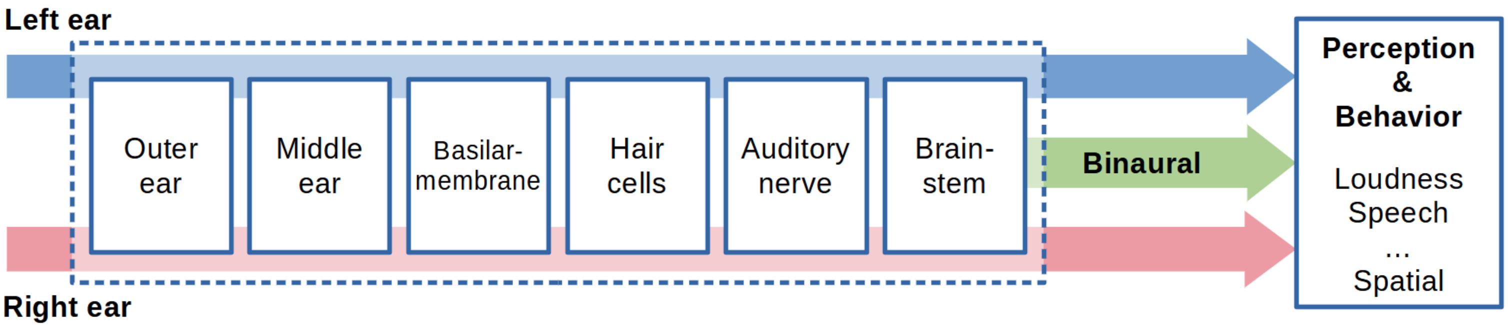


Figure from amtoolbox.org

McKenzie, Thomas, et al. "Predicting the colouration between binaural signals." *Applied Sciences* 2022.
Baumgartner, Robert, and Piotr Majdak. "Decision making in auditory externalization perception:model predictions for static conditions." *Acta Acustica* 2021.
Barumerli, Roberto, et al. "A Bayesian model for human directional localization of broadband static sound sources." *Acta Acustica* 2023.

Experimental Setup

SS2 HRTF Database

- 1625 measurement locations
- 48 kHz sampling rate
- 78 subjects (65 for training, 13 for testing)

Train AI models with spectral reconstruction for HRTF data

↓

Compute pairwise latent distance across subjects

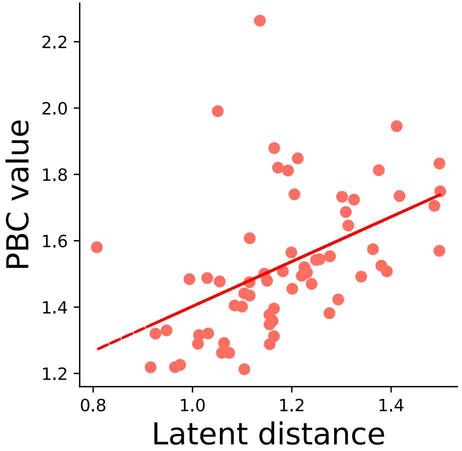
↓

Compute pairwise perceptual distance across subjects

Warnecke, Michaela, et al. "Sound sphere 2: A high-resolution HRTF database." *AES AVAR* 2024.

Alignment Between Latent Space and Perceptual Metrics

Pearson correlation (pairwise latent distances vs. perceptual distances)



$$\rho_{A,B} = \frac{\mathbb{E}[(A-\mu_A)(B-\mu_B)]}{\sigma_A \sigma_B}$$

A higher positive correlation indicates better alignment with human perception.

