Frontiers of VR I

Cinematic VR, spatial sound, and the vestibular system

Gordon Wetzstein
Stanford University

EE 267 Virtual Reality

Lecture 13

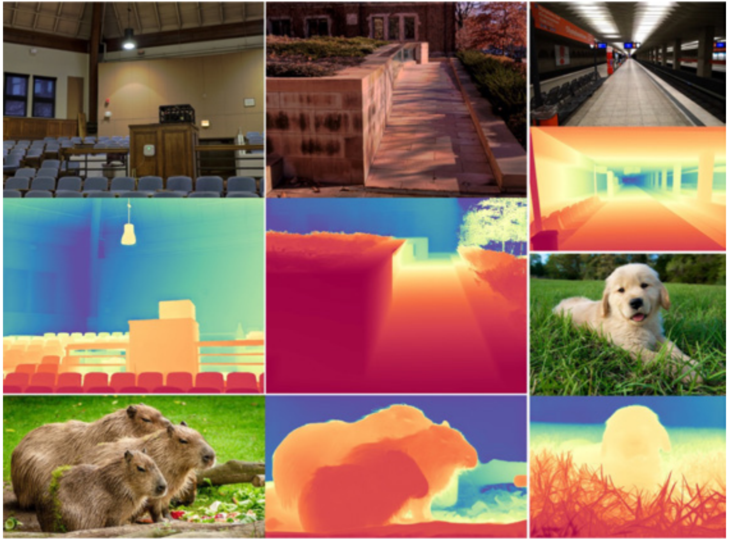
stanford.edu/class/ee267/

Apple Vision Pro – Spatial Video

Apple iPhone 15 pro max

How it probably works

Step 1: Monocular Depth Estimation



- Ranftl et al., "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer", 2019
- Ke et al,, "Marigold: Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation", CVPR 2024
- Gui et al., "DepthFM: Fast Monocular Depth Estimation with Flow Matching", 2024

Step 2: View Extrapolation (warp & inpaint)

- Srinivasan et al., "Pushing the Boundaries of View Extrapolation with Multiplane Images", CVPR 2019

Panoramic Imaging and Cinematic VR

Jaunt VR

Jaunt VR

Lytro

Lytro

Google

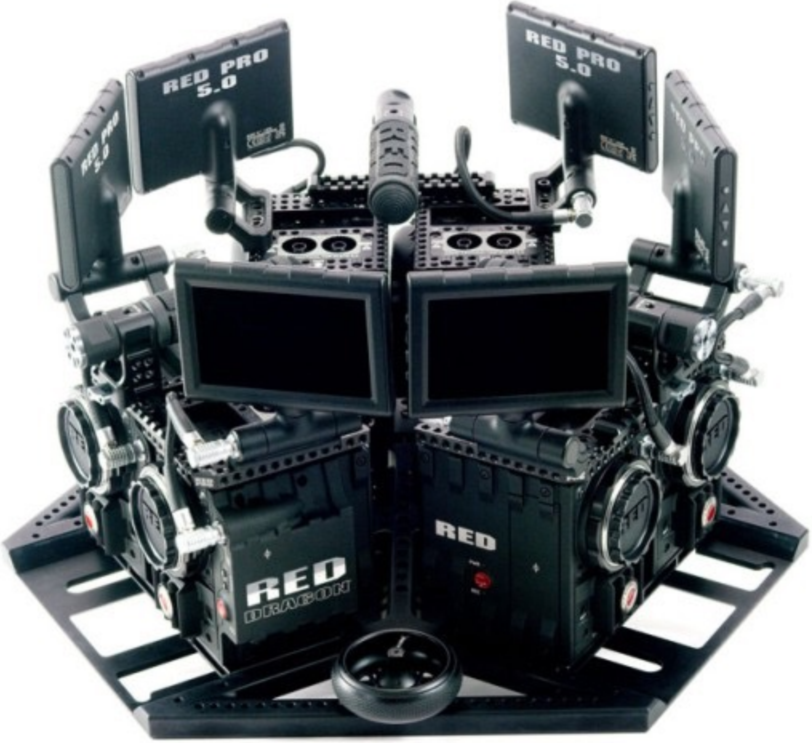
Nokia

Facebook

see Brian Cabral’s SCIEN talk @ talks.stanford.edu


facebook
SURROUND 360
5
161.6°
12
13
-13.6°
14
4
facebook
SURROUND 360

Red

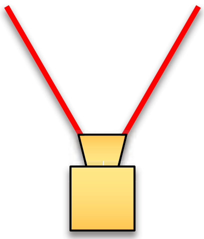


Samsung

Panorama v Stereo Movie v Stereo Panorama

Panorama

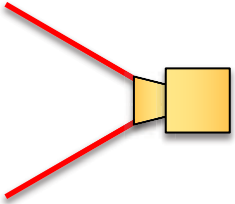
mono & head rotation



Panorama v Stereo Movie v Stereo Panorama

Panorama

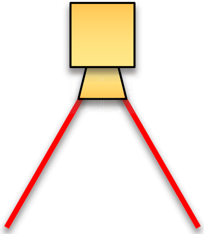
mono & head rotation



Panorama v Stereo Movie v Stereo Panorama

Panorama

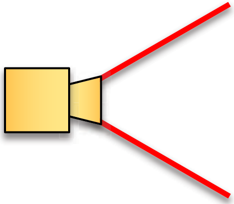
mono & head rotation



Panorama v Stereo Movie v Stereo Panorama

Panorama

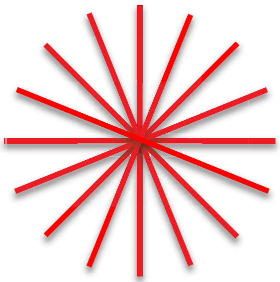
mono & head rotation



Panorama v Stereo Movie v Stereo Panorama

Panorama

mono & head rotation

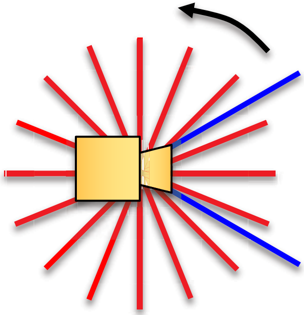


1 center of
projection!

Panorama v Stereo Movie v Stereo Panorama

Panorama

mono & head rotation

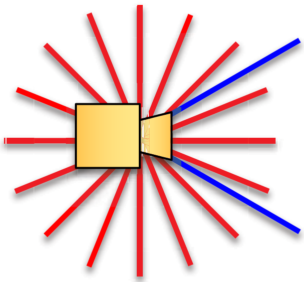


1 center of projection!

Panorama v Stereo Movie v Stereo Panorama

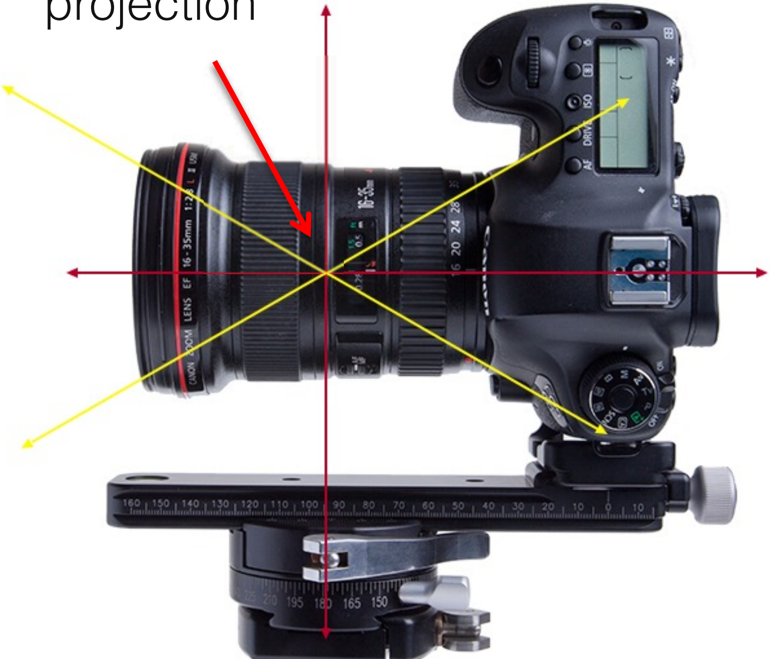
Panorama

mono & head rotation



1 center of projection!

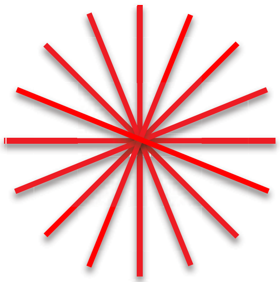
center of projection



Panorama v Stereo Movie v Stereo Panorama

Panorama

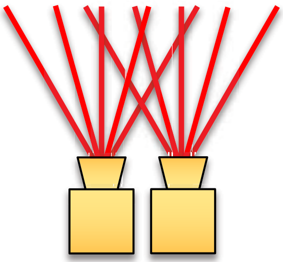
mono & head rotation



1 center of projection!

