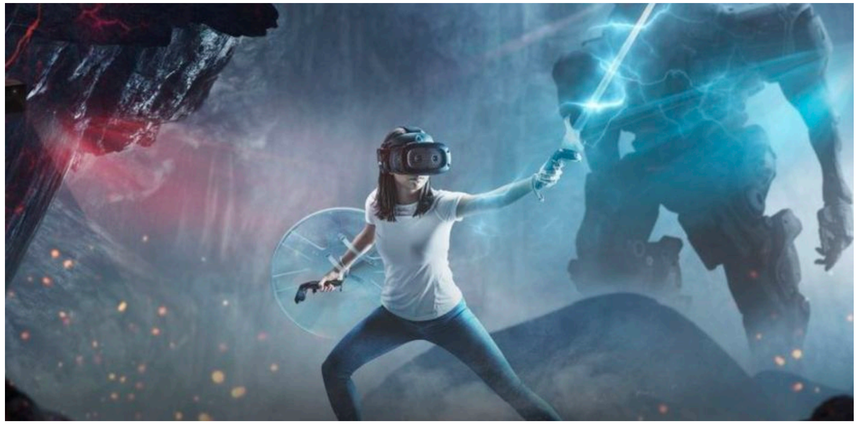
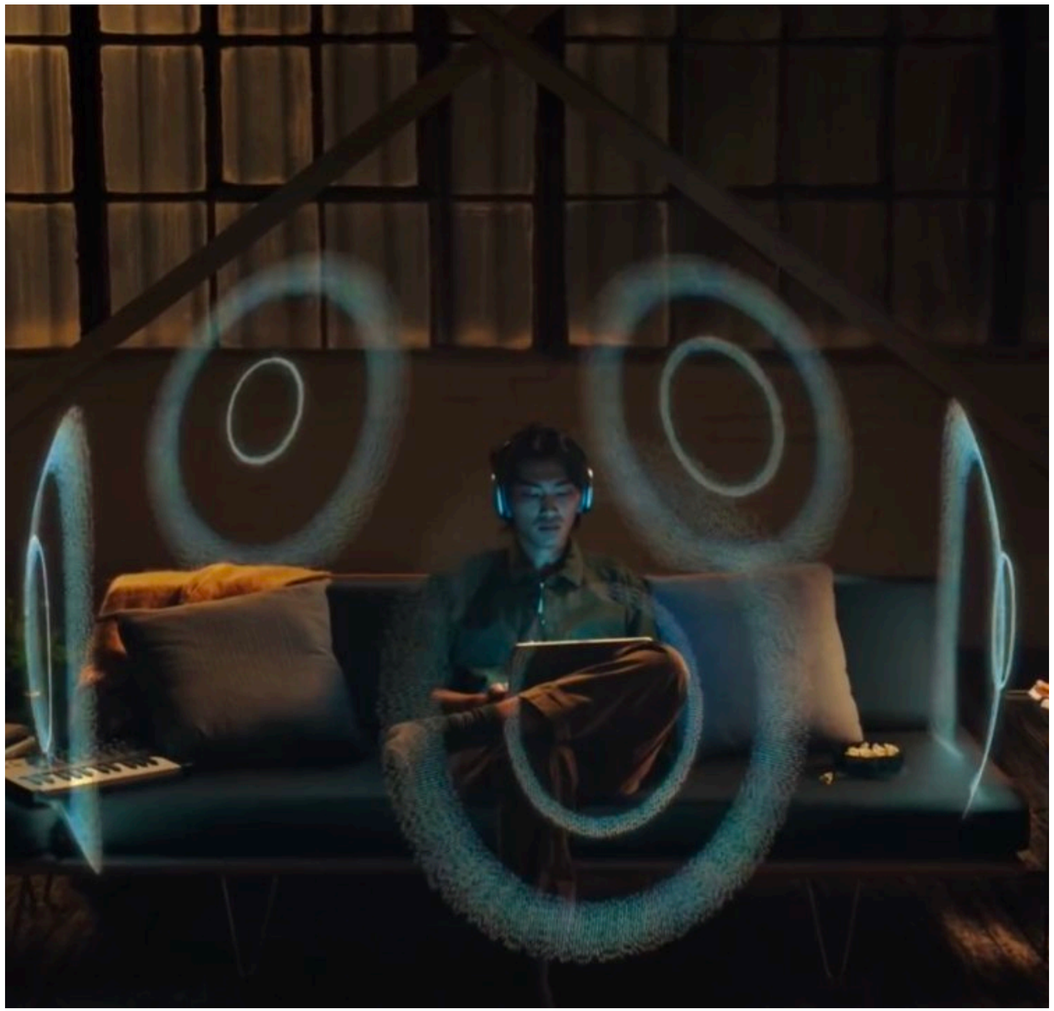
HRTF Estimation in the Wild

