Spatial Audio

Matteo Luperto
Manuel Pezzera

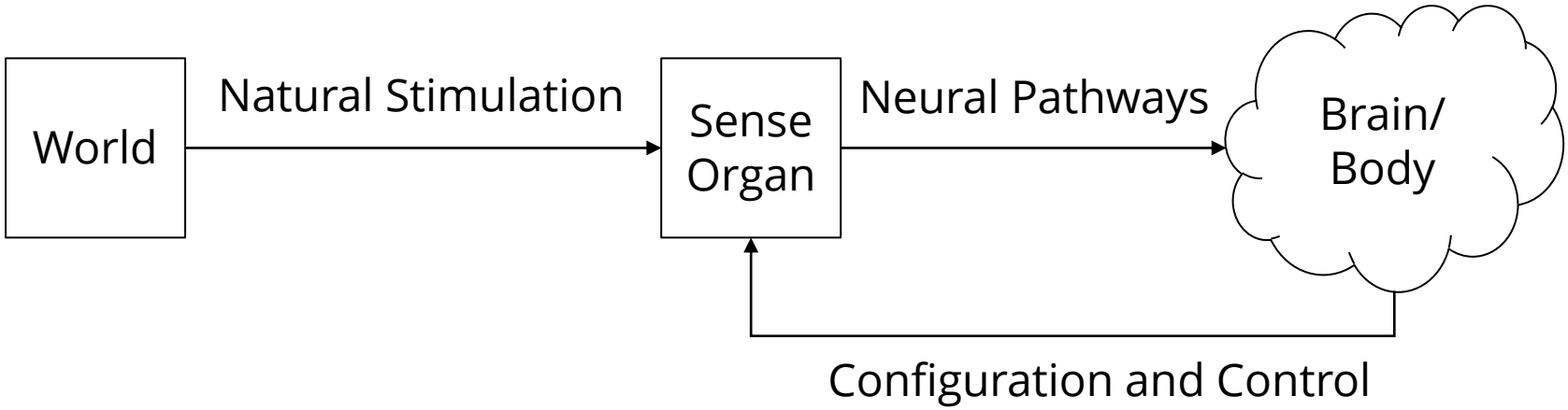
matteo.luperto@unimi.it
manuel.pezzera@unimi.it

