# Room Acoustics and Spatial Audio

Neil Zhang

ECE 477 - Fall 2024

(Some slides adapted from AES AfG tutorial on personalized spatial audio)

# Outline

Room Acoustics

- Room Impulse Response Generation
- Cross-Modal RIR Generation
- Blind Room Acoustics Parameter Estimation

Spatial Audio

- HRTF Interpolation
- HRTF Personalization
- Binaural Synthesis

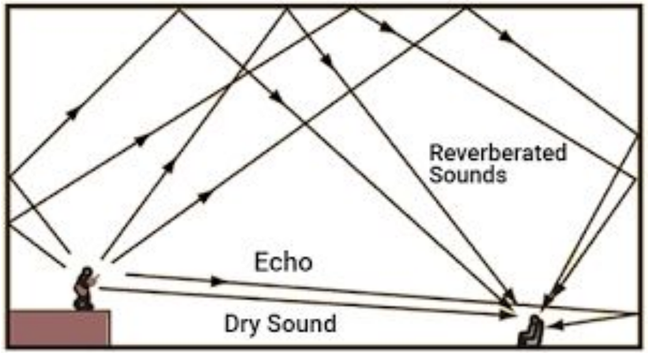
# Room Acoustics

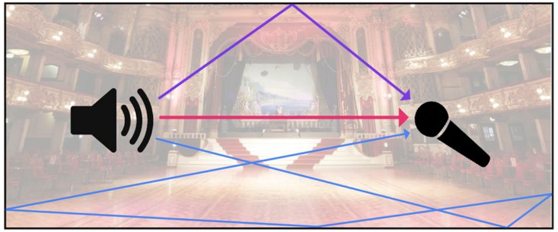
# Reverberation

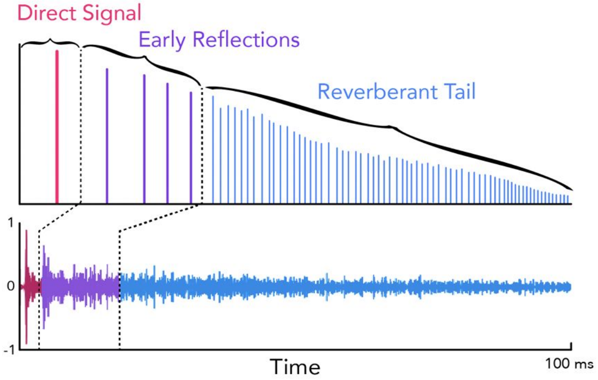
Reverberation is the process of multipath propagation of a sound from its source to one or more receivers.

View room as an LTI system → $x(t) = \mathbf{h(t)} * s(t) + n(t)$

Room Impulse Response



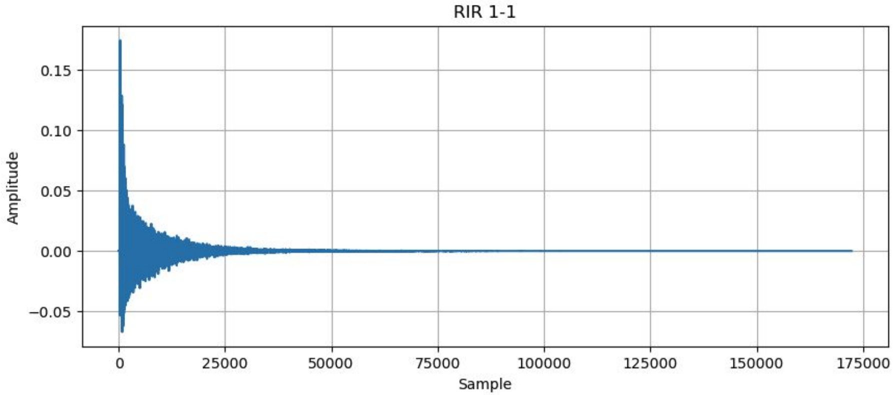



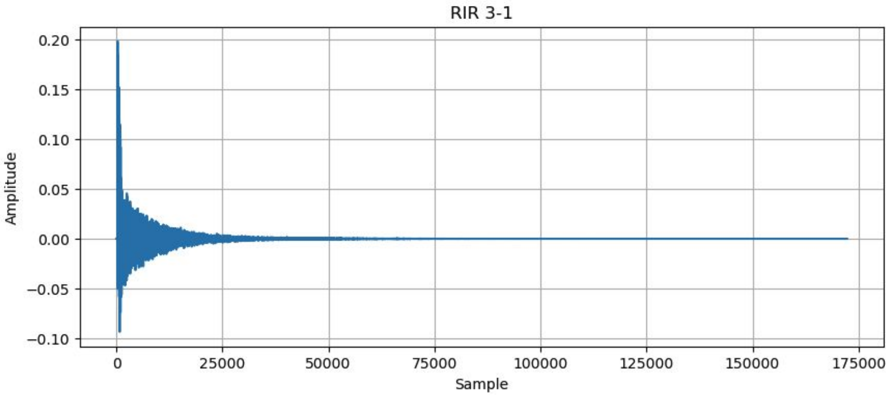


# RIR examples

RIRs are different from location to location

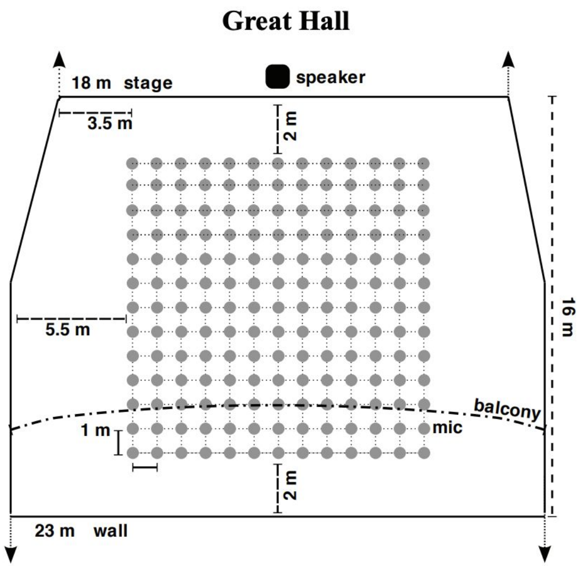
(Figures simulated with Pyroomacoustics)

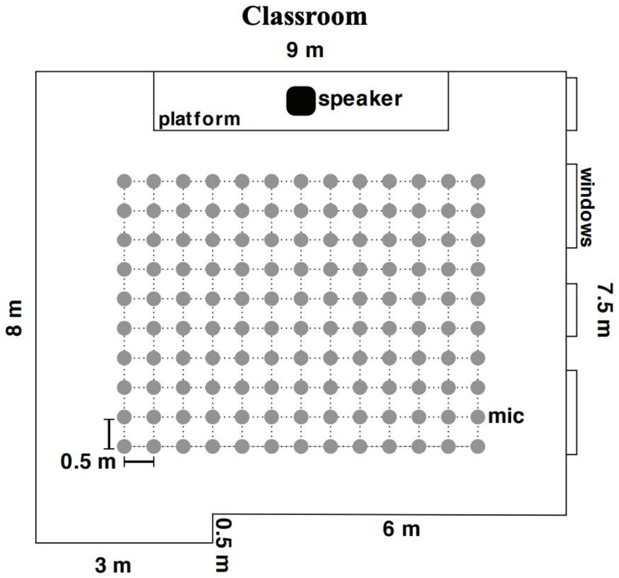


# Measure RIRs

Set up speaker and microphones




Jeub, Marco, Magnus Schafer, and Peter Vary. "A binaural room impulse response database for the evaluation of dereverberation algorithms." *2009 16th International Conference on Digital Signal Processing*. IEEE, 2009.

# Measure RIRs

Apply specific signals with predetermined cross-correlation results, to enable extraction of the room impulse response (RIR) from the output signal.

## Exponential Sine Sweep (ESS)

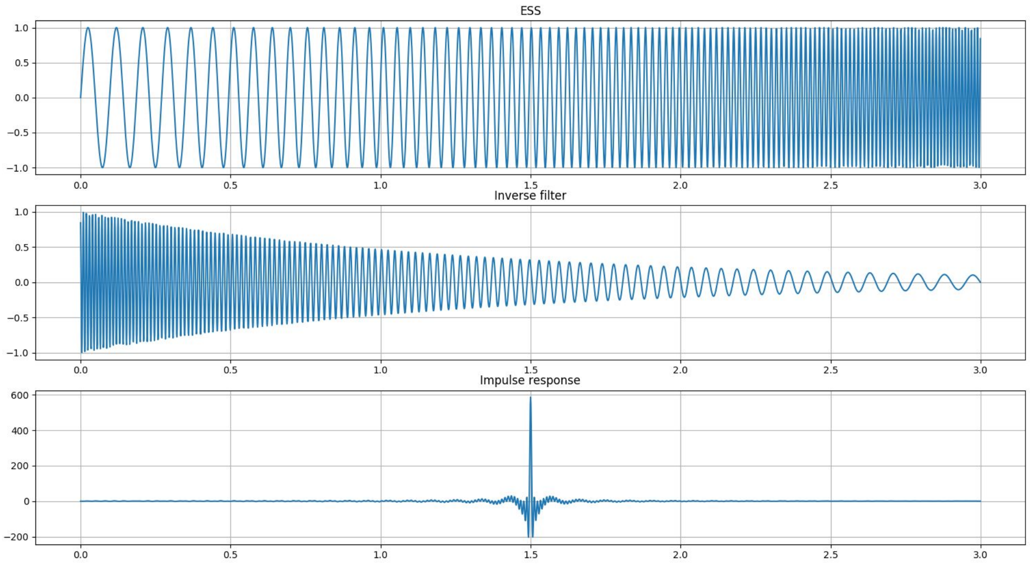
$$y[n] = (h * s)[n].$$

Taking the cross-correlation with respect to $s[n]$ of both sides,

$$\phi_{sy} = h[n] * \phi_{ss}$$

and assuming that $\phi_{ss}$ is an impulse (valid for long sequences)

$$h[n] = \phi_{sy}.$$

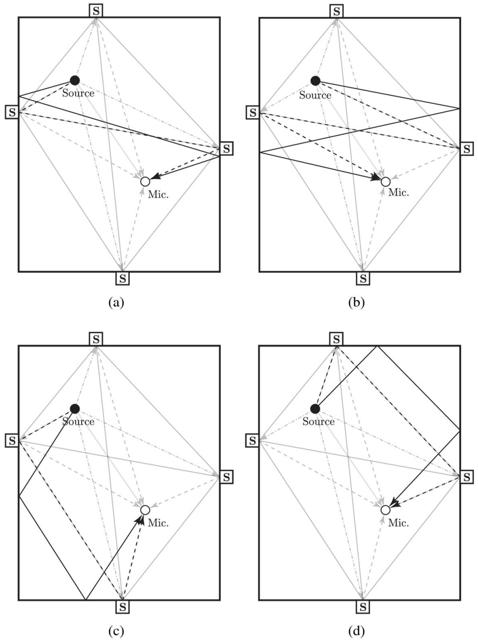



Farina, Angelo. "Simultaneous measurement of impulse response and distortion with a swept-sine technique." *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