Stereo

stereo & no head rotation



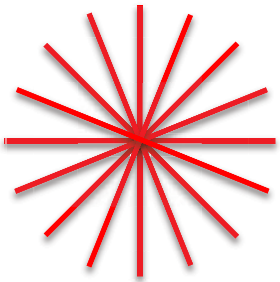
Stereo Panorama

stereo & head rotation

Panorama v Stereo Movie v Stereo Panorama

Panorama

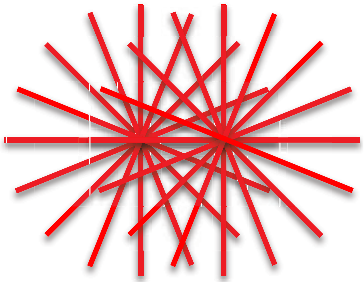
mono & head rotation



1 center of projection!

Stereo

stereo & no head rotation



2 centers of projection!

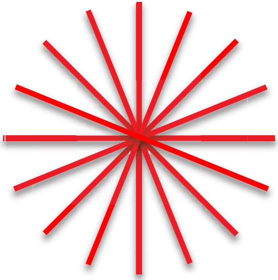
Stereo Panorama

stereo & head rotation

Panorama v Stereo Movie v Stereo Panorama

Panorama

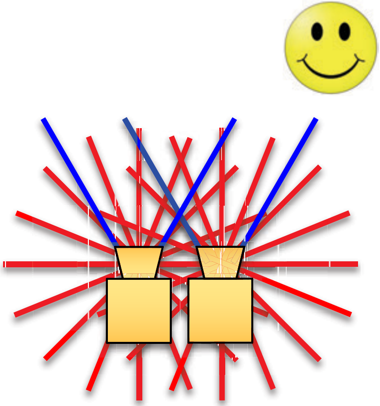
mono & head rotation



1 center of projection!

Stereo

stereo & no head rotation



2 centers of projection!

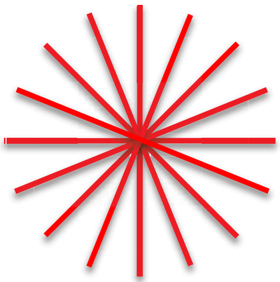
Stereo Panorama

stereo & head rotation

Panorama v Stereo Movie v Stereo Panorama

Panorama

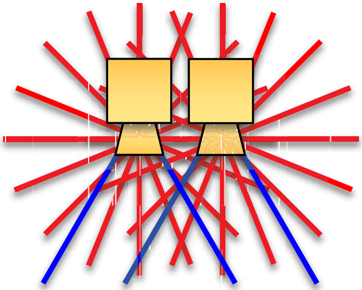
mono & head rotation



1 center of projection!

Stereo

stereo & no head rotation



2 centers of projection!

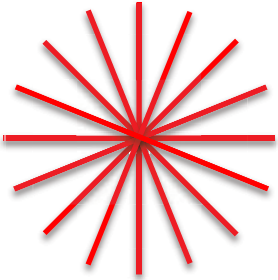
Stereo Panorama

stereo & head rotation

Panorama v Stereo Movie v Stereo Panorama

Panorama

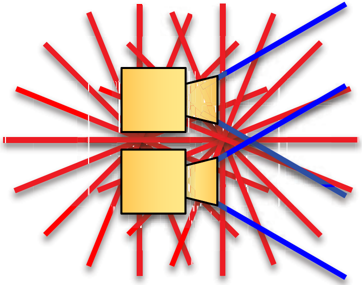
mono & head rotation



1 center of projection!

Stereo

stereo & no head rotation



2 centers of projection!

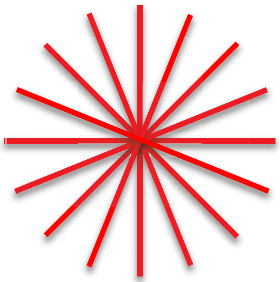
Stereo Panorama

stereo & head rotation

Panorama v Stereo Movie v Stereo Panorama

Panorama

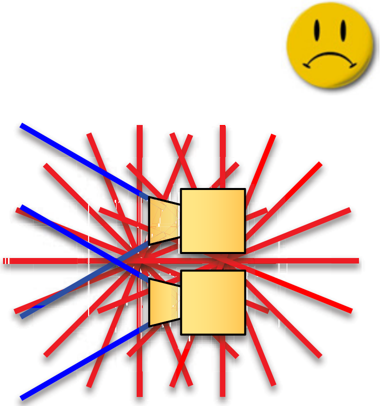
mono & head rotation



1 center of projection!

Stereo

stereo & no head rotation



2 centers of projection!

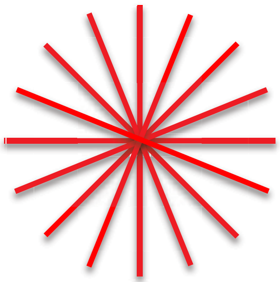
Stereo Panorama

stereo & head rotation

Panorama v Stereo Movie v Stereo Panorama

Panorama

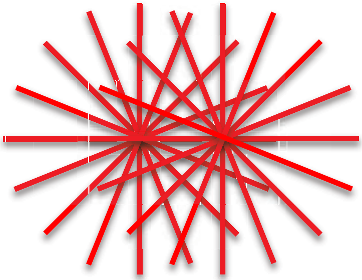
mono & head rotation



1 center of projection!

Stereo

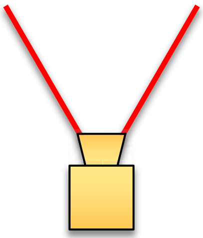
stereo & no head rotation



2 centers of projection!

Stereo Panorama

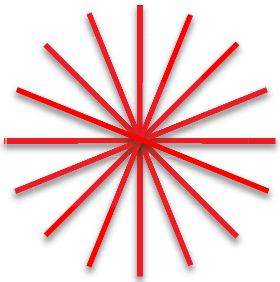
stereo & head rotation



Panorama v Stereo Movie v Stereo Panorama

Panorama

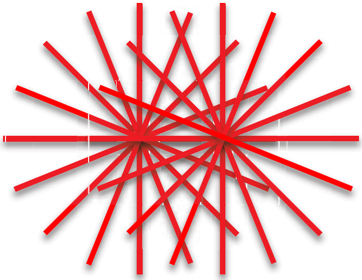
mono & head rotation



1 center of
projection!

Stereo

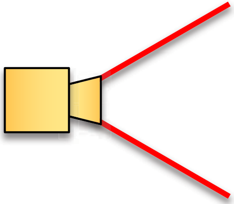
stereo & no head rotation



2 centers of
projection!

Stereo Panorama

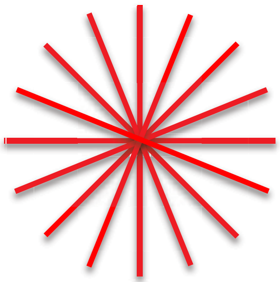
stereo & head rotation



Panorama v Stereo Movie v Stereo Panorama

Panorama

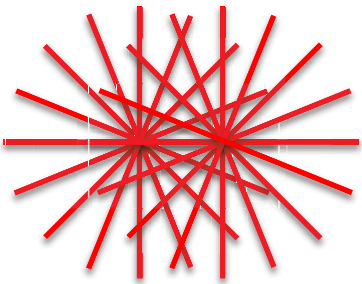
mono & head rotation



1 center of projection!

Stereo

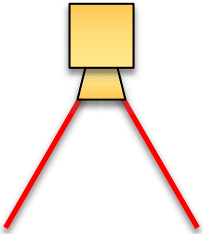
stereo & no head rotation



2 centers of projection!

Stereo Panorama

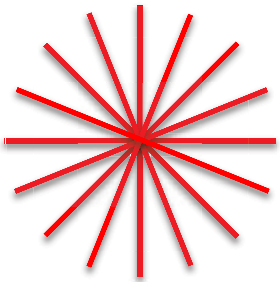
stereo & head rotation



Panorama v Stereo Movie v Stereo Panorama

Panorama

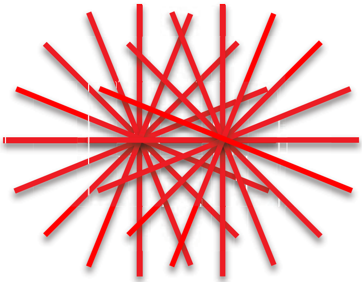
mono & head rotation



1 center of projection!

