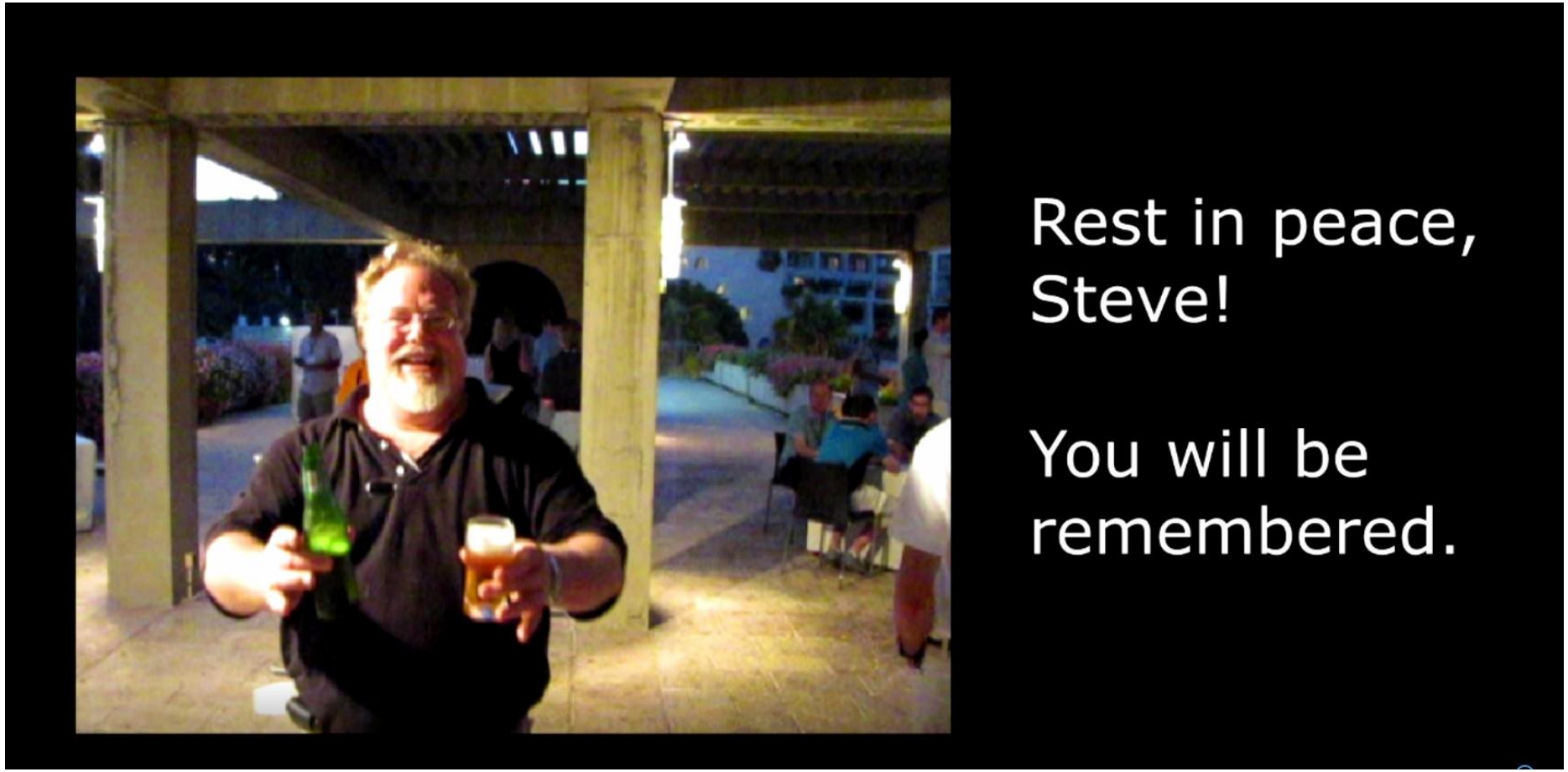
Audio challenges in virtual and augmented reality devices

Dr. Ivan Tashev
Partner Software Architect
Audio and Acoustics Research Group
Microsoft Research Labs - Redmond

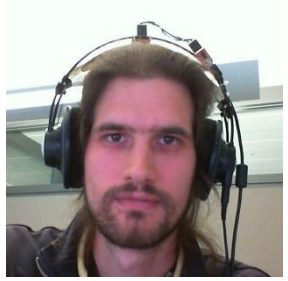
In memoriam: Steven L. Grant



In this talk

- Devices for virtual and augmented reality
- Binaural recording and playback
- Head Related Functions and their personalization
- Object-based rendering of spatial audio
- Modal-based rendering of spatial audio
- Conclusions

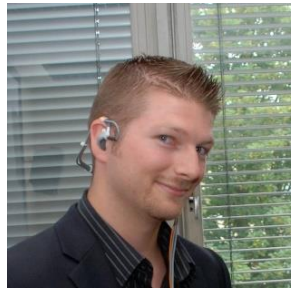
Colleagues and contributors:



Hannes Gamper
Microsoft Research

David Johnston
Microsoft Research

Ivan Tashev
Microsoft Research



Mark R. P. Thomas
Dolby Laboratories

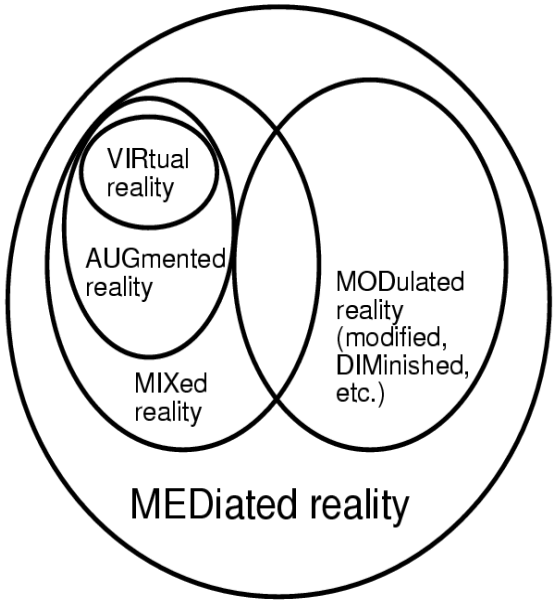
Jens Ahrens
Chalmers University, Sweden

Devices for Augmented and Virtual Reality

They both need good spatial audio

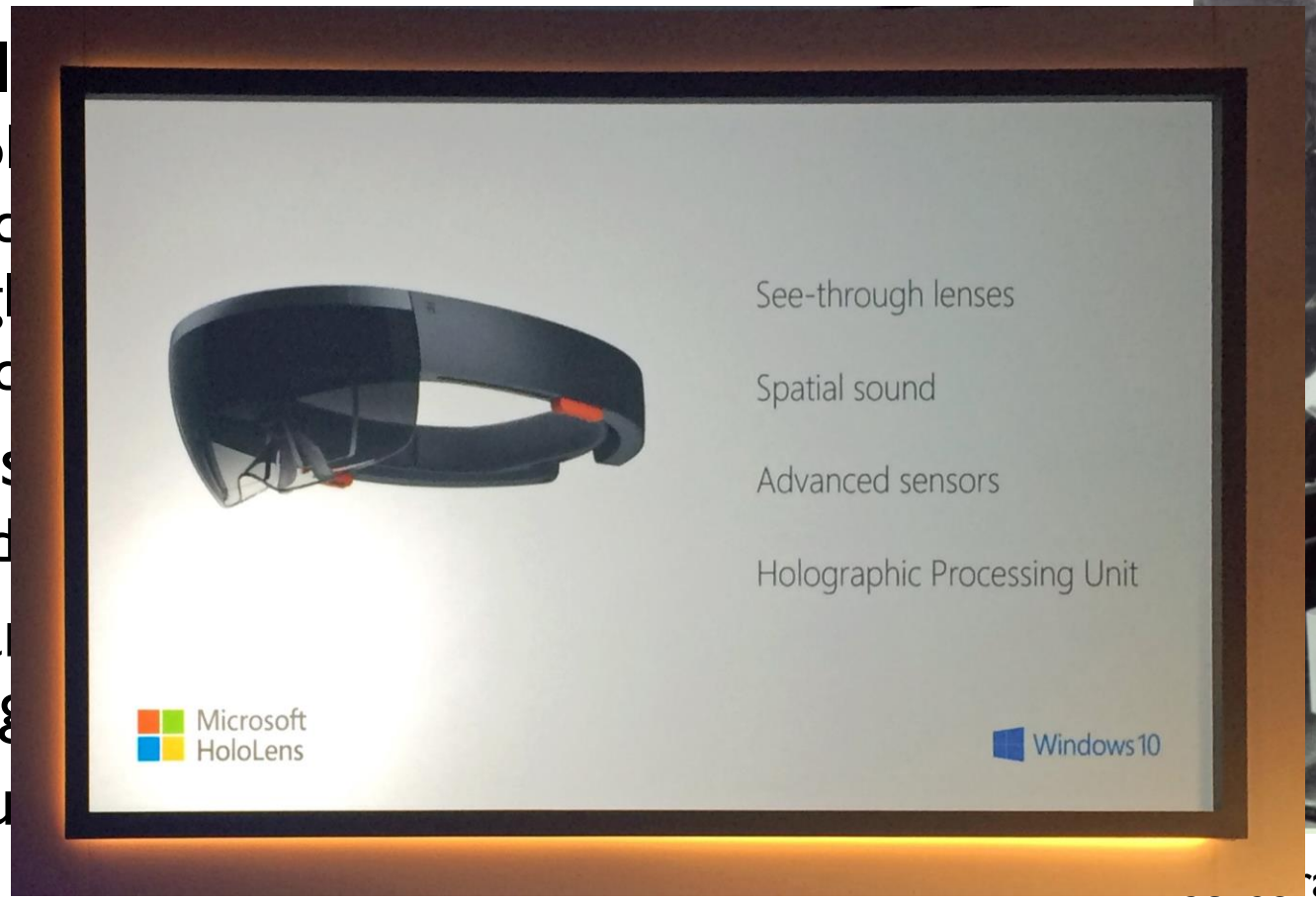
Augmented vs. Virtual Reality

- **Augmented reality (AR)** is a live direct or indirect view of a physical, real-world environment whose elements are augmented (or supplemented) by computer-generated sensory input such as sound, video, graphics
- **Virtual reality (VR)** is a computer technology that replicates an environment, real or imagined, and simulates a user's physical presence in a way that allows the user to interact with it. It artificially creates sensory experience, which can include sight, touch, hearing, etc.