Partitions	**PBC**	**AEP**	**DRMSP**
train	0.60	0.60	0.40

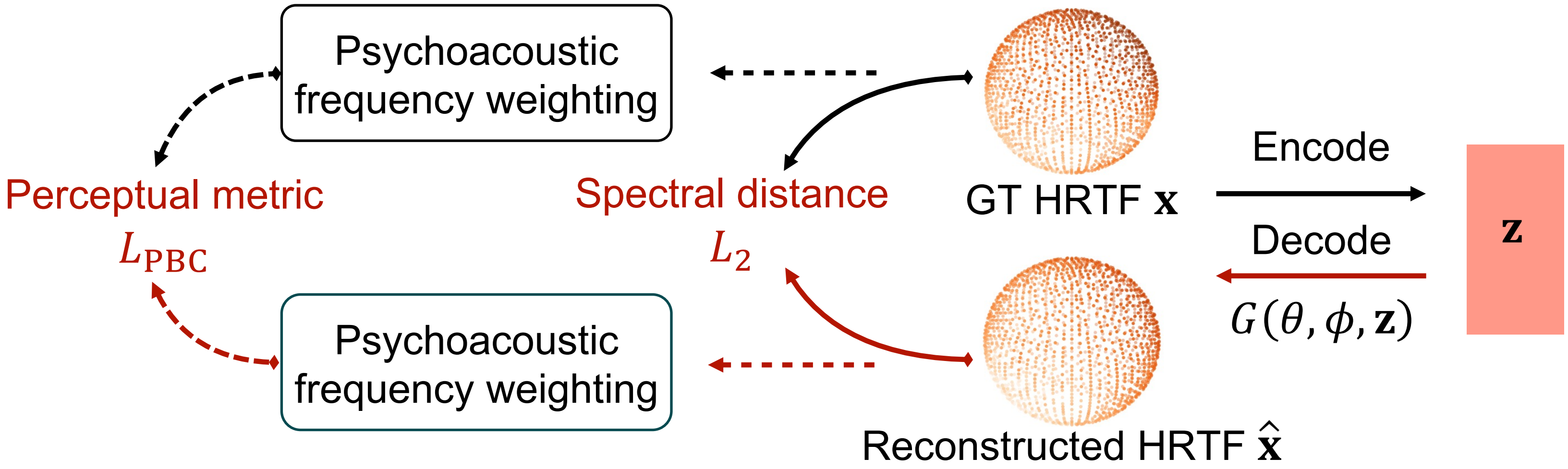
Minimizing spectral distances leads to limited perceptual correlation.

2. How do we **align** latent HRTF representations with **perception-informed space**?

Aligning with Perception-Informed Space

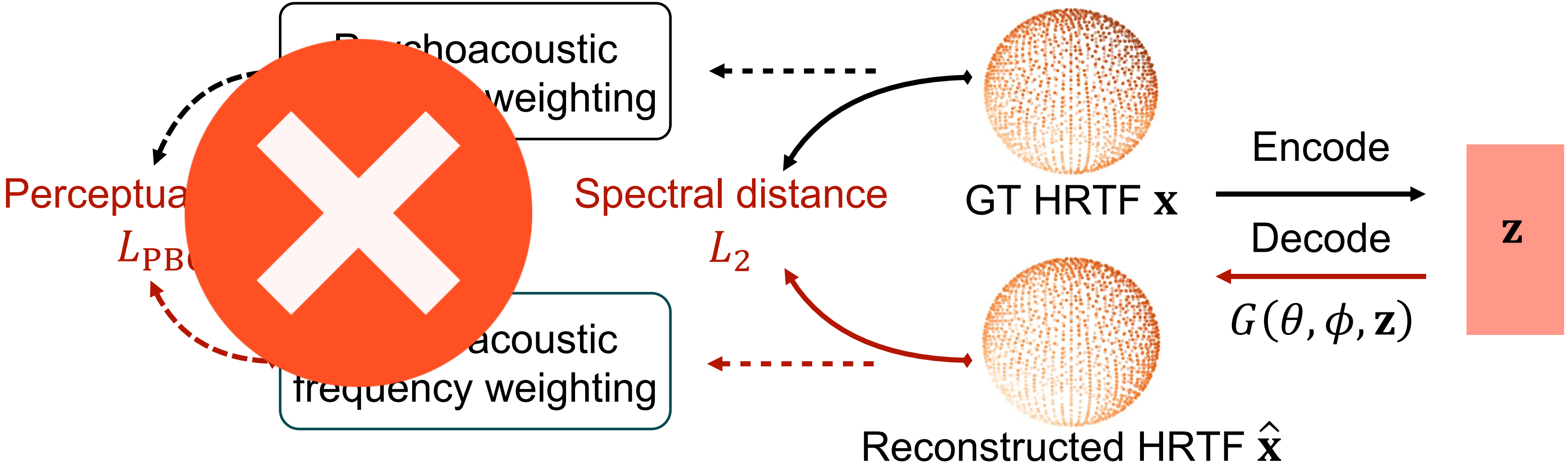
If the perceptual metric is differentiable, just add a straightforward perceptual loss.

- This only applies to PBC, which we reimplemented with PyTorch.