Stereo

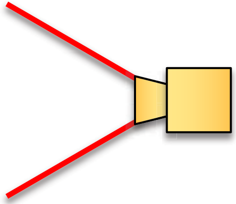
stereo & no head rotation



2 centers of projection!

Stereo Panorama

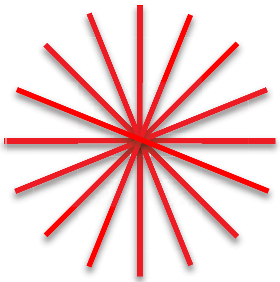
stereo & head rotation



Panorama v Stereo Movie v Stereo Panorama

Panorama

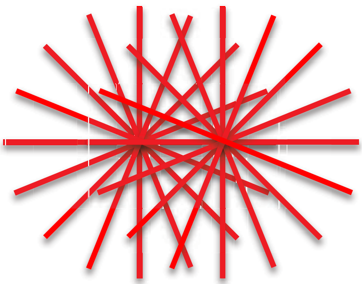
mono & head rotation



1 center of projection!

Stereo

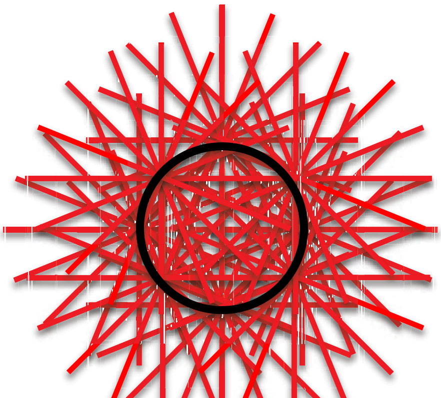
stereo & no head rotation



2 centers of projection!

Stereo Panorama

stereo & head rotation

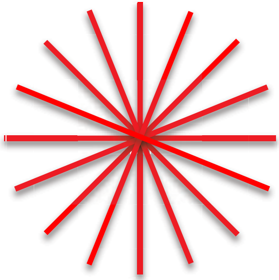


multiple centers of projection

Panorama v Stereo Movie v Stereo Panorama

Panorama

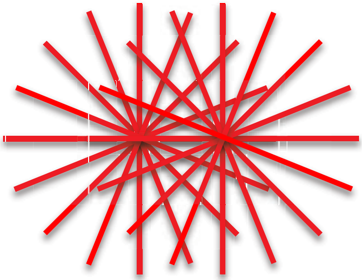
mono & head rotation



1 center of projection!

Stereo

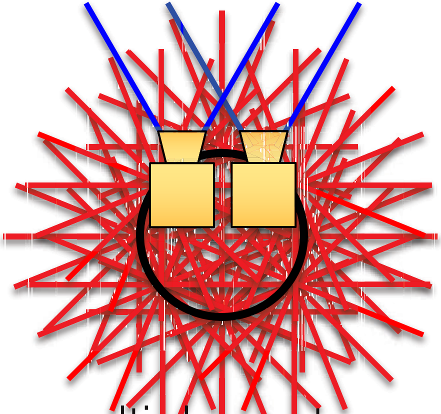
stereo & no head rotation



2 centers of projection!

Stereo Panorama

stereo & head rotation

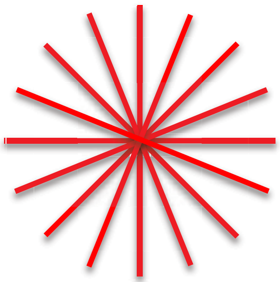


multiple centers of projection

Panorama v Stereo Movie v Stereo Panorama

Panorama

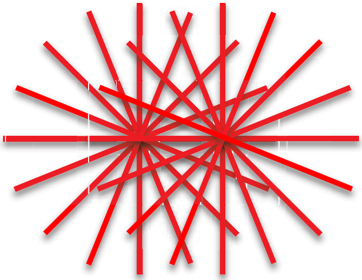
mono & head rotation



1 center of projection!

Stereo

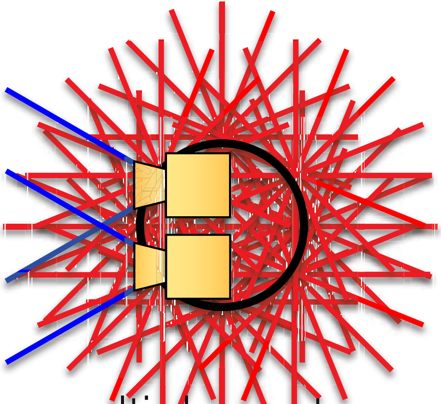
stereo & no head rotation



2 centers of projection!

Stereo Panorama

stereo & head rotation

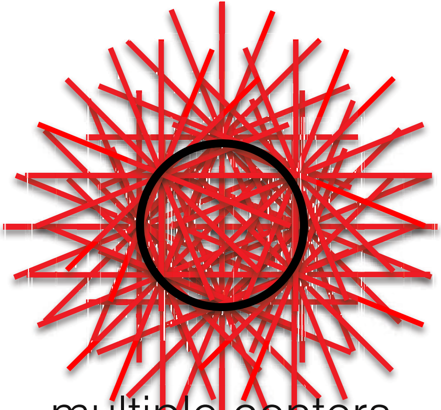


multiple centers of projection

Panorama v Stereo Movie v Stereo Panorama

Stereo Panorama

stereo & head rotation



Light Field!

multiple centers of projection

Panorama v Stereo Movie v Stereo Panorama

Panorama

mono & head rotation

Ricoh Theta

Stereo

stereo & no head rotation

Stereo Panorama

stereo & head rotation

horizontal-only parallax

Introduction to Spatial Sound

Overview

- what is sound? how do we synthesize it?
- the human auditory system
- stereophonic sound
- spatial audio of point sound sources
- surround sound
- ambisonics

What is Sound?

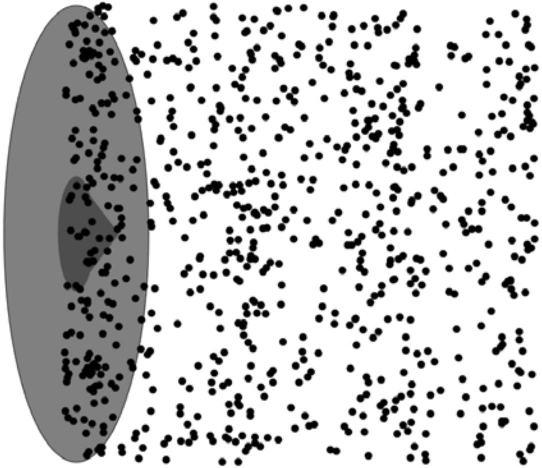
- “sound” is a pressure wave propagating in a medium
- speed of sound is $c = \sqrt{K/\rho}$ where c is velocity, ρ is density of medium and K is elastic bulk modulus
- in air, speed of sound is 340 m/s
- in water, speed of sound is 1,483 m/s

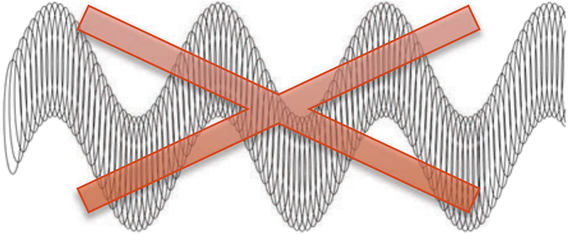
How do we Synthesize Sound?