Vivek Jayaram, Ira Kemelmacher-Shlizerman, Steven M. Seitz
University of Washington

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

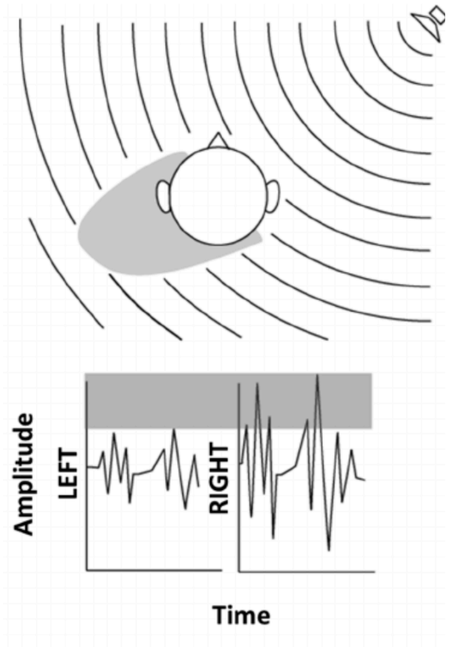
Spatial Audio is Important for Mixed Reality





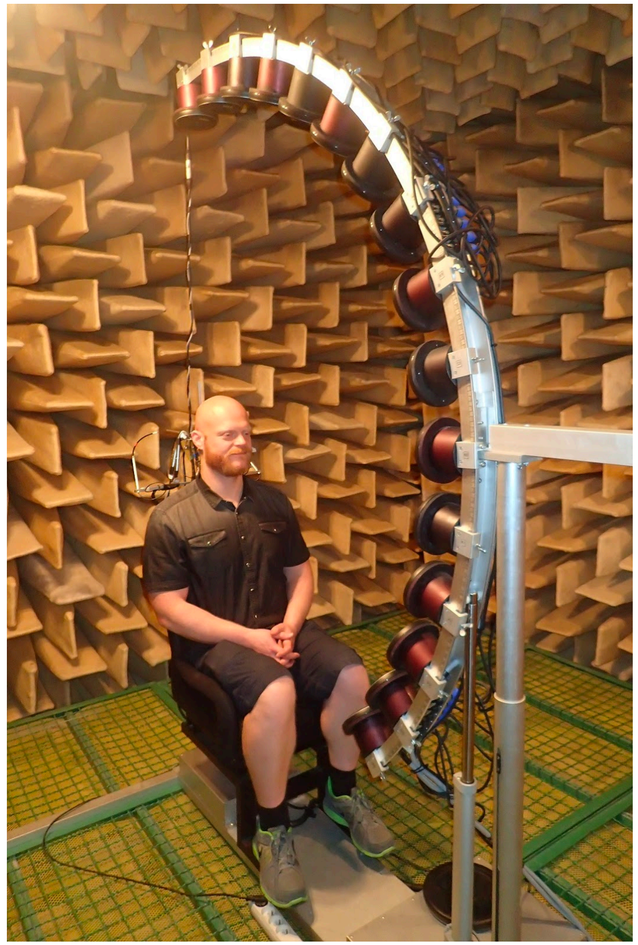
Head-Related Transfer Function Gives Directionality

- HRTF - Direction dependent filtering of sound by ears and head
- What frequencies are filtered, by how much
- Highly personalized and difficult to measure
- Render content in headphones with that user’s HRTF
- Need an easy way to measure individual HRTF