Aligning with Perception-Informed Space (Cont'd)

If the perceptual metric is not differentiable (AEP, DRMSP)



Metric multi-dimensional scaling (MMDS)

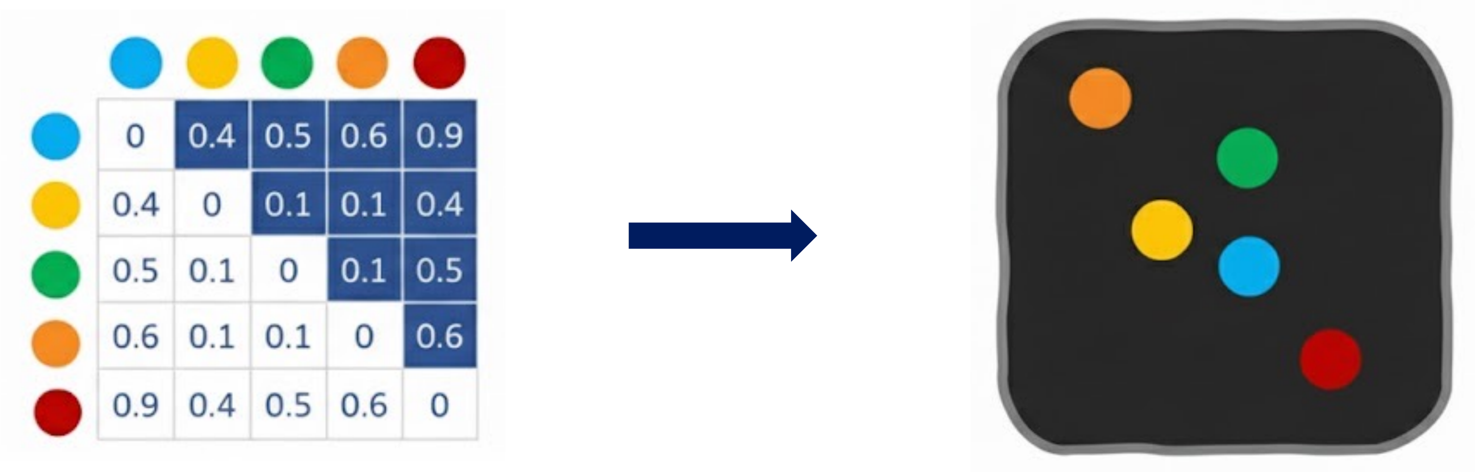
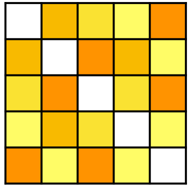


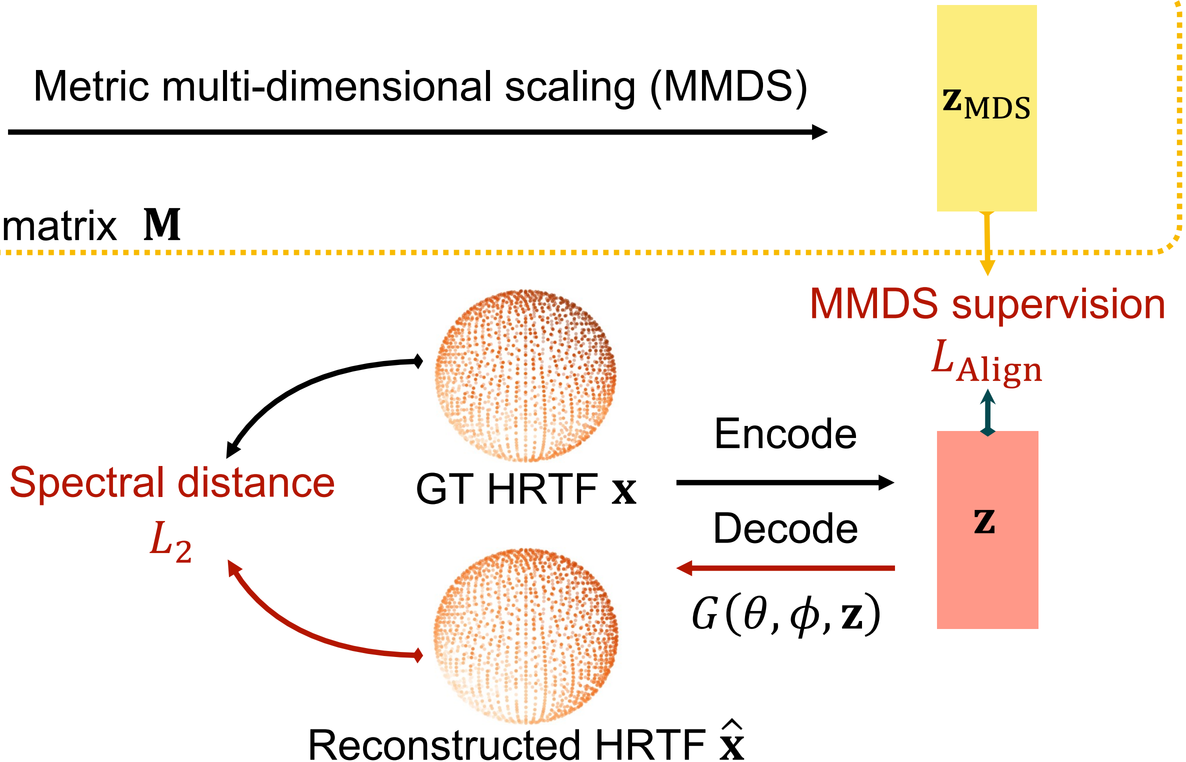
Figure from https://youtu.be/VKSJayDi_IQ post-processed by Gemini