Components of the devices

- Optical
  - Displ
  - Head
  - Eyegl
  - Head
- Sound s
  - Loud
- Head a
  tracking
- Compu



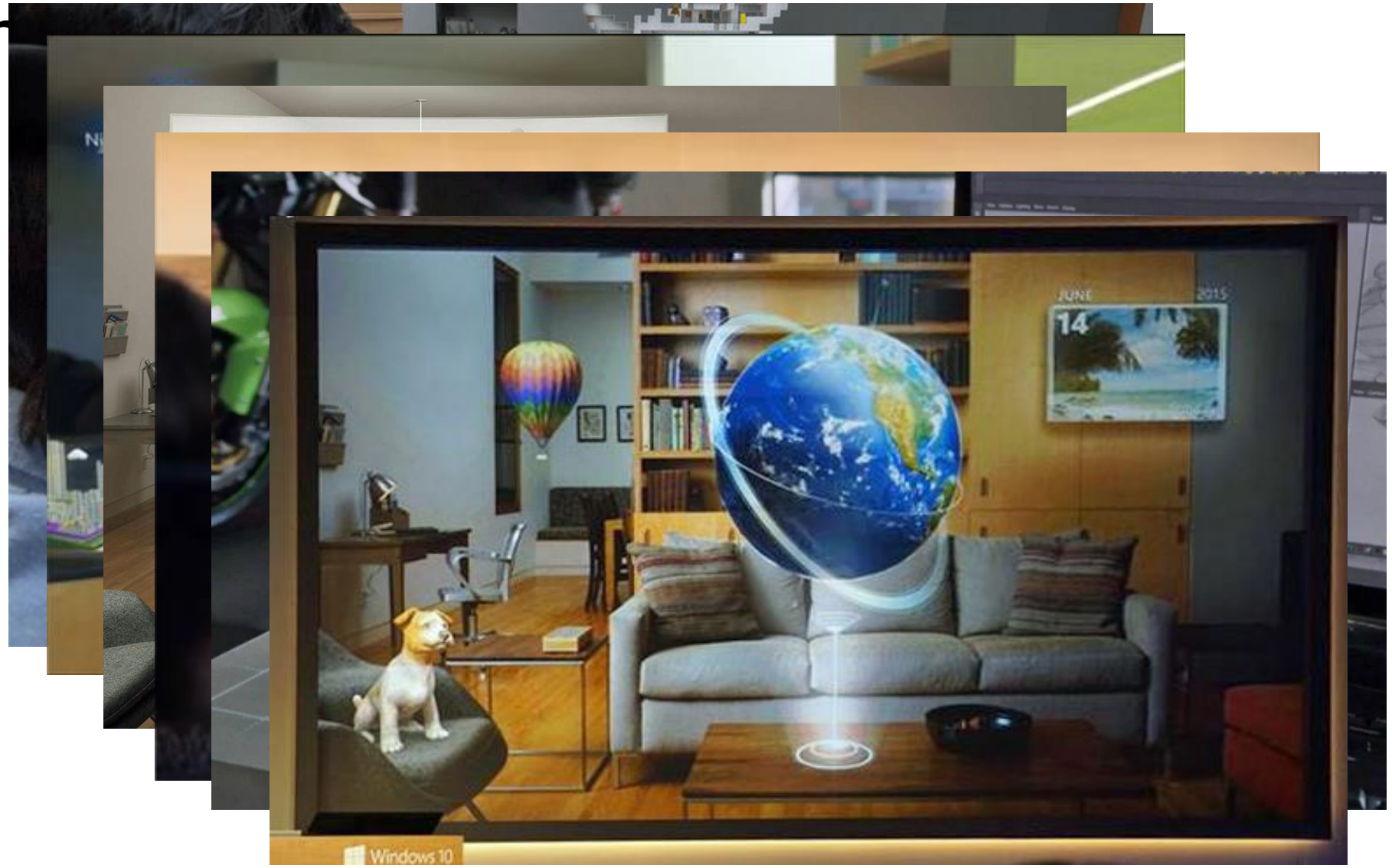
rama virtual reality headset (1950s)
Credit: Wikipedia

Examples for AR/VR devices

- Oculus Rift: VR glasses
  - Head-up display, controller
  - Needs PC to operate
- Smartphone holders
  - Simple low-cost VR solution
  - Limited applicability
- Microsoft HoloLens: AR device
  - Head-up display
  - Autonomous device

Usage scen

- Gaming
- Entertainment
- Productivity
- Science
- Design and art
- Education



Audio components of AR/VR devices

- Capture user's voice
  - Voice input is critical modality in the UI
- Capture the audio environment
  - Especially important for AR devices
  - Adds value for some usage scenarios
- Spatial audio rendering engine
  - To be *en par* with the holographic imaging
- Personalization of the spatial audio engine
  - We all hear differently

In this talk

Binaural recording and playback

A brief history

Binaural recording and reproduction

- Théâtrophone, 1881
- Binaural recordings – mid 50's

BINAURAL RECORDING
USE HEADPHONES

- Problems:
  - Fixed scene
  - HRTFs mismatch

Neuman KU-100

Headphones for Spatial Audio rendering

- Headphones ☺
- Head orientation tracking
  - direction, elevation, roll
- Head position is a plus
  - x, y, z (cameras or Kinect)

OSSIC X

Dysonics Rondo Motion

MSR experimental prototype

Jabra Intelligent headset

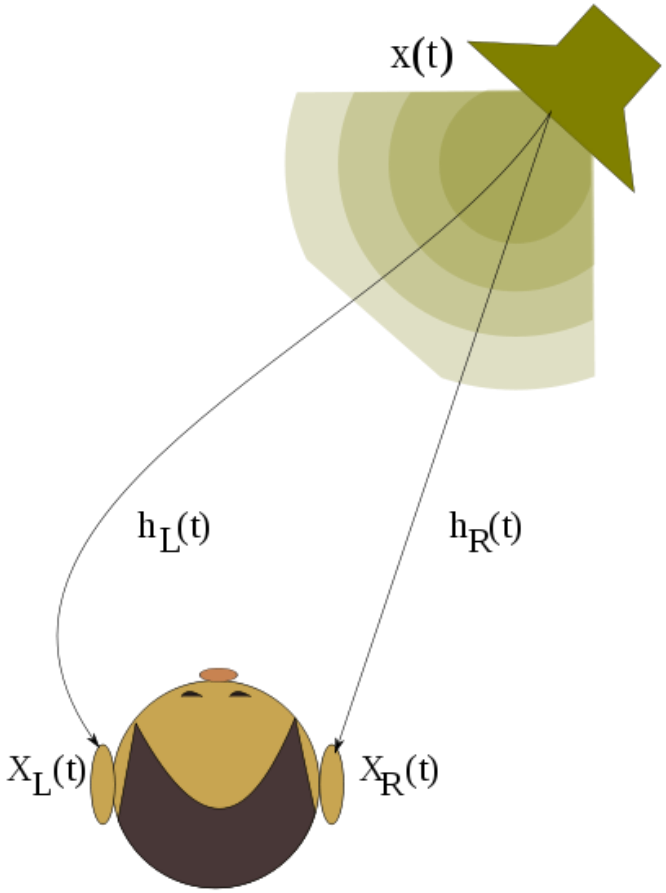
Head Related Transfer Functions and their personalization

They are just directivity patterns

Spatial hearing

- Direction and distance perception
- Within audible frequency and dynamic range
- Delivered to one or both ears
- Contains auditory localisation cues:
  - Interaural time and level differences
  - Spectral cues

  HRTFs
  - Reverberation/reflections
  - Dynamic and multimodal cues
  - (expectation and experience)

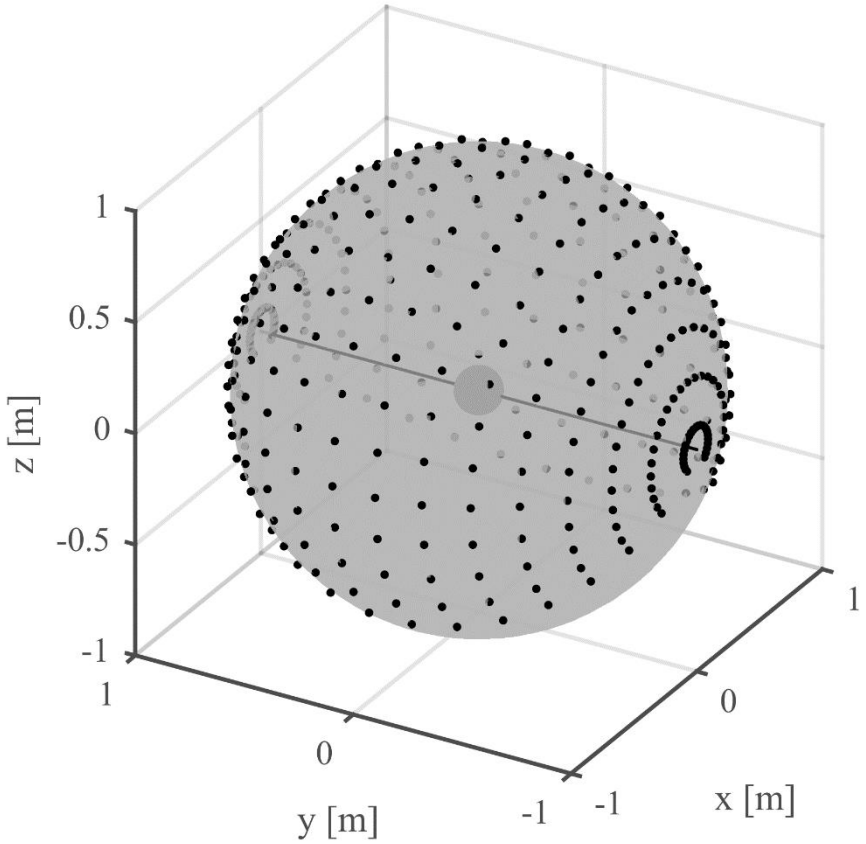


HRTF personalisation

- Describe acoustic path from sound source to ear entrances
  → Contain all interaural and spectral localisation cues
  → Are a function of sound direction
  → Can be considered distance-independent for radii > 1m
- Head and torso geometry affects wave propagation
  → Anthropometric features are individual
  → HRTFs are individual
  → Spatial hearing is individual!

