Microsoft

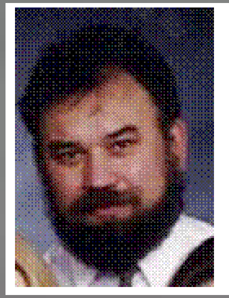
# Physically Accurate Low Latency Audio for Virtual and Augmented Reality


*Ramani Duraiswami*
University of Maryland, and
VisiSonics Corporation

*Joint with Dmitry Zotkin, Zhiyun Li, Elena Grassi, Adam O'Donovan, Nail Gumerov,*



*Work supported by NSF, ONR, DARPA and UMIACS*

# Virtual and Augmented Reality

## VR/AR

Bets by Facebook, Google, and of course Microsoft

Vision of a world where the way you interact with the world is smartly intermediated by technology

Predicted to reach $120B in 5 years

## The next platform:

Personal Computers, 1990s

Internetworked Personal Computers, 200x

Mobile, 201x

# VR/AR

## Virtual Reality

- Create artificial world that the user believes is real
- Simulation, Gaming and Entertainment

## Augmented Reality

- Insert objects and information into the real world
- Training, Surgery, Entertainment

## Enablers

- Hardware: Moore's law, displays, graphics, tracking,
- Software: Engines for creating virtual worlds
- Visual Perception: Improved latency, using persistence

# When rendering doesn’t work: Sickness/Fatigue

## Sickness:

- Most studied for vision/vestibular system interaction
- Use of persistence and improved frame rates mentioned by Valve/Oculus as primary improvements allowing VR
- Smaller fields of view
- Still lots of stories about gamers getting sick


CINEMABLEND NEWS TRAILERS REVIEWS UPCOMING HEROES

GAMES

**VR Games At E3 Were Making People Sick, Get The Details**

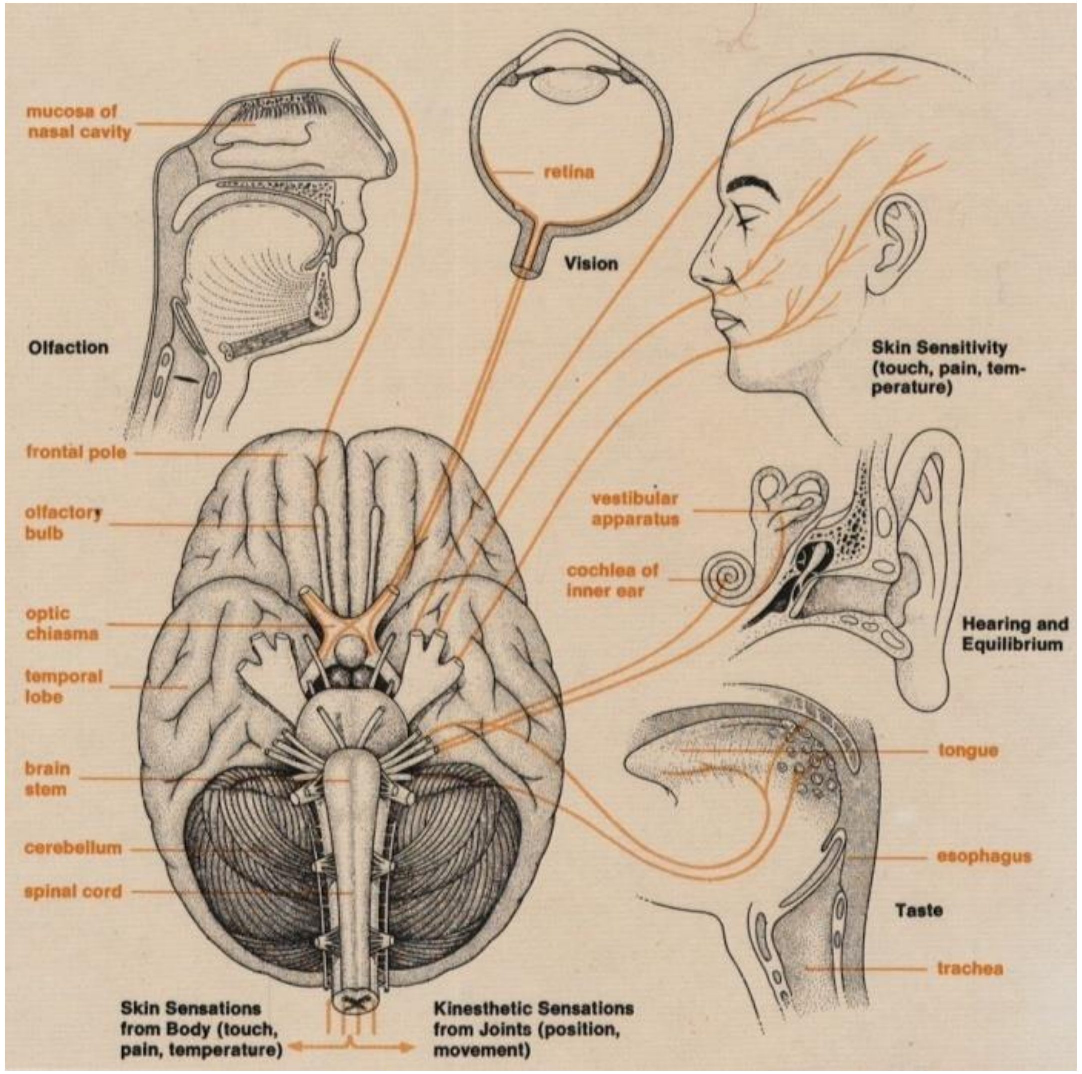
BY WILLIAM USHER 3 WEEKS AGO 16 COMMENTS

## Fatigue:

- Tendency of users to stop the VR/AR experience early
- Leave experience in minutes instead of hours
- Maybe an even bigger problem for the success of VR/AR

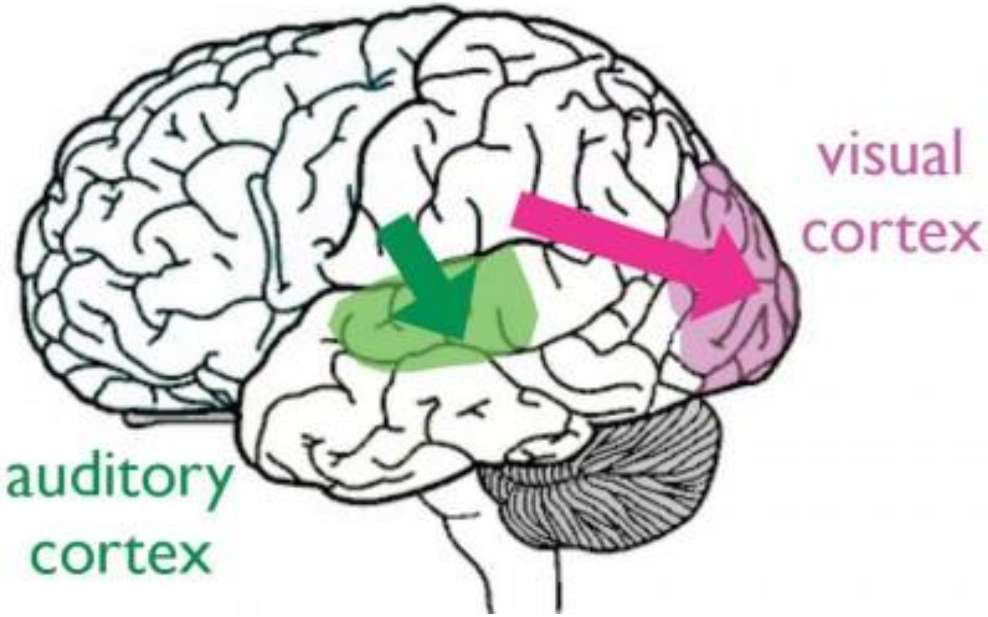
# Fool the Visual System?

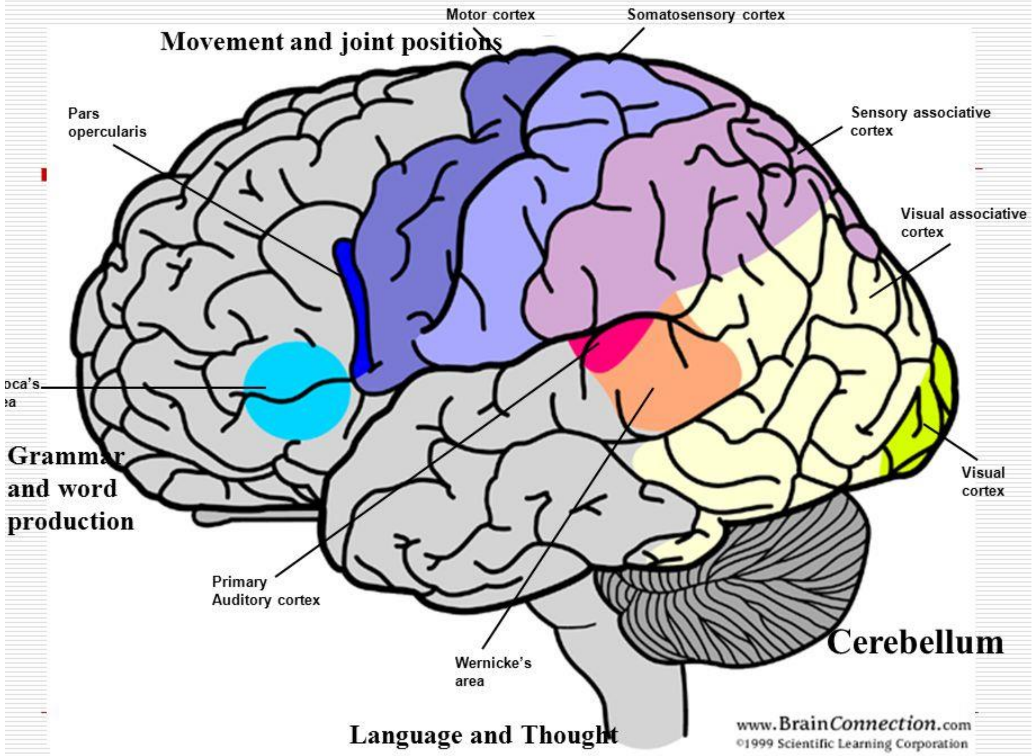
- Visual System part of larger perceptual system, responsible for sense-making
- perceptual system is a sophisticated sensing, measuring and computing system
- Designed by evolution to perform real time measurements and take quick decisions
  - **Fool this system in to believing that it is perceiving an object that is not there**