Aligning with Perception-Informed Space (Cont'd)

If the perceptual metric is not differentiable (AEP, DRMSP)

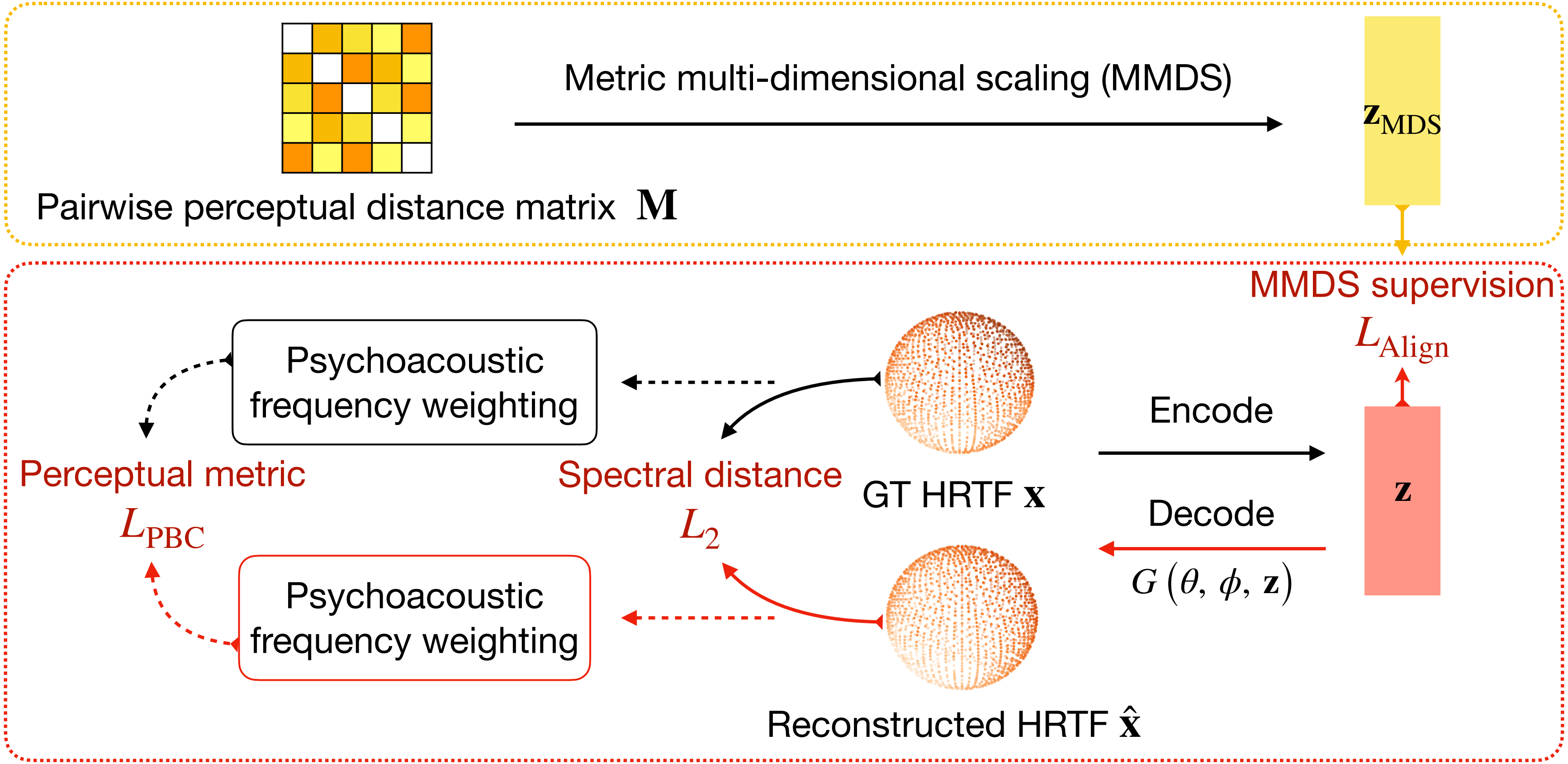


