IEEE ICASSP 2021

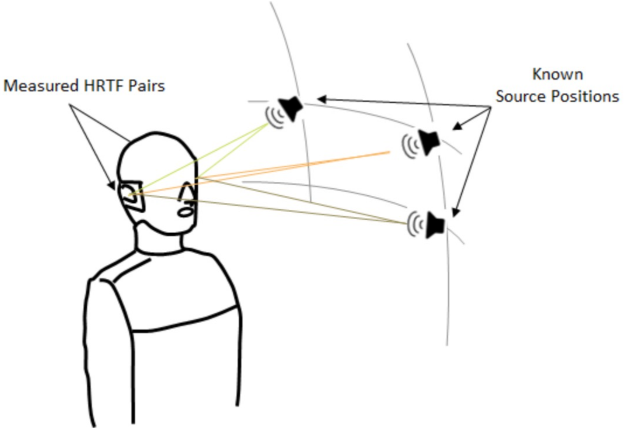
On the Predictability of HRTFs from Ear Shapes Using Deep Networks

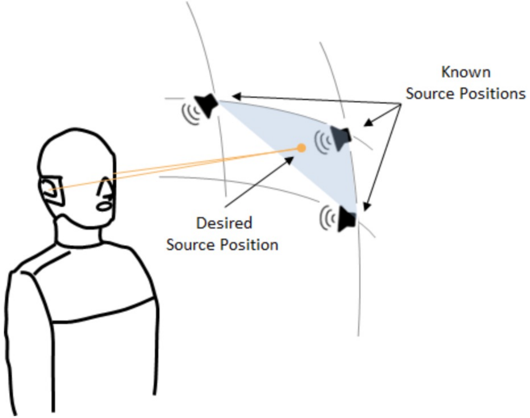
Yaxuan Zhou, Hao Jiang, Vamsi Krishna Ithapu
Facebook Reality Labs Research

yaxuanzh@uw.edu, haojiang@fb.com, ithapu@fb.com

Head-Related Transfer Functions

HRTFs:
- parameterize the transformations for the acoustic signals from source to ear canals.
- key component for spatial audio perception in AR/VR.
- determined by structure of the head (& ear)



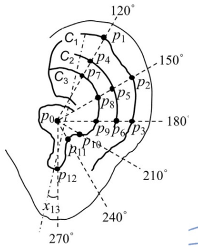


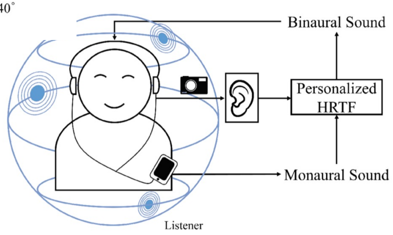
HRTF individualization: get personalized HRTFs for every individual
- Acoustical measurement;
- Numerical simulation using 3D scan of upper body / ear;
- HRTF prediction by data-driven learning-based approaches

Costly, inconvenient and thus non-scalable

Problems in Learning-based HRTF Prediction

Existing learning-based HRTF prediction approaches:





HRTF

Limitations:

- Limited representation of the full structural attributes of ears
- Constraint from generality of the HRTF database

Our Goals

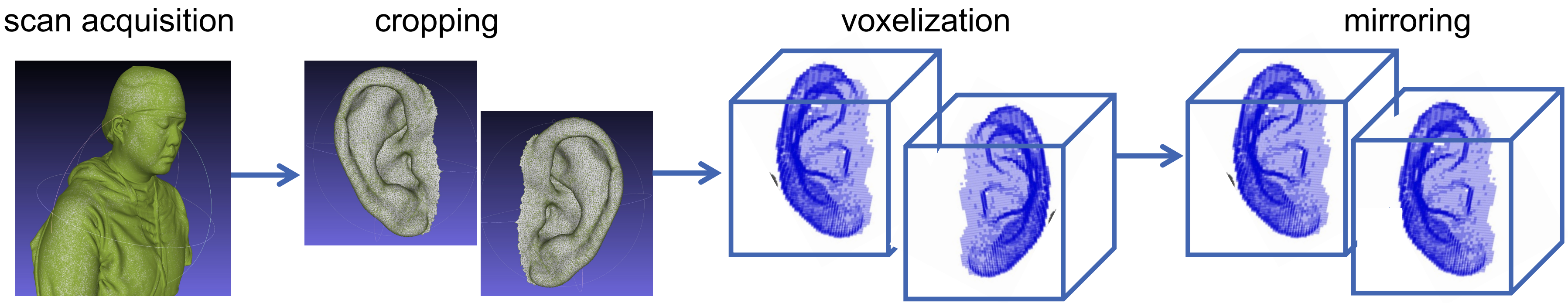
3D ear shape representation from a **large-scale** dataset