# Sensing the world auditorily

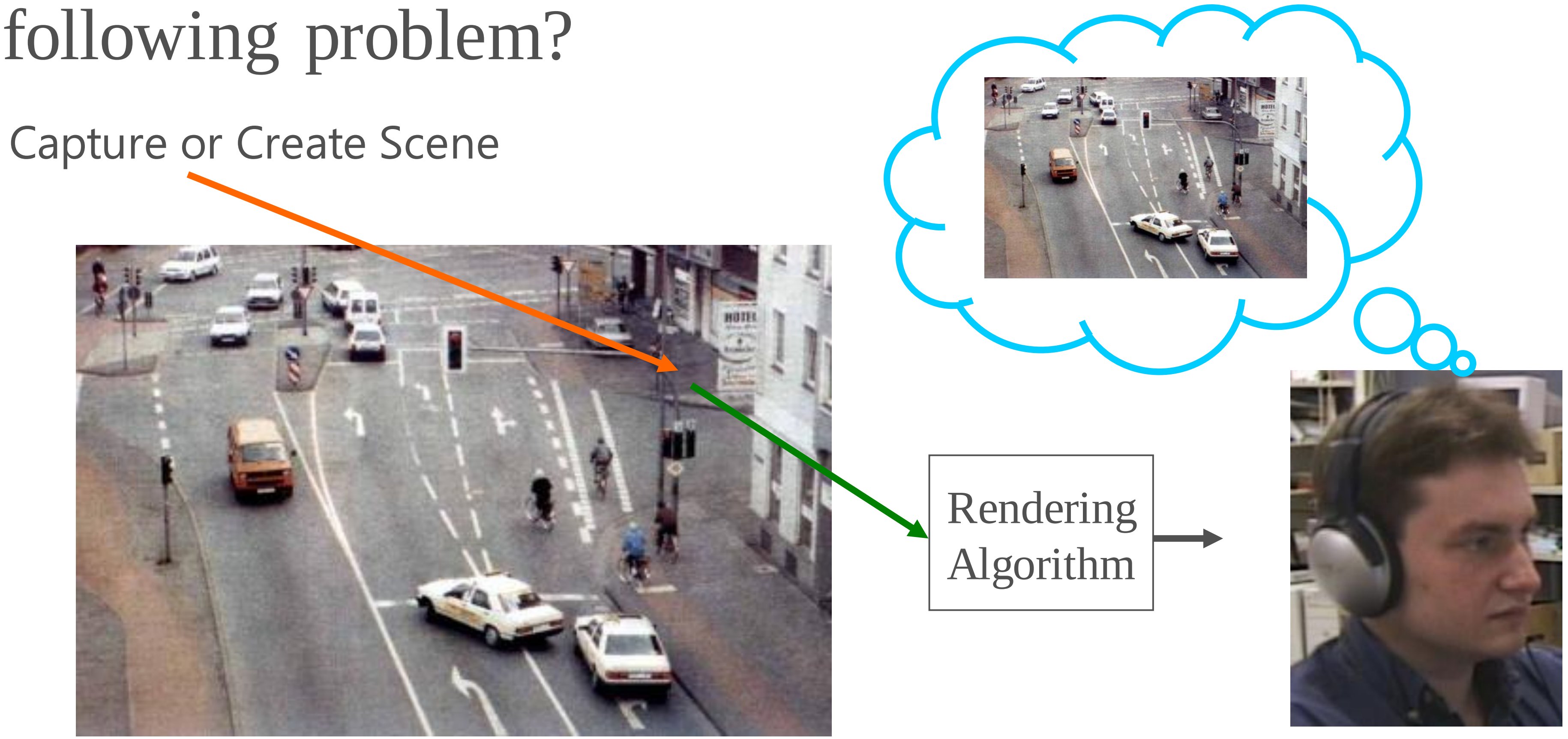
- Vision and audition are stand-off senses
  - Foveated detailed view
  - Broad knowledge of general surroundings
- Occupy nearly same area in the cortical and sensing parts of brain
- Many interconnections, including to the motor areas
- **Our hypothesis: Unless the world is rendered consistently the brain experiences fatigue**



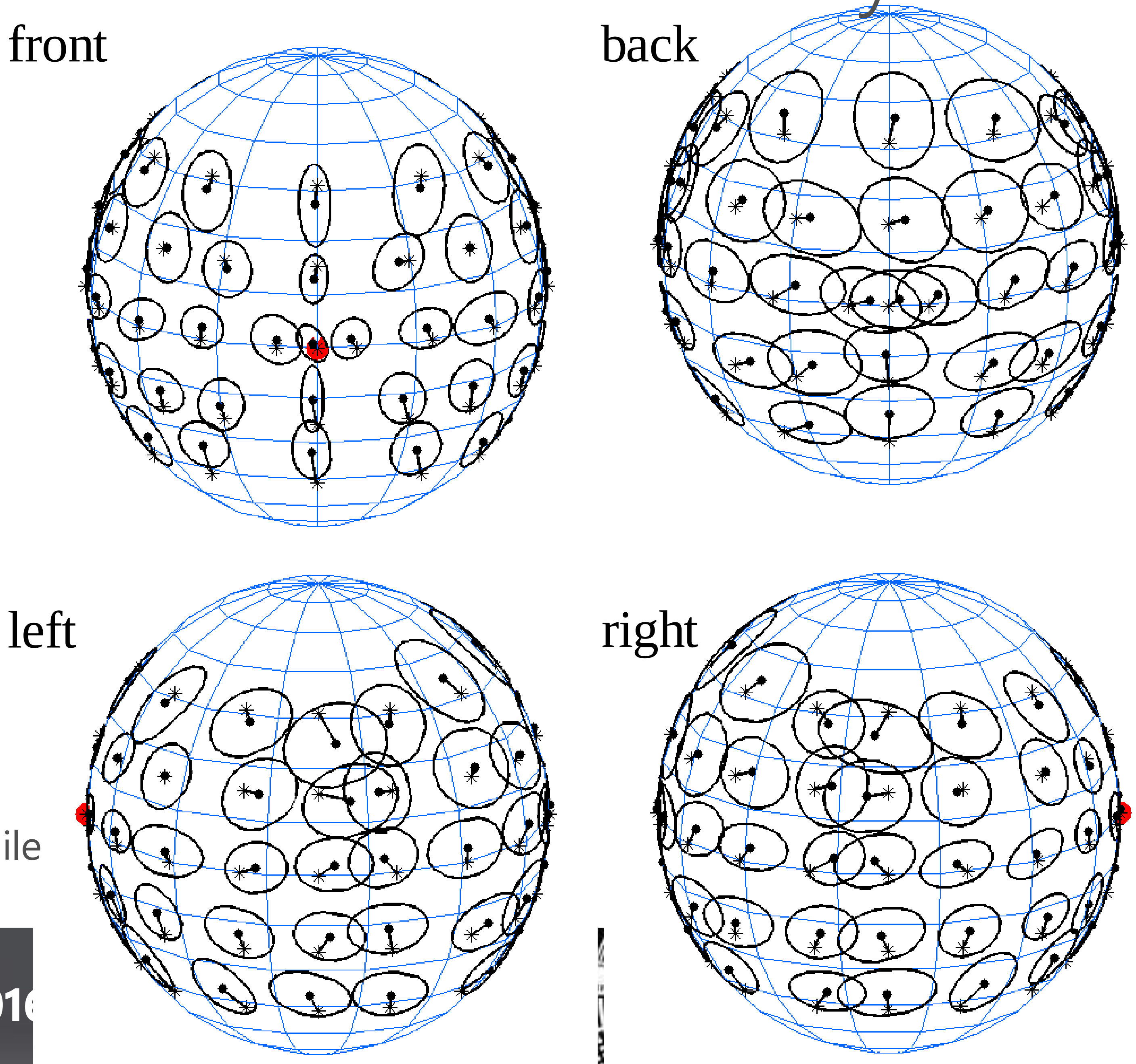
# Problem we have been working on since 2001

What physics/perception based theory can guarantee that we can solve the following problem?




Want to quantify error in measurement and error in reproduction using some theory. Want to do it without knowing the location of the sound sources. Allow interactivity and motion.