Pairwise perceptual distance matrix $\mathbf{M}$



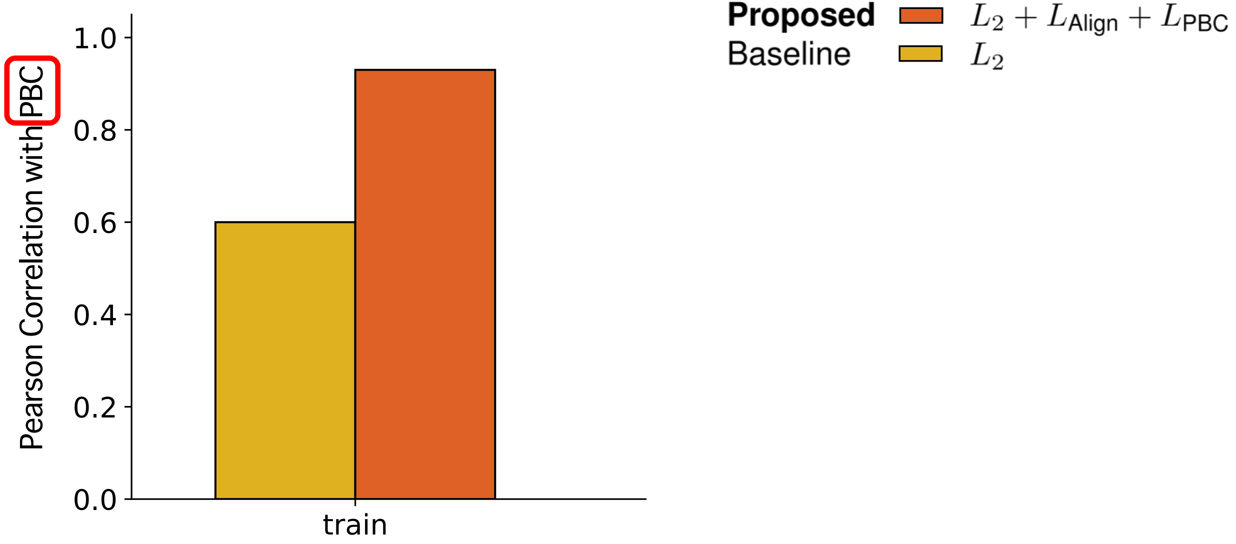
This can also be applied to differentiable metrics.