UNIVERSITÀ DEGLI STUDI
DI MILANO

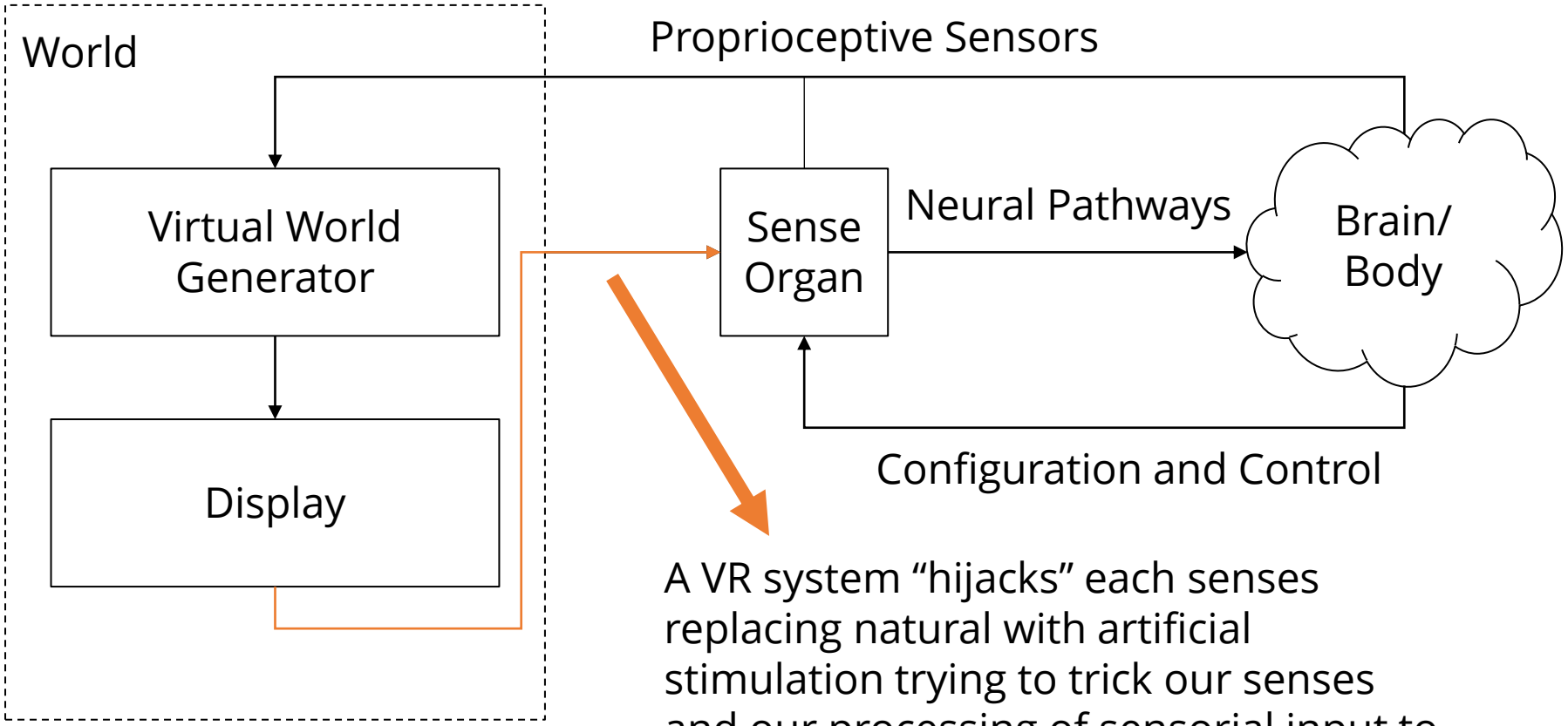
Lab 08

Realtà Virtuale 2020/2021

Perception


World
Natural Stimulation
Sense Organ
Neural Pathways
Brain/ Body
Configuration and Control

Senses and VR



A VR system “hijacks” each senses replacing natural with artificial stimulation trying to trick our senses and our processing of sensorial input to have an immersive experience

Perception + VR

Up to now you have primarily seen how this is done with vision/sight as:

- Vision is the primarily sense that we need to “fool”
- The development in the last 15 years of portable VR kits and tracking algorithms/devices has de-facto enabled VR as a commodity

However, for other senses that are “simpler” than vision, the technology suited for a VR experience was already there

we already experimented how to “trick” those senses into perceiving something different than what is happening

Audio directionality

In music you try to have an immersive experience by recording/mastering tracks so to give the feeling to the listener of being “live” with musicians:

- Guitar on the left,
- Bass on the right,
- Drums behind,
- …

What you do is give a “directionality” to each instrument/sound

Audio directionality

By focusing on a simpler setting (only 1 sense, hearing, fixed subject position) music tried for decades to recreate a “virtual reality” experience

Audio for VR

Spatialized audio and VR-related techniques are perhaps the oldest form of “virtual reality” experience, and the only commonly used and settled technology

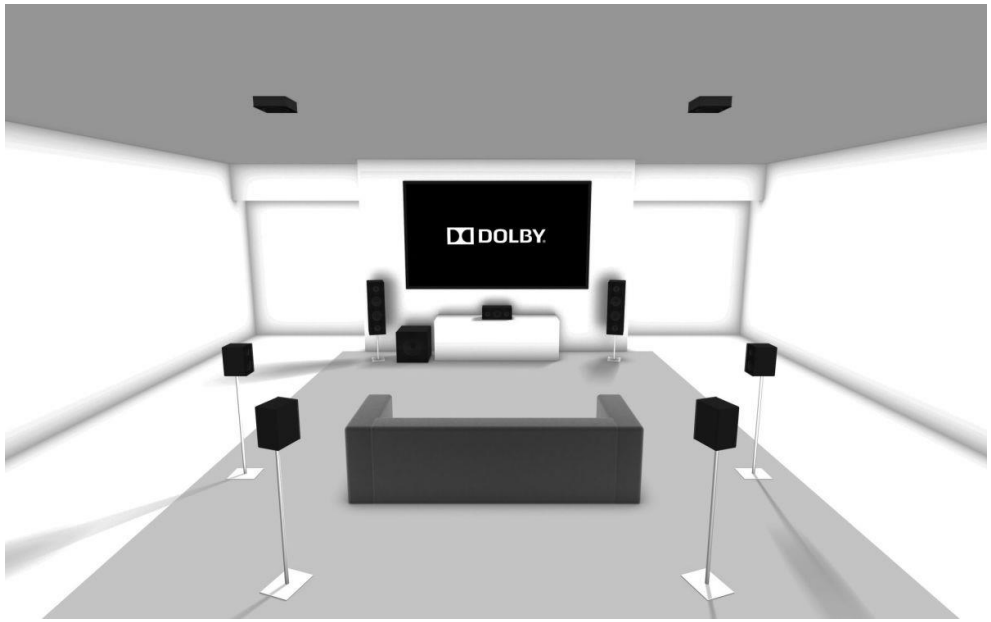
- Stereo audio started in 1930’ – mainstream since 1960’

… but its use is mostly related to:

- Music
- Cinema

Also, audio-VR interfaces can be pretty cheap (and probably you are using one right now)

World- vs user-fixed



audio 7.1

headphones

Headphones ~= Visors

Audio → Spatialized Audio

The main difference between (spatialized) audio in music/movies and VR is that the former assumes a fixed/passive position of the listener.