# Human spatial localization ability



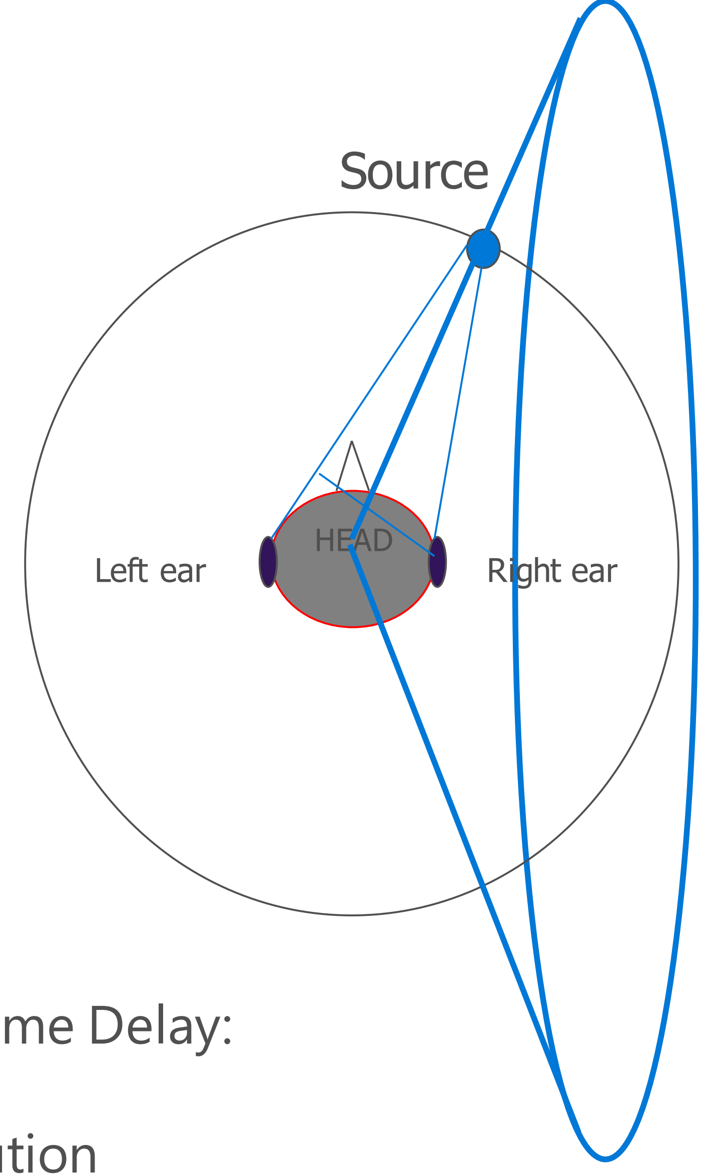

Best & Carlile
2003

# Hypothesis: Render Sound Correctly

- Get the sound right at the entrances to the ear canals
- Approximately solve the audio propagation problem from sources in the scene to the ear canal
- Do what graphics and vision did –
  - Move from emulation to approximate simulation
  - Use physics based models, appropriately simplified
  - Simplify based on knowledge of what is perceptible: focus attention on things that matter
  - Level of detail based on available computing power
  - Capture representations of the real world that allow rendering
- Render not only objects but scenes

# How do we perceive sound location?

- Naïve time and level difference at ears are not sufficient to describe our ability
- Other mechanisms necessary to explain
  - Scattering of sound
    - Off our bodies
    - Off the environment
  - Purposive Motion




Surfaces of constant Time Delay:
$|x-x_L| - |x-x_R| = c\,\delta t$
hyperboloids of revolution
Delays same for points on cone-of-confusion

# Audible Sound Scattering

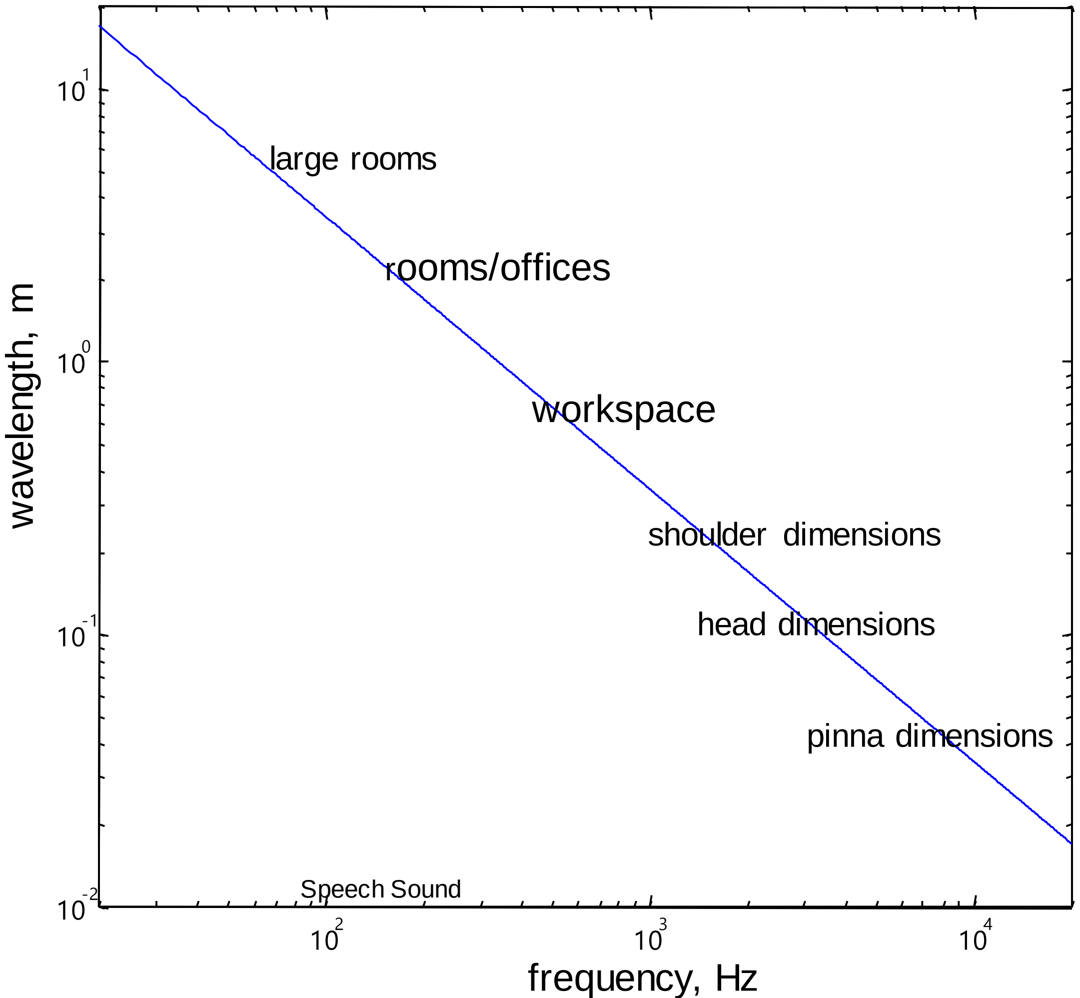
- Sound wavelengths comparable to human dimensions and dimensions of spaces we live in.
- $f\lambda = c$
- When $\lambda >> a$
  wave is unaffected by object

  $\lambda \sim a$
  behavior of scattered wave is complex and diffraction effects are important.

  $\lambda << a$
  wave behaves like a ray

wavelengths are comparable to our rooms, bodies, and features

Not an accident but evolutionary selection!



# Mathematical modeling of scattering

Wave equation: $$\frac{\partial^2 p'}{\partial t^2} = c^2\left(\frac{\partial^2 p'}{\partial x^2} + \frac{\partial^2 p'}{\partial y^2} + \frac{\partial^2 p'}{\partial z^2}\right) = c^2\nabla^2 p'$$

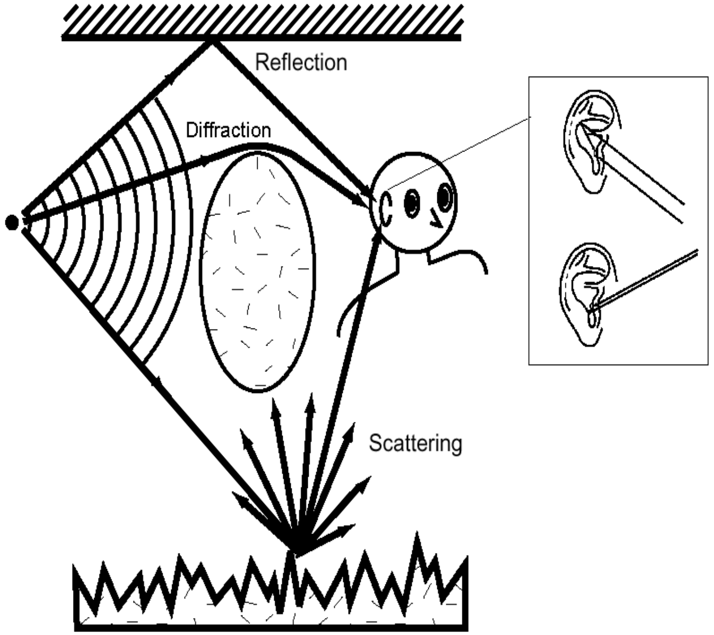
Fourier Transform from Time to Frequency Domain $$P(x,y,z,w) = \int_{-\infty}^{\infty} p'(x,y,z,t)e^{-i\omega t}dt$$

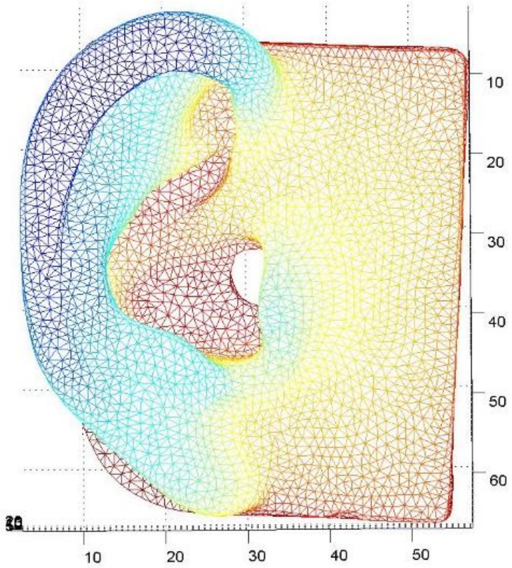
Helmholtz equation: $$\nabla^2 P + k^2 P = s\,\delta(x - x')$$

**Boundary conditions:**

Sound-hard boundaries: $$\frac{\partial P}{\partial n} = 0$$

Sommerfeld radiation condition $$\lim_{r\to\infty} r\left(\frac{\partial P}{\partial r} - ikP\right) = 0$$