3D Deep Neural Network →

HRTF

- Establish a lower bound of HRTF prediction error from different ear-related input
- Explore possibility of using deep learning as a computationally-efficient alternative to numerical simulation

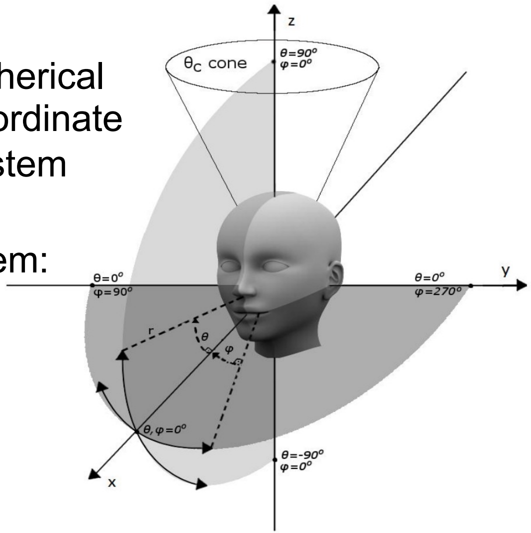
Ear Data



HRTF Data

- Simulated far-field HRTFs
- 30 frequency bins ranging from 1kHz-12kHz
- 360 directions parameterized by spherical coordinate system:
  - 36 azimuths: [0º : 10º : 360º)
  - 10 elevations: [-30º : 10º : 60º]

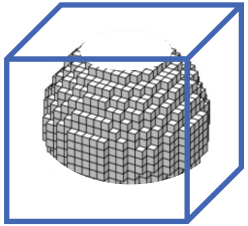
Spherical coordinate system



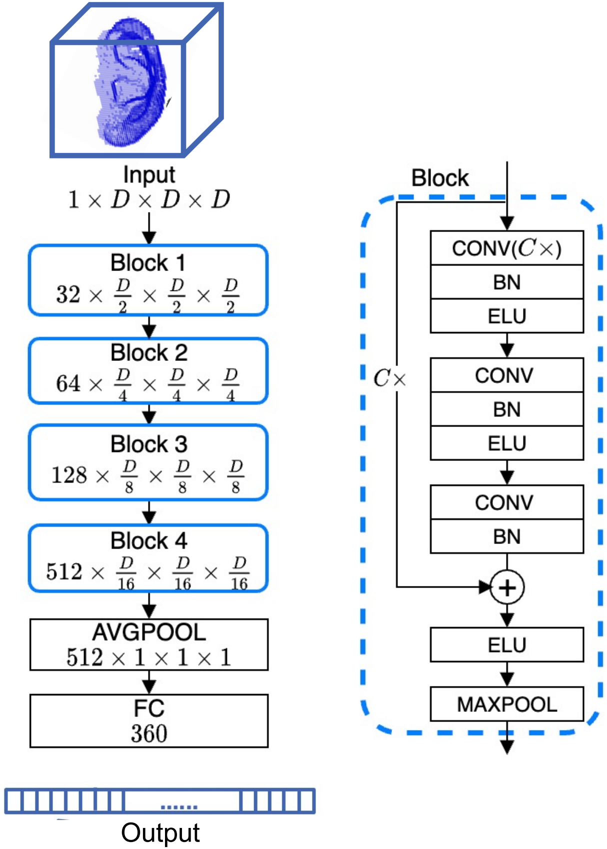
HRTF representations:

HRTF vector: a flat vector with 360 magnitudes.
→ Simple and space-efficient.

HRTF tensor: 3D tensor with 360 HRTF magnitudes embedded based on spatial coordinates.
→ Retains spatial information of HRTF.



Convolutional Neural Network Regression model



CNN-Reg:

Design choice:

- Train 30 CNN-Reg models, each predicting HRTF magnitudes across 360 directions on 1 frequency bin.

Design considerations:

- Response on different frequency may rely on different set of features;
- Dimension of output vector influences the size of fully-connected layer which is a major bottleneck for network footprint.