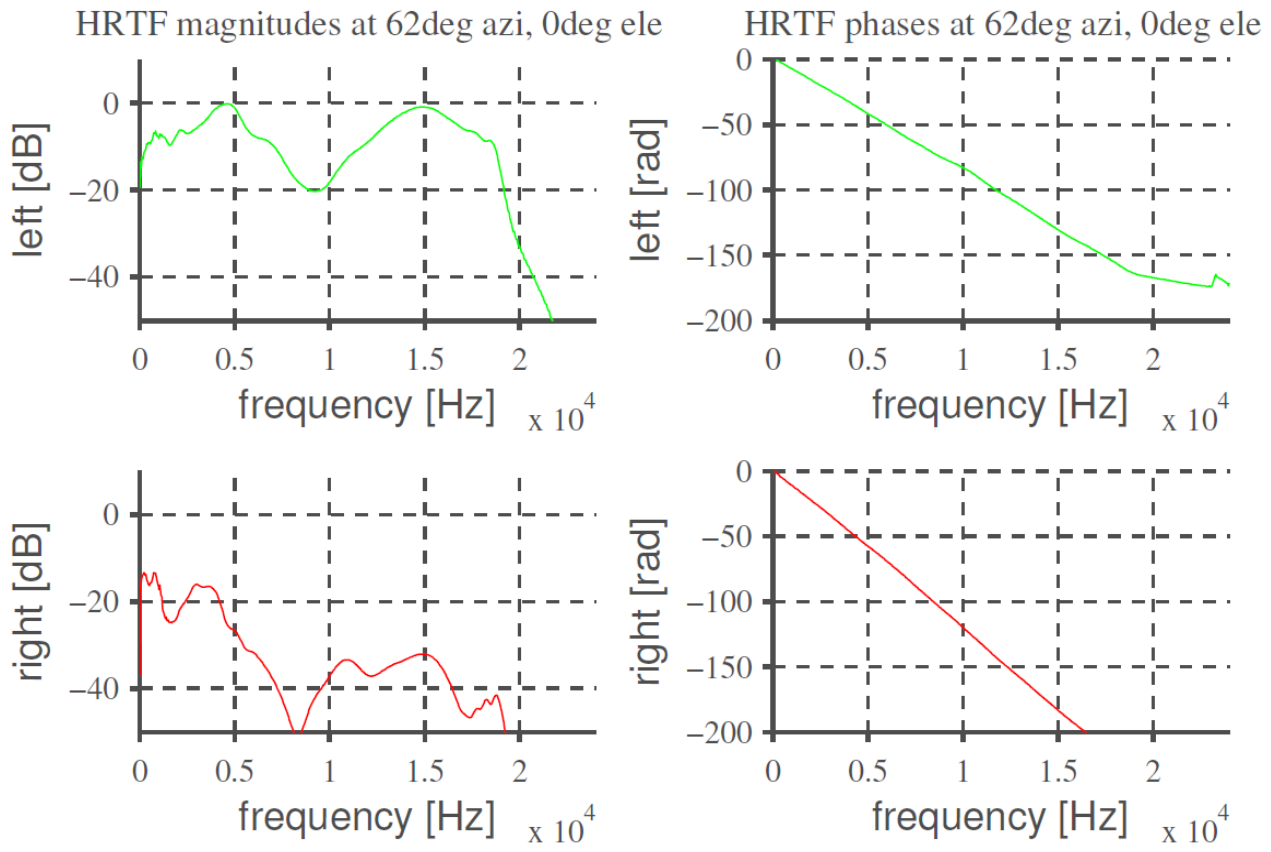
Measuring HRTFs

HRTF measurement rig



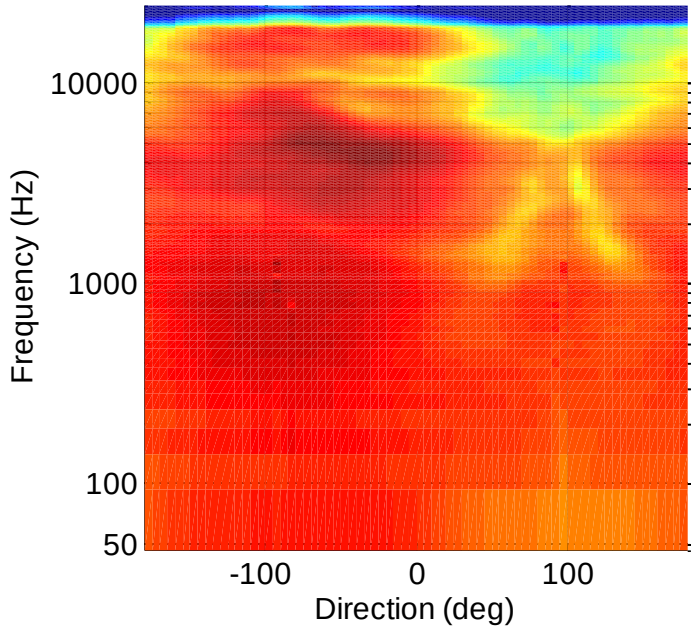
Measurement locations

Example measurements

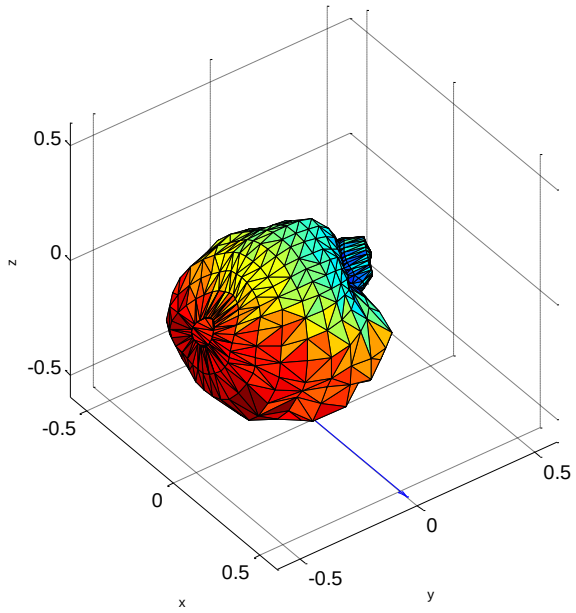


Head-related transfer function (HRTF)

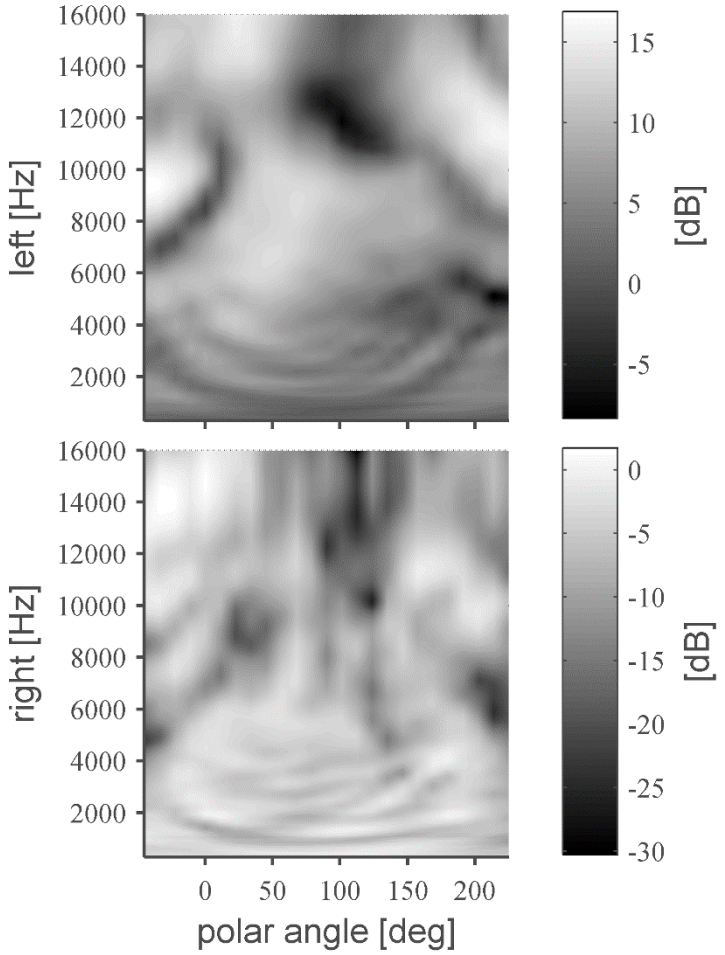
Example measurements (2)