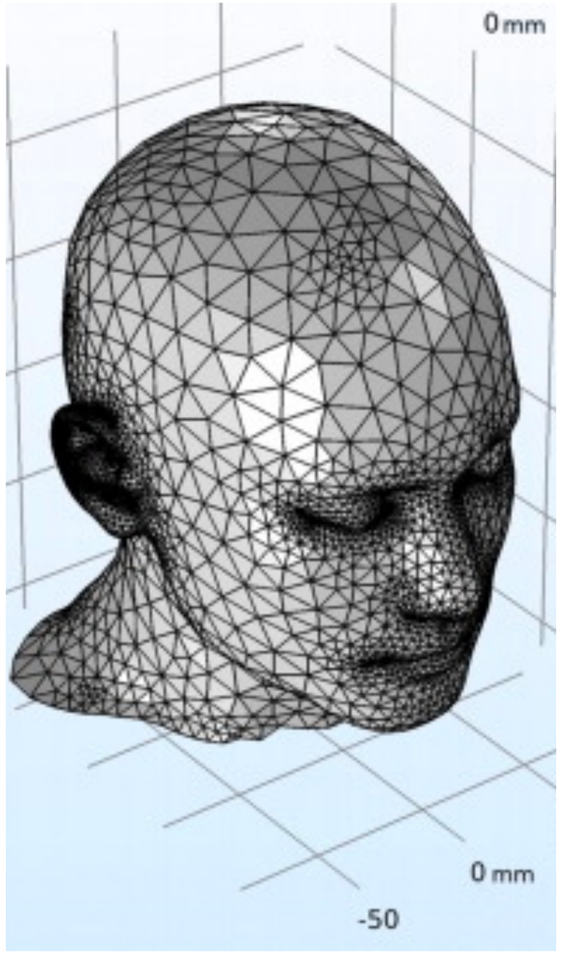


erence (ILD).

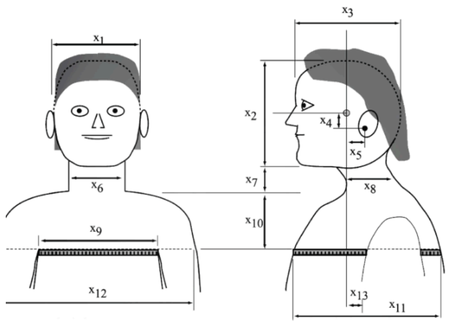
Existing Methods of HRTF Estimation Involve Complex Measurements

