Towards Perception-Informed Latent HRTF Representations

IEEE WASPAA 2025

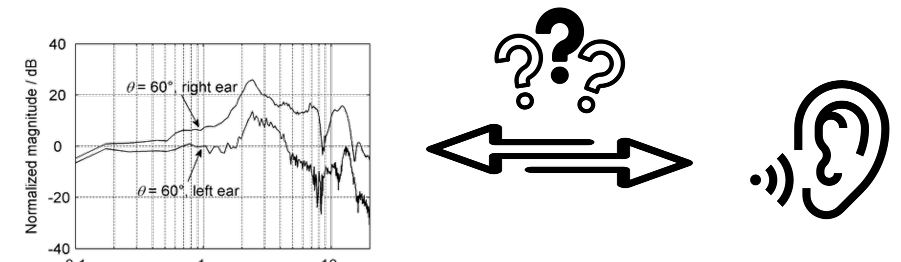
You (Neil) Zhang [1,2], *Andrew Francl* [2], *Ruohan Gao* [3], *Paul Calamia* [2], ***Zhiyao Duan*** [1], ***Ishwarya Ananthabhotla*** [2]

1 University of Rochester
2 Meta
3 University of Maryland
Reality Labs

TL;DR

Beyond spectral reconstruction, we learn a perception-informed HRTF latent space by preserving perceptual relations among HRTFs.

Research question:
- We investigate: how well do **existing** learned HRTF representations **preserve** perceptual relations.
- We improve: the latent HRTF representations to **align** them with human perception.

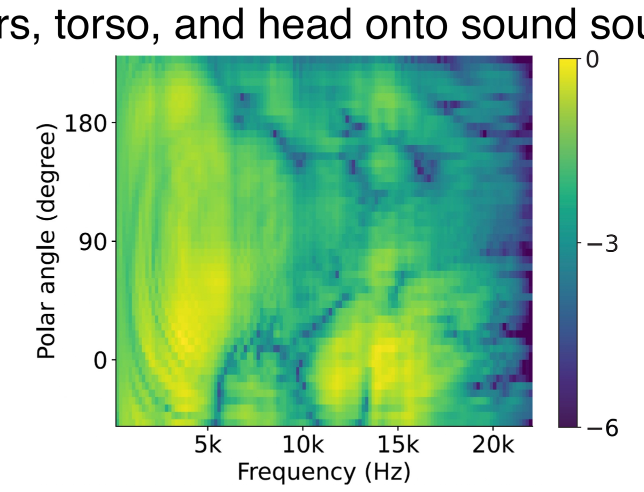


Proposed solution:
- Perceptual metric-based loss function
- Supervision via Metric Multidimensional Scaling (MMDS)

Application:
HRTF personalization

PRELIMINARIES

Head-related transfer functions (HRTFs) are a set of functions of **frequency** at different **azimuth** and **elevation** angles, describing the spatial filtering effect of the ears, torso, and head onto sound sources.



Spectral distance: Spectral Difference Error (**SDE**)

$$\mathrm{SDE}_k(\mathrm{H}, \hat{\mathrm{H}}) = \frac{1}{L}\sum_{\theta,\phi}\left|20 \cdot \log_{10}\left(\frac{\mathrm{H}(\theta,\phi,k)}{\hat{\mathrm{H}}(\theta,\phi,k)}\right)\right|$$

Computational Auditory Modeling
- **Coloration**: Predicted Binaural Coloration (**PBC**) [1]
- **Externalization**: Auditory Externalization Perception (**AEP**) [2]
- **Localization**: Difference of Root Mean Square Error in Polar Angles (**DRMSP**) [3]