HRTF in horizontal plane, right ear



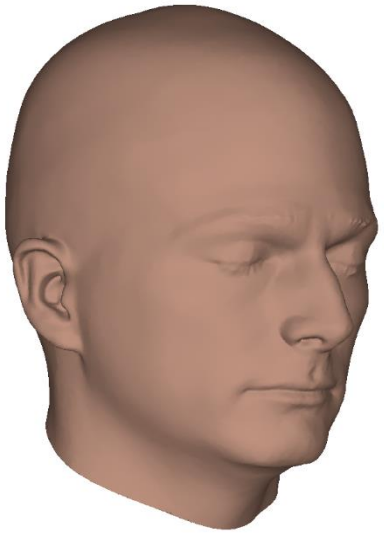
HRTF for 1000 Hz, right ear



HRTF for azimuth, 90°, both ears

MSR HRTF database

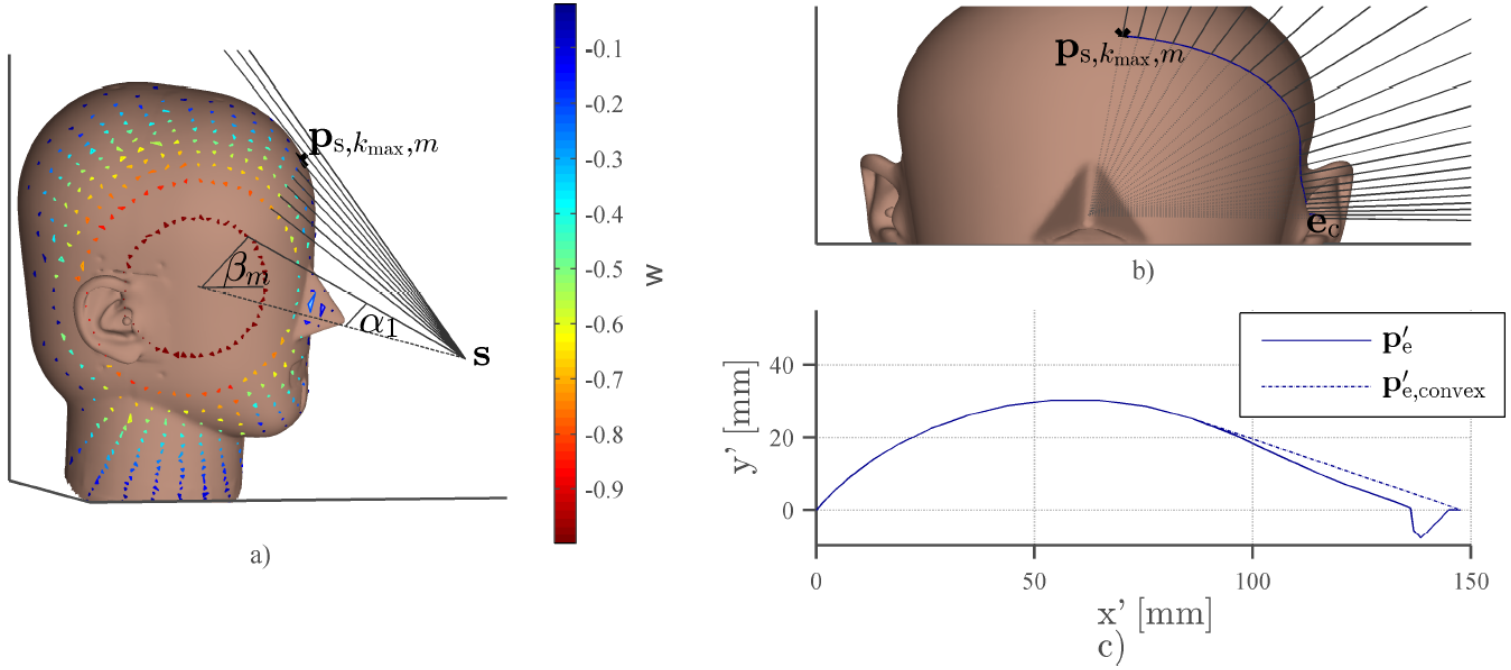
- ~300 subjects
- HRTFs measured at 400 locations
- High resolution 3D head scans
- Direct anthropometrics measurements
  - Head width, depth, height, etc.
- Questionnaire
  - Hat size, shirt size, jeans size, etc.



3-D head scan

HRTF personalization: direct estimation

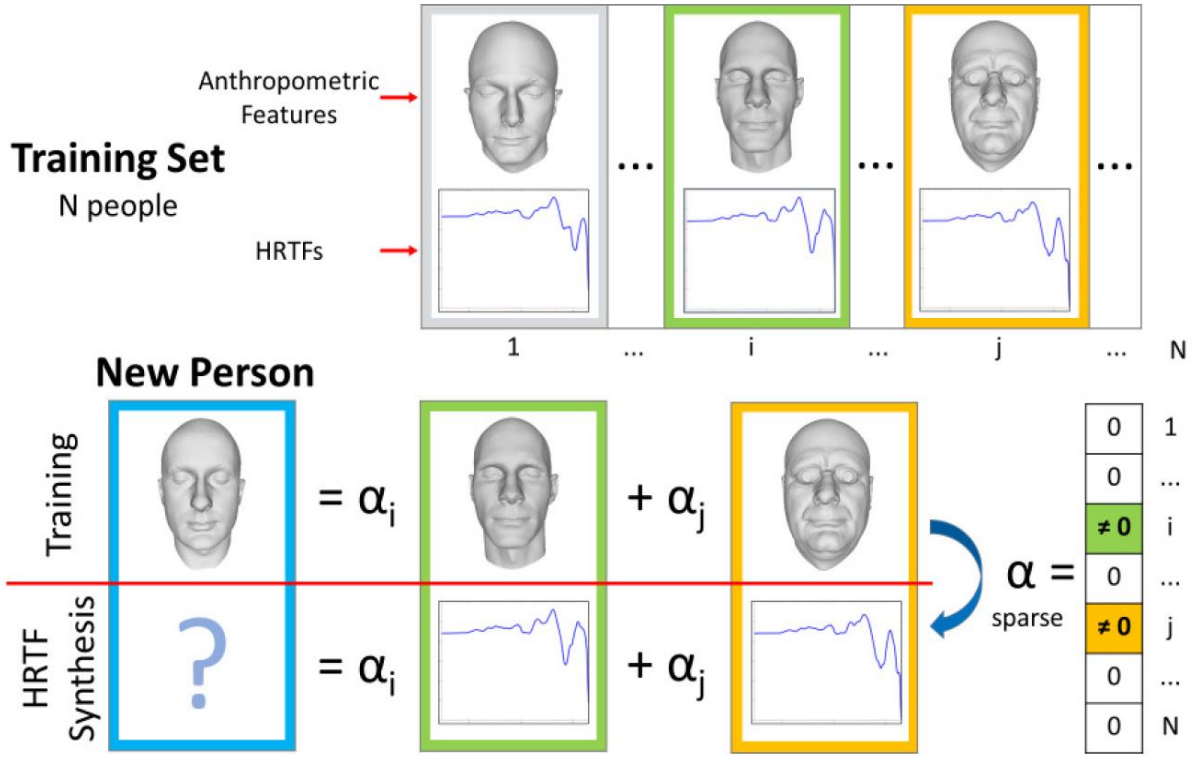
- Given 3-D head scan
- Trace acoustic propagation from source positions to ear entrances
- Good results with high-resolution scan



Gamper, H.; Thomas, M. R. P. & Tashev, I. J. (2015). "Estimation of multipath propagation delays and interaural time differences from 3-D head scans." *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP).*

Anthropometrics-based HRTF personalization: Magnitude Synthesis

1. Measure anthropometric features on a large database of people.
2. Represent a new candidate's anthropometric features as a **sparse combination** α of people in the database.
3. Combine HRTFs with **same weights** α to synthesize personalized HRTF.

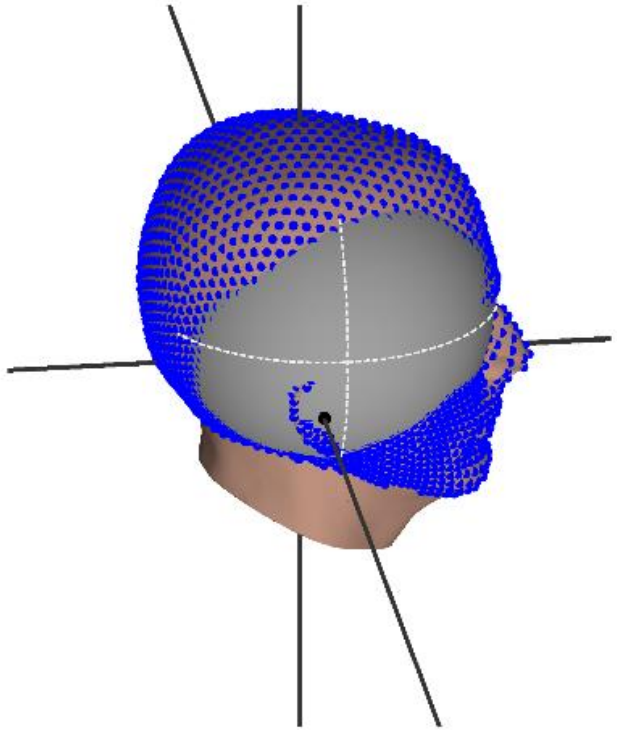


P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," *ICASSP,* 2014.