U-shaped Network Regression model

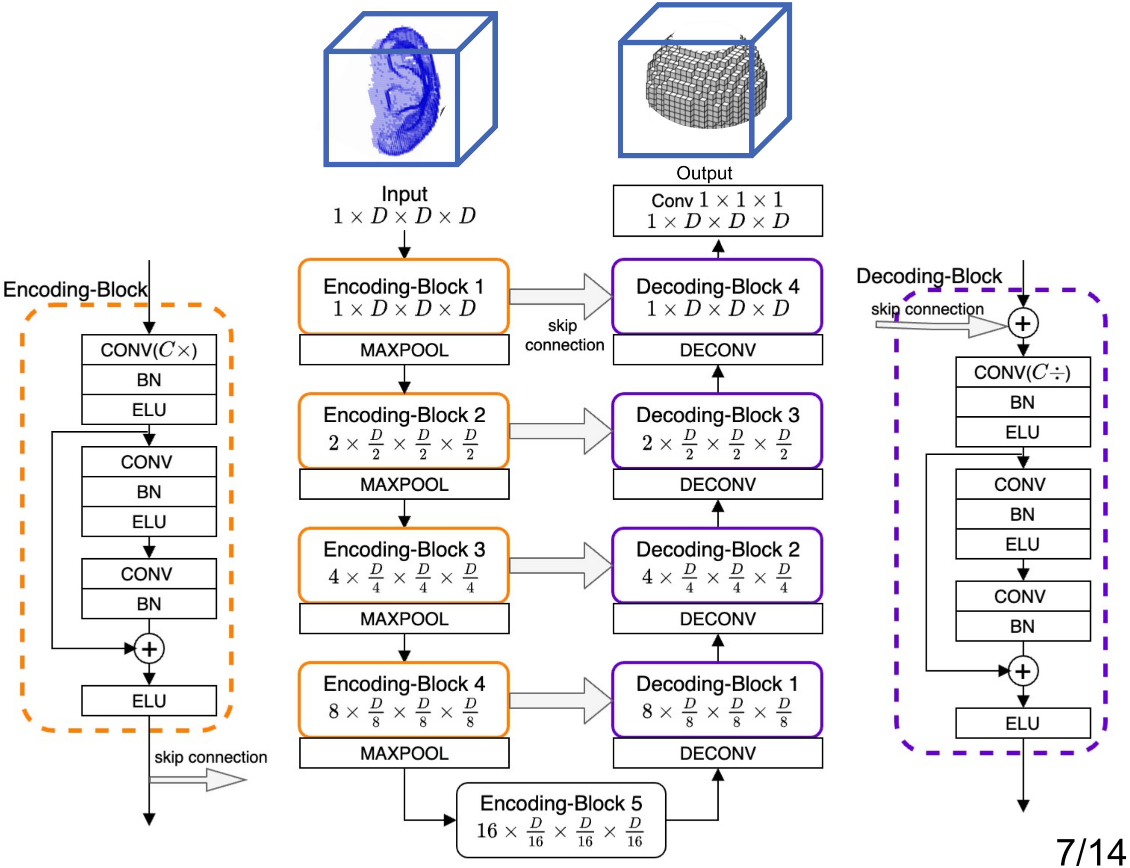
UNet-Reg:

Design choice:

- Train 30 UNet-Reg models.
- Domain-inspired design: Use UNet architecture to allow for spatial correspondence between ear shape and HRTF tensor.

Advantages:

- Scalability to denser HRTF spatial grid.
- Scalability to near-field HRTF prediction.
- Fewer network parameters: 35k vs. 17m(CNN-Reg)



Experiment Methodology

Loss function / evaluation:

- Spectral distance error (SDE) in dB: the lower the better

$$\mathrm{SDE}(f) = \frac{1}{N_d} \sum_{\theta,\varphi} \left| 20 \log \frac{\hat{h}(\theta,\varphi,f)}{h(\theta,\varphi,f)} \right|$$