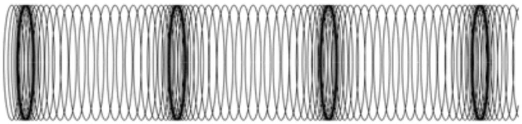
https://www.youtube.com/watch?v=aDrs6EieFCM

Producing Sound

- Sound is longitudinal vibration of air particles
- Speakers create wavefronts by physically compressing the air, much like one could a slinky

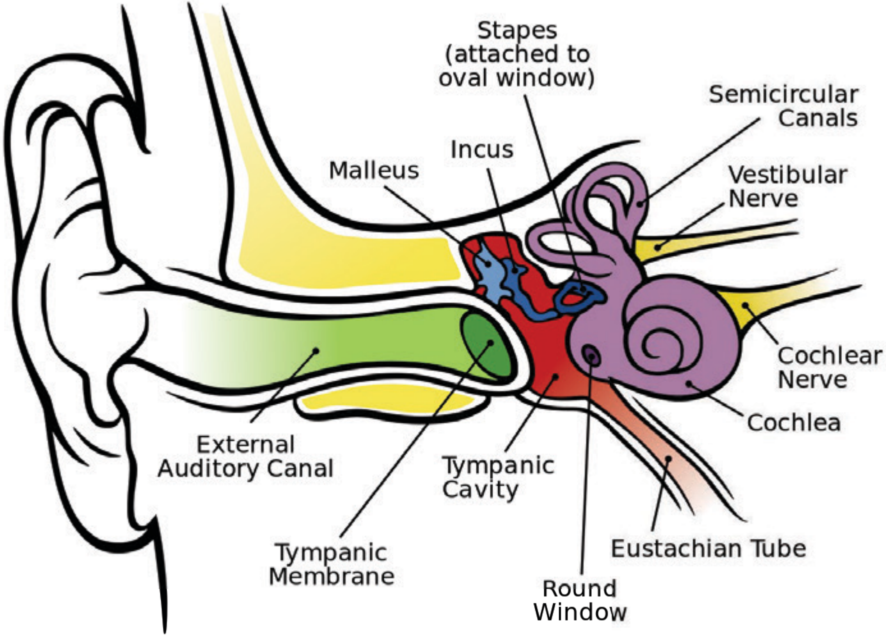






The Human Auditory System

pinna

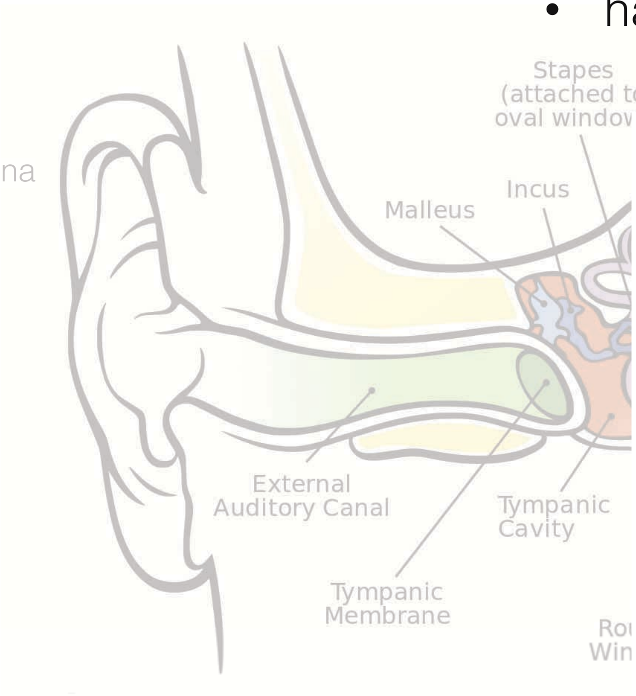


wikipedia

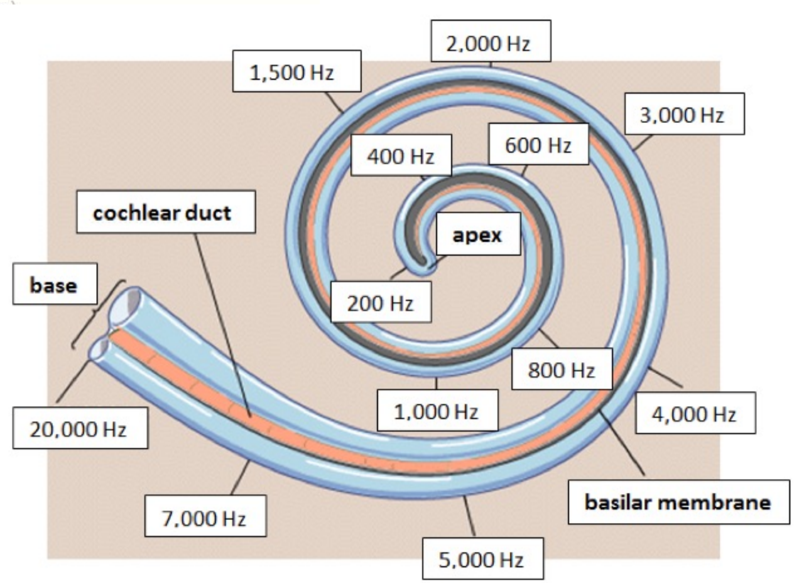
The Human Auditory System

- hair receptor cells pick up vibrations

pinna



cochlea

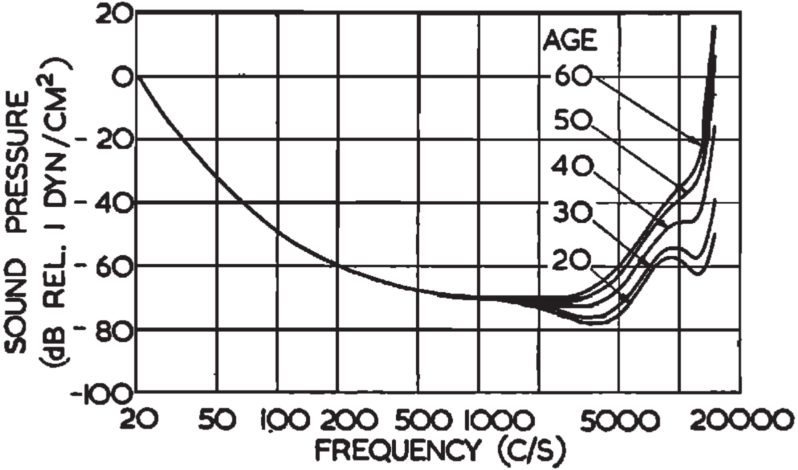


wikipedia

The Human Auditory System

- Human hearing range: ~20–20,000 Hz
- Variation between individuals
- Degrades with age

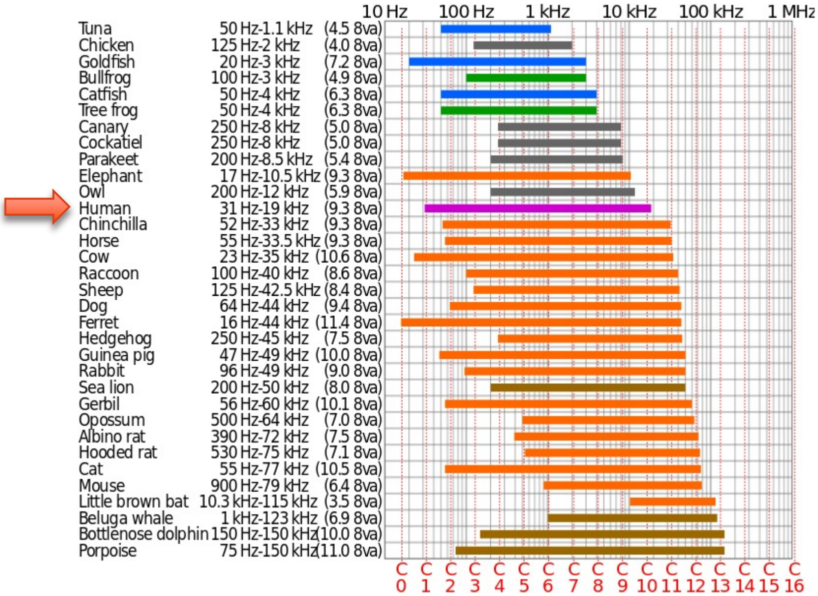
Hearing Threshold in Quiet



D. W. Robinson and R. S. Dadson, 1957

The Human Auditory System

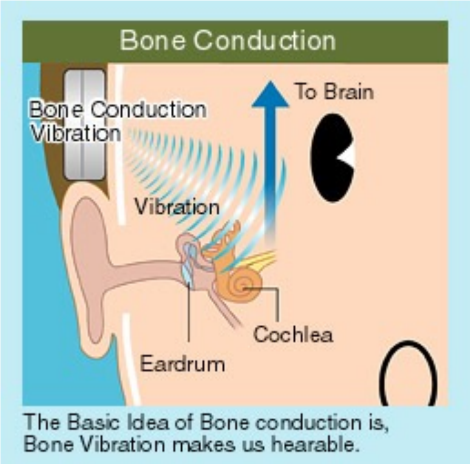
- human hearing range: ~20 – 20,000 Hz
- variation between individuals and changes with age



wikipedia

Bone Conduction

- can stimulate eardrum mechanically to create the illusion of audio, e.g. with bone conduction