I. J. Tashev, "HRTF phase synthesis via sparse representation of anthropometric features," *ITA Workshop,* 2014.

HRTF personalization: parametrization

- Given 3-D head scan
- Fit sphere to scan
- Parameterize ITD models
- Works with noisier (e.g., Kinect 1) scans

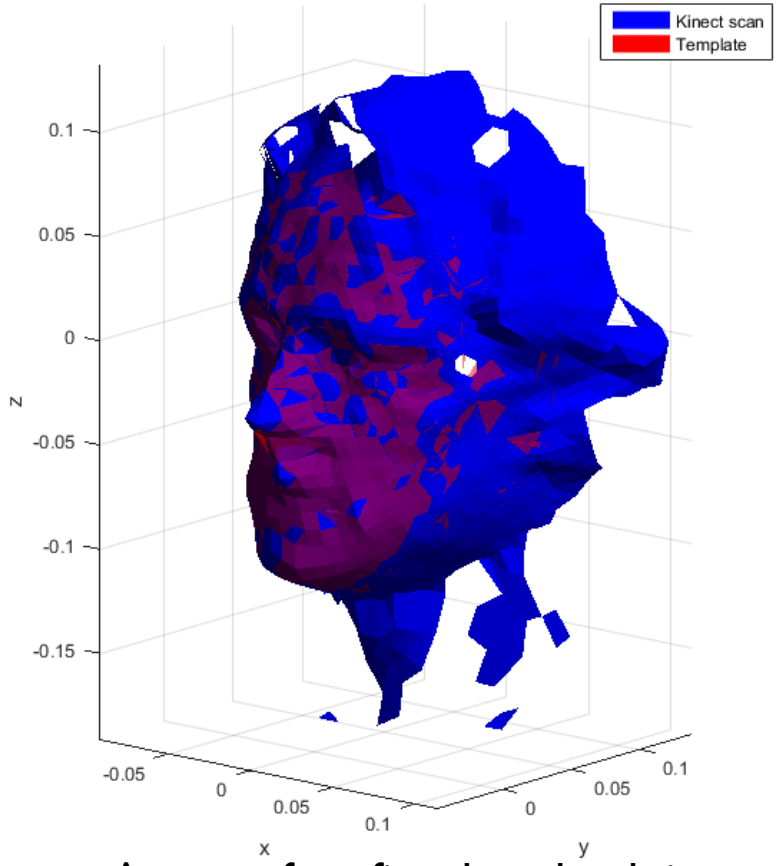


Sphere fitted to 3-D head scan.

Gamper, H.; Thomas, M. R. P. & Tashev, I. J. (2015). "Anthropometric parameterisation of a spherical scatterer ITD model with arbitrary ear angles." *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).*

HRTF personalization: Eigen-face approach

- Given single (Kinect 2) depth image
- Fit average face to scan
- Use deformations as a features in ML-based estimator
- HRTF personalization in less than 10 seconds!

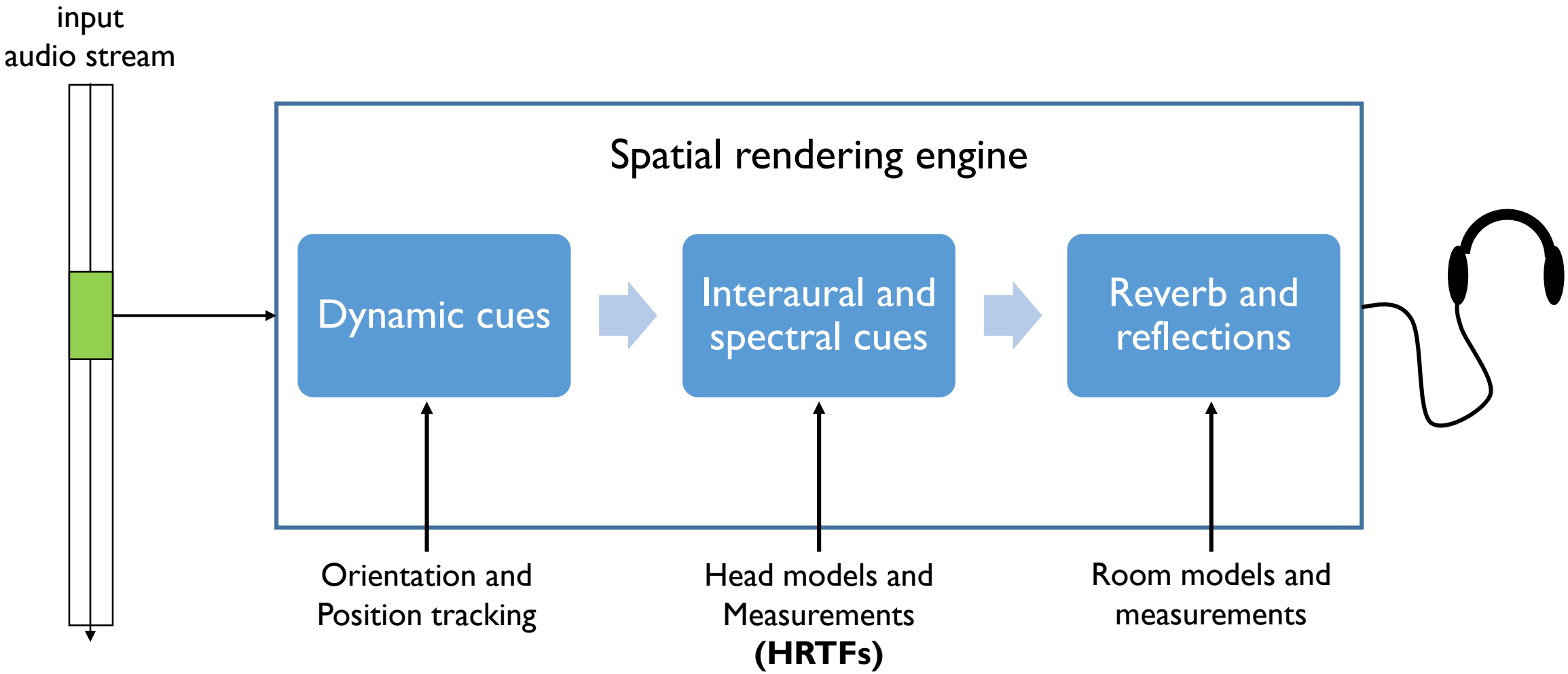


Average face fitted to depth image.

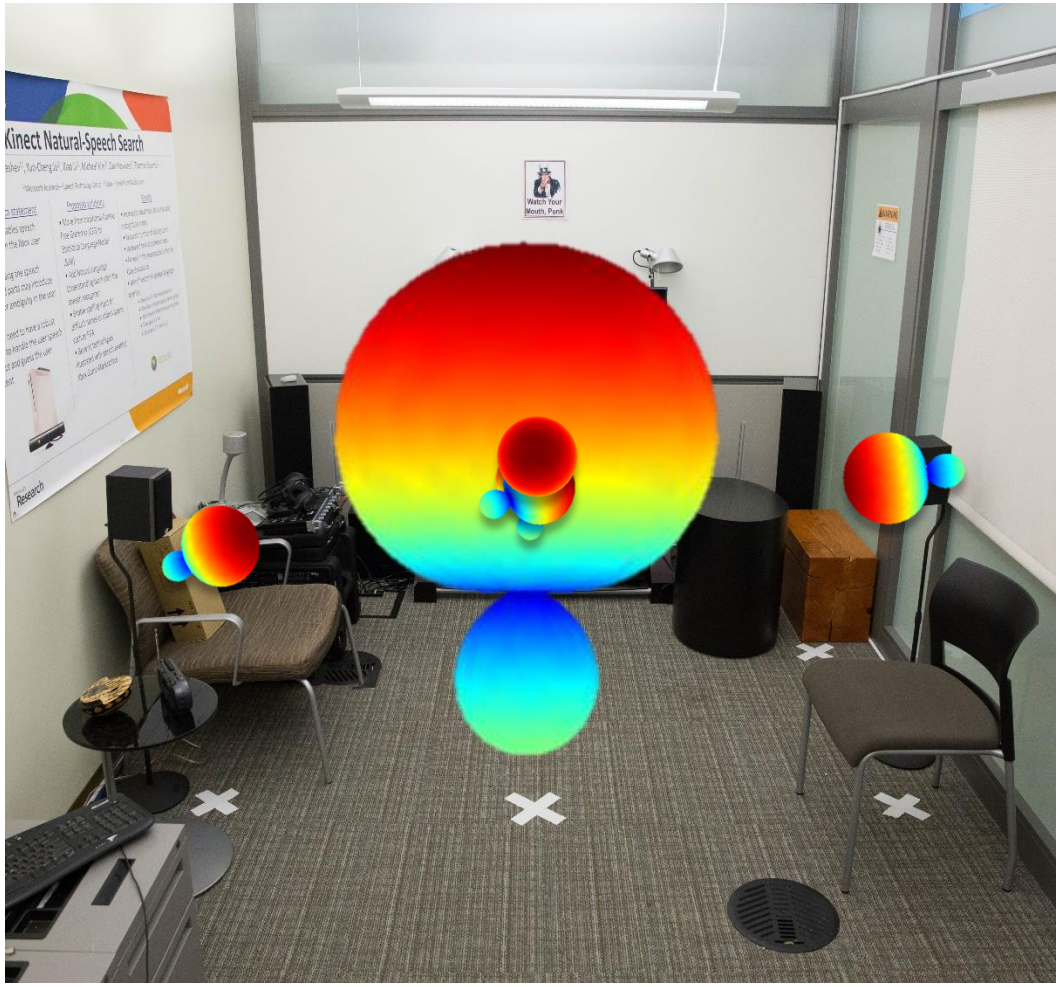
Object-based Rendering of Spatial Audio

Two ears – three dimensions?