With VR we introduce a feedback involving the user's movements to dynamically adapt the sound.

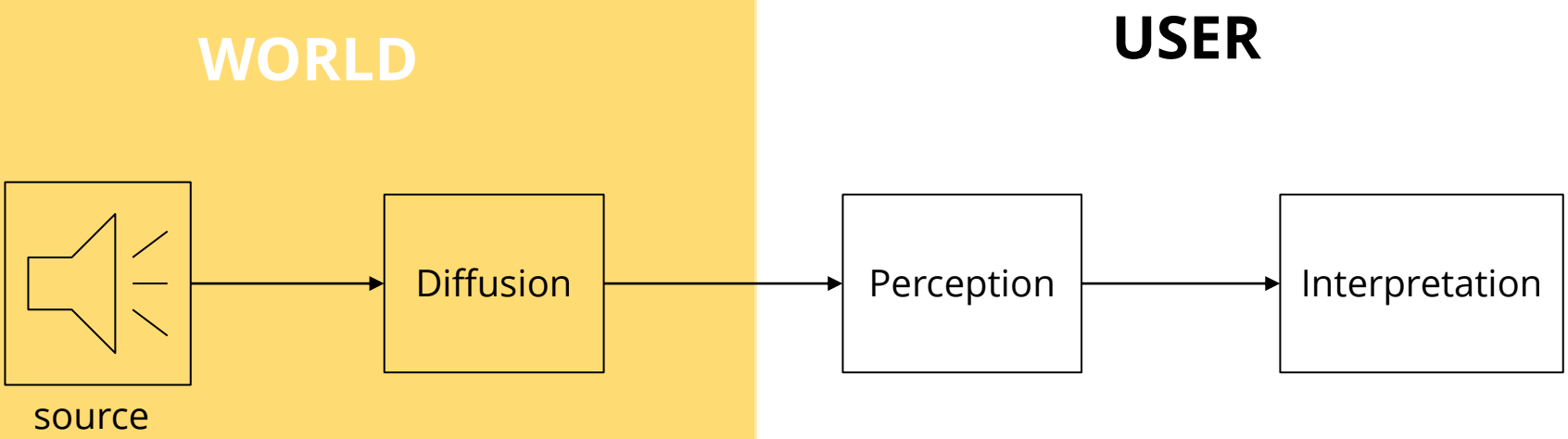
For that we rely on headphones.

However, we lose the immersivity of surround system due to "feeling" certain frequencies.

Today’s outline

- Audio 101 – some basics about how sound works
- Perception
- Spatial audio modeling
- Available SDKs
  Resonance Audio + Unity
- Design Tips