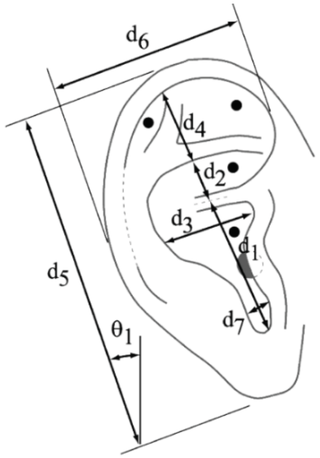


Anechoic Chamber Recordings



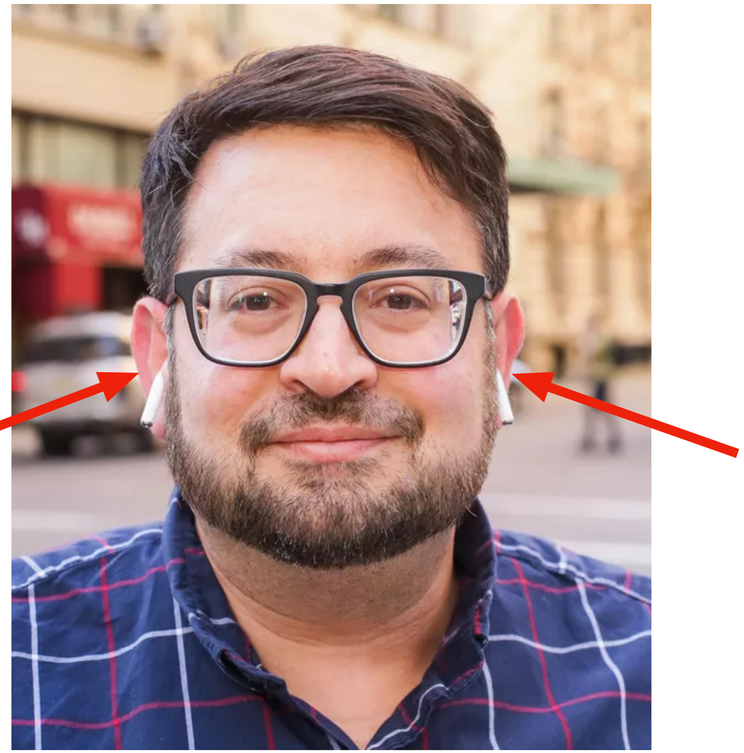
3D Head Scans and Meshes

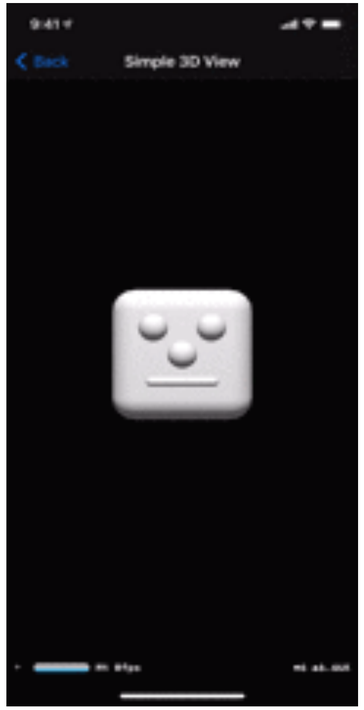




Anthropometric Measurements

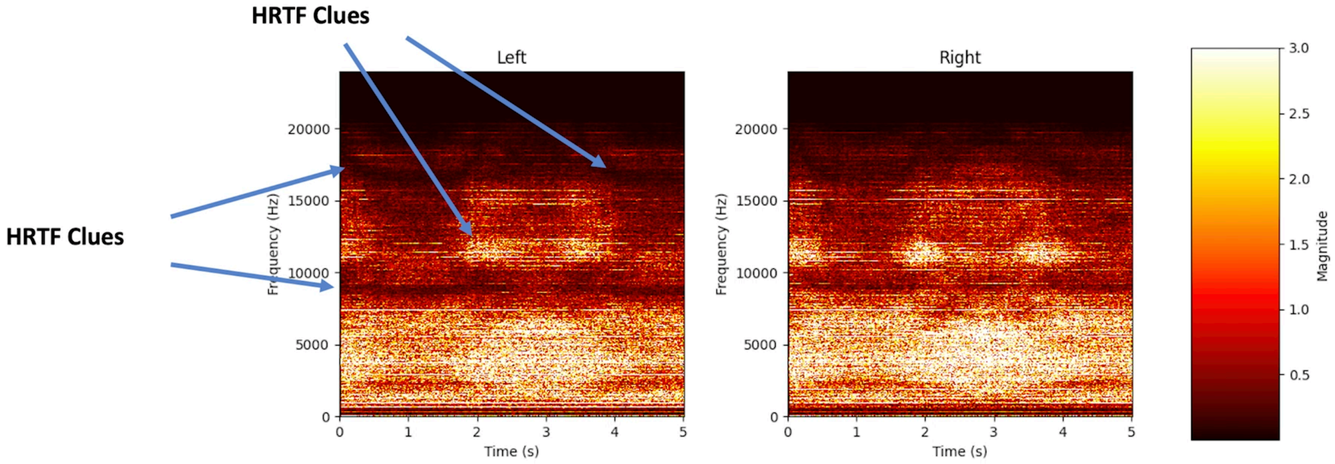
Observation 1: Millions of people wearing earbuds with microphones and gyroscopes





3D head tracking from gyroscopes

Observation 2: Sound captured as you rotate your head provides clues about your HRTF