# Simulating RIRs

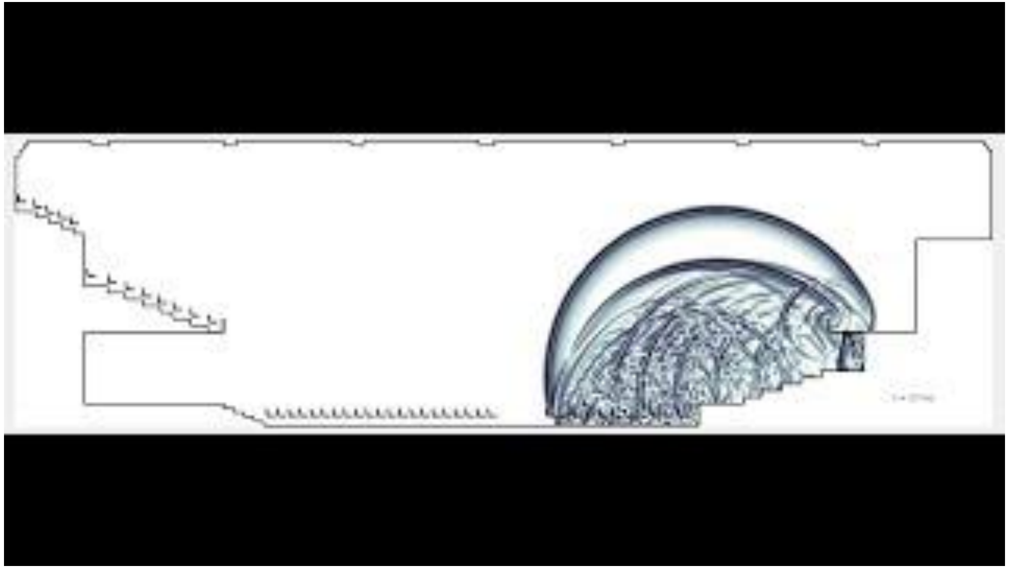
## Ray-based method

**Image-Source Method (ISM), Ray Tracing (RT), …**




## Wave-based method

**Finite-Difference Time-Domain (FDTD), …**



Figures from: https://arxiv.org/pdf/1502.05751
https://www.brianhamilton.co/

# Limitations

Measuring RIRs:

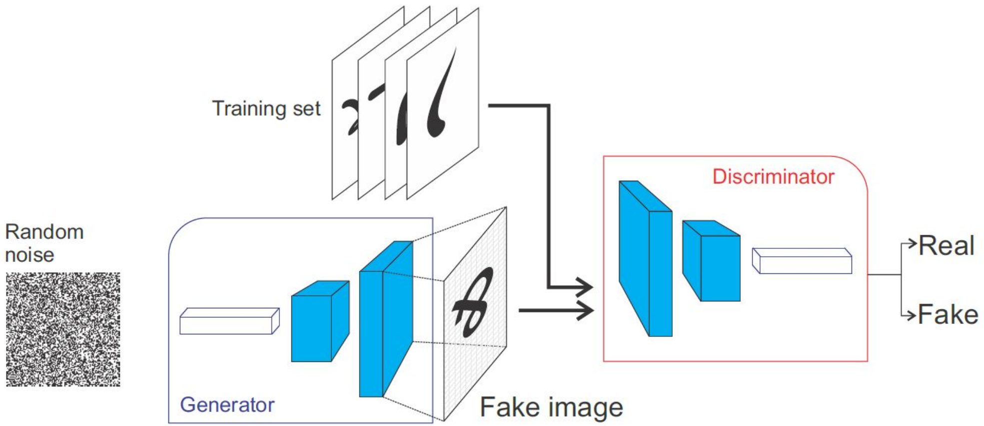
- Time-intensive and expensive
- Infeasible for inaccessible locations

Simulating RIRs:

- Shoebox empty room
- Strong physical assumptions

# Generative Adversarial Network (GAN)

Learn a mapping from a low-dimensional vector space to a high-dimensional space where the data is represented.




$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

# IR-GAN [Ratnarajah+2021]

AUDIO INFORMATION RESEARCH

Use Generative Adversarial Network (GAN) to generate RIRs. Constrained RIR Generation with key acoustic parameters to avoid noisy RIRs.

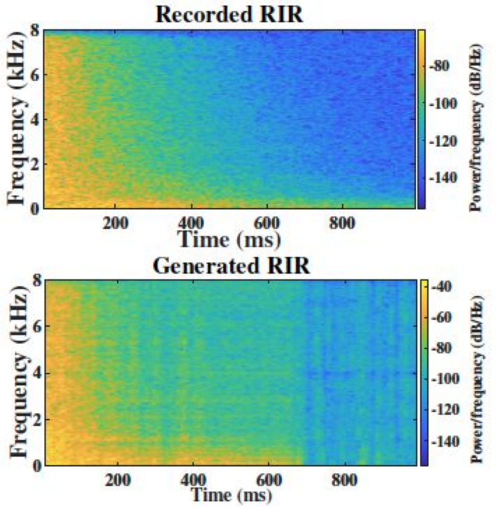



Figure 1: *Spectrogram of real RIR and RIR generated using our GAN-based approach. We can see both spectrograms have similar energy distributions.*

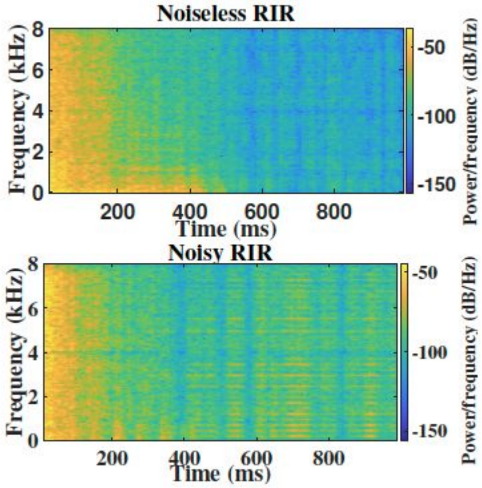



Figure 2: *Spectrogram of noiseless RIR and noisy RIR. The noiseless RIR has a $T_{60}$ value of around 1, and the noisy RIR has a $T_{60}$ value of around 3. In the noisy spectrogram, we can see many horizontal artifacts around 700ms.*

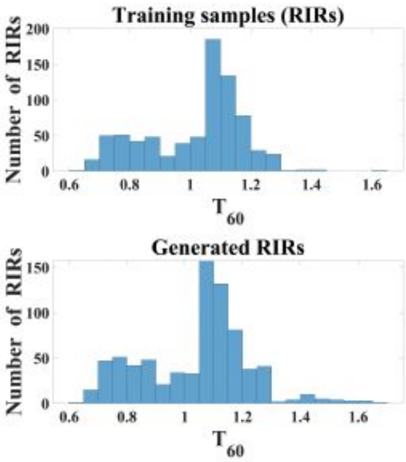
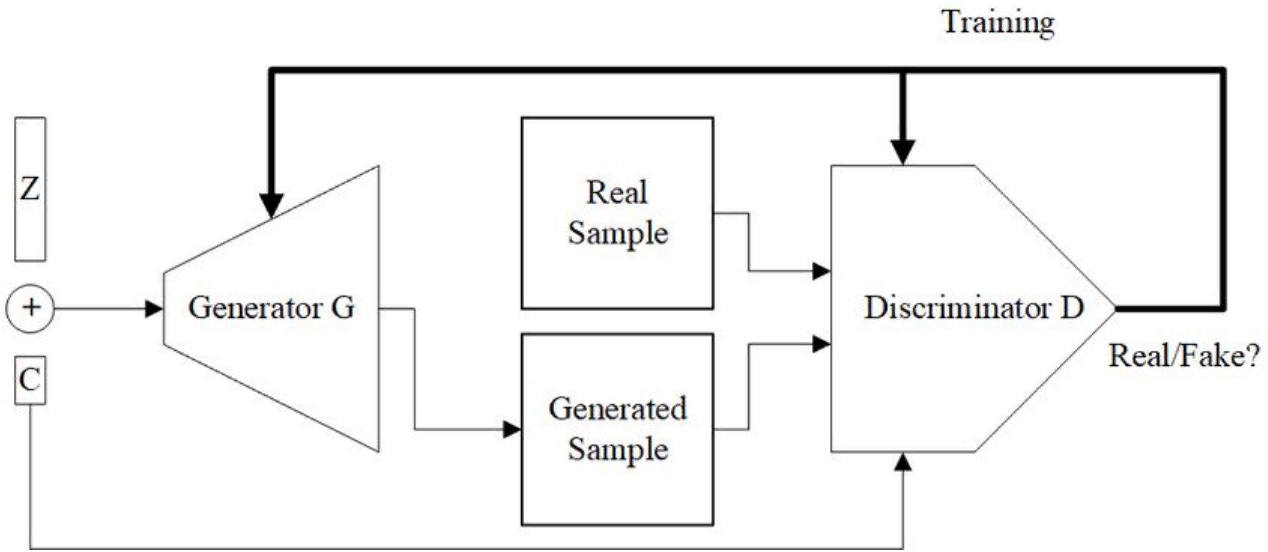



Figure 3: *$T_{60}$ distribution of training samples and $T_{60}$ d*
*tion of RIRs generated using our IR-GAN with the const*



Ratnarajah, Anton, Zhenyu Tang, and Dinesh Manocha. "IR-GAN: Room impulse response generator for far-field speech recognition." *Proc. Interspeech* 2021.