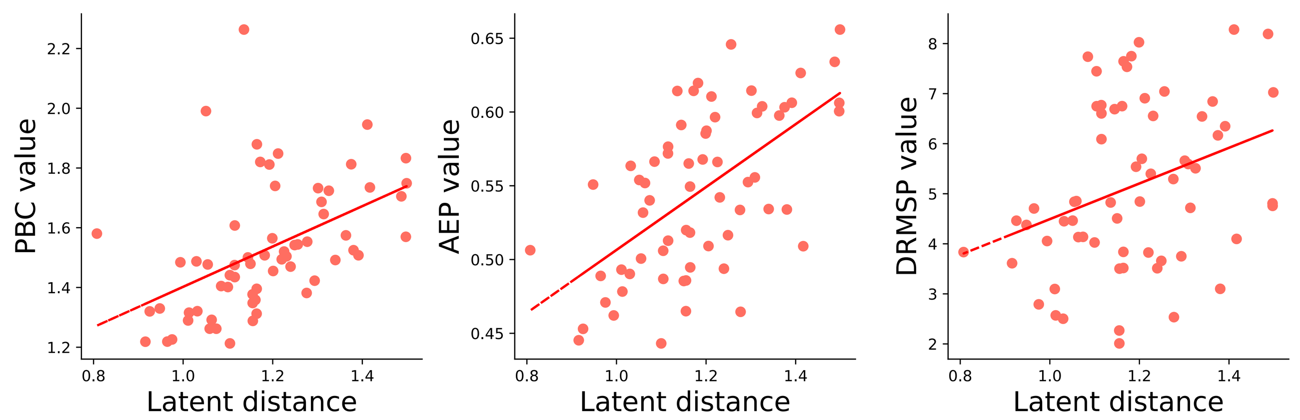
Pearson Correlation

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

CASE STUDY: Do Existing Learned HRTF Representations Preserve Perceptual Relations?

Dataset: SS2 HRTF dataset [4]
Setup: 1) Learning with spectral reconstruction
2) Compute pairwise latent distance across subjects
3) Compute pairwise perceptual distance across subjects

Correlation between latent space and the perceptual metrics
Model: Implicit Neural Representations; Anchor: one subject



Pearson correlation results for three perceptual metrics

Models	Partitions	PBC	AEP	DRMSP
Convolutional Autoencoder [5]	train	$0.60_{\pm 0.11}$	$0.71_{\pm 0.08}$	$0.43_{\pm 0.13}$
	test	$-0.15_{\pm 0.21}$	$0.07_{\pm 0.31}$	$-0.10_{\pm 0.27}$
Implicit Neural Representations [6]	train	$0.60_{\pm 0.09}$	$0.60_{\pm 0.14}$	$0.40_{\pm 0.15}$
	test	$0.71_{\pm 0.22}$	$0.55_{\pm 0.23}$	$0.41_{\pm 0.27}$
Correlation with SDE:		0.78	0.73	0.37

Minimizing spectral distance leads to **limited** perceptual correlation.

EXPERIMENTS: Improving Latent Representation Alignment with Perception-Informed Space

Comparing Pearson correlation and reconstruction error for the proposed methods and the baseline.
PBC metric; **Both losses** applied; **SS2** dataset

	Methods	Pearson Correlation ↑ Ground-truth (GT)		Reconstructed		Reconstruction Error SDE (dB) ↓		PBC ↓	
		train	test	train	test	train	test	train	test
Proposed	$L_2 + L_{\text{Align}} + L_{\text{PBC}}$	$0.93_{\pm 0.02}$	$\mathbf{0.80_{\pm 0.14}}$	$\mathbf{0.95_{\pm 0.01}}$	$\mathbf{0.86_{\pm 0.13}}$	0.87	1.58	**0.56**	1.04
Baseline	L_2	$0.60_{\pm 0.09}$	$0.71_{\pm 0.22}$	$0.78_{\pm 0.06}$	$0.80_{\pm 0.14}$	**0.82**	**1.51**	0.67	1.09
Ablation study	$L_2 + L_{\text{Align}}$	$\mathbf{0.96_{\pm 0.01}}$	$0.78_{\pm 0.14}$	$0.87_{\pm 0.04}$	$0.82_{\pm 0.13}$	1.00	1.58	0.79	1.11
	$L_2 + L_{\text{PBC}}$	$0.64_{\pm 0.10}$	$0.71_{\pm 0.21}$	$0.77_{\pm 0.08}$	$0.83_{\pm 0.17}$	1.03	1.58	0.64	**1.02**

- Our proposed method **achieves better alignment** with perception-informed space.
- The perceptual correlation learned in training transfer to test subjects (unseen).
- L_{Align} and L_{PBC} complement each other, and MMDS supervision (L_{Align}) dominates.