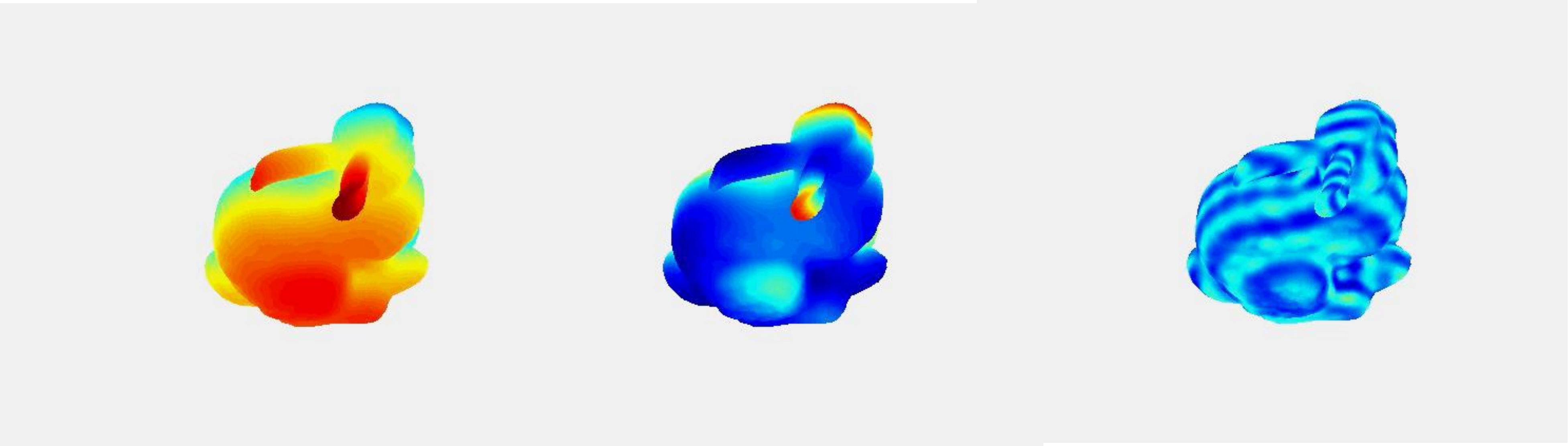


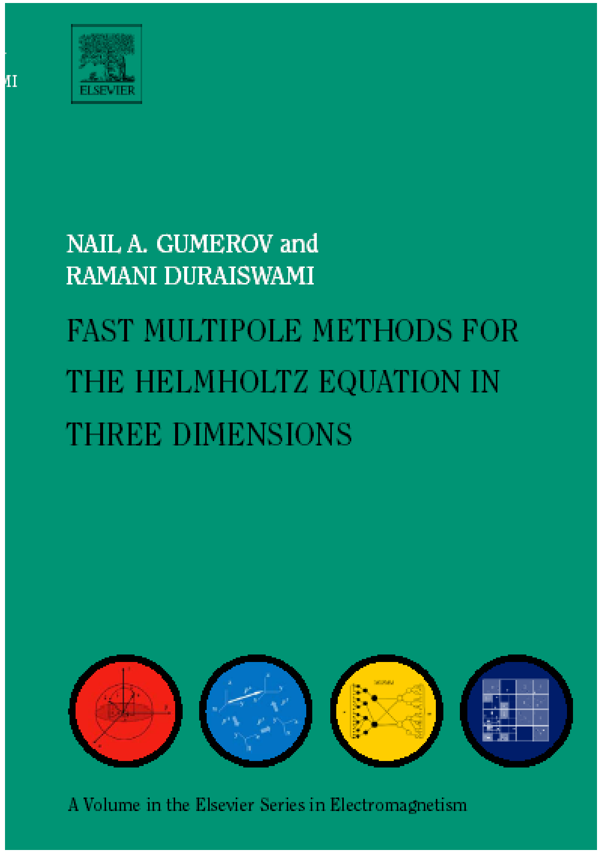
# Fast Multipole Accelerated Solver for Helmholtz equation

## $O(kD)^2$ *instead of* $O(kD)^6$

Sound pressure



| kD=0.96 | kD=9.6 | kD=96 |
|---|---|---|
| (250 Hz) | (2.5 kHz) | (25 kHz) |



# Accurate Approximate Scattering

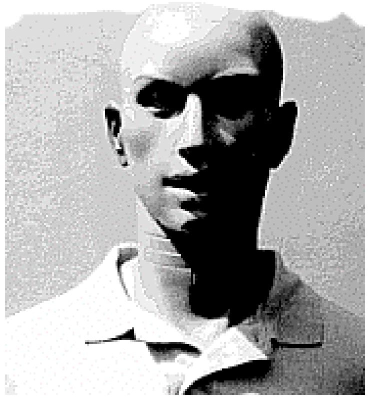
- Linear systems can be characterized by impulse response (IR)
  - Knowing IR, can compute response to general source by convolution
- Response to impulsive source at a particular location
  - Scattering off person by  Head Related Impulse Response (HRIR)
  - Room scattering by Room Impulse Response (RIR)
- Response differs according to source and receiver locations
  - Thus encodes source location
- HRTF and RTF are Fourier transforms of the Impulse response
  - Convolution is cheaper in the Fourier domain (becomes a multiplication)
- Motion is slow enough that a quasi-static model works

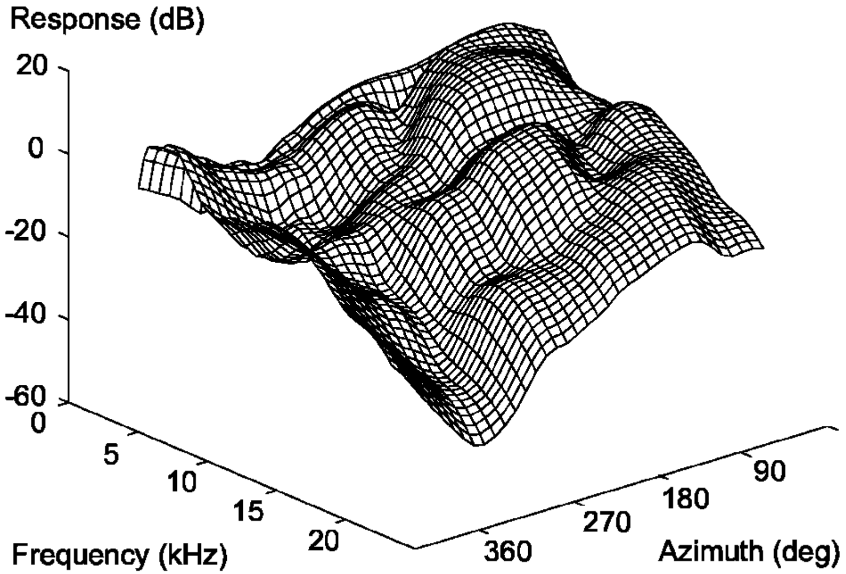
# Creating Auditory Reality

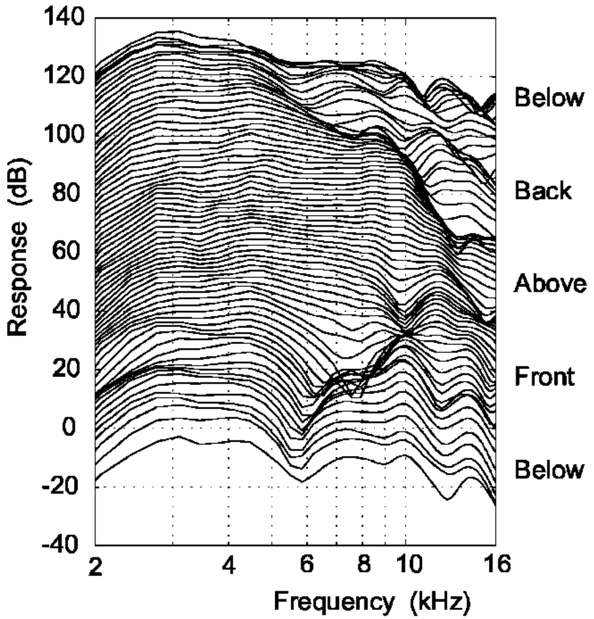
- VR/Gaming: Given a sound source and an environment build an engine that reproduces the cues
- Augmented Reality: Capture sound remotely and rerender it by reintroducing cues that exist in the real world
- Scattering of sound off the human
  - Head Related Transfer Functions
- Scattering off the Environment
  - Room Models
- Head motion
  - Head/Body Tracking

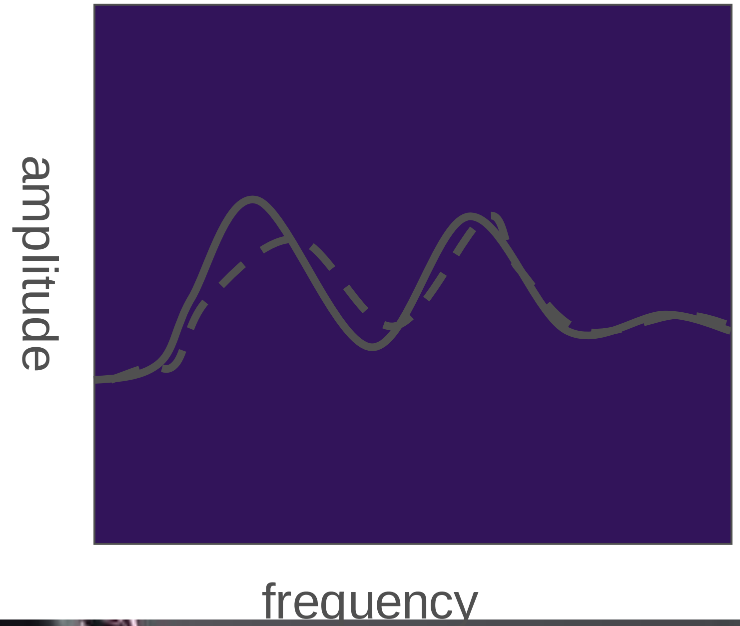
# Head Related Transfer Function

- Scattering causes frequency dependent amplification/attenuation
  - Effects can be of the order of tens of dB
  - Encodes location