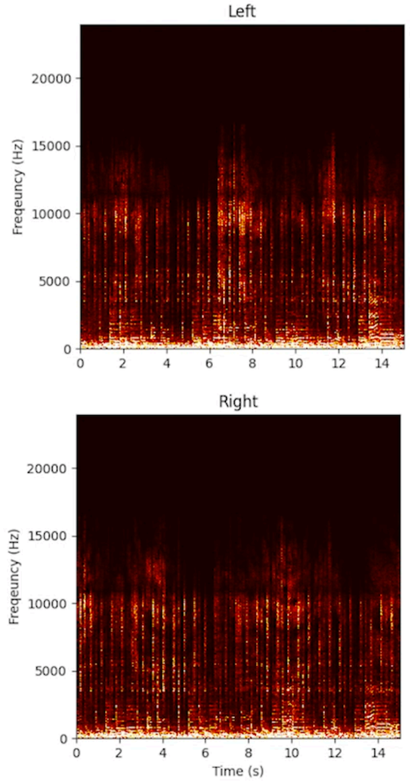
Can we estimate your HRTF as you move your head around in everyday environments?

Our Method

Method Overview

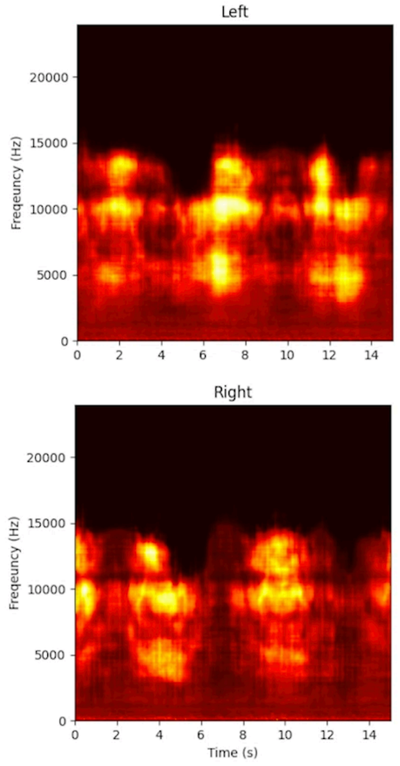
Our Method of HRTF Estimation

What Does HRTF Prediction Look Like



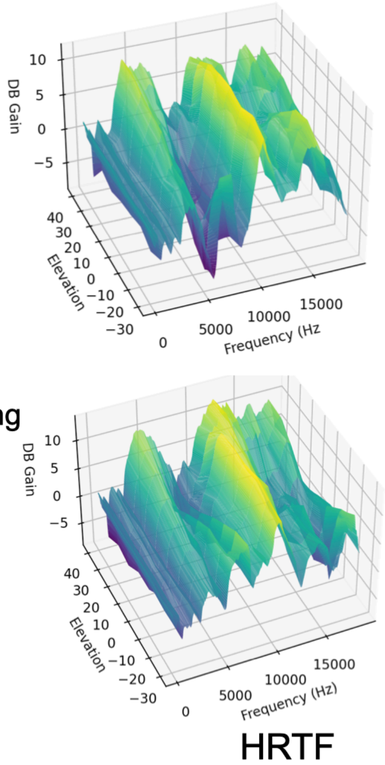
Binaural Recording of Ambient Music (R)