Overall Pipeline to Align Latent HRTF Representations



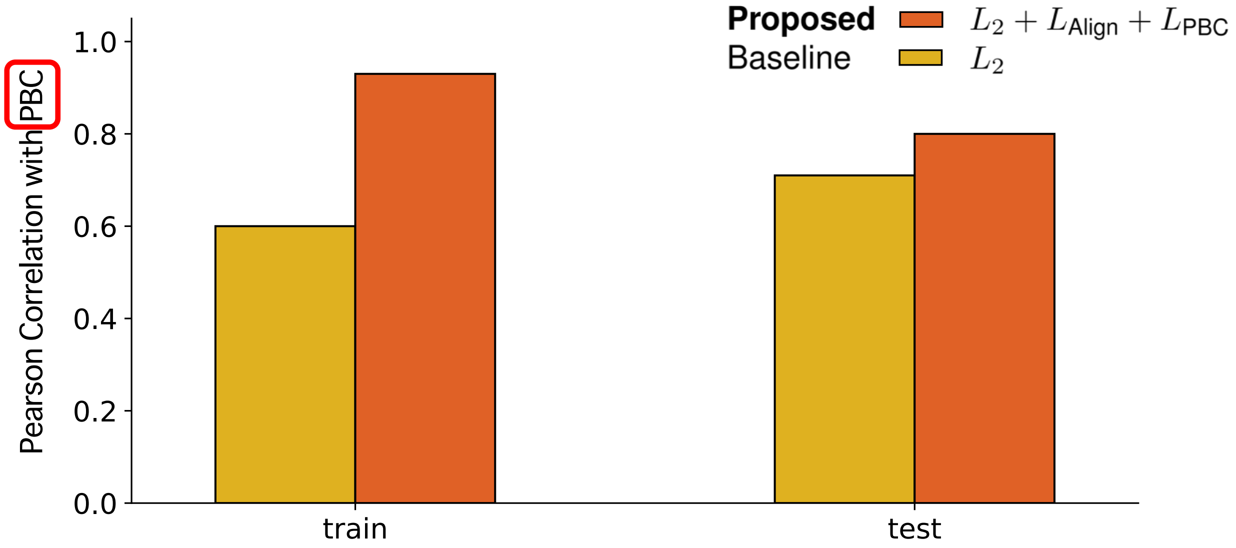
Results: Objective Perceptual Correlation Evaluation on PBC

- Our proposed method **achieves better alignment** with perception-informed space.
- The perceptual correlation learned in training transfer to test subjects (unseen).



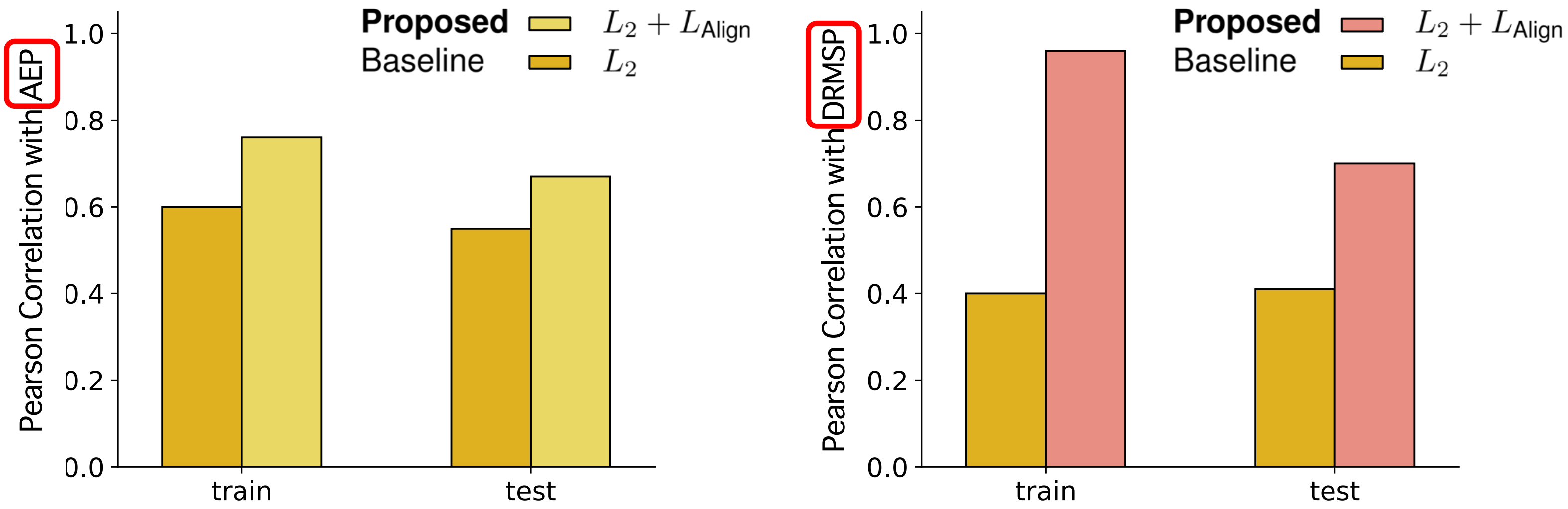
Results: Objective Perceptual Correlation Evaluation on PBC

- Our proposed method **achieves better alignment** with perception-informed space.
- The perceptual correlation learned in training transfer to test subjects (unseen).