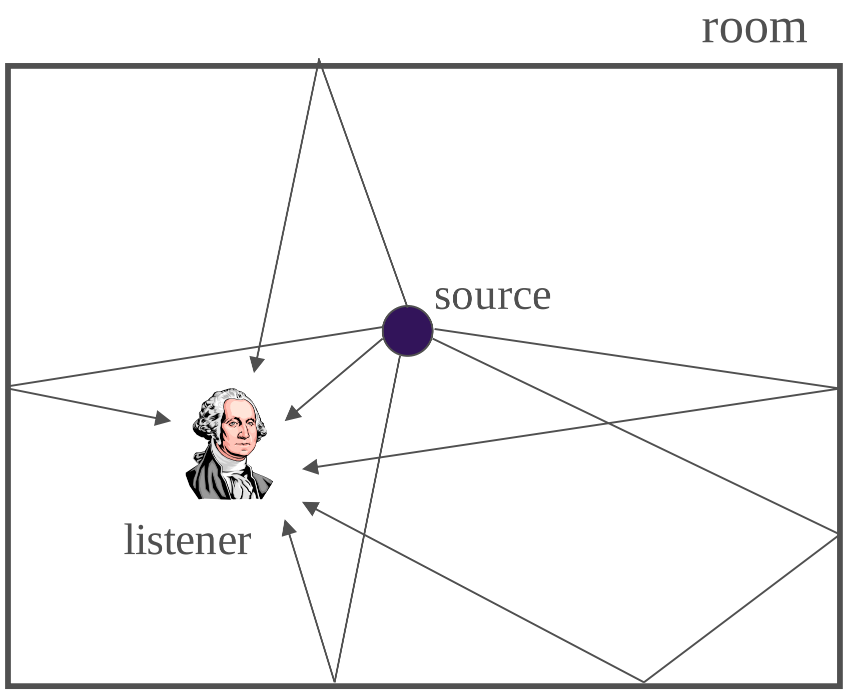



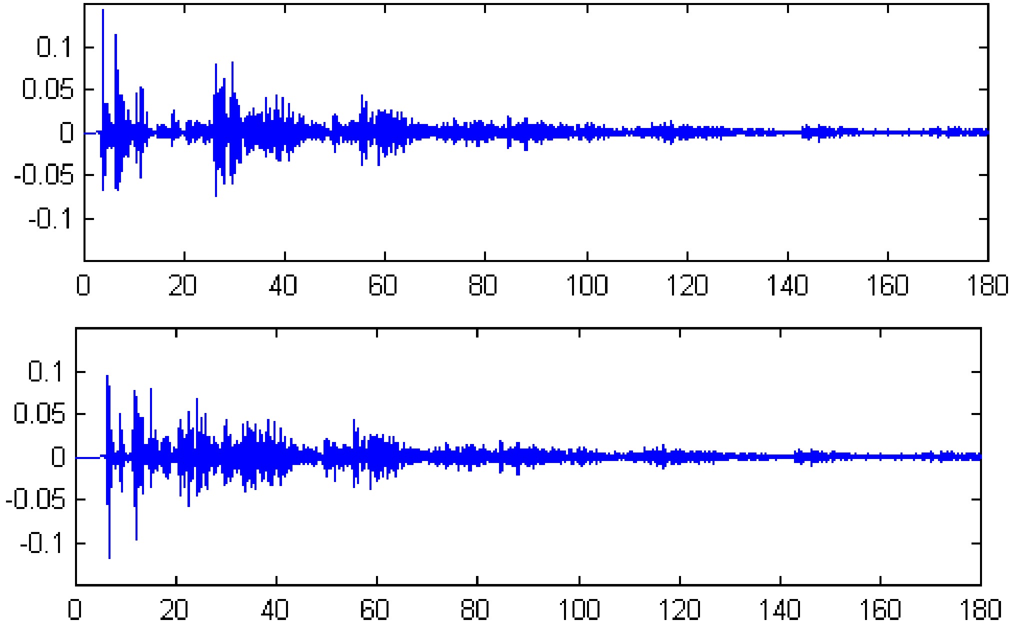



# Breaking up the Filter

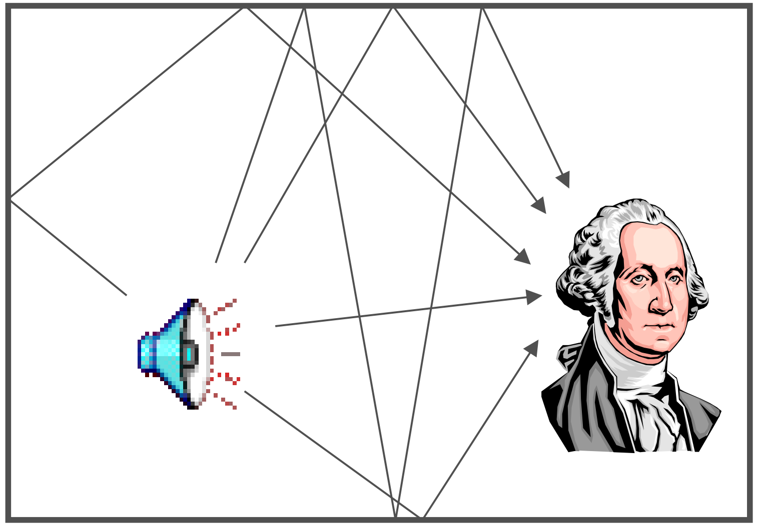
- Convolution is linear
- Early reflections are more important and time separated
  - Important for determining range
- Later reflections are a continuum
  - important for “spaciousness,” “envelopment,” “warmth,” etc.
- Create early reflections filter on the fly
  - reflections of up to 5th or 6th order (depending on computational resources)
  - These are convolved with their HRTF
- Tail of room impulse response is approximated depending on room size

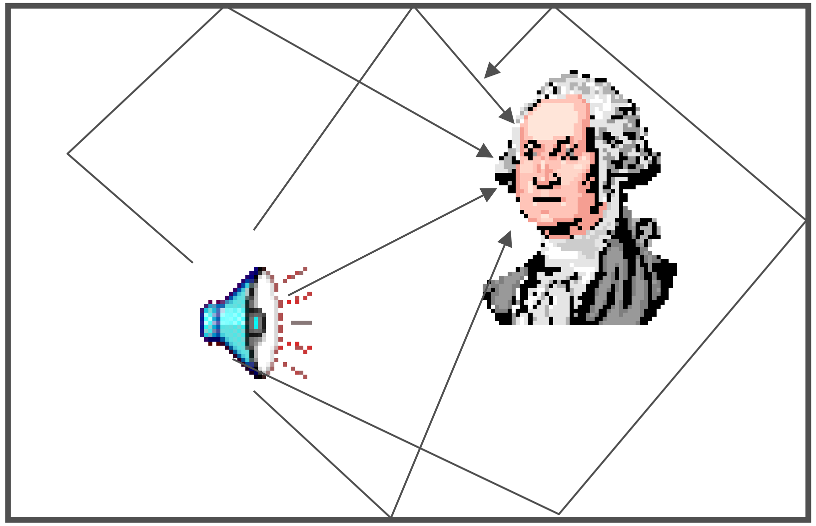
# Room response and HRTF



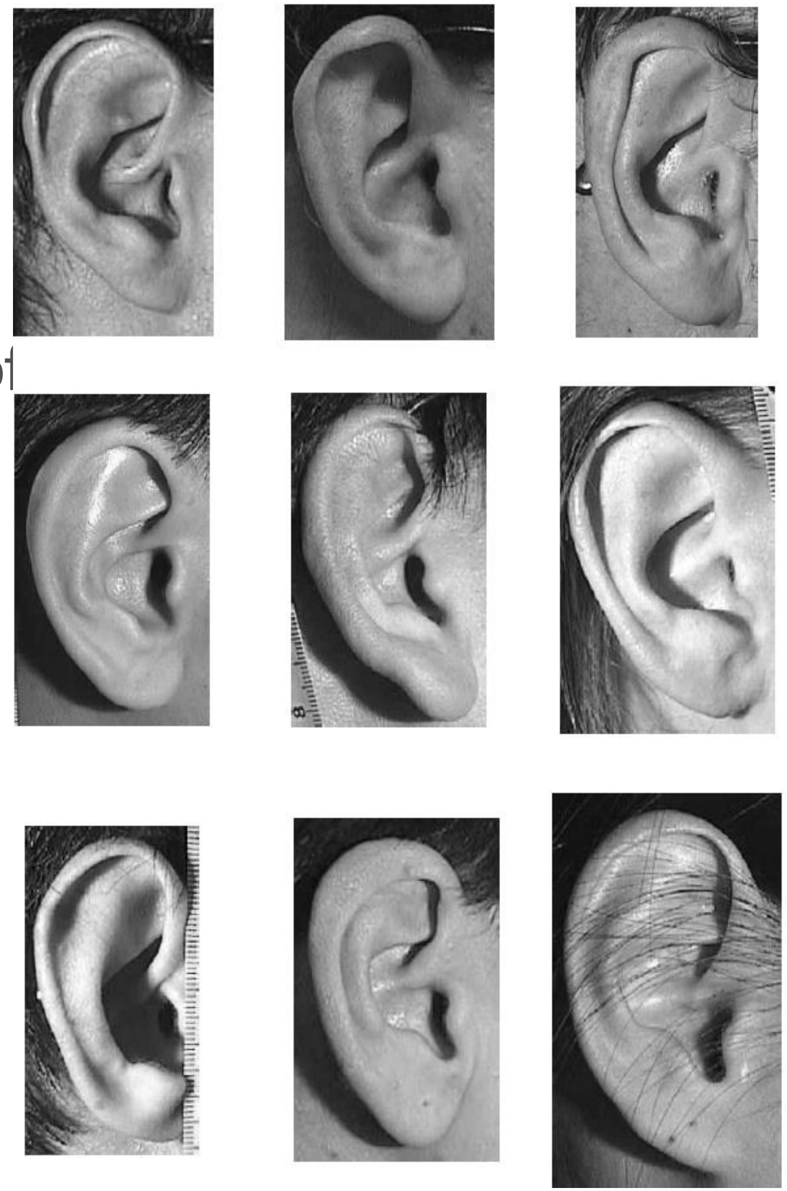

- Six to eight orders have perceptive live effect
- 30 orders influence the room ambience