# Conditional GAN




$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x}|\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))].$$

# Fast-RIR [Ratnarajah+2022]

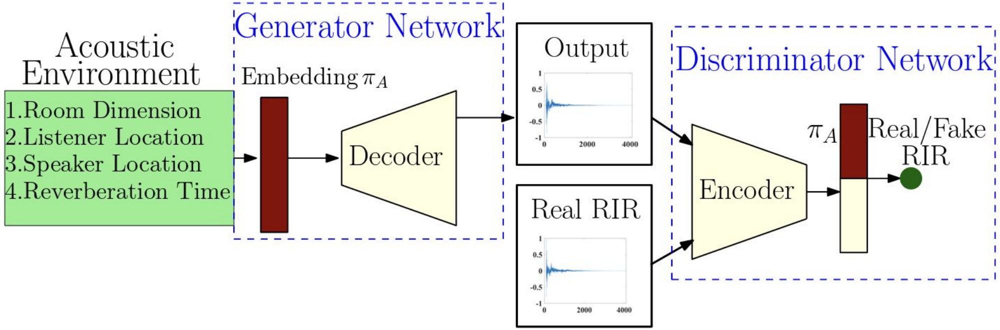



**Fig. 1**. The architecture of our FAST-RIR. Our Generator network takes acoustic environment details as input and generates corresponding RIR as output. Our Discriminator network discriminates between the generated RIR and the ground truth RIR for the given acoustic environment during training.


Ratnarajah, Anton, et al. "FAST-RIR: Fast neural diffuse room impulse response generator." *Proc. ICASSP 2022*.


**Table 1**. The runtime for generating 30,000 RIRs using image method, gpuRIR, DAS, and our FAST-RIR. Our FAST-RIR significantly outperforms all other methods in runtime.

| RIR Generator | Hardware | Total Time | Avg Time |
|---|---|---|---|
| DAS [7] | CPU | $9.01\text{x}10^5$s | 30.05s |
| Image Method [5] | CPU | $4.49\text{x}10^3$s | 0.15s |
| **FAST-RIR(Batch Size 1)** | **CPU** | **$2.15\text{x}10^3$s** | **0.07s** |
| gpuRIR [13] | GPU | 16.63s | $5.5\text{x}10^{-4}$s |
| FAST-RIR(Batch Size 1) | GPU | 34.12s | $1.1\text{x}10^{-3}$s |
| **FAST-RIR(Batch Size 64)** | **GPU** | **1.33s** | **$4.4\text{x}10^{-5}$s** |
| FAST-RIR(Batch Size 128) | GPU | 1.77s | $5.9\text{x}10^{-5}$s |

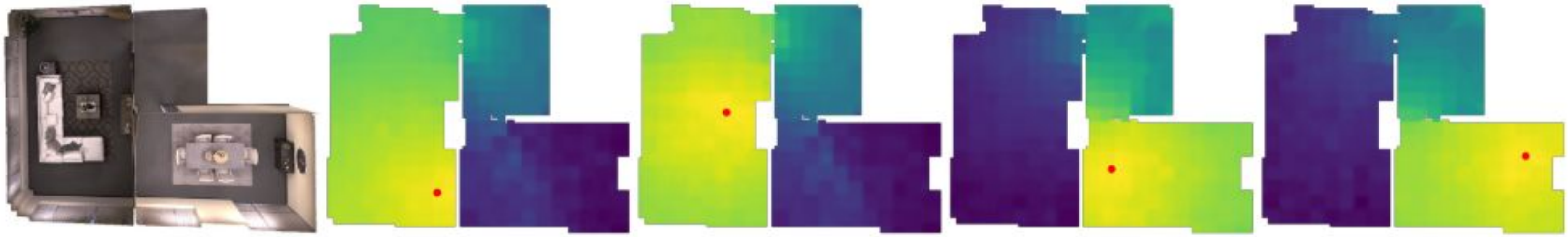
**Table 2**. $T_{60}$ error of our FAST-RIR for 30,000 testing acoustic environments. We report the $T_{60}$ error for RIRs cropped at $T_{60}$ and full RIRs. We only crop RIRs with $T_{60}$ below 0.25s.

| $T_{60}$ Range | Crop RIR at $T_{60}$ | $T_{60}$ Error |
|---|---|---|
| 0.2s - 0.25s | No | 0.068s |
| **0.2s - 0.25s** | **Yes** | **0.033s** |
| 0.25s - 0.7s | - | 0.021s |
| 0.2s - 0.7s | No | 0.029s |
| **0.2s - 0.7s** | **Yes** | **0.023s** |

# Learning Neural Acoustic Field (NAF) [Luo+2022]

Render spatial audio for arbitrary emitter and listener locations

Capture sound propagation in a scene



(a) (b) (c) (d) (e)

Figure 1: Neural Acoustic Field (NAF) learns an implicit representation for acoustic propagation. **(a)** A 3D top-down view of the house with two rooms. **(b)-(e)** The loudness of acoustic field as predicted by our NAF is visualized for an emitter located at the red dot. Notice how sound does not leak through walls, and the portaling effect open doorways can have. Louder regions are shown in yellow.



Luo, A., Du, Y., Tarr, M., Tenenbaum, J., Torralba, A., & Gan, C. (2022). Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, *35*, 3165-3177.

# Learning Neural Acoustic Field (NAF) [Luo+2022]

Key idea: Condition the network on a shared geometric feature grid

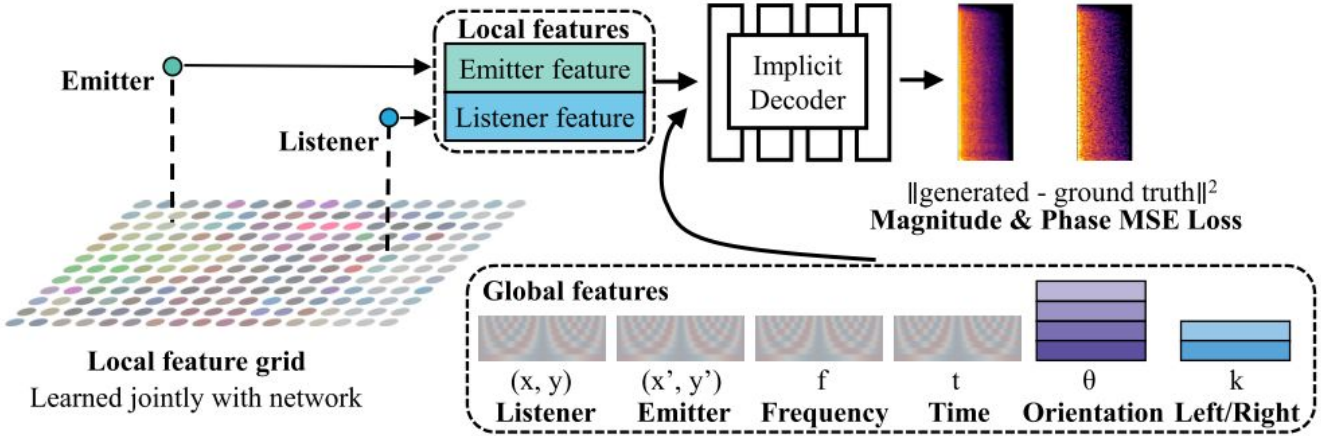


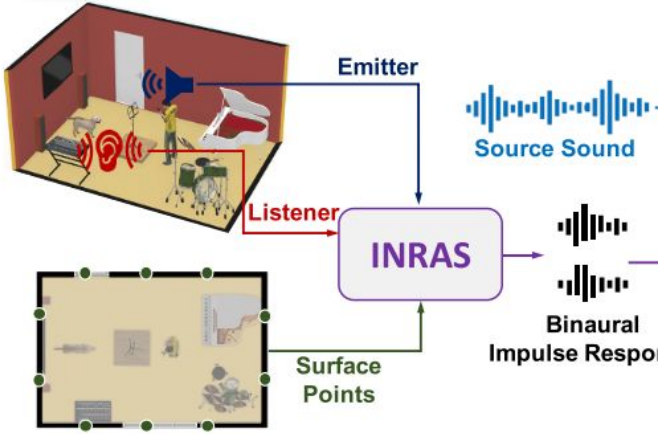

Figure 2: Overview of our NAF architecture where listener and emitter share a feature grid. Given a listener position and an emitter location, we first query a grid for local features which are learned together with the network during training. We compute the sinusoidal embedding of the positions, frequency, and time, and query a discrete embedding matrix using the orientation and left/right ear. Our method predicts magnitude and phase.



Luo, A., Du, Y., Tarr, M., Tenenbaum, J., Torralba, A., & Gan, C. (2022). Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, *35*, 3165-3177.

# INRAS [Su+2022]

## Implicit Neural Representation for Audio Scenes



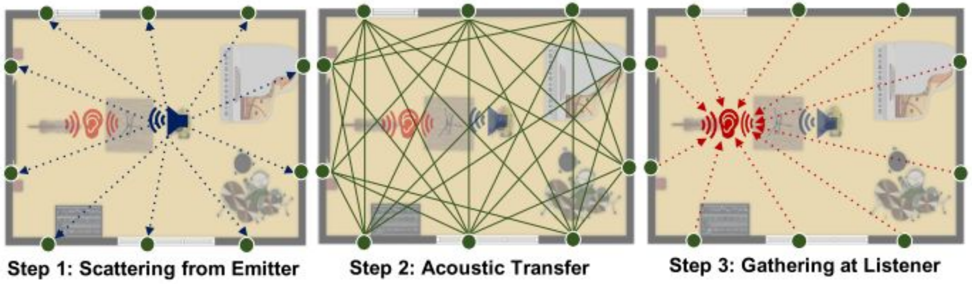




Figure 2: Acoustic radiance transfer steps overview.

Figure 1: INRAS learns an implicit neural representation for audio scenes such that given the geometry of a scene, emitter and listener positions, INRAS renders the sound perceived by the listener. See supplementary video of demonstration examples of spatial sound rendering.



Su, K., Chen, M., & Shlizerman, E. (2022). Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, *35*, 8144-8158.