AEP / DRMSP metric; **MMDS supervision** loss; **SS2** dataset

	Methods	Pearson correlation↑ train GT	test GT	SDE (dB)↓ train	test
AEP	$L_2 + L_{\text{Align}}$	$\mathbf{0.76_{\pm 0.09}}$	$\mathbf{0.67_{\pm 0.16}}$	1.09	1.65
	L_2	$0.60_{\pm 0.14}$	$0.55_{\pm 0.23}$	0.82	1.51
DRMSP	$L_2 + L_{\text{Align}}$	$\mathbf{0.96_{\pm 0.02}}$	$\mathbf{0.70_{\pm 0.20}}$	0.91	1.74
	L_2	$0.40_{\pm 0.15}$	$0.41_{\pm 0.27}$	0.82	1.51

- Our proposed correlation improvement method generalizes to externalization and localization.

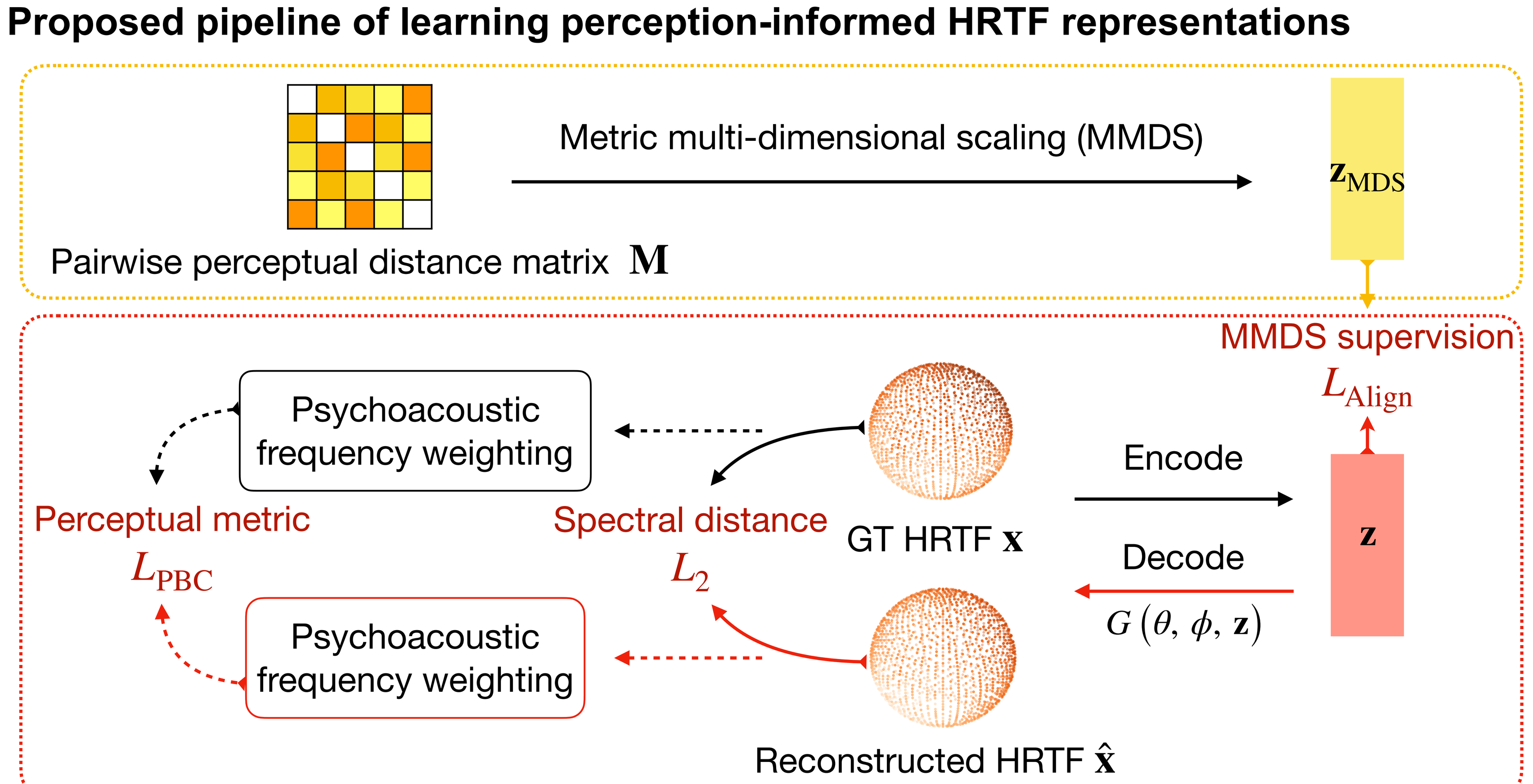
PBC metric; **Both losses** applied; **HUTUBS** dataset