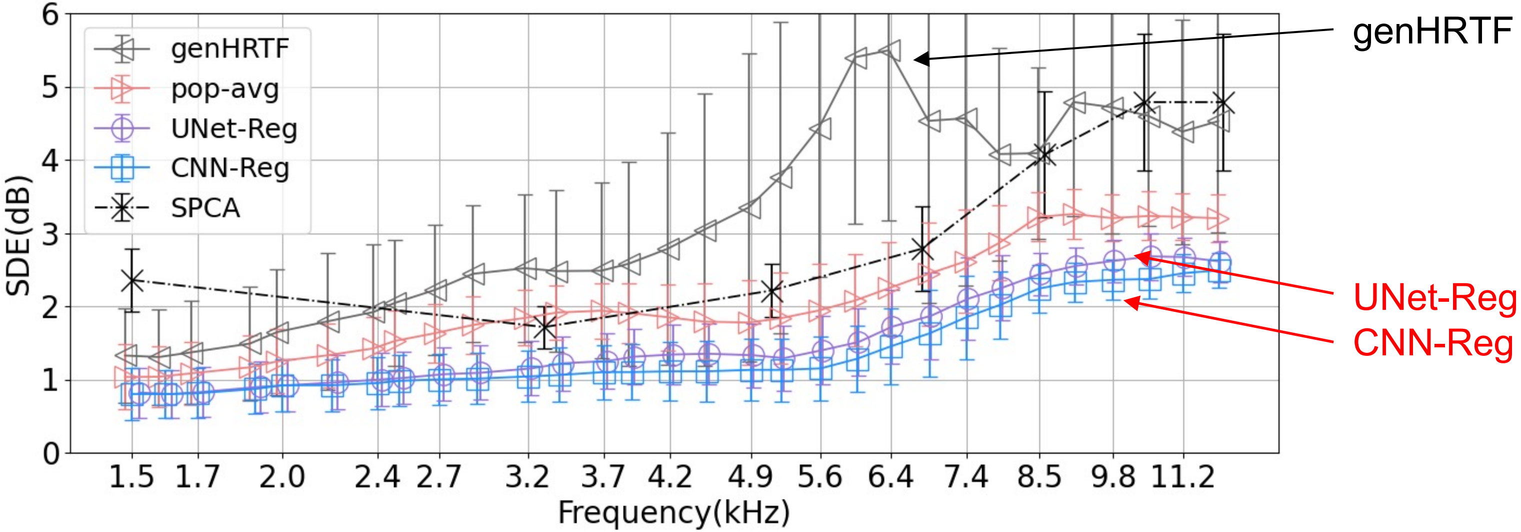
Training scheme:

- 1290 ear-HRTF data divided into 1000 for training and 290 for evaluation.
- 5-fold cross validation.

Baselines for evaluation:

1. **genHRTF**: KEMAR simulated HRTF
2. **pop-avg**: population average of HRTFs in training set.