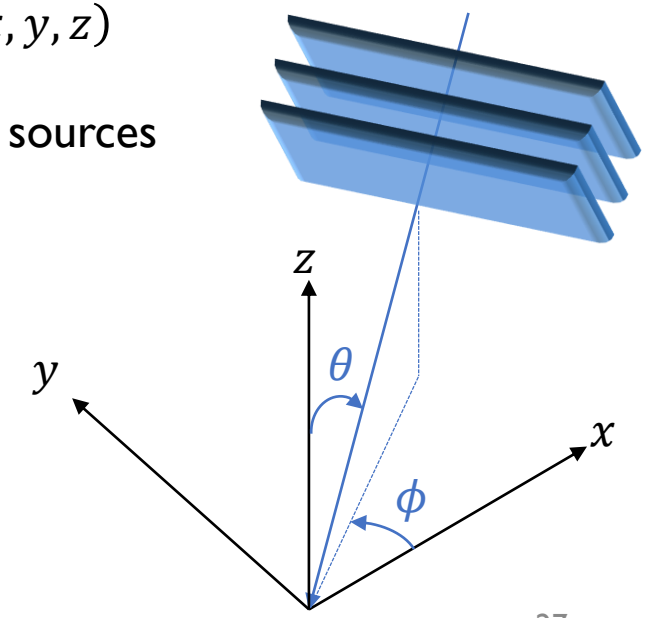
Object-based spatial audio rendering



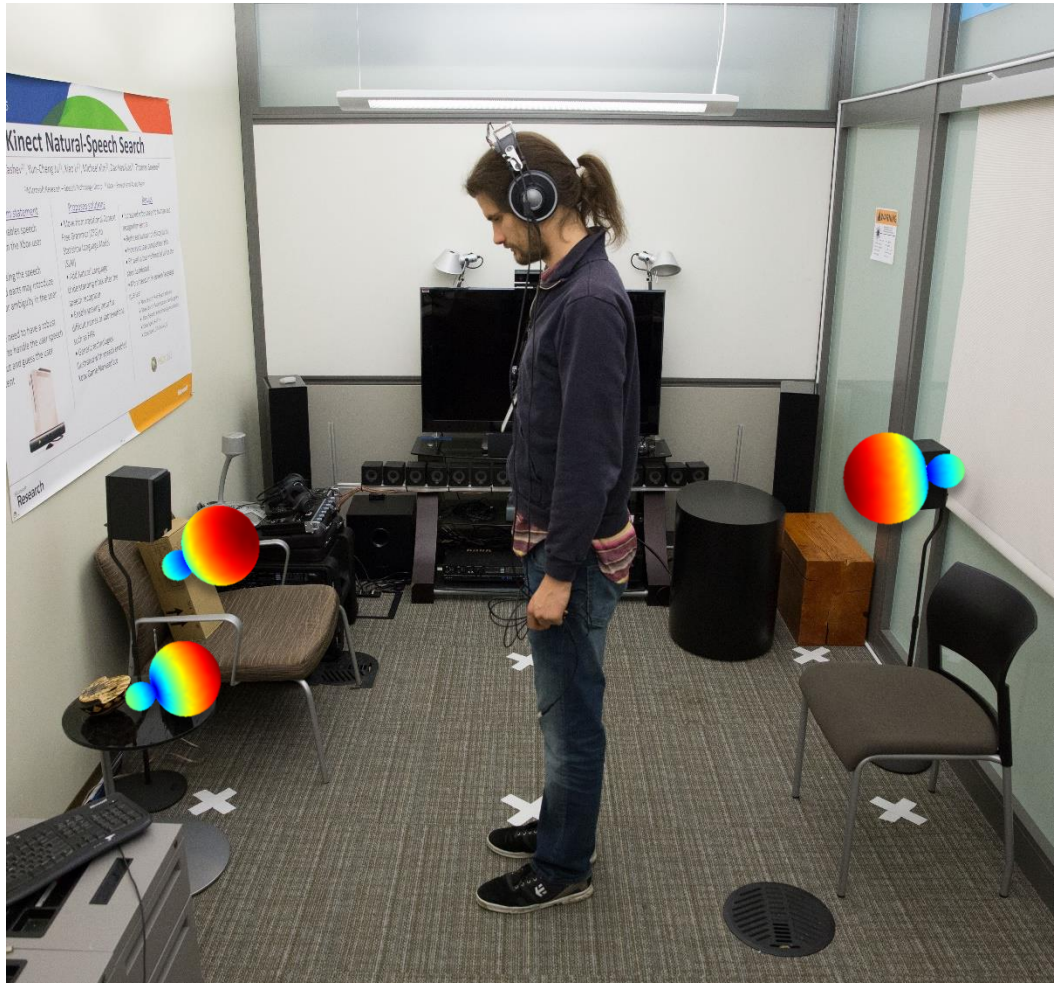
Composing a Sound Scene - World



1. Source radiation pattern $f(\theta, \phi, \omega)$
2. Scaling (gain)
3. Rotation (α, β, γ)
4. Translation (x, y, z)
5. Repeat for all sources

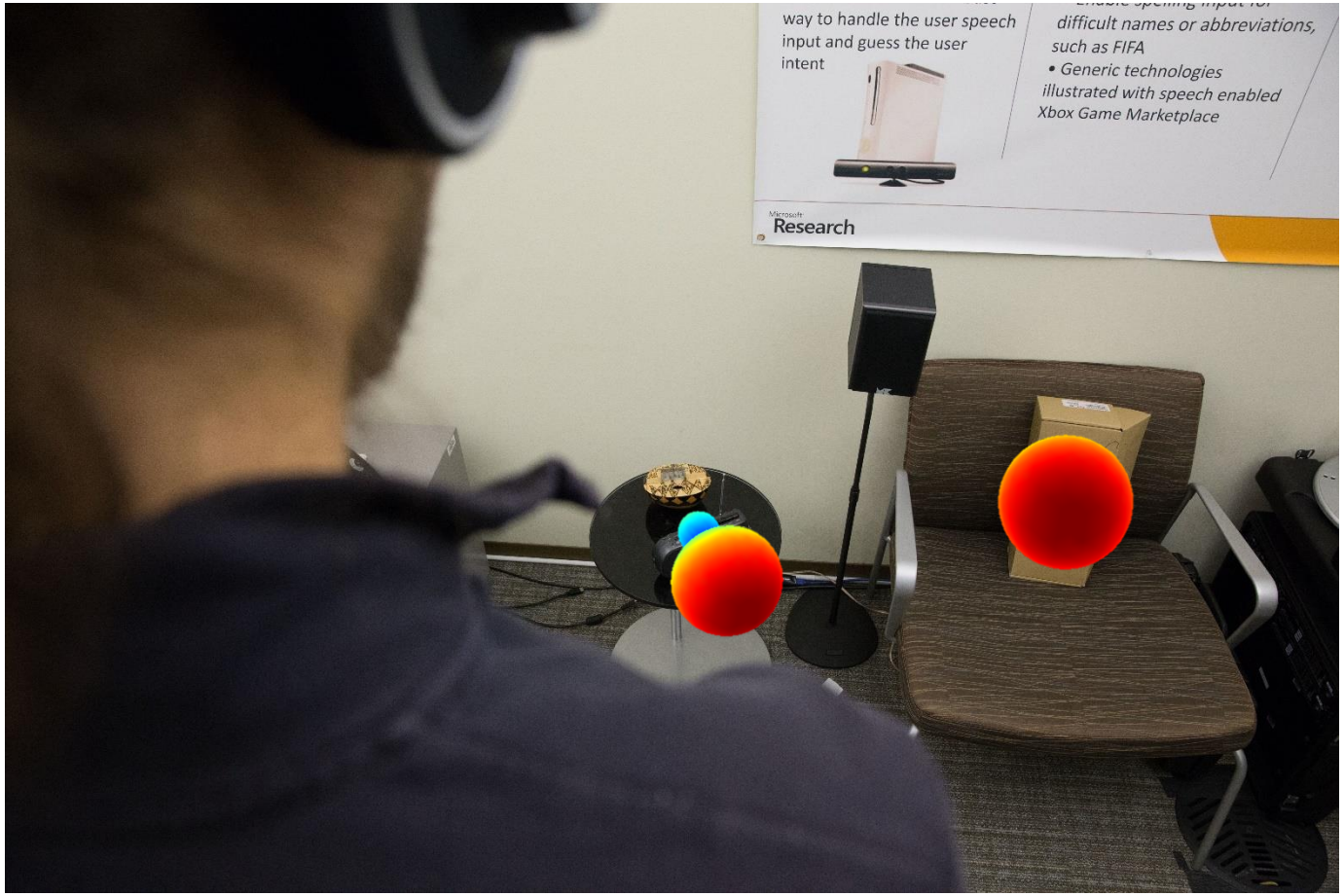


Composing a Sound Scene - View



1. Source radiation pattern $f(\theta, \phi, \omega)$
2. Scaling (gain)
3. Rotation (α, β, γ)
4. Translation (x, y, z)
5. Repeat for all sources