UNet



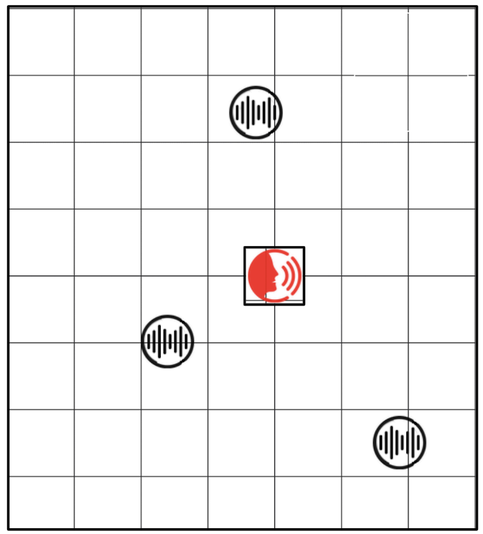
Predicted Filtering ($\widehat{H}$

Head Tracking



HRTF

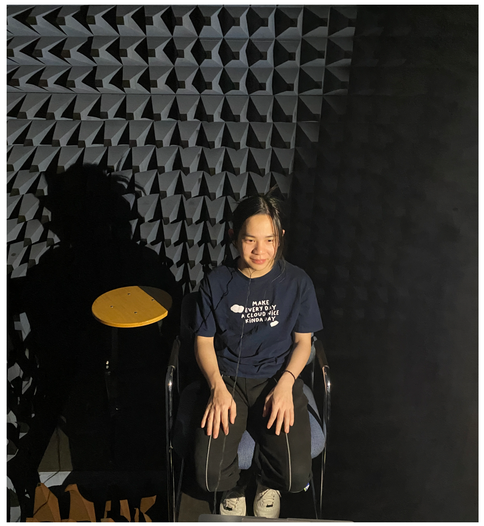
Training the Network



Spatially Rendered Data

Easiest to Collect

Least Representative



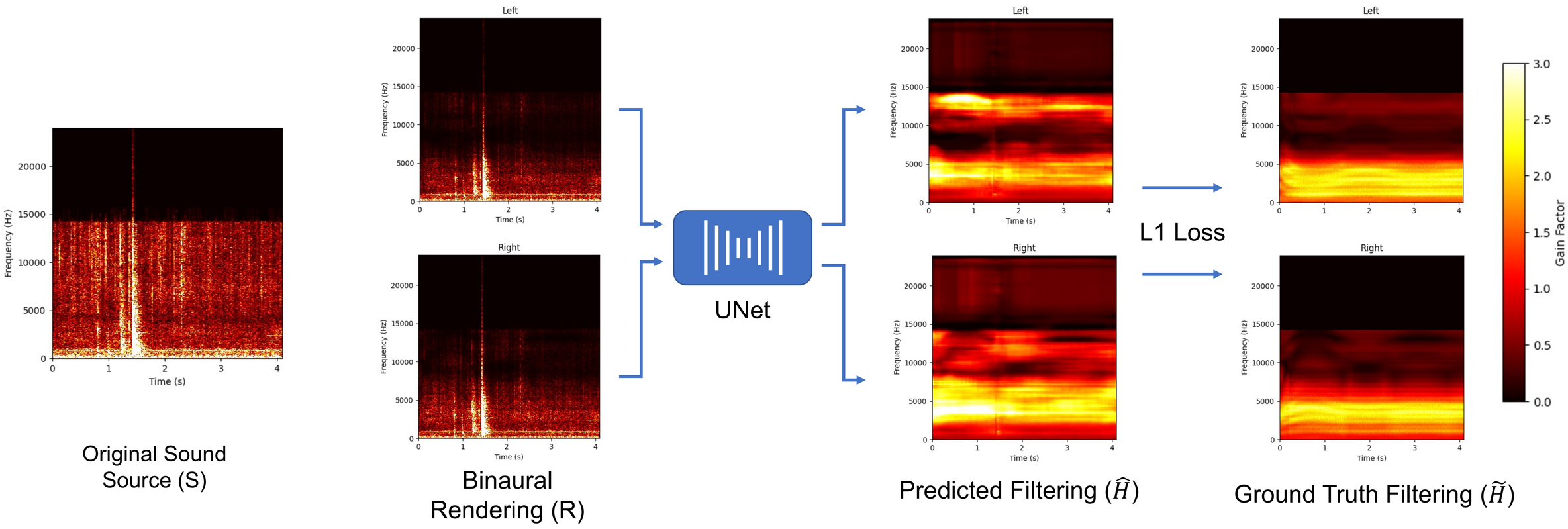
Ground Truth HRTF + In-the-wild Recordings