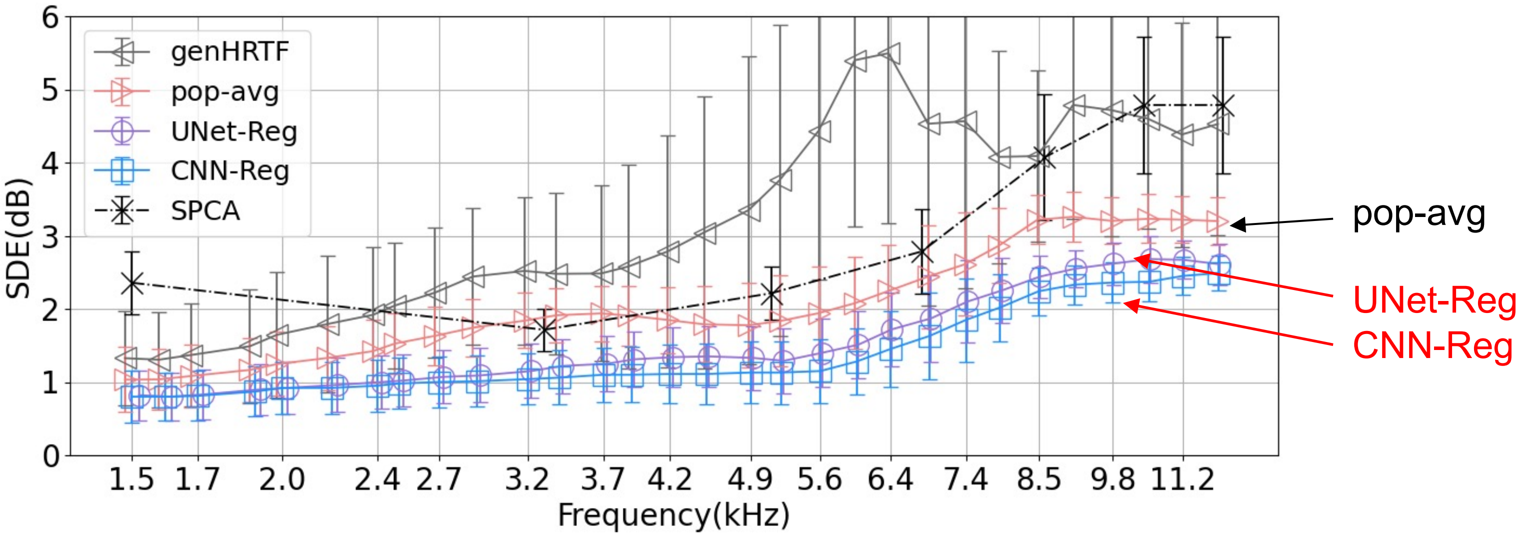
Comparison with baseline genHRTF



CNN/UNet-Reg vs. genHRTF:

Our methods significantly outperform genHRTF at all frequencies, proving the need for individualization.

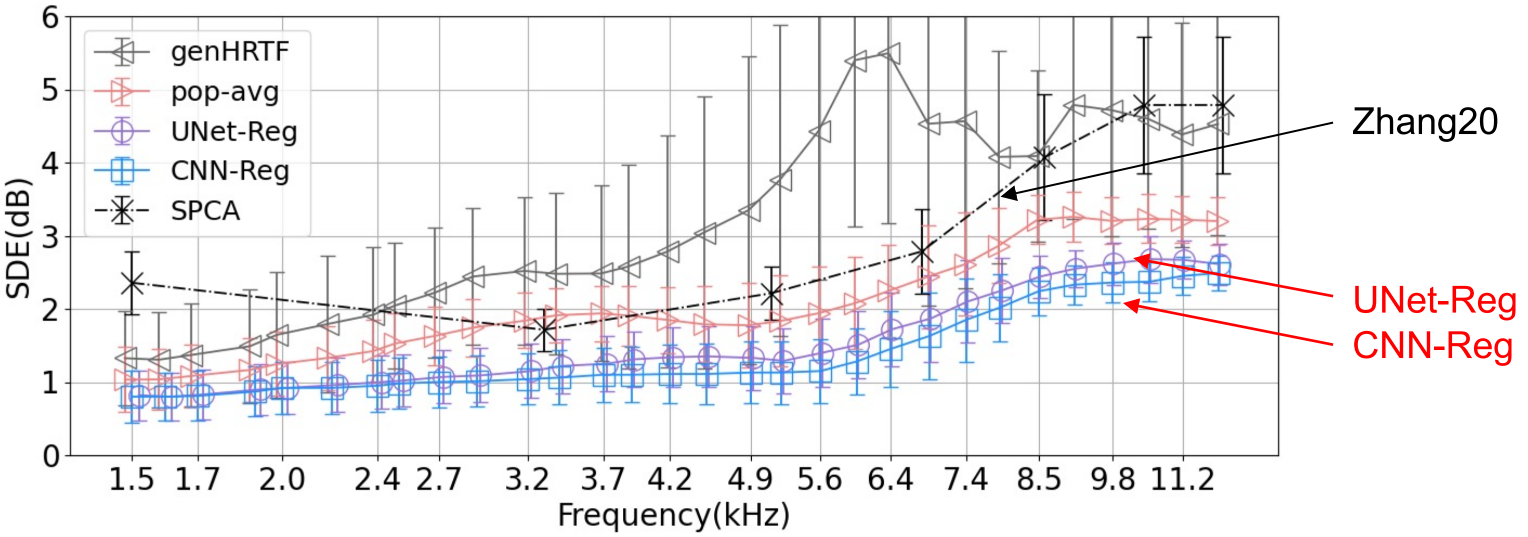
Comparison with baseline pop-avg



CNN/UNet-Reg vs. pop-avg:

Our methods outperform pop-avg by ~1dB.