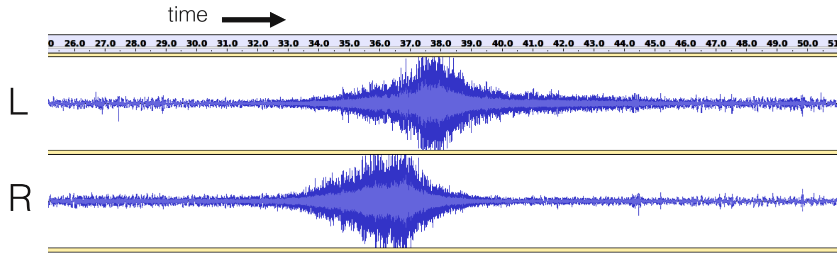


http://www.goldendance.co.jp/English/boneconduct/01.html

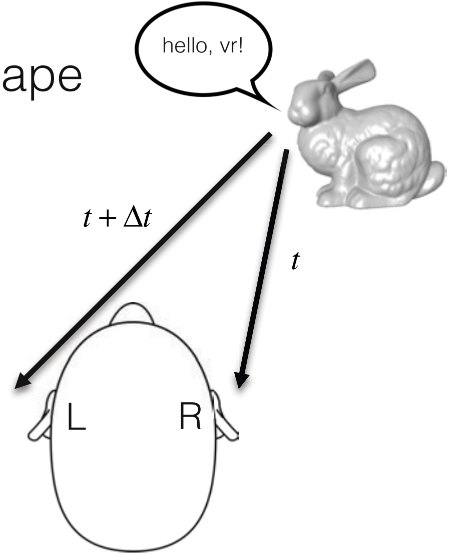
the verge

Stereophonic Sound

- mainly captures differences between the ears:
  - interaural time difference
  - amplitude differences from body shape (nose, head, neck, shoulders, …)



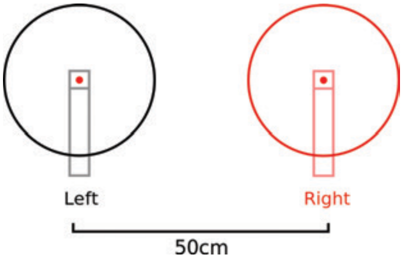
wikipedia



Stereophonic Sound Recording

- use two microphones
- A-B techniques captures differences in time-of-arrival
- other configurations work too, capture differences in amplitude

Olympus



wikipedi

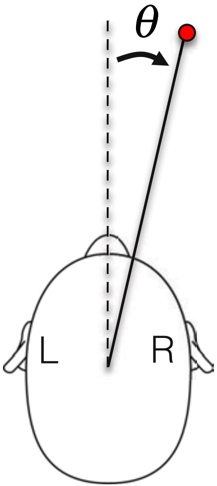
Rode

X-Y technique

Head-related Impulse Response (HRIR)

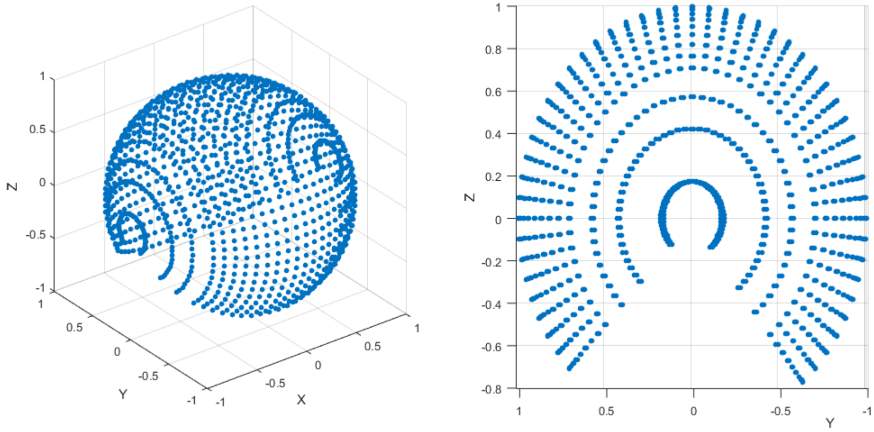
- models phase and amplitude differences for all possible sound directions parameterized by azimuth θ and elevation ϕ
- can be measured with two microphones in ears of mannequin & speakers all around

Zhong and Xie, “Head-Related Transfer Functions and Virtual Auditory Display”



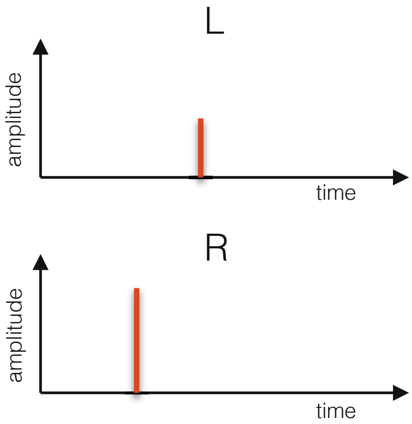
Head-related Impulse Response (HRIR)

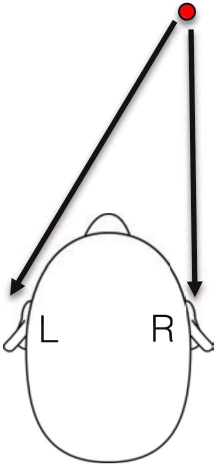
- CIPIC HRTF database: http://interface.cipic.ucdavis.edu/sound/hrtf.html
- elevation: -45° to 230.625°, azimuth: -80° to 80°
- need to interpolate between discretely sampled directions



Head-related Impulse Response (HRIR)

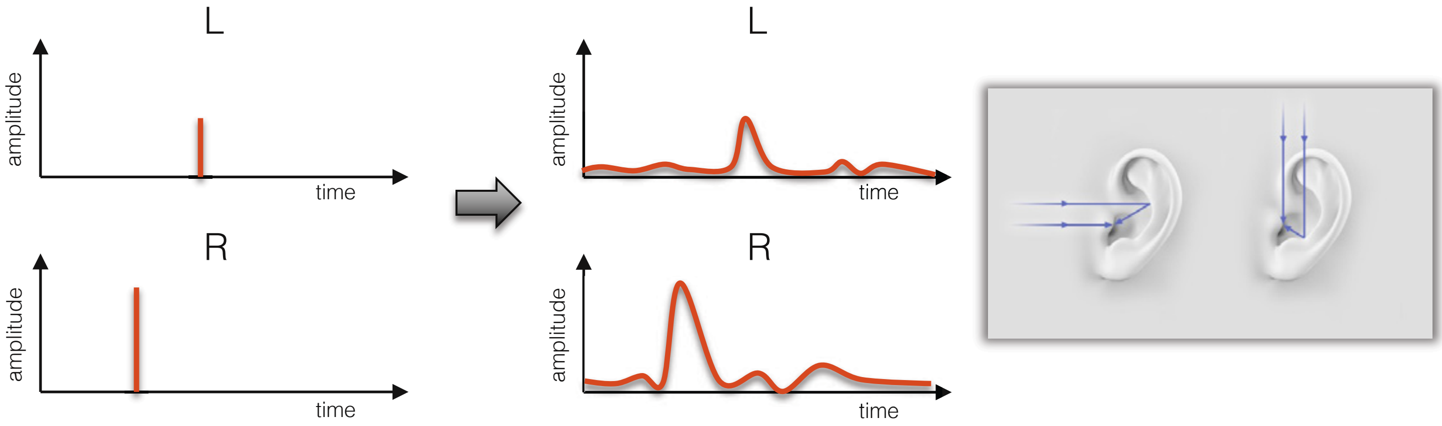
- measuring the HRIR
  - ideal case: scaled & shifted Dirac peaks





Head-related Impulse Response (HRIR)

- measuring the HRIR
  - ideal case: scaled & shifted Dirac peaks
  - in practice: more complicated, includes scattering in the ear, sholders etc.



Head-related Impulse Response (HRIR)

- measuring the HRIR
    - need one temporally-varying function for each angle
    - total of $2 \cdot N_\theta \cdot N_\phi \cdot N_t$ samples, where $N_{\theta,\phi,t}$ is the number of samples for azimuth, elevation, and time, respectively

$$hrir\_l(\theta,\phi,t)$$

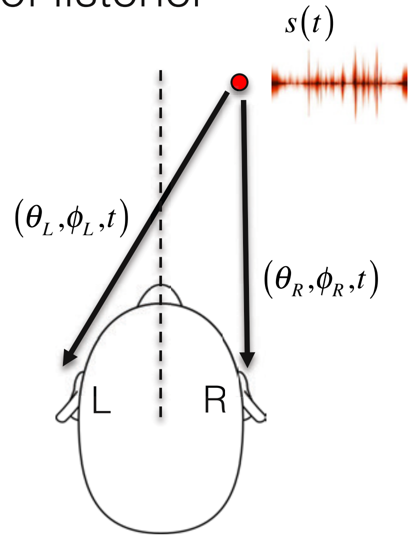
$$hrir\_r(\theta,\phi,t)$$

Head-related Impulse Response (HRIR)

applying the HRIR:

- given a mono sound source $s(t)$ and it's 3D position

1. compute (θ_L, ϕ_L) and (θ_R, ϕ_R) relative to center of listener

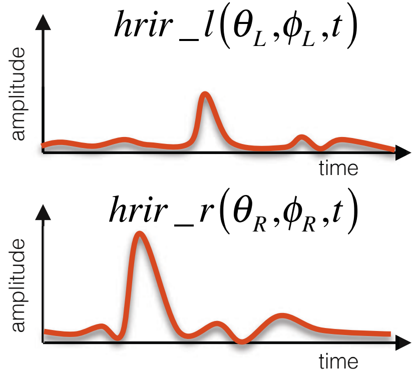


Head-related Impulse Response (HRIR)

applying the HRIR:

- given a mono sound source $s(t)$ and it's 3D position

1. compute (θ_L, ϕ_L) and (θ_R, ϕ_R) relative to center of listener
2. look up measured HRIR for left and right ear at these angles

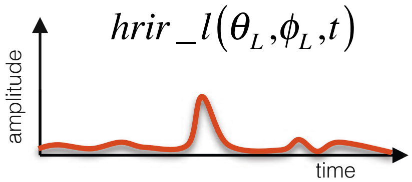


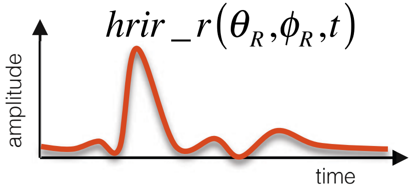
Head-related Impulse Response (HRIR)

applying the HRIR:

- given a mono sound source $s(t)$ and it's 3D position

1. compute (θ_L, ϕ_L) and (θ_R, ϕ_R) relative to center of listener
2. look up measured HRIR for left and right ear at these angles
3. convolve signal with HRIRs to get response for each ear as





$$s_L(t) = hrir\_l(\theta_L, \phi_L, t) * s(t)$$

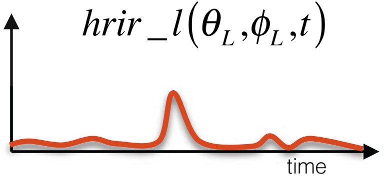
$$s_R(t) = hrir\_r(\theta_R, \phi_R, t) * s(t)$$

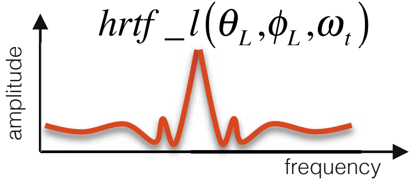
Head-related Transfer Function (HRTF)

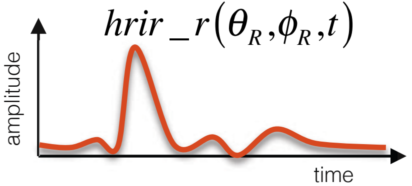
- HRTF is Fourier transform of HRIR! (you'll find the term HRTF more often that HRIR)

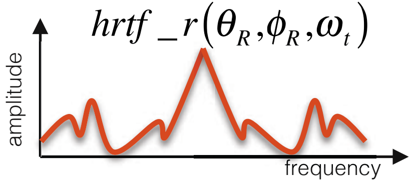
$$s_L(t) = hrir\_l(\theta_L, \phi_L, t) * s(t)$$
$$s_R(t) = hrir\_r(\theta_R, \phi_R, t) * s(t)$$

$$s_L(t) = F^{-1}\left\{hrtf\_l(\theta_L, \phi_L, \omega_t) \cdot F\{s(t)\}\right\}$$
$$s_R(t) = F^{-1}\left\{hrtf\_r(\theta_R, \phi_R, \omega_t) \cdot F\{s(t)\}\right\}$$









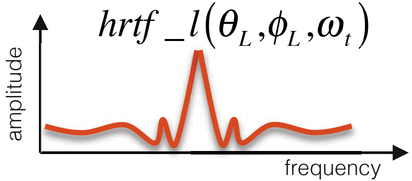
Head-related Transfer Function (HRTF)

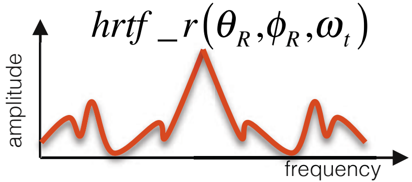
- HRTF is Fourier transform of HRIR! (you'll find the term HRTF more often that HRIR)

$$s_L(t) = hrir\_l(\theta_L, \phi_L, t) * s(t)$$
$$s_R(t) = hrir\_r(\theta_R, \phi_R, t) * s(t)$$

$$s_L(t) = F^{-1}\{hrtf\_l(\theta_L, \phi_L, \omega_t) \cdot F\{s(t)\}\}$$
$$s_R(t) = F^{-1}\{hrtf\_r(\theta_R, \phi_R, \omega_t) \cdot F\{s(t)\}\}$$

convolution theorem



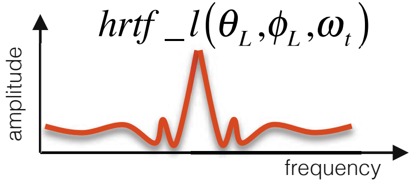


Head-related Transfer Function (HRTF)

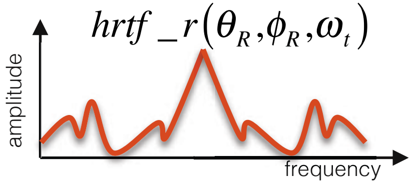
- HRTF is Fourier transform of HRIR! (you'll find the term HRTF more often that HRIR)

$$s_L(t) = hrir\_l(\theta_L, \phi_L, t) * s(t)$$
$$s_R(t) = hrir\_r(\theta_R, \phi_R, t) * s(t)$$

$$s_L(t) = F^{-1}\{hrtf\_l(\theta_L, \phi_L, \omega_t) \cdot F\{s(t)\}\}$$
$$s_R(t) = F^{-1}\{hrtf\_r(\theta_R, \phi_R, \omega_t) \cdot F\{s(t)\}\}$$



- properties of HRTF:
  - complex-valued
  - symmetric (because HRIR is real-valued)



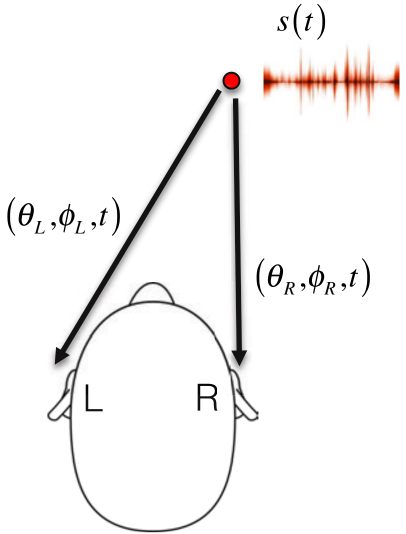
Head-related Transfer Function (HRTF)

$$s_L(t) = F^{-1}\left\{hrtf\_l(\theta_L, \phi_L, \omega_t) \cdot F\left\{s(t)\right\}\right\}$$

$$s_R(t) = F^{-1}\left\{hrtf\_r(\theta_R, \phi_R, \omega_t) \cdot F\left\{s(t)\right\}\right\}$$

Spatial Sound of 1 Point Sound Source

- given *s(t)* and 3D position, follow instructions from last slides by convolving Fourier transform of *s* with HRTFs for each each

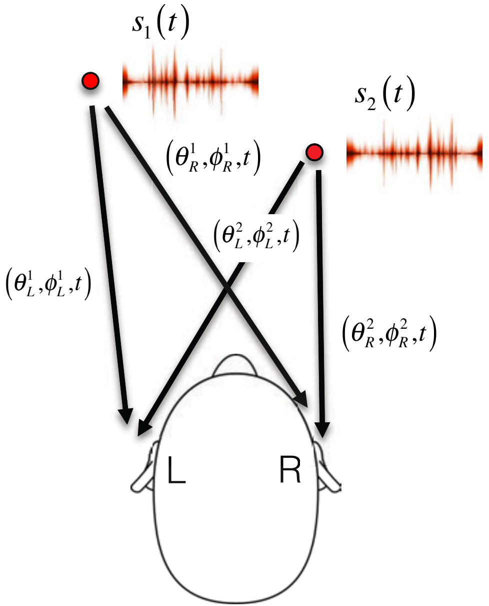


Spatial Sound of N Point Sound Sources

- superposition principle holds, so just sum the contributions of each

$$s_L(t) = \sum_{i=1}^{N} F^{-1}\left\{hrtf\_l\left(\theta_L^i, \phi_L^i, \omega_t\right) \cdot F\left\{s_i(t)\right\}\right\}$$

$$s_R(t) = \sum_{i=1}^{N} F^{-1}\left\{hrtf\_r\left(\theta_R^i, \phi_R^i, \omega_t\right) \cdot F\left\{s_i(t)\right\}\right\}$$