Hardest to Collect

Most Representative

Anechoic Chamber Needed for GT

Training the Network

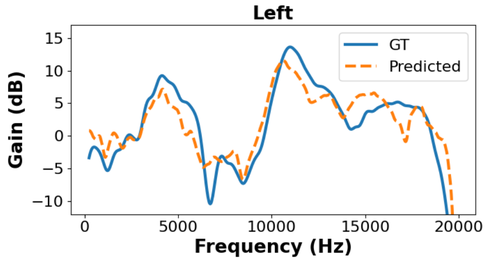


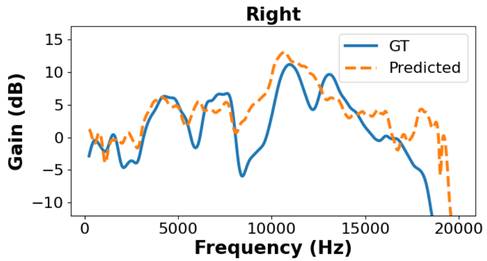
Results

User Study

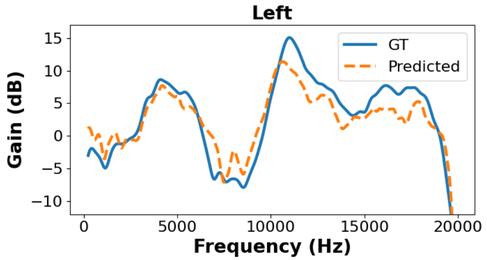
- 8 Users (unseen during training) in regular indoor environments (background noise level ~50db)
- 15 minutes of sounds from AudioSet played over a speaker as users rotated their heads through various positions
- Ground Truth HRTF also collected in anechoic chamber
- Quantitative and qualitative comparisons

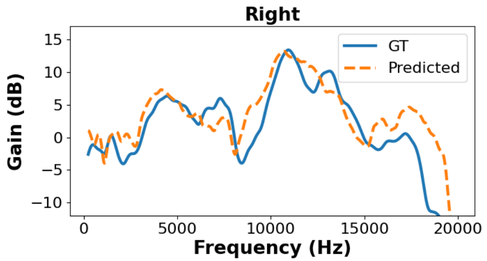
Result 1 - How well does our HRTF match the GT



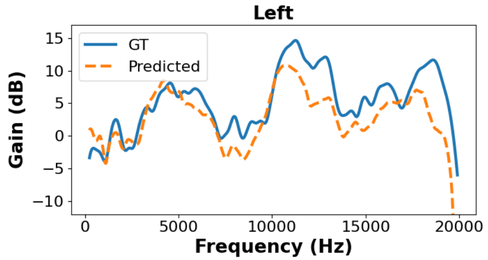


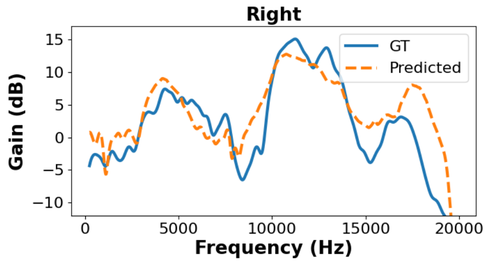
φ = 0°



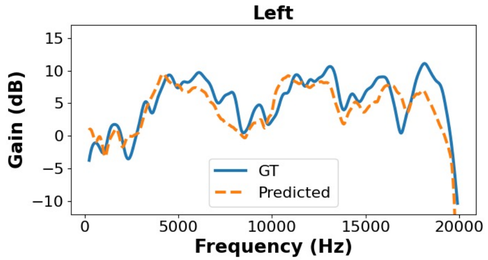


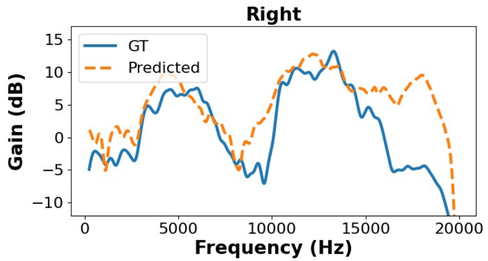
φ = 10°





φ = 20°





φ = 30°

Method	LSD (dB)
Random RIEC Subject	8.23
Generic HRTF	7.32
Zandi et. al [50]	4.5
Ours	**4.38**
Hu et. al [15]	3.5

(Lower is Better)