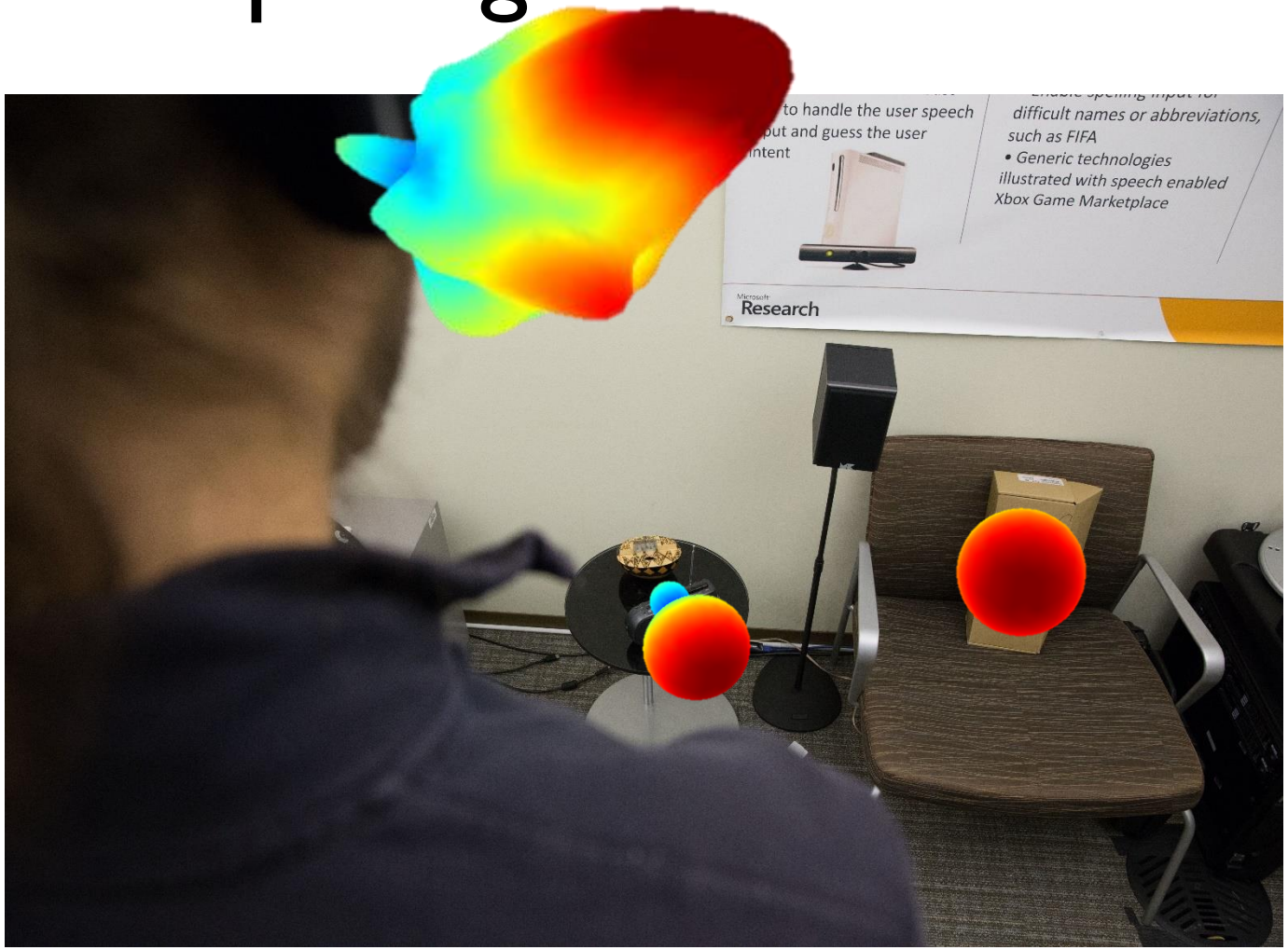
Composing a Sound Scene – View



View transform into user space

$(x_0, y_0, z_0, \alpha_0, \beta_0, \gamma_0)$

Composing a Sound Scene – Projection



View transform into user space

$$(x_0, y_0, z_0, \alpha_0, \beta_0, \gamma_0)$$

'Project' onto user's head-related transfer function (HRTF)

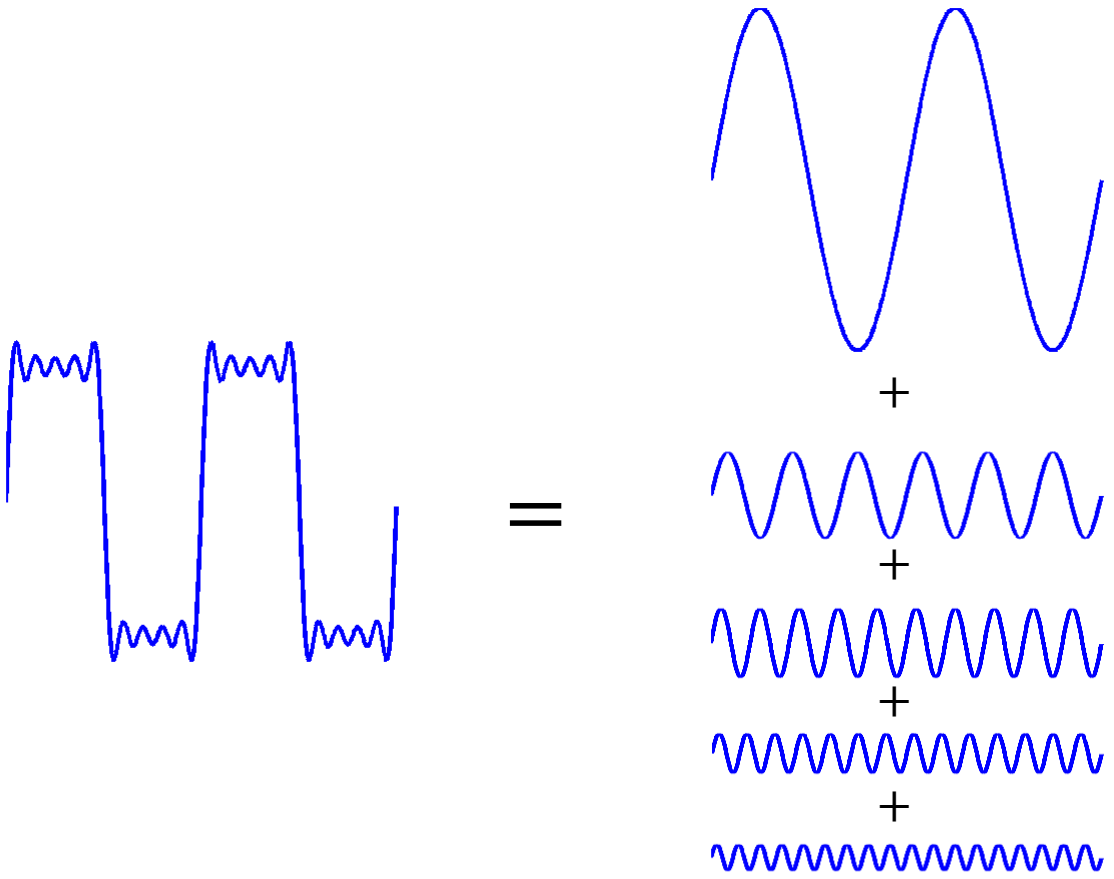
Entire scene becomes two signals that are fed to the headphones.

Modal-based Rendering of Spatial Audio

Can we model the sound field?

Detour: Fourier Analysis of a 1D Signal

- Consider a signal $x(t)$.
- Any time series can be decomposed as a weighted sum of sin() and cos() functions.
- These weights are its *Fourier spectrum* $X(\omega)$.



Plane Wave Density Function

- *Plane Wave Density Function*, or *Signature Function*

Time ↔ Frequency

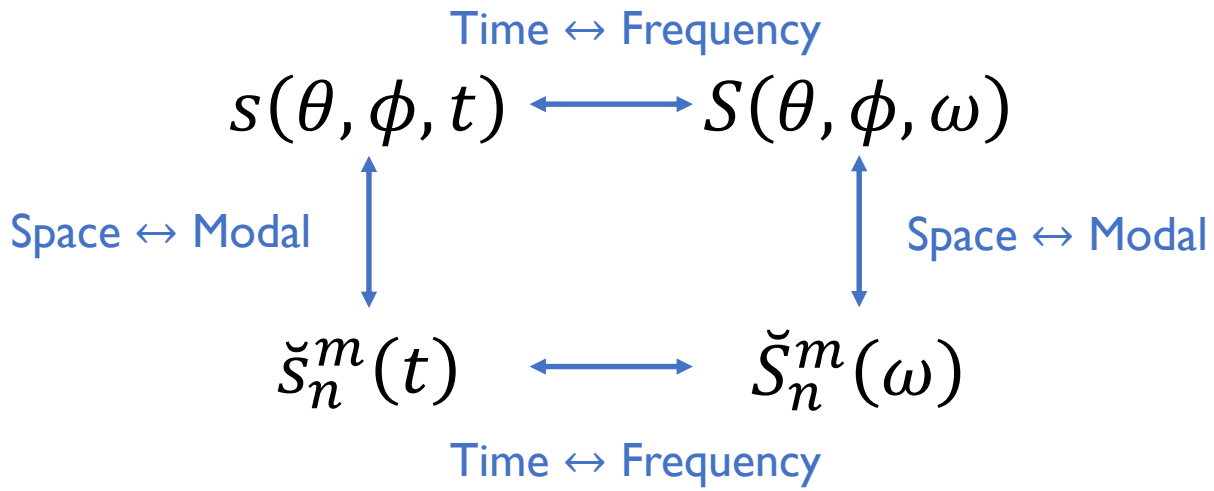
$$s(\theta, \phi, t) \leftrightarrow S(\theta, \phi, \omega)$$

- Plane wave density function is very powerful
  - Continuous representation of a sound field at a single point*
  - Can be used to place a user at the same location; an audio panorama in which the user can freely rotate in (α, β, γ).

*Under certain conditions, we can theoretically extrapolate the pressure field to any point in space (though extremely difficult to do in practice).