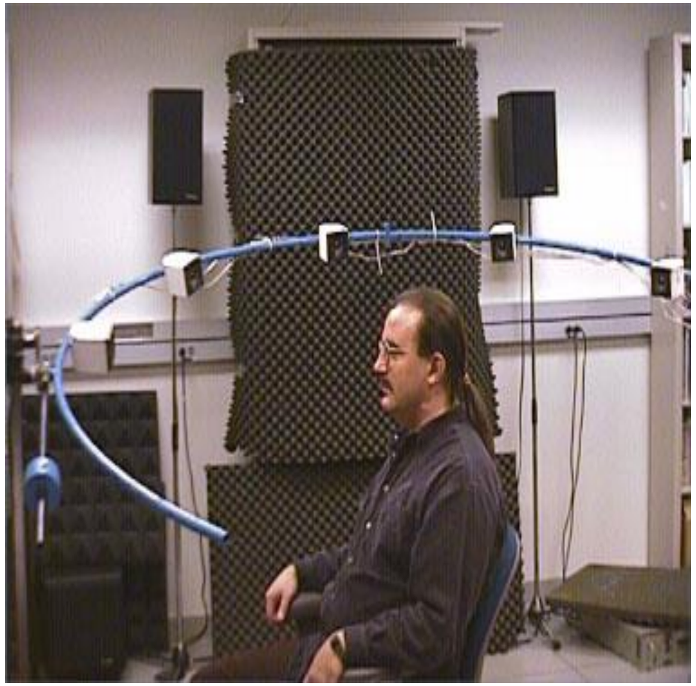
# HRTFs are very individual

- Humans have different sizes and shapes
- Ear shapes are very individual as well
  - **Before fingerprints, Alphonse Bertillon used a system of identification of**
- Even today ear shots are part of
  - **Mugshots & INS photographs**
- If ear shapes and body sizes are different
  - **Properties of scattered wave are different**
  - **HRTFs will be very individual**
- Need individual HRTFs for creating accurate virtual audio



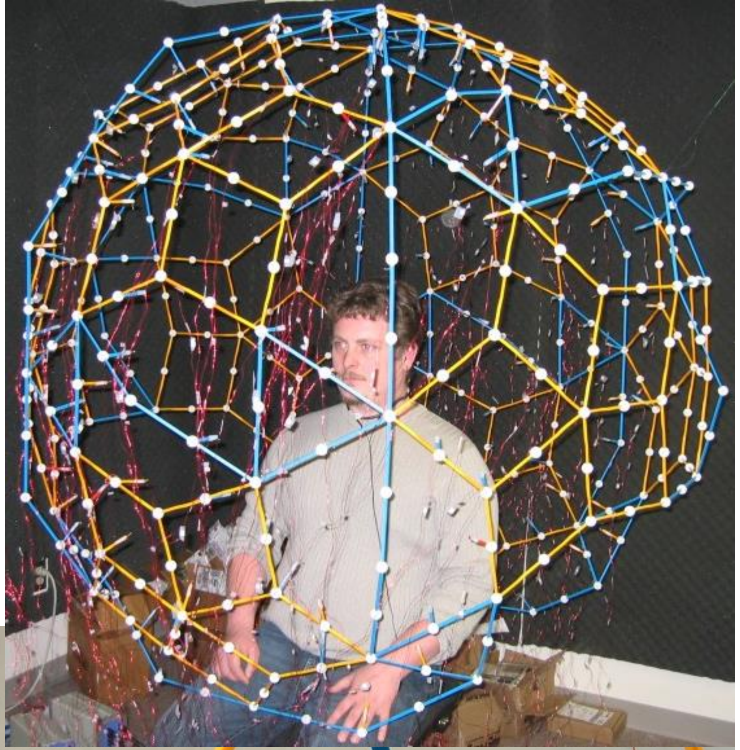
# Typically measured

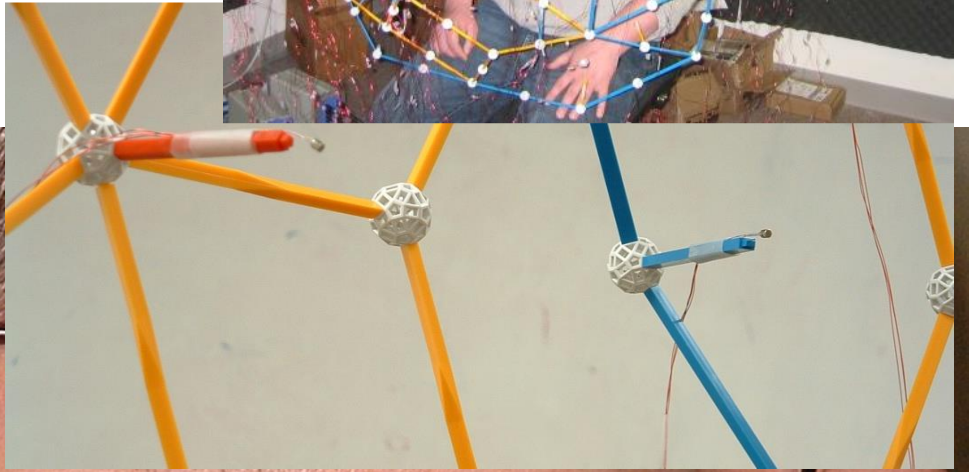
- Sound presented via speakers
- Speaker locations sampled
- Takes  10 minutes to several hours
- Subject given feedback to keep pose relatively steady
- Hoop is usually >1m away (no range data)

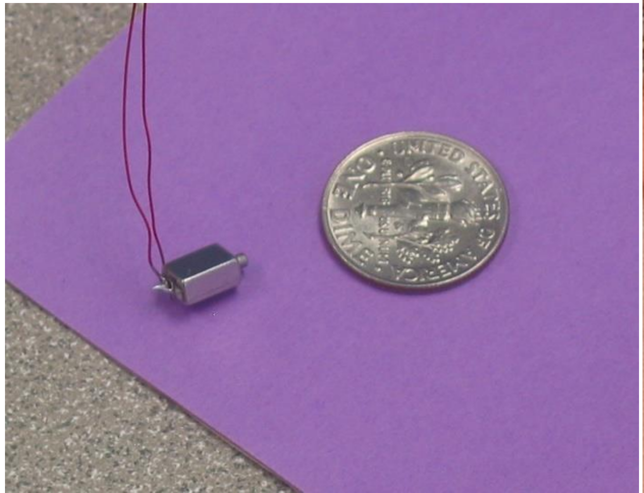


# Fast Approach

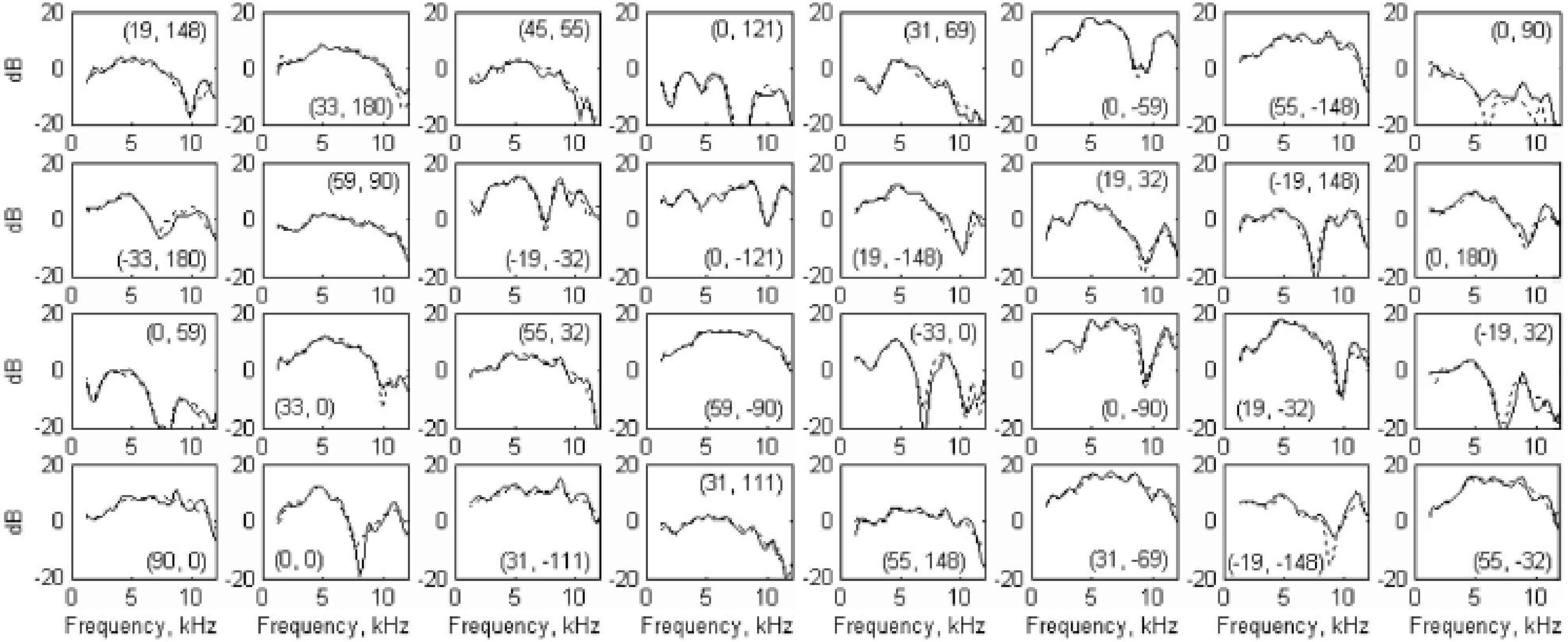
- Turned out headphone drivers
- Array of tiny microphones
- Send out a highpass signal and measure received signal
- Use analytical anthropometric representation for low frequencies
  and compose
- Extrapolate range