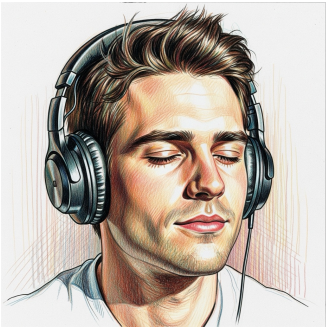
Generalization to AEP and DRMSP Metrics

- YES, our proposed correlation improvement method generalizes to externalization and localization.



Application: Personalized HRTF Selection

For each of the 13 test (unseen) subjects, we select the nearest HRTFs from the 65 training subjects, based on the learned latent representations.

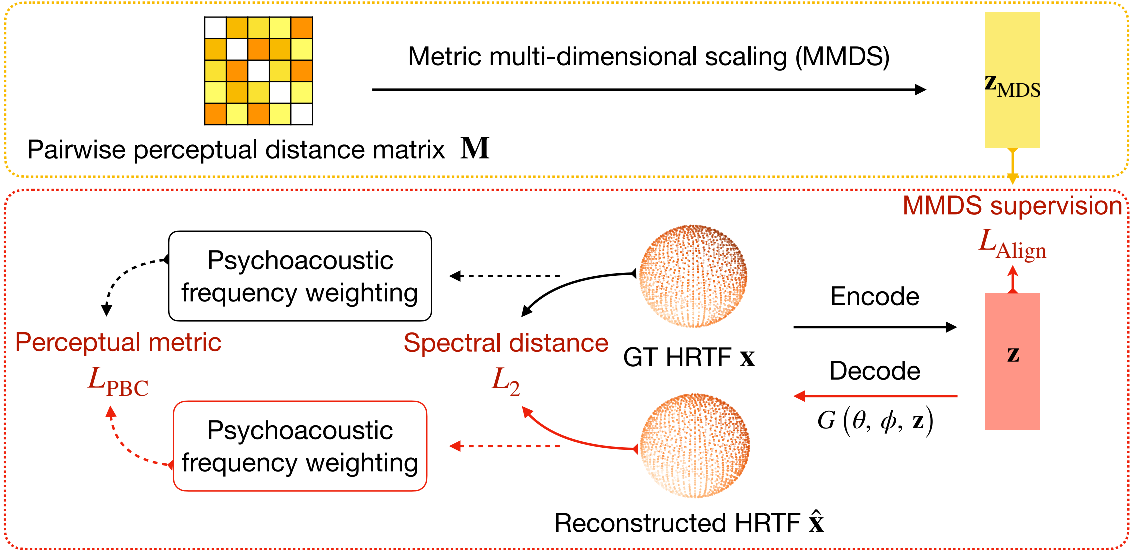


Methods	Best candidate	
	DRMSP↓	SDE (dB)↓
$L_2 + L_{\text{Align}}$	**3.20**	2.12
L_2	4.21	**2.07**

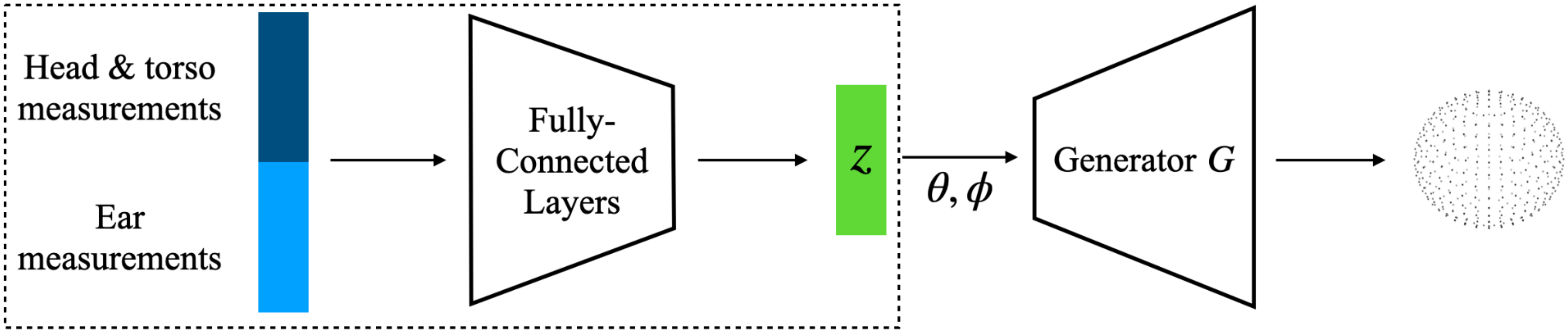
HRTFs selected by our proposed method yield lower perceptual distances with slightly higher Spectral Difference Error (SDE).