# INRAS [Su+2022]

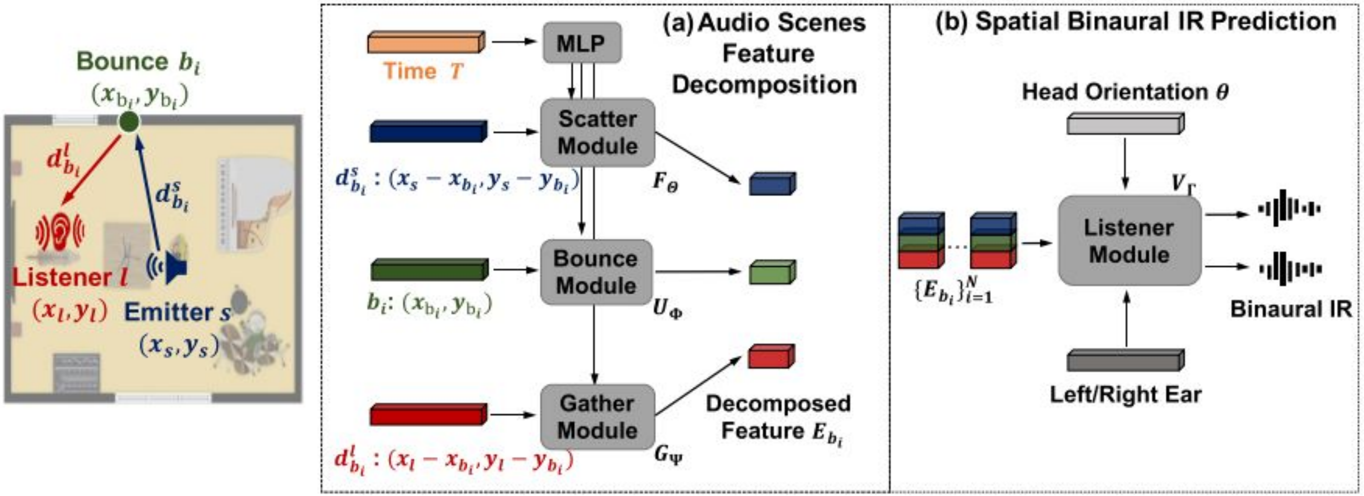



Figure 3: System overview of INRAS. (a) Audio Scenes Feature Decomposition: inputs to scatter/gather module are the relative distances between the emitter/listener locations and bounce points. The bounce module takes all bounce points to generate scene-dependent features. (b) Spatial Binaural IR Prediction: in this stage, the decomposed features are stacked and fed to the Listener module which generates the spatial binaural impulse responses.



Su, K., Chen, M., & Shlizerman, E. (2022). Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, *35*, 8144-8158.

# Takeaways for RIR generation

GAN-based methods

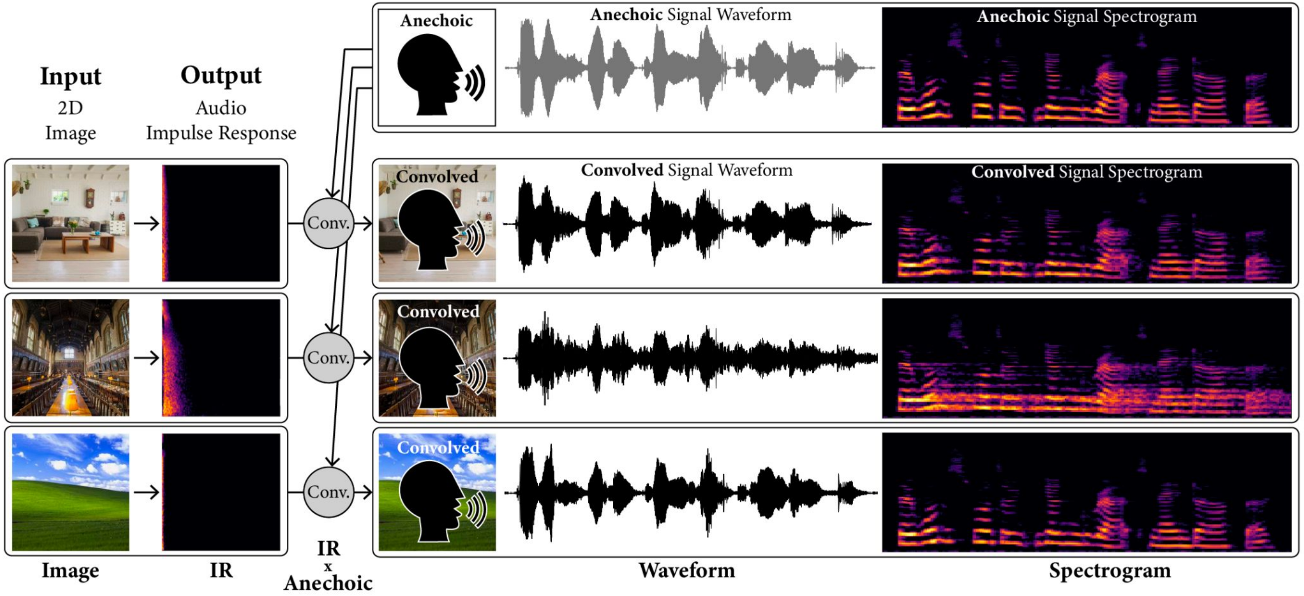
- Synthesized RIR can be used to augment the speech data for far-field ASR
- They are not designed for accurate spatialization

Neural-field based methods

- More accurate acoustic modeling
- Features can be decoded for acoustic scenes

# Cross-modal RIR Synthesis

Image2Reverb [Singh+2021]: Generate plausible audio IRs from single images of acoustic environments.






Singh, N., Mentch, J., Ng, J., Beveridge, M., & Drori, I. (2021). Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 286-295).

# Image2Reverb [Singh+2021]

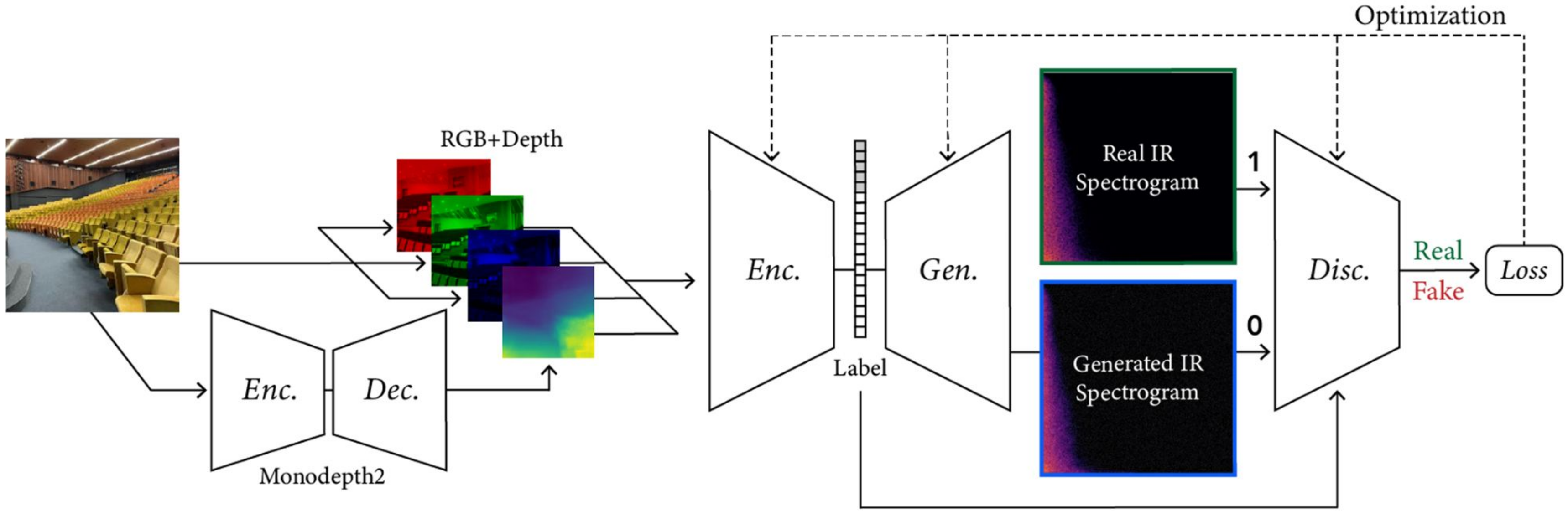



Figure 4. System architecture. Our system consists of autoencoder and GAN networks. Left: An input image is converted into 4 channels: red, green, blue and depth. The depth map is estimated by Monodepth2, a pre-trained encoder-decoder network. Right: Our model employs a conditional GAN. An image feature encoder is given the RGB and depth images and produces part of the Generator's latent vector which is then concatenated with noise. The Discriminator applies the image latent vector label at an intermediate stage via concatenation to make a conditional real/fake prediction, calculating loss and optimizing the Encoder, Generator, and Discriminator.


Singh, N., Mentch, J., Ng, J., Beveridge, M., & Drori, I. (2021). Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 286-295).

# Visual Acoustic Matching [Chen+2022]

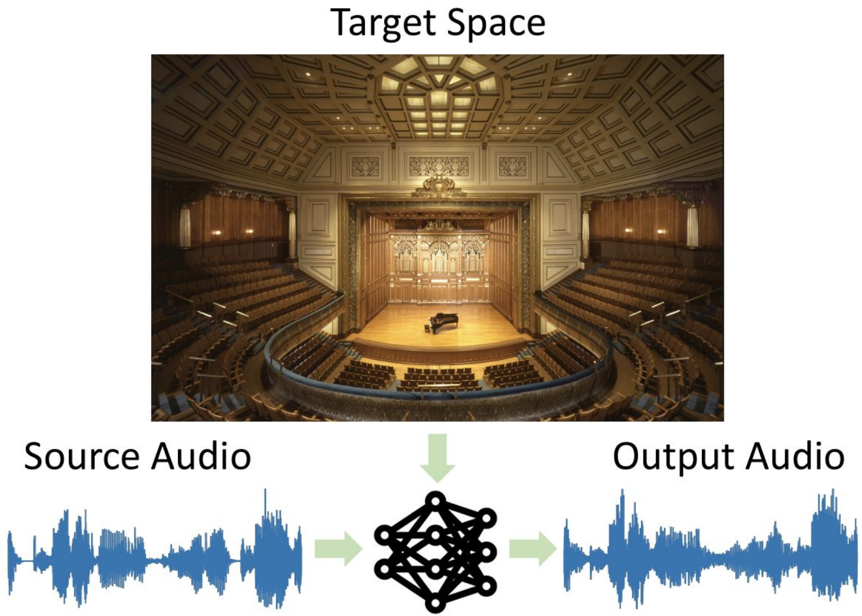



Figure 1. Goal of visual acoustic matching: transform the sound recorded in one space to another space depicted in the target visual scene. For example, given source audio recorded in a studio, re-synthesize that audio to match the room acoustics of a concert hall.