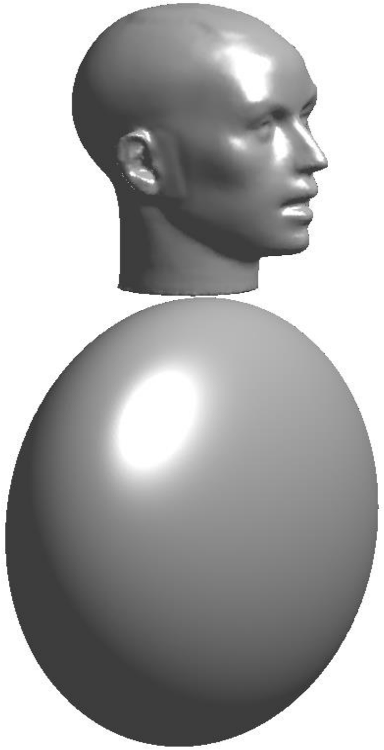
- Direct vs. Reciprocal (Zotkin et al. 2006, JASA)
- Currently reduced to under 30 s



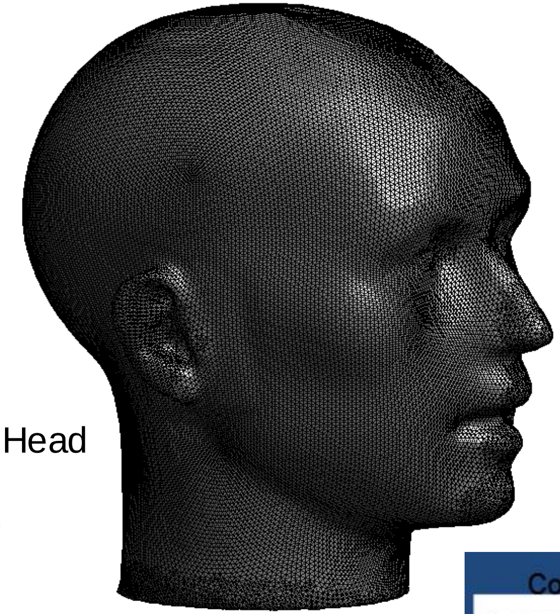

D.N. Zotkin, R. Duraiswami, E. Grassi, and N.A. Gumerov, "Fast head-related transfer function measurement via reciprocity," J. Acoust. Soc. Am., 120:2202-14, 2006

R. Duraiswami and N.A. Gumerov, "Representation, interpolation and measurement of head-related transfer functions," US Patent 7,720,229, 2010

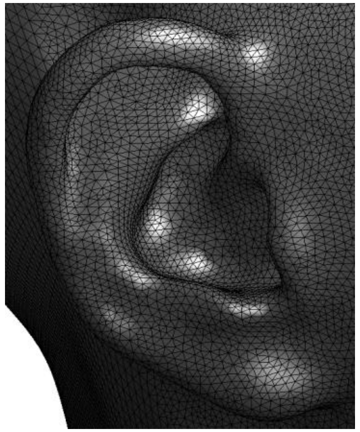
# Compute HRTFs via Fast Multipole Acelerated BEM



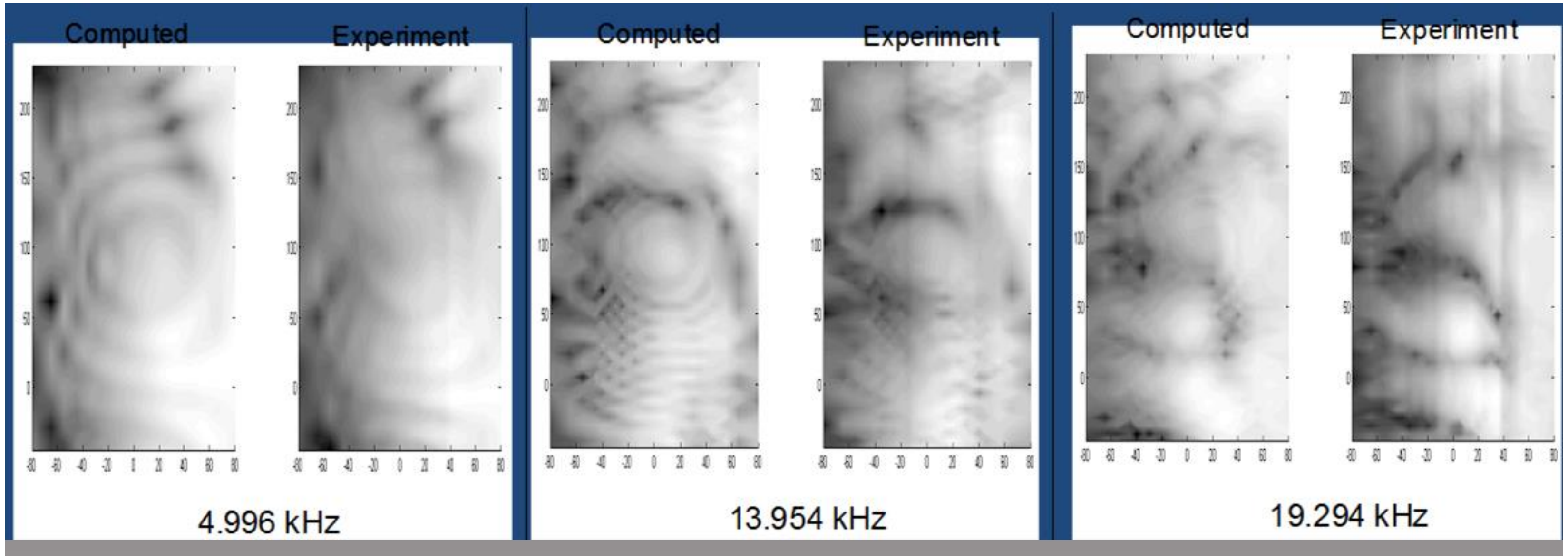
Head+Torso



Head



Pinnae



# Best Seat in the House: Telepresence

**RECORDING**

**PLAYBACK**

- Place microphones at a remote location (e.g. concert hall)
- Replay spatialized audio at a remote location
- Must play it for many users
- Use rendering algorithms/ representatons

# Representation via spherical wavefunctions

- sound at a point
  - So we can represent the sound at a point in terms of the local point-eigenfunctions of the Helmholtz equation

$$\psi_{in}(k;\mathbf{r}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} A_n^m R_n^m(k;\mathbf{r}),$$

$$R_n^m(k;\mathbf{r}) = j_n(kr) Y_n^m(\theta,\varphi),$$

  - Expand solutions in series, but truncate at *p* terms causing an error $\varepsilon_p$

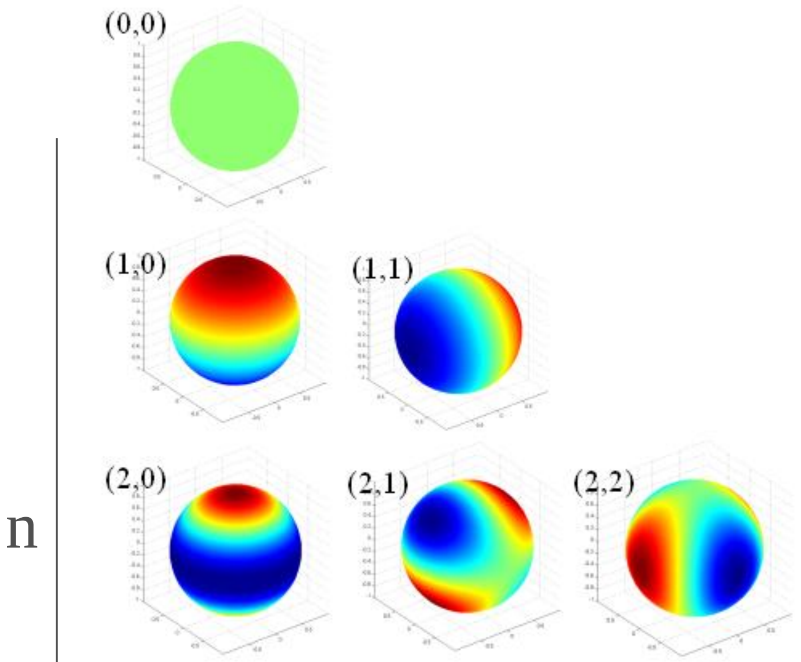
$$|\epsilon_p(\mathbf{s},\mathbf{r})| \lesssim \exp\left\{-\frac{1}{3}\left[2\frac{p-kR}{(kR)^{1/3}}\right]^{3/2}\right\} = \delta_p, \quad kR \gg 1.$$