Comparison with prior works

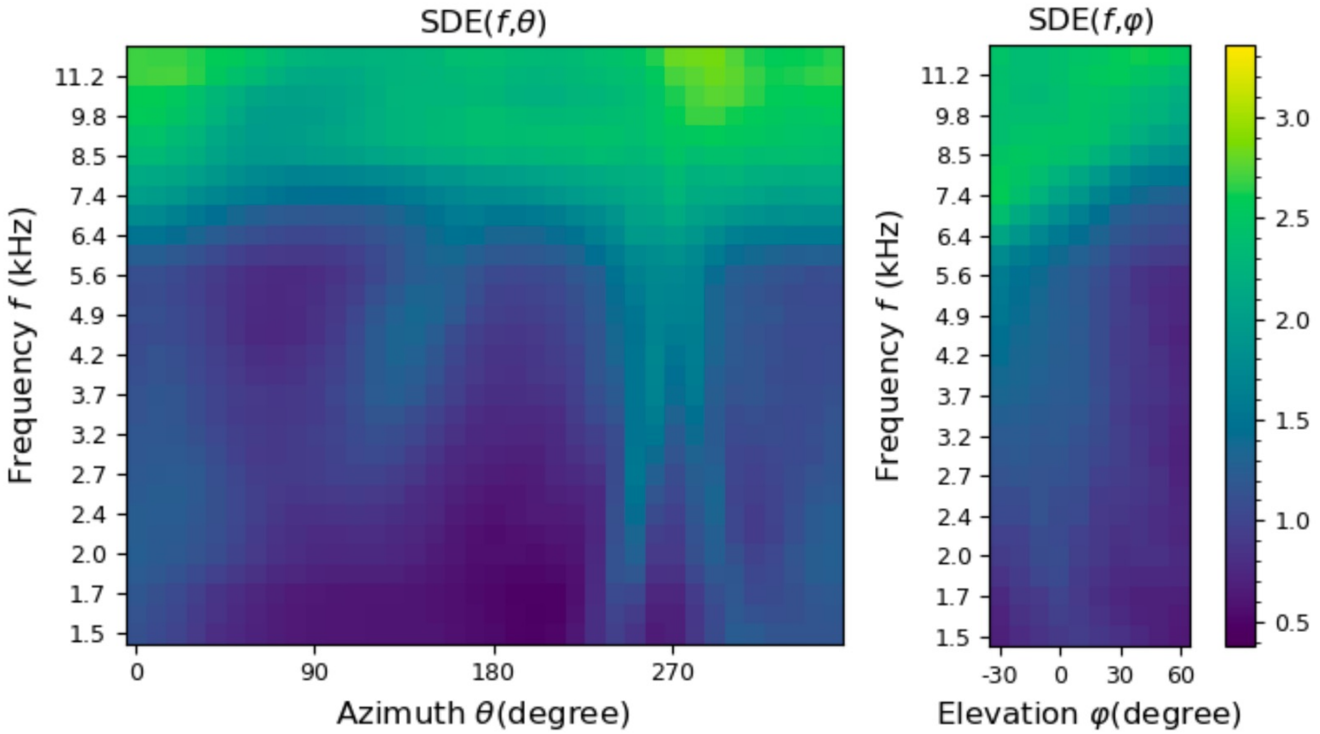


CNN/UNet-Reg vs. SPCA[Zhang20] & [Chen19]:

Our methods outperform both prior works.

	CNN-Reg	UNet-Reg	Zhang20*	Chen19*	genHRTF
SDE↓	**1.67**	1.84	3.24	3.43	3.63

Visualization of prediction SDE from CNN-Reg



Other results

Effect of voxelization

Input Grid	$16 \times 16 \times 16$	$32 \times 32 \times 32$	$64 \times 64 \times 64$
CNN-Reg	1.49 ± 0.36	$\mathbf{1.38 \pm 0.38}$	1.57 ± 0.43
UNet-Reg	1.61 ± 0.45	1.53 ± 0.38	$\mathbf{1.52 \pm 0.41}$

Comparison with numerical simulation

	SDE ↓	Speed ↓
CNN/UNet-Reg	1.38 dB / 1.52dB	3-8 ms/ear
Numerical simulation	-	20-30 min/ear

Summary

Our contributions:

- We proposed two DNN models that predict HRTFs from 3D ear tensors.
- We trained the models with a large-scale ear-HRTF dataset and achieved highest HRTF prediction accuracy in efforts to identify the lower bound of error in learning-based HRTF prediction.
- We've shown the potential and bottleneck of using learning-based HRTF prediction as a computationally efficient alternative to numerical simulation.

Future works:

- Include perceptual loss functions during DNN training.
- Further improve model design in terms of computational efficiency and representational capability.