Chen, C., Gao, R., Calamia, P., & Grauman, K. (2022). Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18858-18868).

# Novel-View Acoustic Synthesis [Chen+2023]

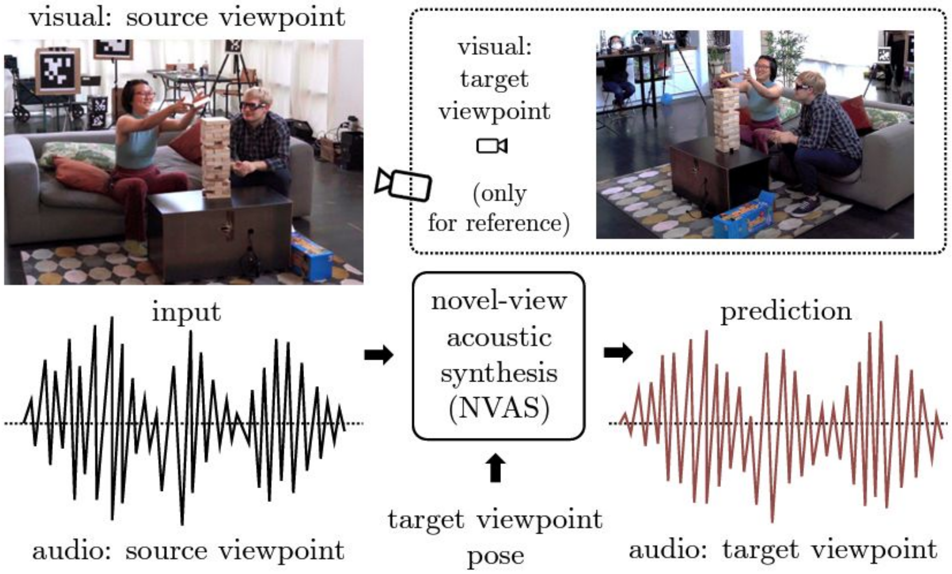
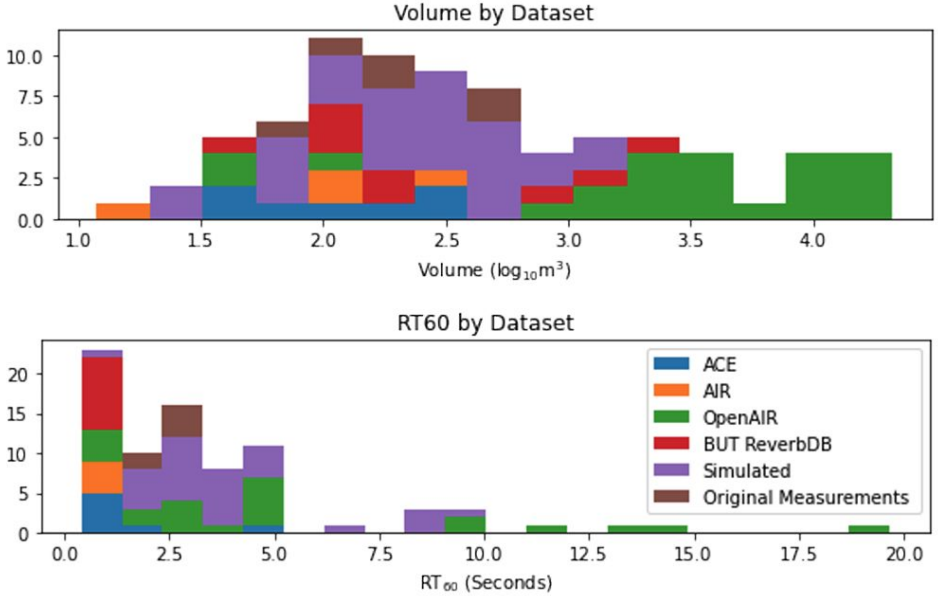


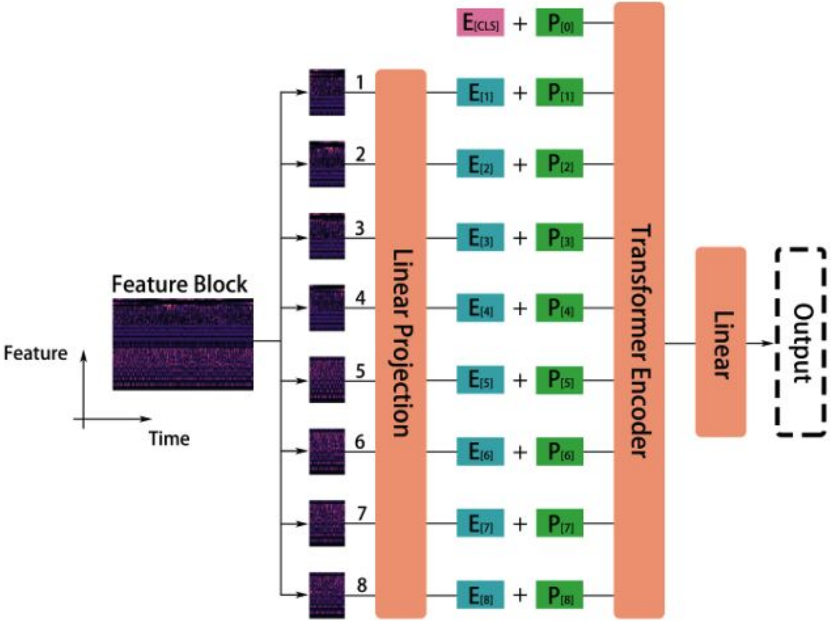

Figure 1. **Novel-view acoustic synthesis task.** Given audio-visual observations from one viewpoint and the relative target viewpoint pose, render the sound received at the target viewpoint. Note that the target is expressed as the desired pose of the microphones; the image at that pose (right) is neither observed nor synthesized.

Chen, C., Richard, A., Shapovalov, R., Ithapu, V. K., Neverova, N., Grauman, K., & Vedaldi, A. (2023). Novel-view acoustic synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6409-6419).

# Blind Room Parameter Estimation

Room acoustics parameters can be predicted given RIRs.







C. Ick, A. Mehrabi and W. Jin, "Blind Acoustic Room Parameter Estimation Using Phase Features," *Proc. ICASSP 2023*
C. Wang, M. Jia, M. Li, C. Bao and W. Jin, "Attention Is All You Need For Blind Room Volume Estimation," Proc. *ICASSP 2024*, pp. 1341-1345

# Spatial Audio

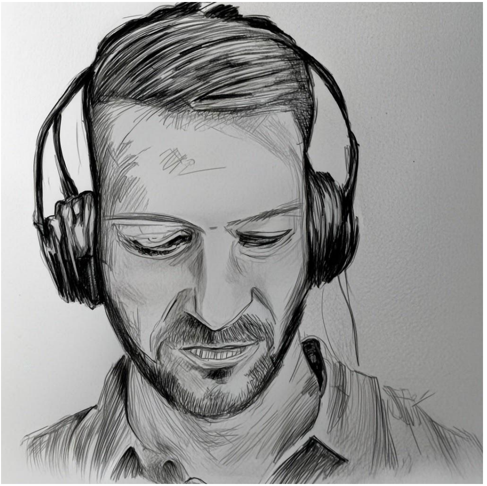
(Some slides adapted from AES AfG tutorial on personalized spatial audio)

# Immersive Audio Environment

Spatial Audio!

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.