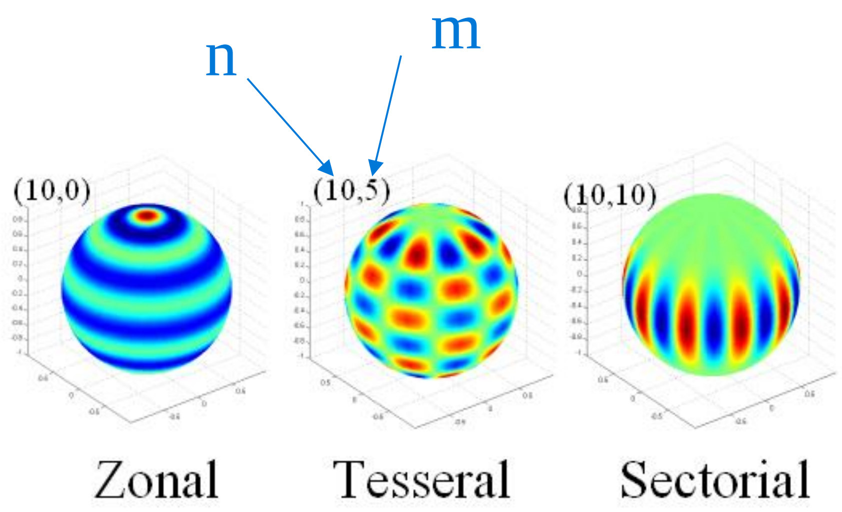
  - Error depends on frequency
    - For a given sound of wavenumber *k* this gives us minimum order for sensible representation

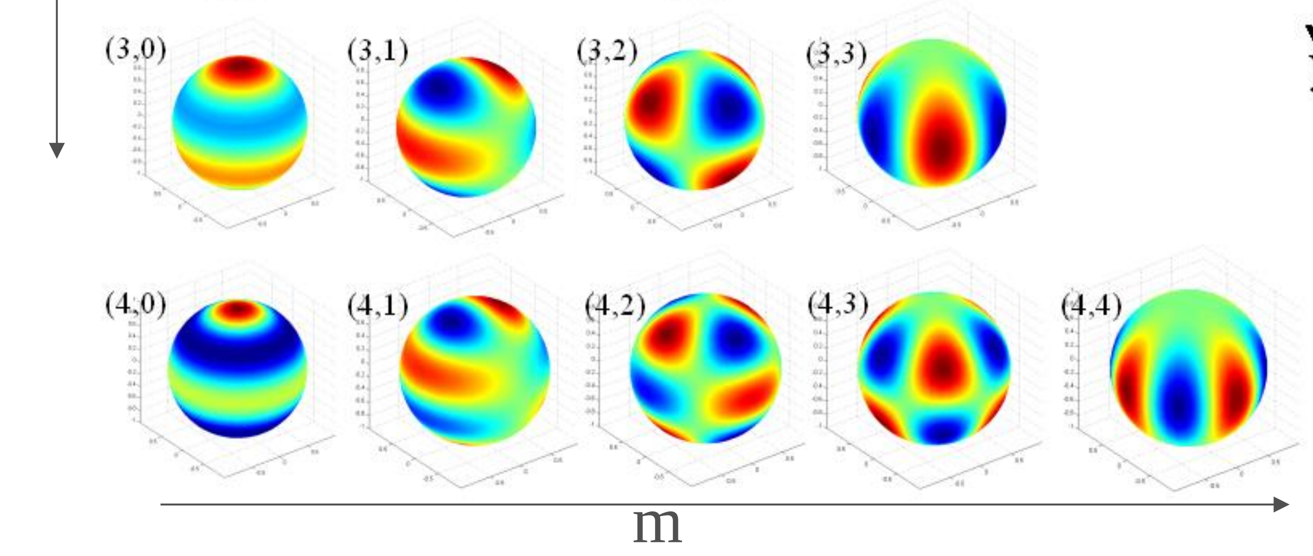
# Spherical Harmonics

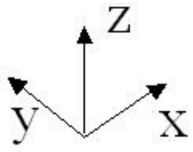
$$Y_n^m(\theta,\varphi) = (-1)^m \sqrt{\frac{2n+1}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}\, P_n^{|m|}(\cos\theta)e^{im\varphi},$$

$$n = 0, 1, 2, \ldots; \qquad m = -n, \ldots, n.$$




$$Y_n^{-m}(\theta,\varphi) = \overline{Y_n^m(\theta,\varphi)}.$$

$$Y_0^0(\theta,\varphi) = \text{const} = \sqrt{\frac{1}{4\pi}}.$$

# Yet another representation (Plane Waves)

- any soundfield in regular region can be expressed as an integral form of plane waves.
  - Integral over a unit sphere at the point
  - Decomposes any sound field in to a set of planewaves of various strengths

$$\psi_{in}(\mathbf{r}) = \frac{1}{4\pi} \int_{S_u} e^{ik\mathbf{s}\cdot\mathbf{r}} \mu_{in}(\mathbf{s})\, dS(\mathbf{s}),$$

Plane waves

Coeffs

  - Connected to spherical representation

$$e^{ik\mathbf{s}\cdot\mathbf{r}} = 4\pi \sum_{n=0}^{\infty} \sum_{m=-n}^{n} i^n Y_n^{-m}(\mathbf{s})\, R_n^m(\mathbf{r}), \quad R_n^m(\mathbf{r}) = \frac{i^{-n}}{4\pi} \int_{S_u} e^{ik\mathbf{s}\cdot\mathbf{r}} Y_n^m(\mathbf{s}) dS(\mathbf{s}),$$

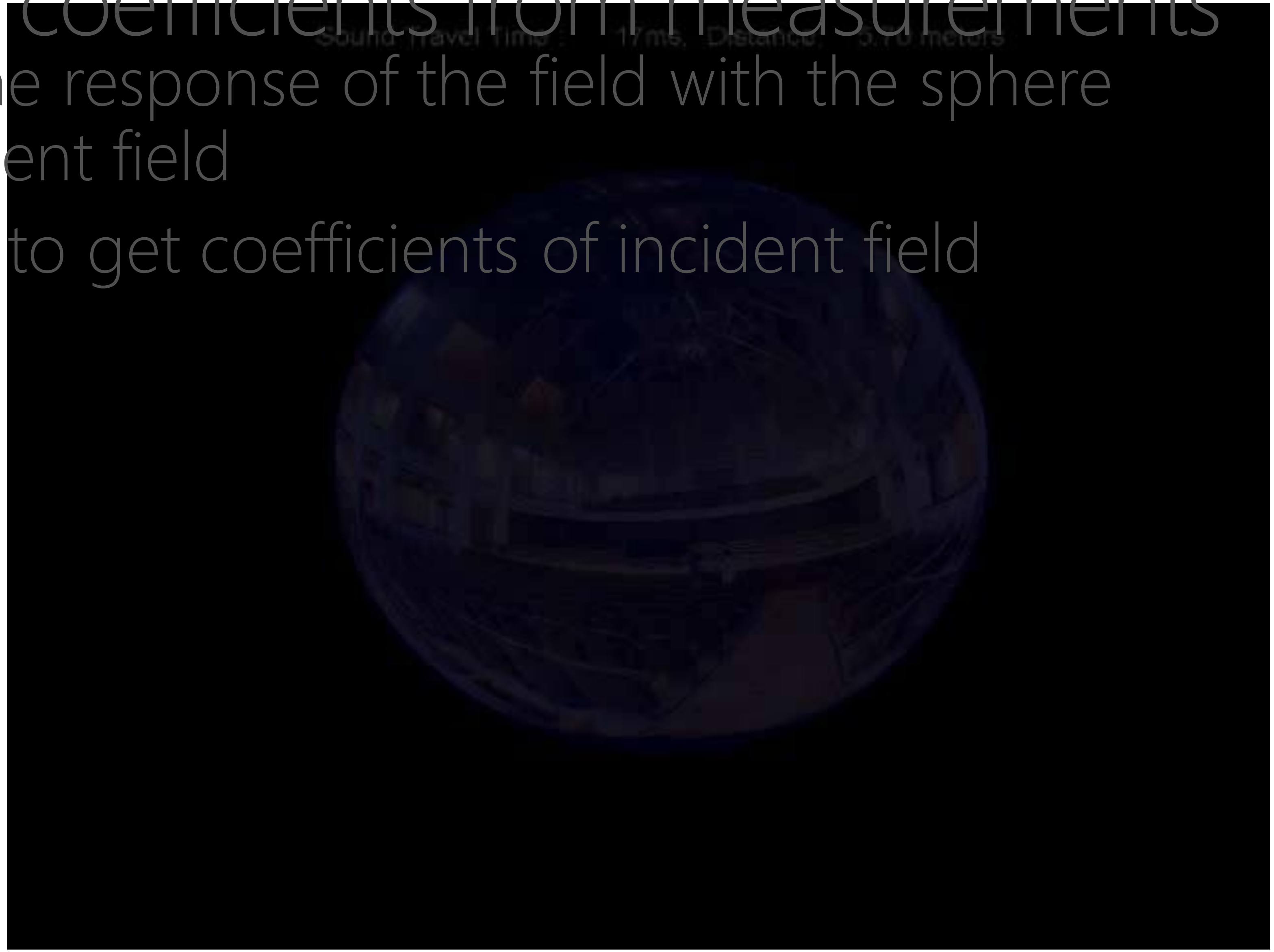
  - In practice these integrals are evaluated via quadrature

$$\int_{S_u} F(\mathbf{s})\, dS = \sum_{j=0}^{L_Q-1} F(\mathbf{s}_j)\, w_j, \; F(\mathbf{s}) = \sum_{n=0}^{p-1} \sum_{m=-n}^{n} C_n^m Y_n^m(\mathbf{s}),$$

  - Approximation error in this case is related to error in the quadrature
  - Quadrature error formula relates $L_Q$ to $p$

# Issues: Reconstruct coefficients from measurements

- What we measure is the response of the field with the sphere present – not the incident field
- Developed algorithms to get coefficients of incident field



# VisiSonics RealSpace3D Engine