Surround Sound

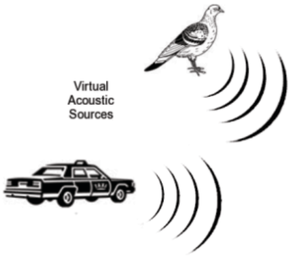
- approximate continuous wave field with discrete set of speakers

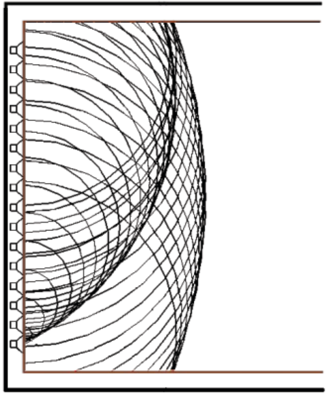
- most common:
  5.1 surround sound =
  5 (channels) . 1 (bass)

→ 6 channels total

Surround Sound

- approximate continuous wave field with discrete set of speakers
- can also use more speakers for “wave field synthesis” (i.e. audio hologram)





http://spatialaudio.net/

ucsb

Surround Sound

- approximate continuous wave field with discrete set of speakers
- can also use more speakers for “wave field synthesis” (i.e. audio hologram)
- for wave field synthesis, phase of speakers needs to be synchronized, i.e. a phased array!

Surround Sound & HRTF

- for all speaker-based (surround) sound, we don’t need an HRTF because the ears of the listener will apply them!
- speaker setup usually needs to be calibrated

Spatial Audio for VR

- VR/AR requires us to re-think audio, especially spatial audio!

- could use 5.1 surround sound and set up “virtual speakers” in the virtual environment – can use existing content, but not super easy to capture new content; also doesn’t capture directionality from above/below

Spatial Audio for VR

Two primary approaches:

1. Real-time sound engine
    - render 3D sound sources via HRTF in real-time, just as discussed in the previous slides
    - used for games and synthetic virtual environments
    - a lot of libraries available: FMOD, OpenAL, …

Spatial Audio for VR

Two primary approaches:

2. Spatial sound recorded from real environments
    - most widely used format now: ambisonics
    - simple microphones exist
    - relatively easy mathematical model
    - only need 4 channels for starters
    - used in YouTube VR and many other platforms

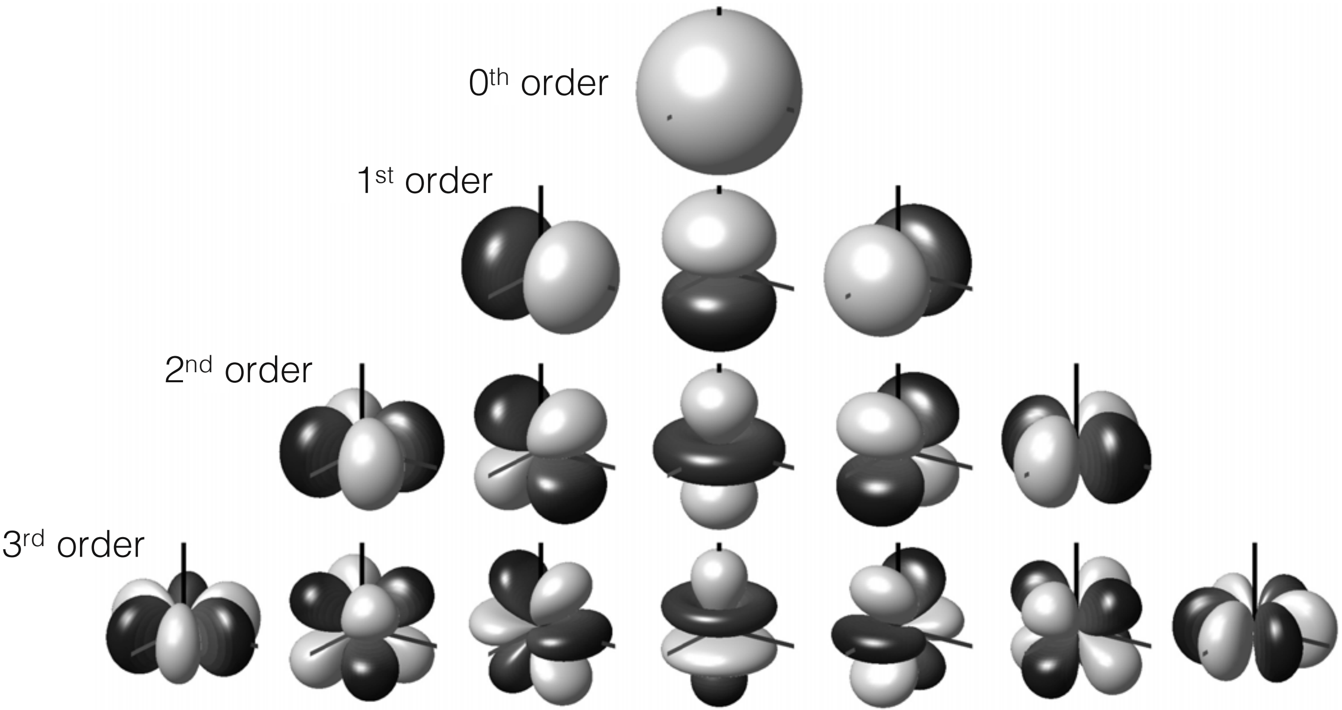
Ambisonics

- idea: represent sound incident at a point (i.e. the listener) with some directional information
- using all angles θ,ϕ is impractical – need too many sound channels (one for each direction)
- some lower-frequency (in direction) components may be sufficient → directional basis representation to the rescue!

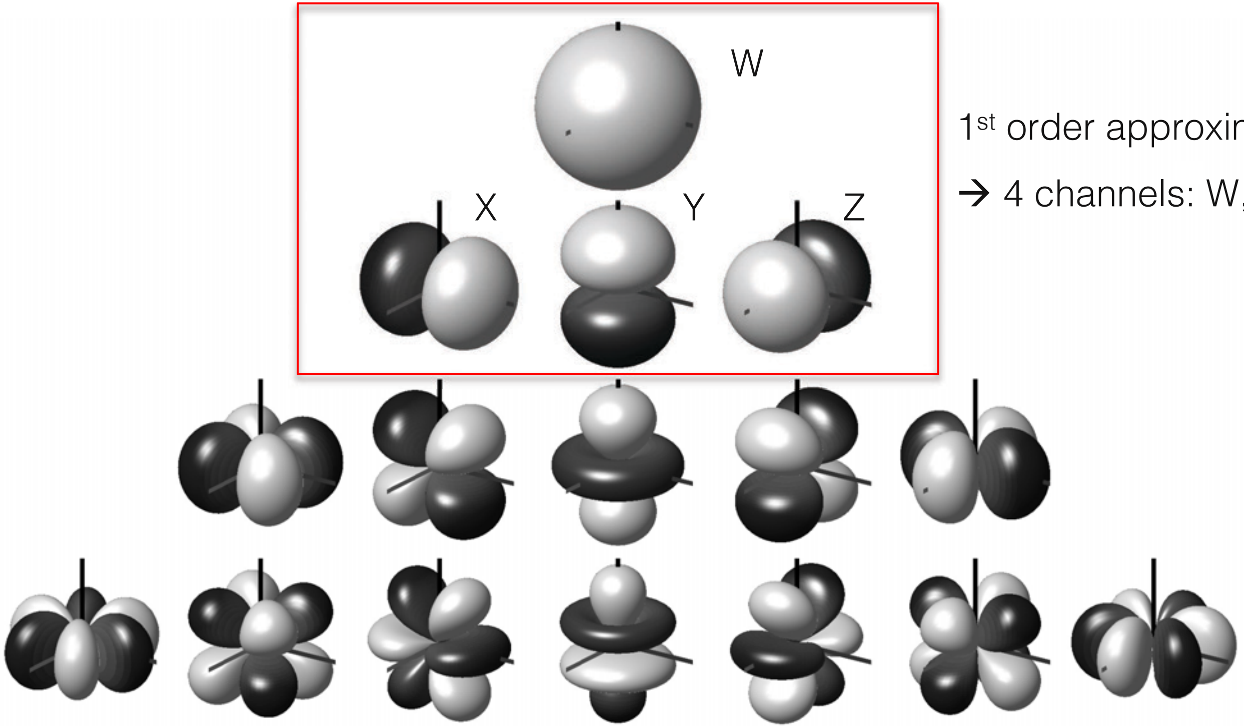
Ambisonics – Spherical Harmonics

- use spherical harmonics!
- orthogonal basis functions on a sphere, i.e. full-sphere surround sound
- think Fourier transform acting on the directions of a sphere