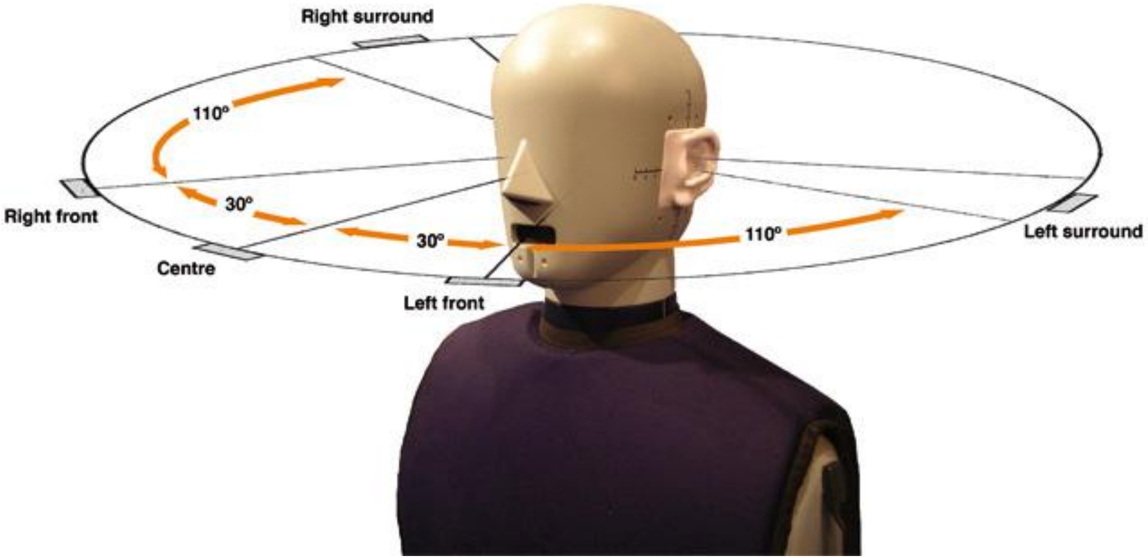



Figure from https://www.soundonsound.com/reviews/mp3-surround

# 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.

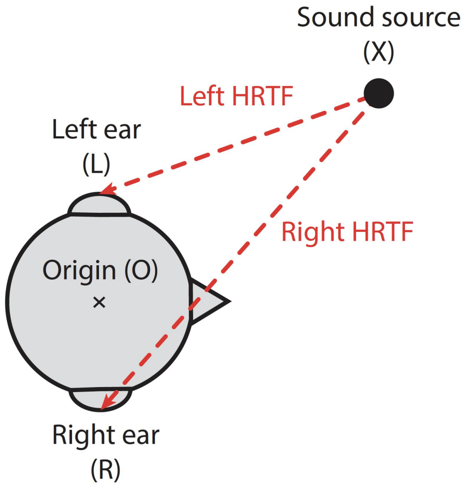



Figure from Isaac Engel's thesis

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

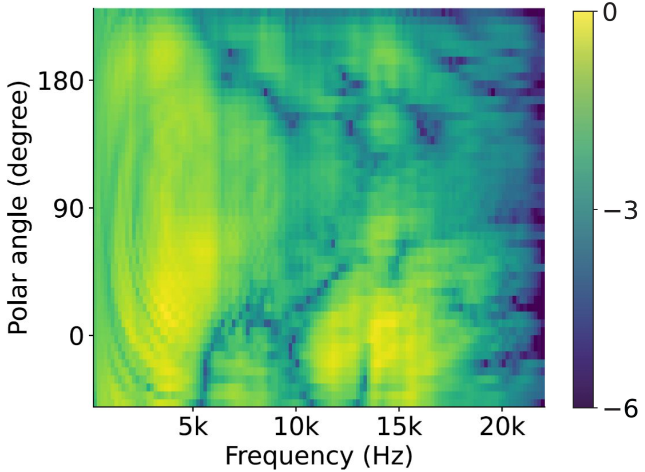


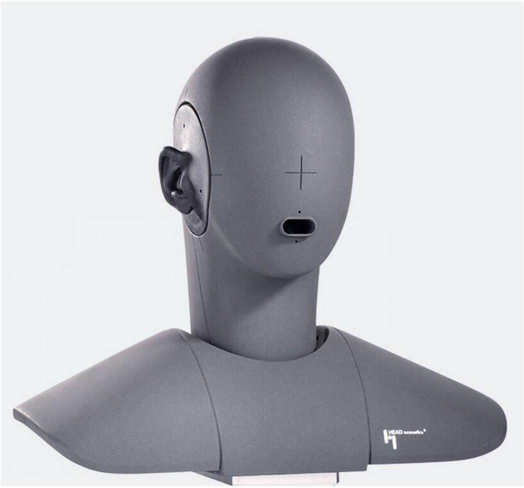

Figure from [Zhang+2023]

# 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


2:27 94%

360 Reality Audio Setup

Take a photo of your left ear

On the next screen, turn the front camera to your face. Shooting is performed automatically when your ears are detected.

1. Hold your mobile device with both hands in front of your face.
2. When your face is detected, turn your head to the right following the voice guidance.
3. Stand still until shooting is performed automatically

TAKE A PHOTO

## 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.

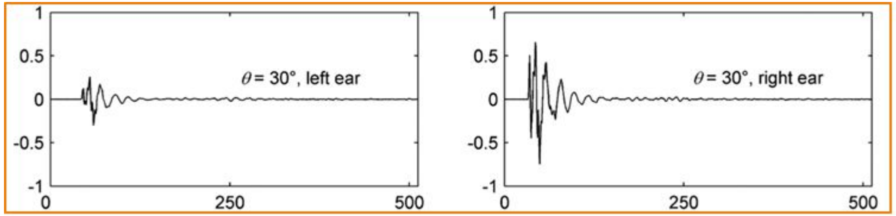
# Measure Personalized HRTFs

- 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

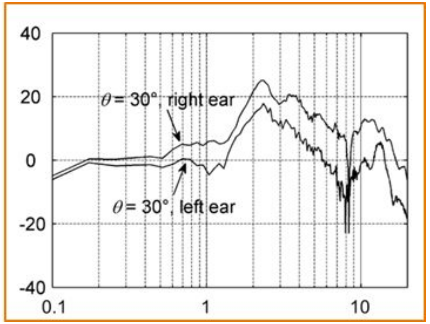
Figure from https://ieeexplore.ieee.org/document/7099223

# 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]

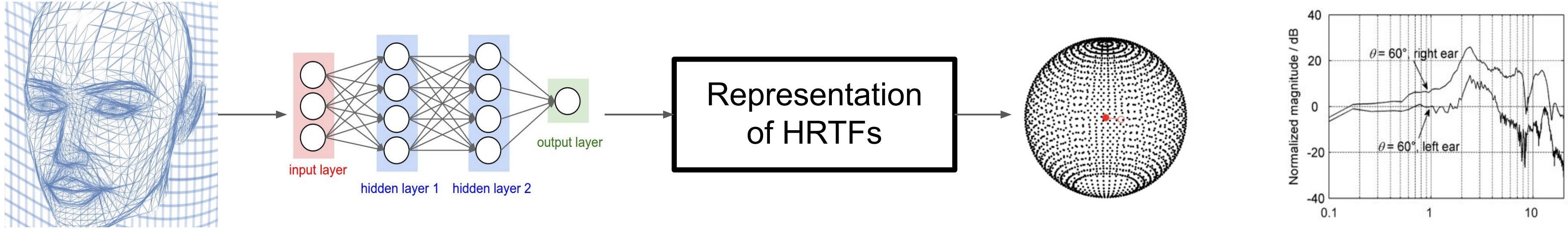
Drawbacks:

- 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

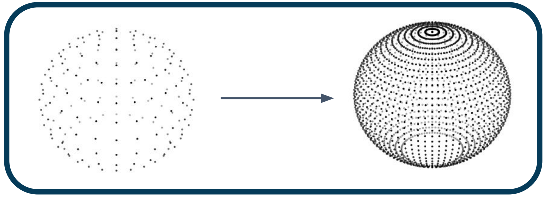
**Assumption**: Many things are common across people (captured by the model), and other effects are personalized (captured by adapting the input).

# Personalized HRTF Modeling

**Two research tasks:**

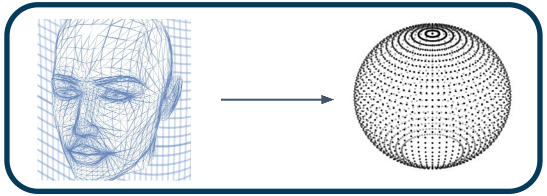
- HRTF Upsampling / Interpolation

  (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

predicted linear-scale magnitude

# spatial locations

# frequency bins

frequency index

# Evaluation Metric (Cont'd)

Subjective evaluation

- Auditory models

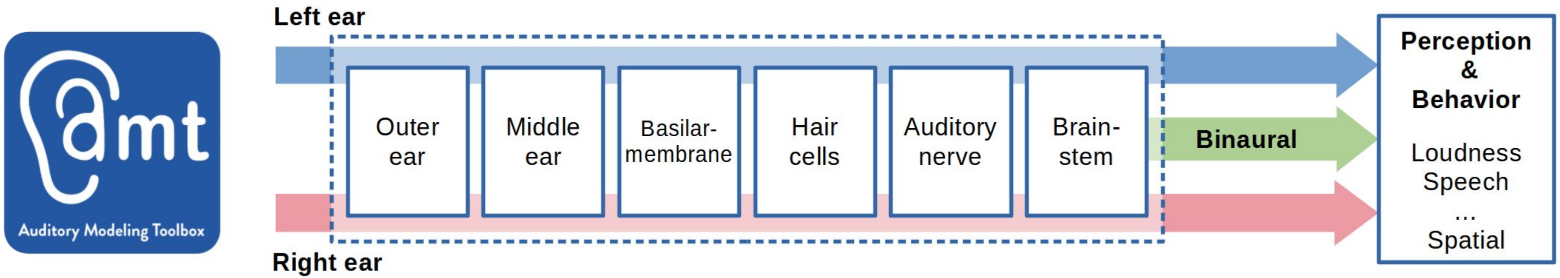



Figure from https://amtoolbox.org/

- 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]

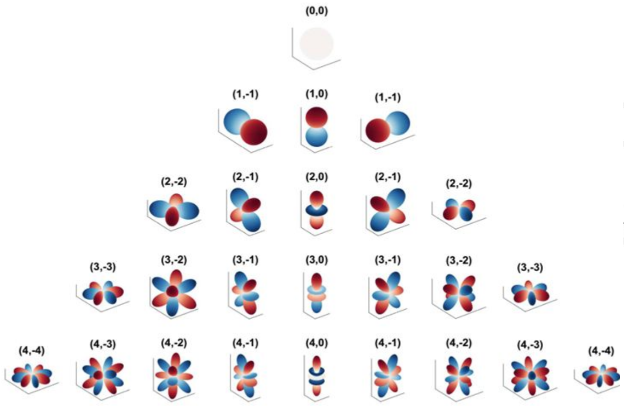



Figure from [Wang+2020]

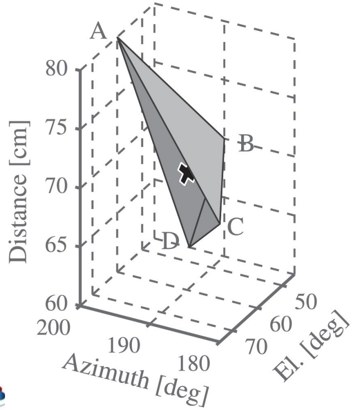



Figure from [Gamper2013]