Limitations

- Objective metrics vs. listening experience
- MMDS assumes symmetric dissimilarity
- Ignoring phase information



Potential Extension on Personalization



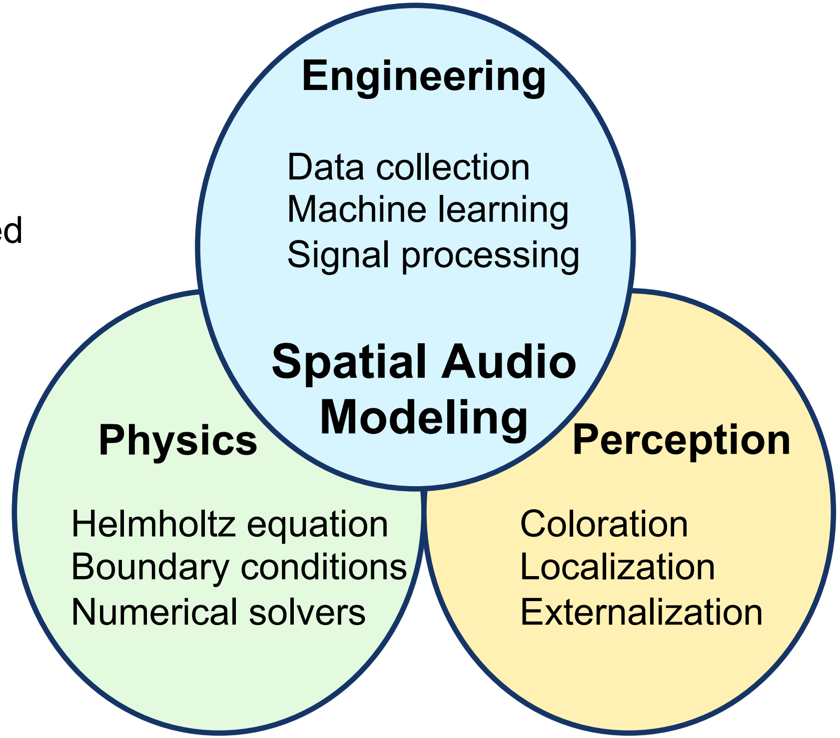
- An encoder to bridge representation learning to personalization
- Extend to SONICOM data
- Subjective validation

Wang, Yuxiang, **You Zhang**, Zhiyao Duan, and Mark Bocko. "Global HRTF personalization using anthropometric measures." In *Audio Engineering Society Convention* 2020.

Future Directions: Toward Unified Spatial Audio Modeling

Gap between simulated data / measured data

Physics-informed NN



Perception-informed binaural synthesis and reproduction

From scalable representation → to perceptually grounded and physics-aware synthesis.

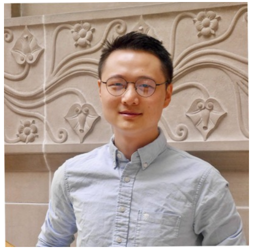
Acknowledgment

Funding

MELIORA
University of Rochester
Goergen Institute for Data Science & Artificial Intelligence

A New York State Center of Excellence

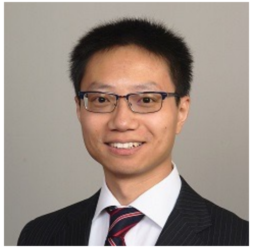
UR



Yuxiang Wang
(now working at Embody)

Yutong Wen
(now PhD student at UIUC)

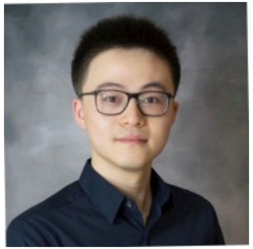
Mark Bocko



Zhiyao Duan

Meta

Andrew Francl



Ruohan Gao
(now Assit. Prof. at UMD)

Paul Calamia

Ishwarya Ananthabhotla

Summary

A Recipe for Scalable and Perceptually Grounded HRTF Personalization:

- **Model**: Neural fields for HRTF modeling
- **Data**: Position-dependent normalization
- **Perception**: Perceptual loss + MMDS supervision

Thank you! Questions?

Representation and dataset harmonization provide the foundation, but perception alignment defines quality.