Ambisonics – Spherical Harmonics



Ambisonics – Spherical Harmonics



1^{st} order approximation

→ 4 channels: W, X, Y, Z

Ambisonics – Spherical Harmonics

- can easily convert a point sound source to the 4-channel ambisonics representation
- given azimuth and elevation θ, ϕ , compute W,X,Y,Z as

$$W = S \cdot \frac{1}{\sqrt{2}}$$ ← omnidirectional component (angle-independent)

$$X = S \cdot \cos\theta \cos\phi$$ ← "stereo in x"

$$Y = S \cdot \sin\theta \cos\phi$$ ← "stereo in y"

$$Z = S \cdot \sin\phi$$ ← "stereo in z"

Ambisonics – Spherical Harmonics

- can also record 4-channel ambisonics via special microphone
- same format supported by YouTube VR and other platforms

http://www.oktava-shop.com/

Ambisonics – Spherical Harmonics

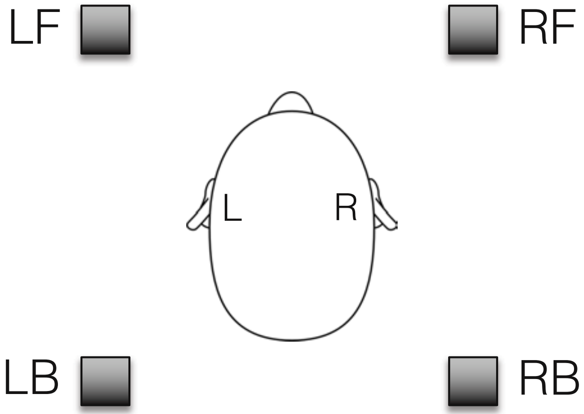
- easiest way to render ambisonics: convert W,X,Y,Z channels into 4 virtual speaker positions
- for a regularly-spaced square setup, this results in

$$LF = (2W + X + Y)\sqrt{8}$$

$$LB = (2W - X + Y)\sqrt{8}$$

$$RF = (2W + X - Y)\sqrt{8}$$

$$RB = (2W - X - Y)\sqrt{8}$$

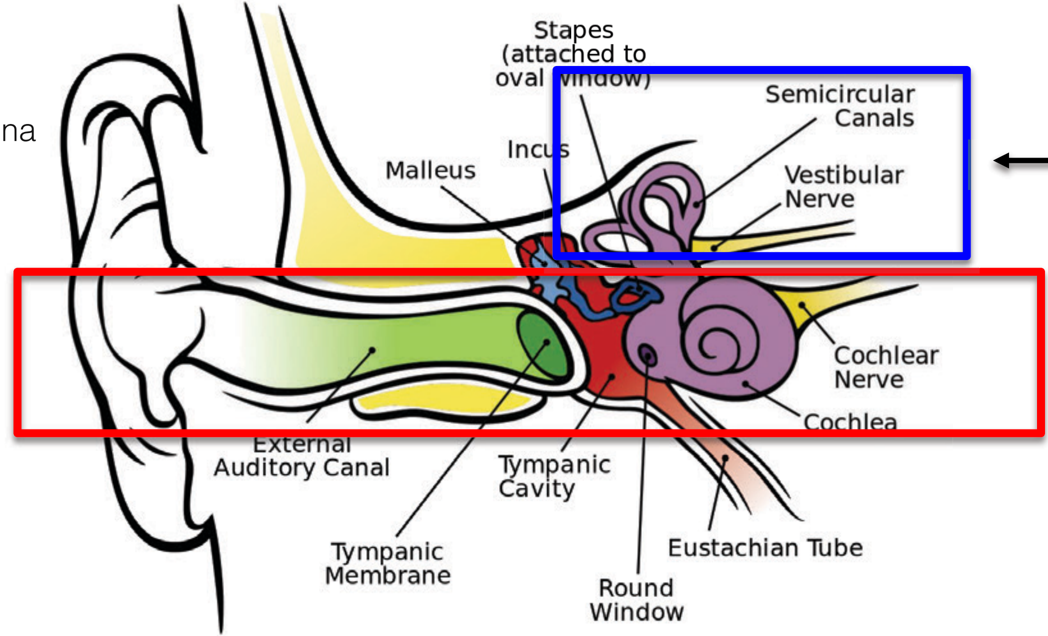


The Vestibular System

or “What else is happening in the inner ear?”

The Inner Ear

pinna



what's this?

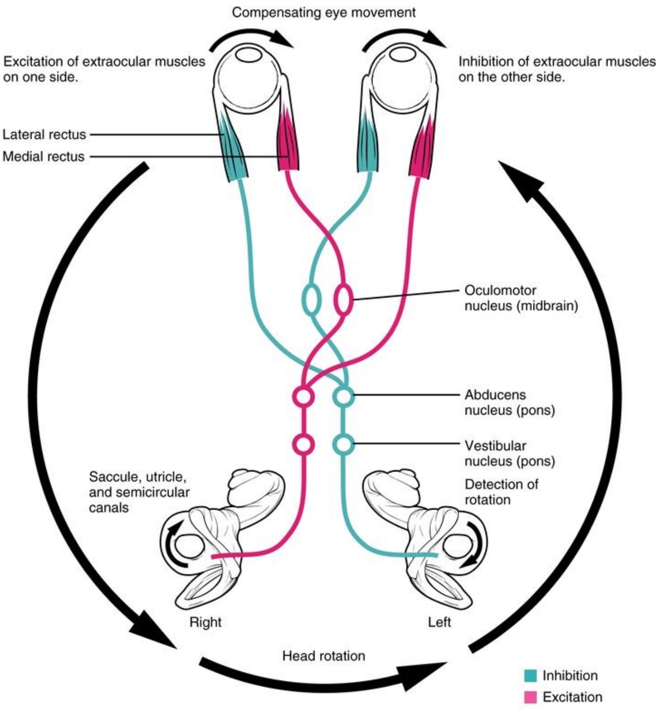
hearing

wikipedia

Brief Overview of the Vestibular System

- provides sense of balance & gravity
- like IMUs – one in each ear!
- in each ear, sense linear (3 dof from otolithic organs) and angular (3 dof from 3 semicircular canals) acceleration via hair cells

Vestibulo-Ocular Reflex (VOR)



- vestibular system and ocular system are directly coupled in a feedback system
- enables low-latency "optical image stabilization" of the visual system with head motion

Motion Sickness

3 types of motion sickness (all related to visual-vestibular conflict theory):

1. Motion sickness caused by motion that is felt but not seen
2. Motion sickness caused by motion that is seen but not felt
3. Motion sickness caused when both systems detect motion but they do not correspond.

Motion Sickness

3 types of motion sickness (all related to visual-vestibular conflict theory):

1. Motion sickness caused by motion that is felt but not seen
2. Motion sickness caused by motion that is seen but not felt
3. Motion sickness caused when both systems detect motion but they do not correspond.

Example: car and sea sickness

Motion Sickness

3 types of motion sickness (all related to visual-vestibular conflict theory):

1. Motion sickness caused by motion that is felt but not seen
2. Motion sickness caused by motion that is seen but not felt
3. Motion sickness caused when both systems detect motion but they do not correspond.

Example: VR sickness or visually-induced motion sickness (VIMS)

Motion Sickness

3 types of motion sickness (all related to visual-vestibular conflict theory):

1. Motion sickness caused by motion that is felt but not seen
2. Motion sickness caused by motion that is seen but not felt
3. Motion sickness caused when both systems detect motion but they do not correspond.

Example: motion in low gravity

References and Further Reading

Panoramic Imaging and VR

- M. Brown, D. Lowe "Automatic Panoramic Image Stitching using Invariant Features", IJCV 2007
- autostitch: http://matthewalunbrown.com/autostitch/autostitch.html
- S. Peleg, M. Ben-Ezra, Y. Pritch "Omnistereo: Panoramic Stereo Imaging" IEEE PAMI 2001
- Konrad et al. "SpinVR: Towards Live Streaming VR Video", ACM SIGGRAPH Asia 2017

References and Further Reading - Spatial Sound

- Google's take on spatial audio: https://developers.google.com/vr/concepts/spatial-audio

HRTF:

- Algazi, Duda, Thompson, Avendado "The CIPIC HRTF Database", Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics
- download CIPIC HRTF database here: http://interface.cipic.ucdavis.edu/sound/hrtf.html

Resources by Google:

- https://github.com/GoogleChrome/omnitone
- https://developers.google.com/vr/concepts/spatial-audio
- https://opensource.googleblog.com/2016/07/omnitone-spatial-audio-on-web.html
- http://googlechrome.github.io/omnitone/#home
- https://github.com/google/spatial-media/