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

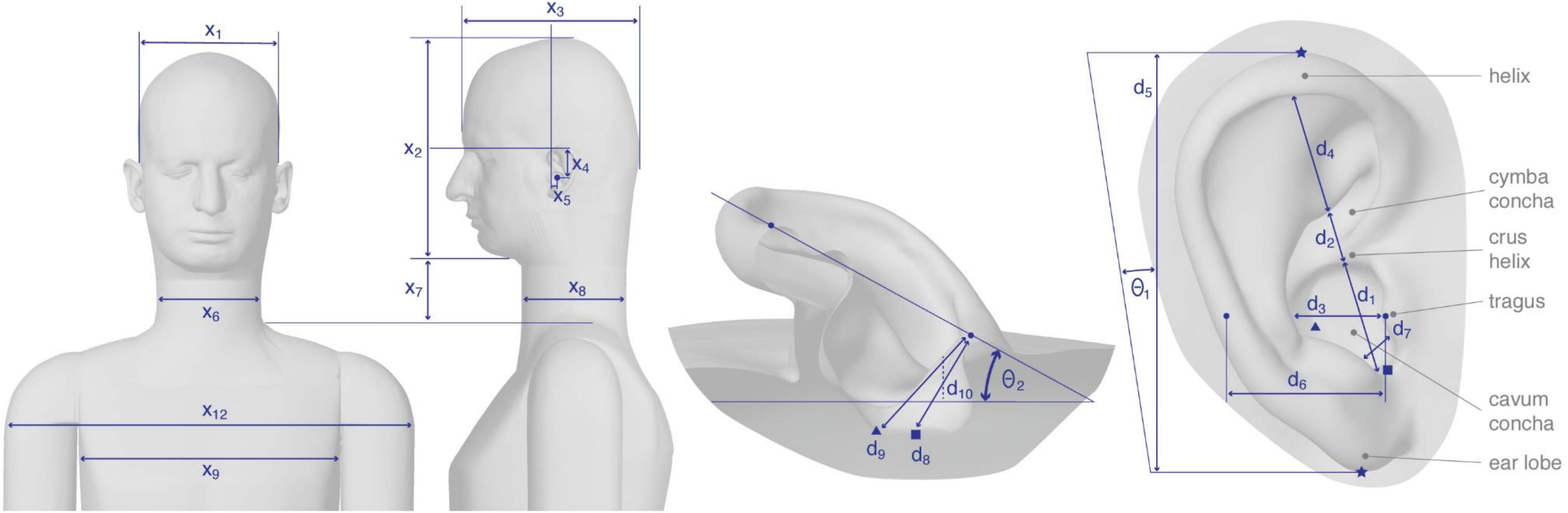
Use datasets to train machine learning models to capture the prior

- 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

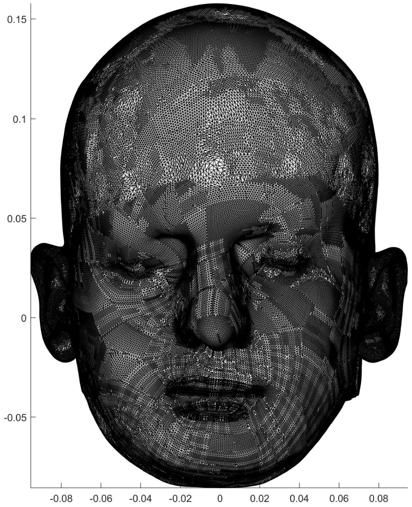



Figure from [Wang+2022]

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]

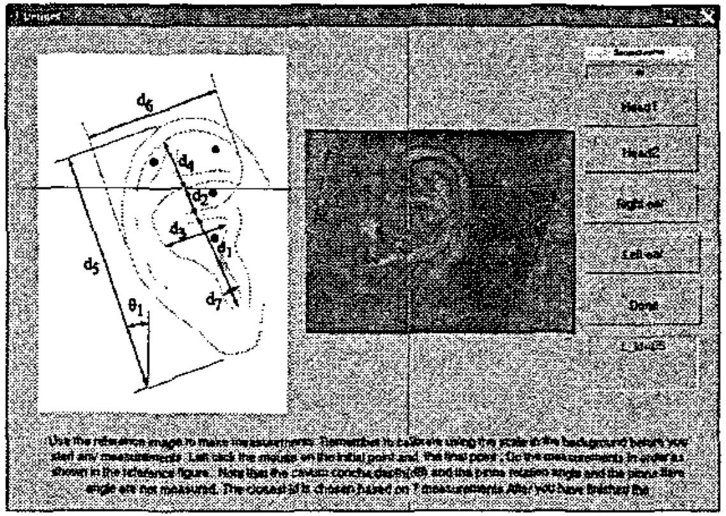



Figure from [Zotkin+2003]


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

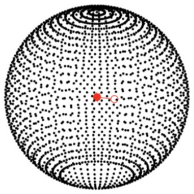
- Principal component analysis (PCA) [Hu+2008]
- Deep neural network (DNN) [Chun+2017]
- Autoencoder [Chen+2019]
- Variational Autoencoder (VAE) [Miccini&Spagnol2020]
- Spatial principal component analysis (SPCA) [Zhang+2020]
- Spherical harmonics transform (SHT) [Wang+2020]

*Can handle arbitrary directions!*

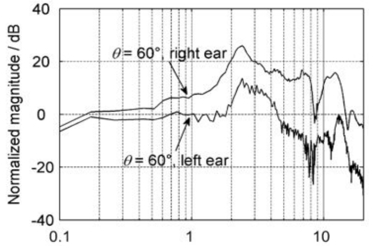
Hu, Hongmei, et al. "HRTF personalization based on artificial neural network in individual virtual auditory space." *Applied Acoustics* 2008.
Chun, Chan Jun, et al. "Deep neural network based HRTF personalization using anthropometric measurements." *AES Convention* 2017.
Chen, Tzu-Yu, Tzu-Hsuan Kuo, 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.

# 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?*

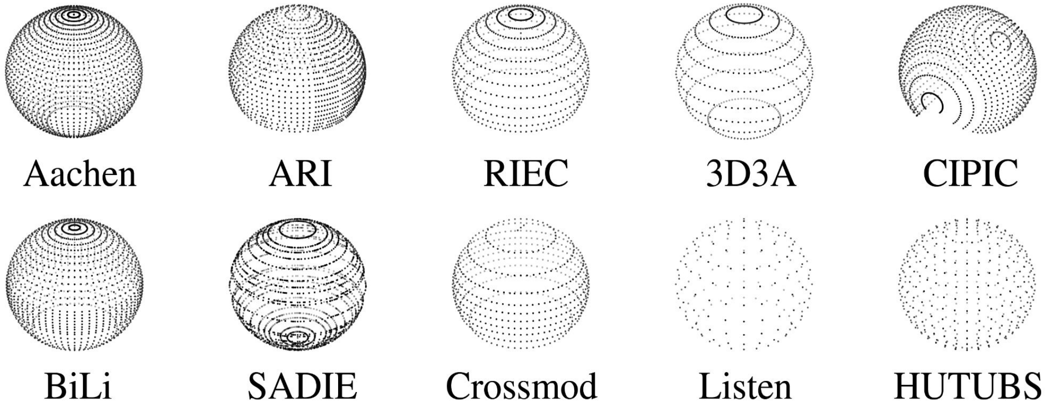
Most of the work trains and evaluates the model on the same database, and it is hard to tell the generalization ability.

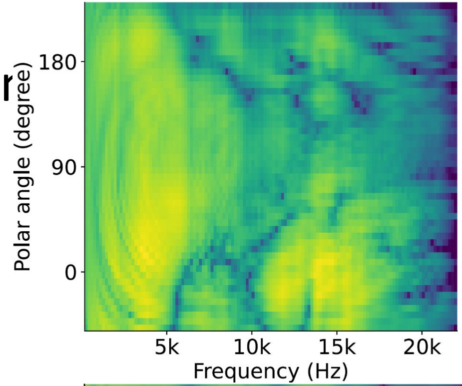
- Leave-one-out validation
- Cross-validation

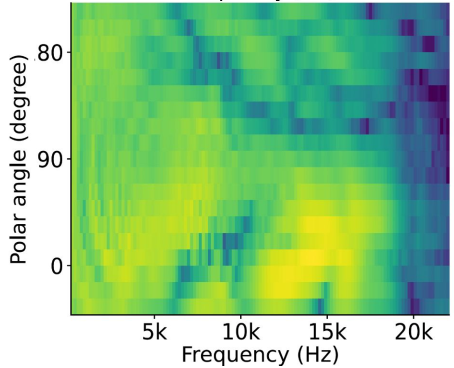
*Open question: Can we merge the existing datasets? If so, how?*

# Challenge2: Spatial Sampling Schemes

The source location grids used in HRTF databases differ fror making cross-dataset learning difficult.



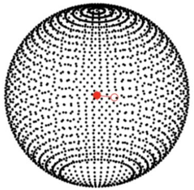




Figures from [Zhang+2023]

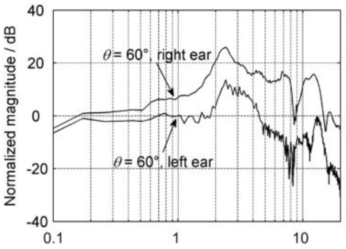
Zhang, You, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.

# 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

azimuth angle → $\theta$

elevation angle → $\phi$

SIREN

Magnitude Spectrum

# frequency bins

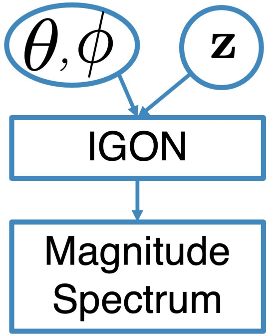
$$f : \mathbb{R}^2 \rightarrow \mathbb{R}^K$$