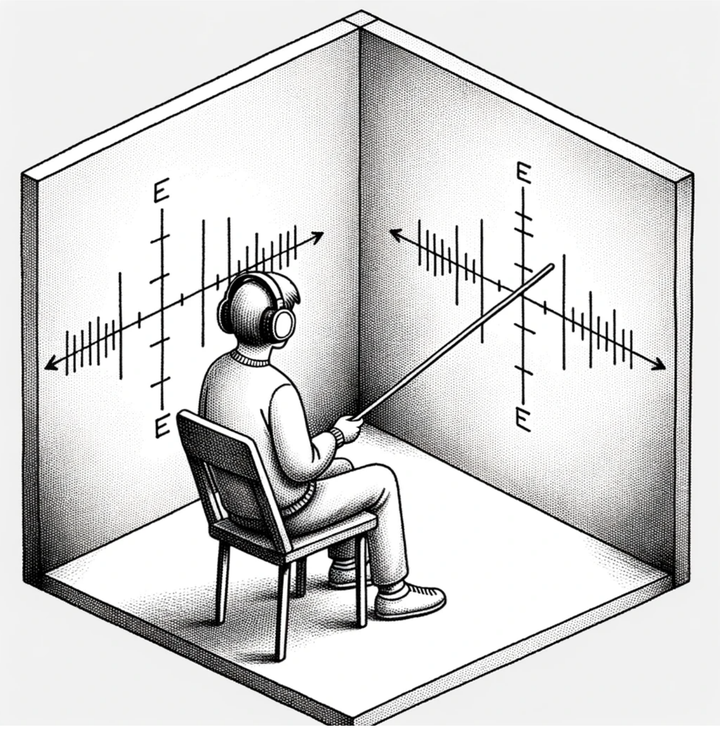
Result - Front Back Confusion With Virtual Sound

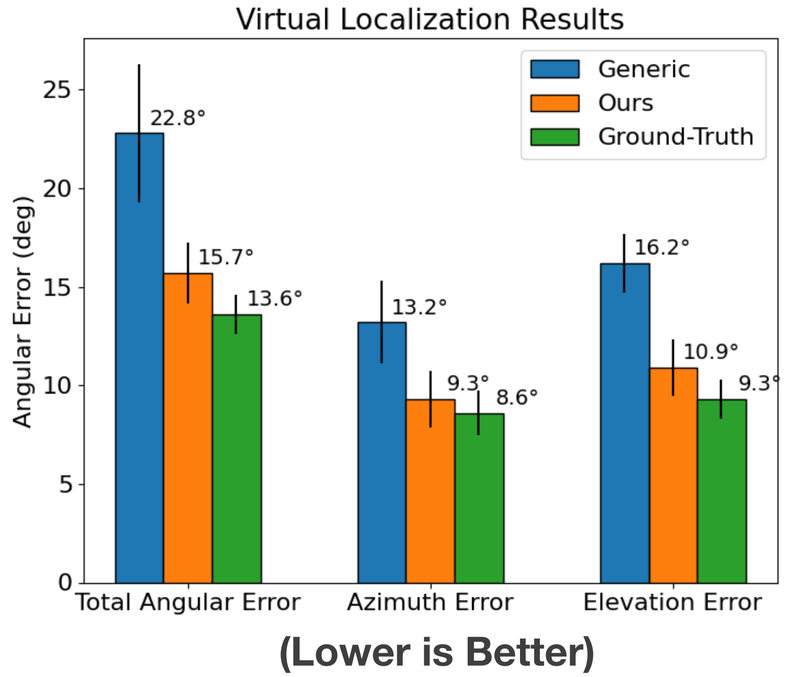
- Render a spatial sound through headphones at a random location
  - User predicts front or back

Method	Front-back confusion rate
Generic	29.0% ± 5.4
Ours	**14.8% ± 4.6**
GT HRTF	9.6% ± 4.2

Result 3 - Localization in Virtual Auditory Display

- User asked to point to direction of virtual sound





Future Work / Improvements

- Predict Interaural Time Differences as well as Level Differences
- Use multiple, non-stationary non sources
- Reduce the amount of recording time to produce a good estimate

Thank You!