Methods	Pearson correlation↑ train GT	test GT	SDE (dB)↓ train	test
$L_2 + L_{\text{Align}} + L_{\text{PBC}}$	$\mathbf{0.98_{\pm 0.01}}$	$\mathbf{0.71_{\pm 0.13}}$	0.29	1.60
L_2	$0.58_{\pm 0.12}$	$0.62_{\pm 0.14}$	0.42	1.45

- Our proposed correlation improvement method generalizes to HUTUBS dataset.

METHOD: Aligning with Perception-Informed Space

Proposed pipeline of learning perception-informed HRTF representations



Loss functions

$$L = L_2 + \alpha L_{\text{Align}} + \beta L_{\text{PBC}}$$

PBC loss (only when the metrics is differentiable)

$$L_{\text{PBC}} = \mathrm{PBC}(\mathbf{x}, \hat{\mathbf{x}})$$

Metric Multidimensional Scaling (**MMDS**) **supervision** (can be applied to every metric)

$$L_{\text{Align}} = \|\mathbf{z} - \mathbf{z}_{\text{MDS}}\|_2$$

APPLICATION: Personalized HRTF Selection

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

	Methods	Best candidate Metrics↓	SDE (dB)↓	Top 5 candidates Metrics↓	SDE (dB)↓
PBC	$L_2 + L_{\text{Align}} + L_{\text{PBC}}$	**1.21**	2.11	**1.31**	**2.17**
	L_2	1.30	**2.07**	1.38	2.19
AEP	$L_2 + L_{\text{Align}}$	0.49	2.17	**0.50**	2.27
	L_2	**0.48**	**2.07**	0.51	**2.19**
DRMSP	$L_2 + L_{\text{Align}}$	**3.20**	2.12	**3.61**	2.26
	L_2	4.21	**2.07**	4.42	**2.19**

HRTFs selected by our methods consistently yield lower perceptual distances.

Full paper

SS2 dataset

REFERENCES

[1] McKenzie, Thomas, et al. "Predicting the colouration between binaural signals." *Applied Sciences* 2022.
[2] Baumgartner, Robert, and Piotr Majdak. "Decision making in auditory externalization perception: model predictions for static conditions." *Acta Acustica* 2021.
[3] Barumerli, Roberto, et al. "A Bayesian model for human directional localization of broadband static sound sources." *Acta Acustica* 2023.
[4] Warnecke, Michaela, et al. "Sound Sphere 2: A high-resolution HRTF database." *AES AVAR* 2024.
[5] Zhao, Jiale, Dingding Yao, and Junfeng Li. "Head-Related Transfer Function Upsampling With Spatial Extrapolation Features." *IEEE TASLP* 2025.
[6] Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *IEEE ICASSP* 2023.