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



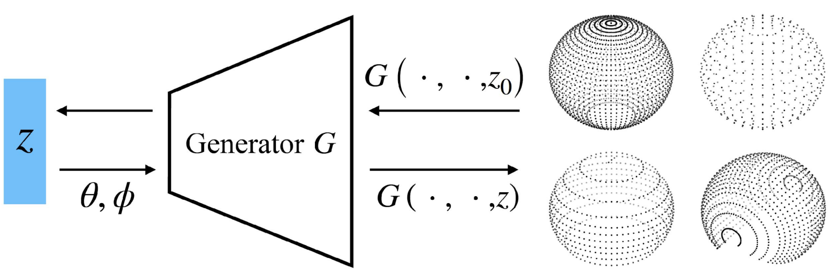

latent code for a subject

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

# frequency bins

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

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



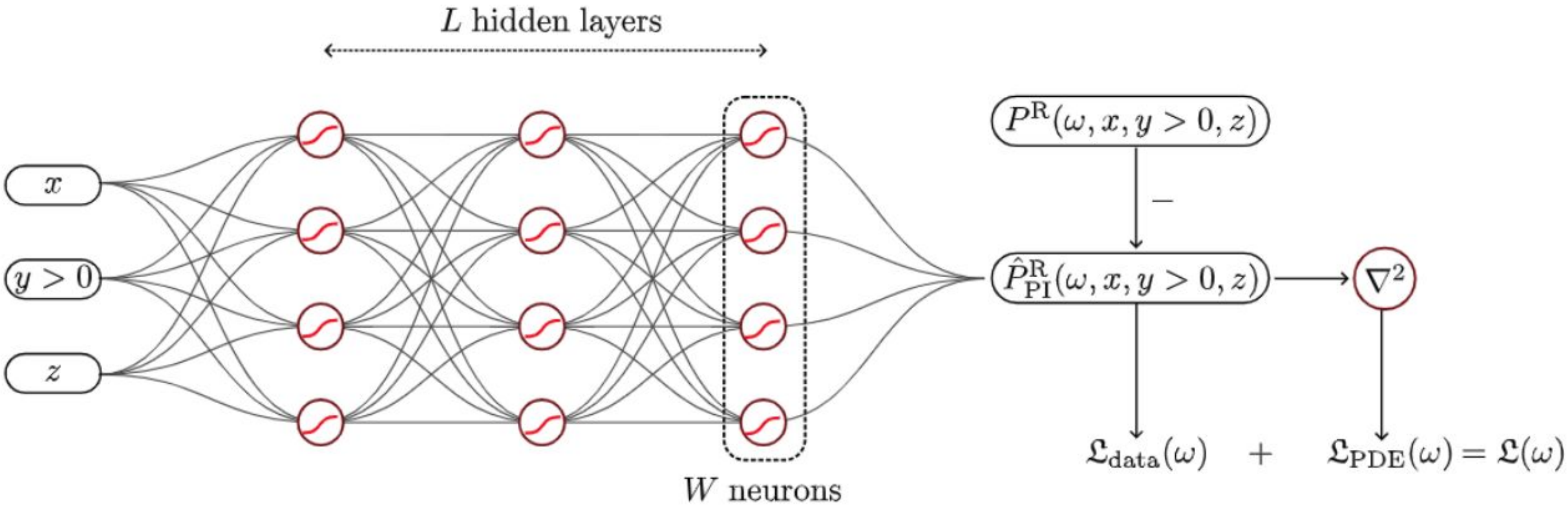

$\mathcal{L} = \mathcal{L}_{\mathrm{MSE}} \left(\mathbf{x}, G\left(\cdot, \cdot, \mathbf{z}\right)\right)$

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.

# Direction1: Regularize the Model with Priors

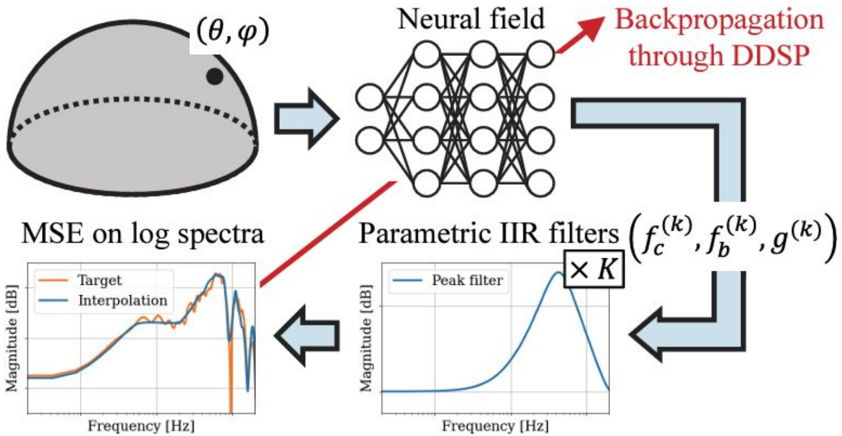
Physics prior: Physics-informed neural network for spatial upsampling
[Ma+2023]

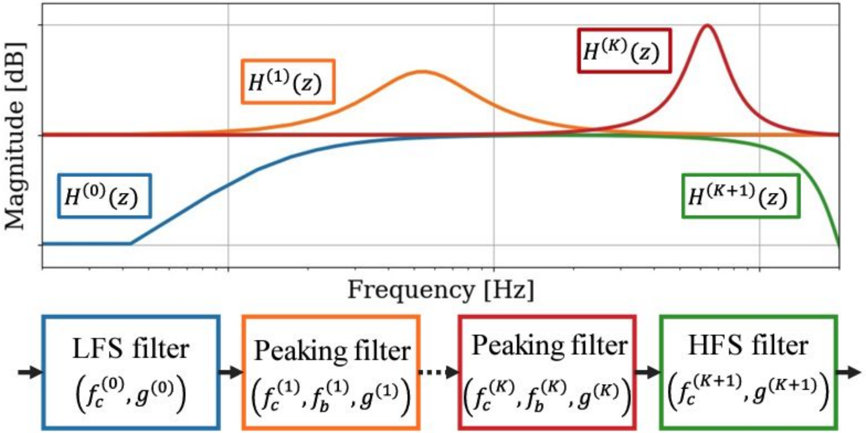



Ma, Fei, et al. "Physics informed neural network for head-related transfer function upsampling." *arXiv* 2023.

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

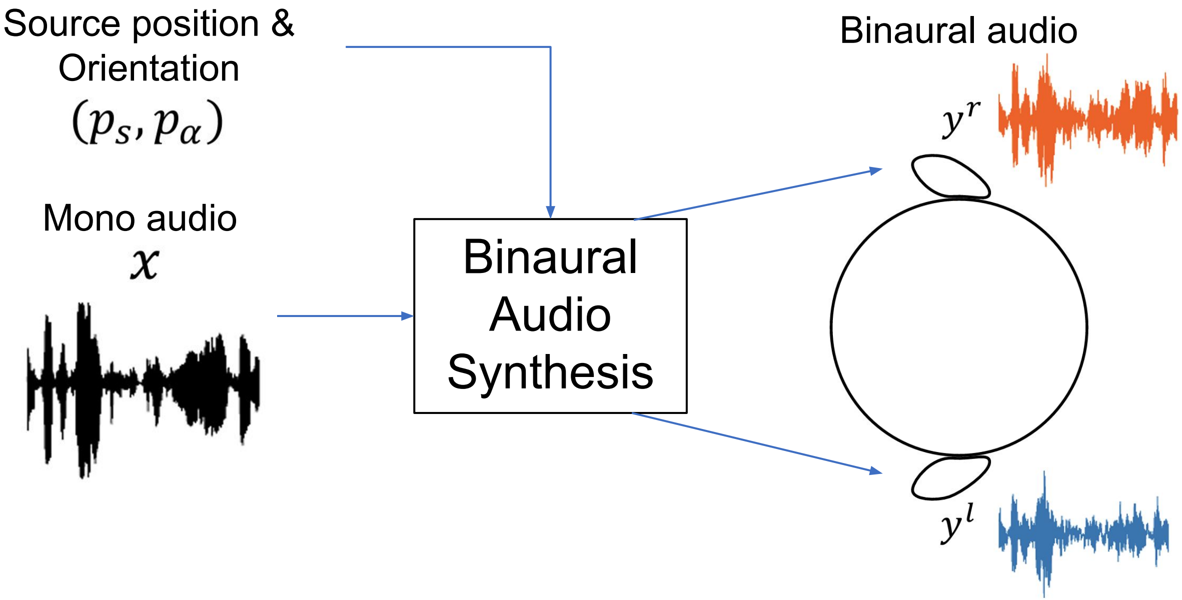
DSP prior: Model HRTF as IIR filters -- Neural IIR filter field (NIIRF) [Yoshiki+2024]




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

# Direction2: Binaural Audio Synthesis




Existing methods:

WarpNet [Richard+2020]

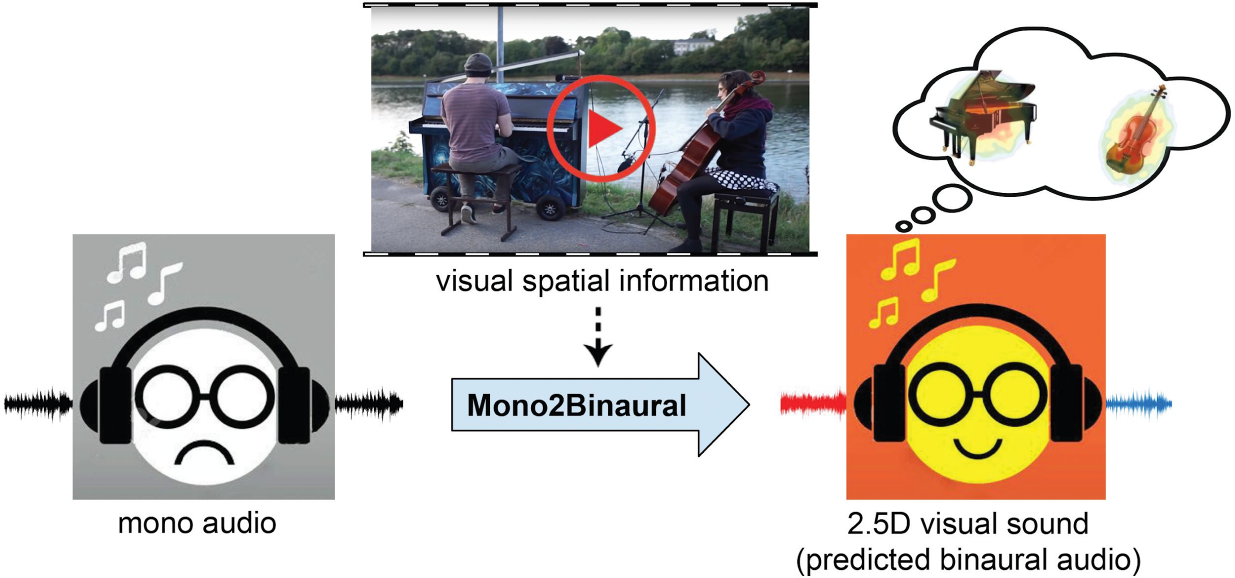
BinauralGrad [Leng+2022]

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.

# Direction2: 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.

# Takeaway messages

- Machine learning methods have been evolving quite a lot for solving room acoustics and spatial audio problems.
- Important problems include:
  - Room impulse response generation
  - Acoustic parameters estimation
  - Personalized HRTF modeling
  - Binaural audio synthesis
  - Cross-modal acoustics synthesis

Thank you! Questions?