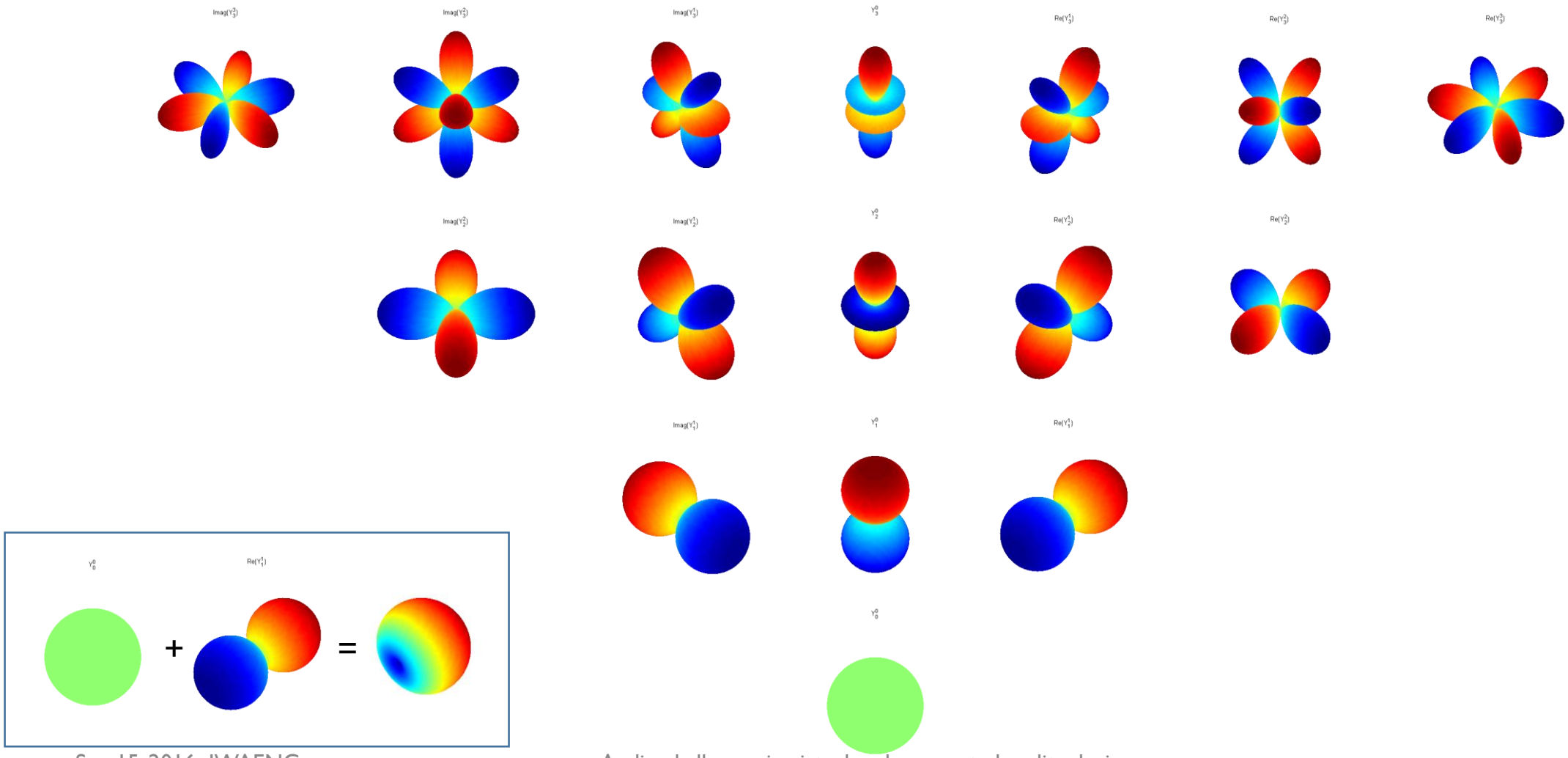
Fourier Transforms with Respect to Space

- Notice Fourier Transform is with respect to time; spatial component (θ, ϕ) is unaffected.
- We can also take Fourier Transforms with respect to space using the *spherical harmonics* as basis functions.



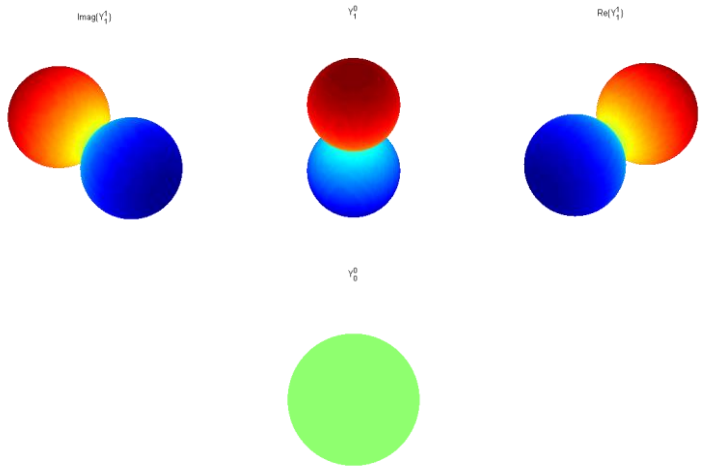
- Notice continuous space maps to discrete order m and degree n.

Spherical Harmonics



First-Order Sound field Decomposition

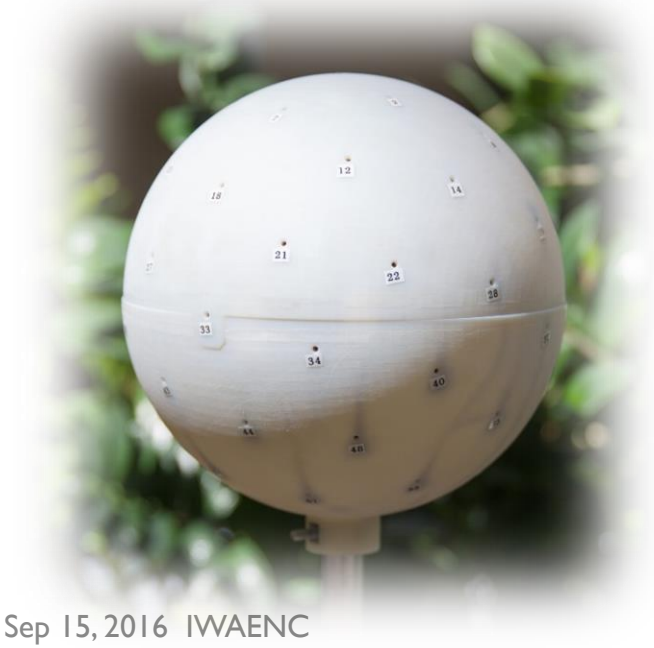
- B-Format
  - Three fig-8 (pressure gradient) microphones and one omnidirectional microphone.
  - Arranged to be near-coincident.
- Provides *first order* decomposition of plane wave density function
- Used in Ambisonics

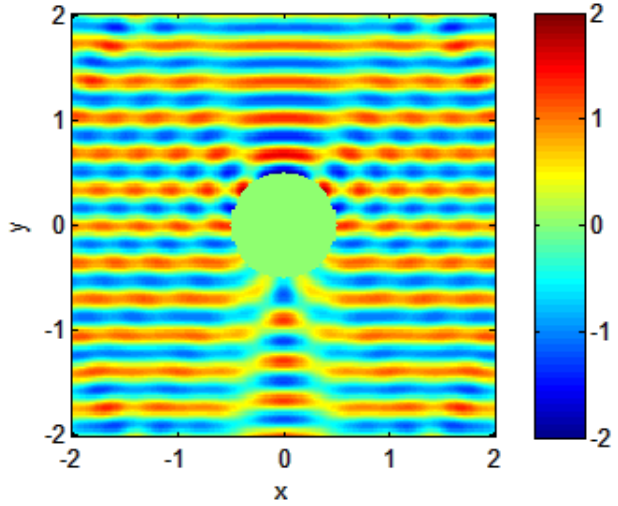


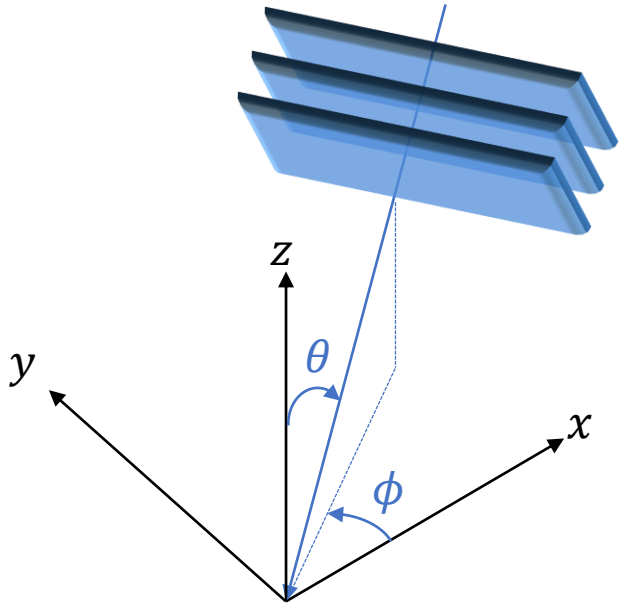
Courtesy ambisonia.com

Higher Orders: Spherical Microphone Array

- Higher orders: sample the pressure on a sphere $p(\theta, \phi, t)$
- Plane waves are scattered by the rigid surface
  - Spherical wave reflects outward







Devices for sound field decomposition

16-ch. 115 mm spherical mic. array

64-ch. 200mm spherical mic. array

16-ch. 115 mm cylindrical mic. array

Modal-based spatial audio rendering

- Decomposition of the sound field
  - Spherical harmonics: representation of the sound field as a sum of basis functions of degree *n* and order *m*
  - Cylindrical harmonics – the 2D version in horizontal plane only
- Properties
  - Linear – allows summing of independently processed signals
  - The sound field can be rotated – allows compensation for head movement
  - Convolution is multiplication – allows fast application of the HRTFs
- Same steps:
  - Reflect the head rotation
  - Apply HRTFs (decomposed in spherical harmonics)
  - Add reverberation

Conclusions

What exactly were the challenges?

Spatial Audio

- Technique to make the listener to perceive the sound coming from any desired position. Also known as 3D audio.
- Challenges:
  - Low latency head orientation tracking
  - HRTF personalization
  - Realistic reverberation and distance cues
- Applications go way beyond augmented and virtual reality devices ☺
  - Listening to stereo music
  - Watching movies with surround sound
  - Gaming

Available today or coming soon:

- HoloLens Development Kit (March 2016):
  - Spatial sound engine in Direct X
  - Spatial sound in Unity plug-in (Microsoft HRTF extension)
- Full spatial audio support in Windows 10
  - Object-based spatial audio engine is part of Windows SDK
- Virtual Surround Sound support in Windows 10
  - Wrapper around the spatial audio engine

- Proven in real applications architecture and algorithms

Shameless plug

- Immediately precedes ICASSP
- More at http://www.hscma2017.org/

Finally …

Questions?