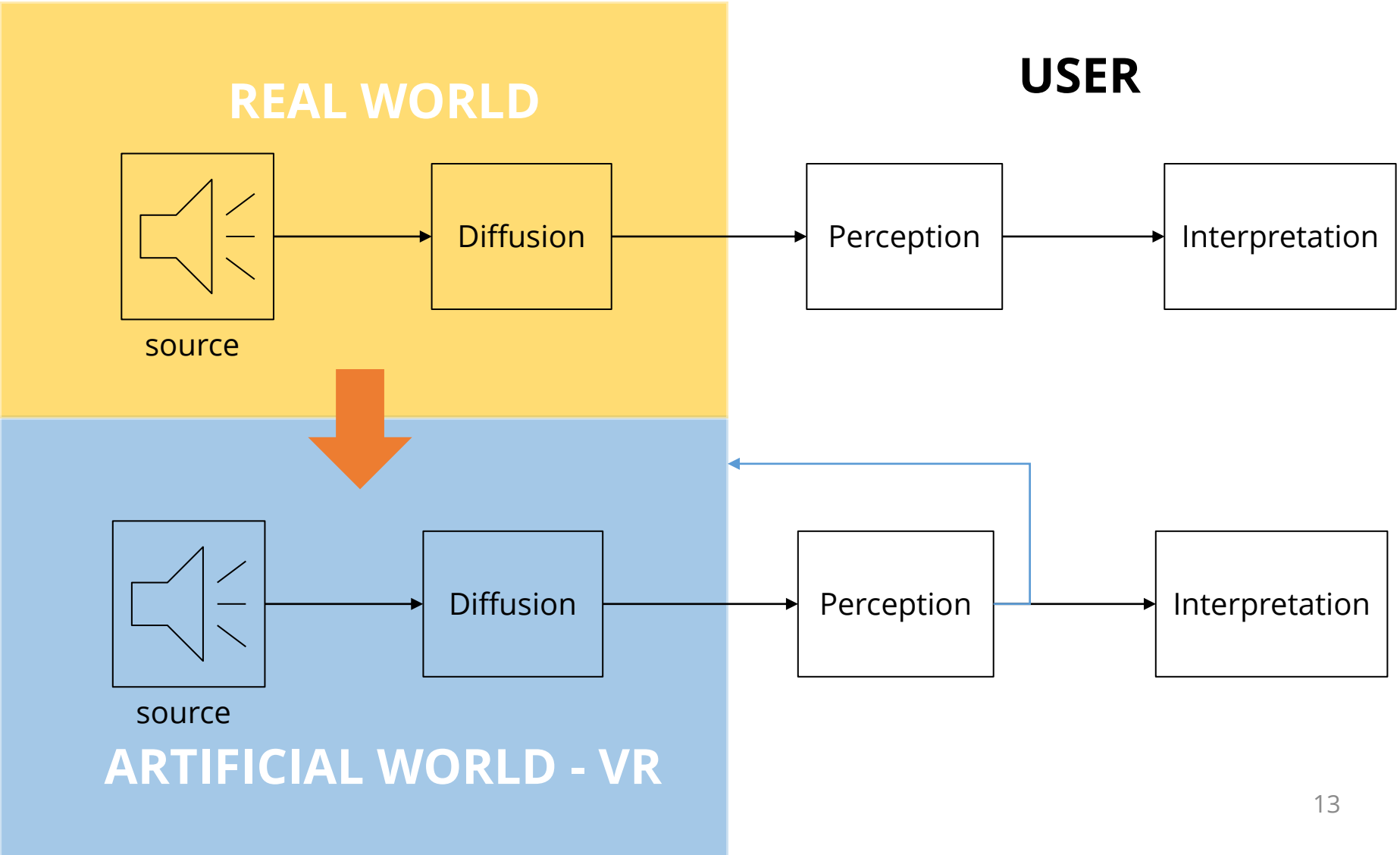
Audio modelling for VR



We need to understand how sound works in our world so to replicate its dynamics in VR.

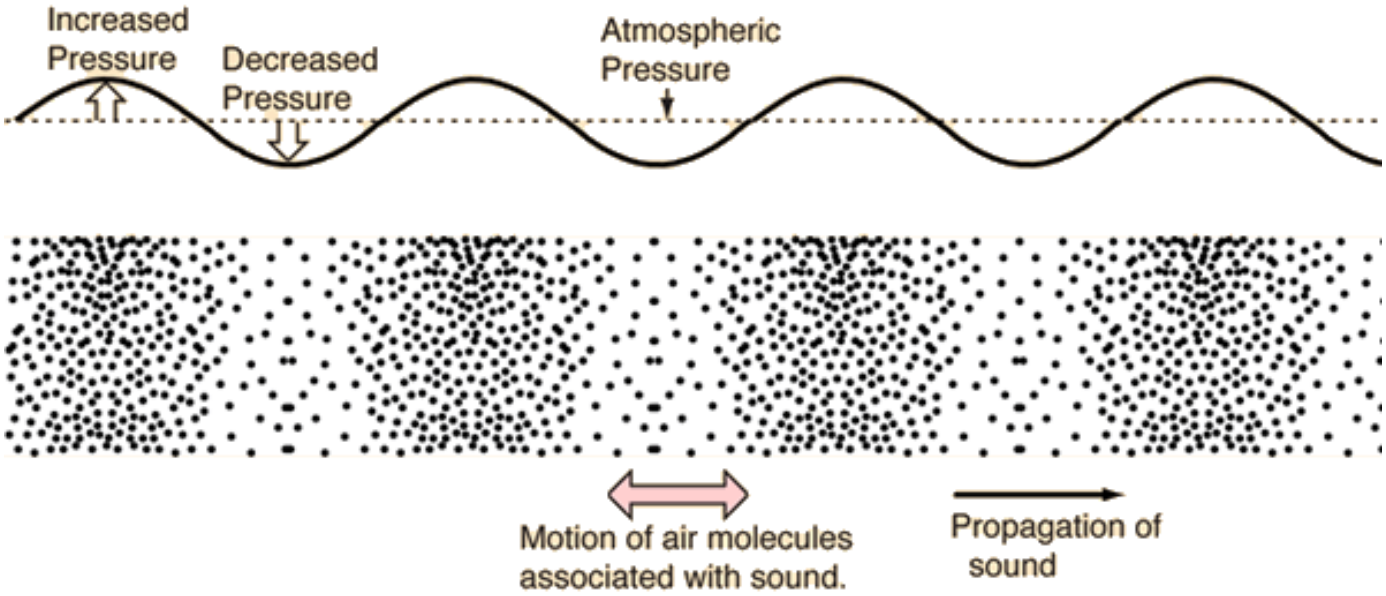
We need to understand how we perceive and process audio so to “fool” our senses

Audio modelling for VR



Audio waves

Sound is a longitudinal wave of compression and rarefaction of air molecules



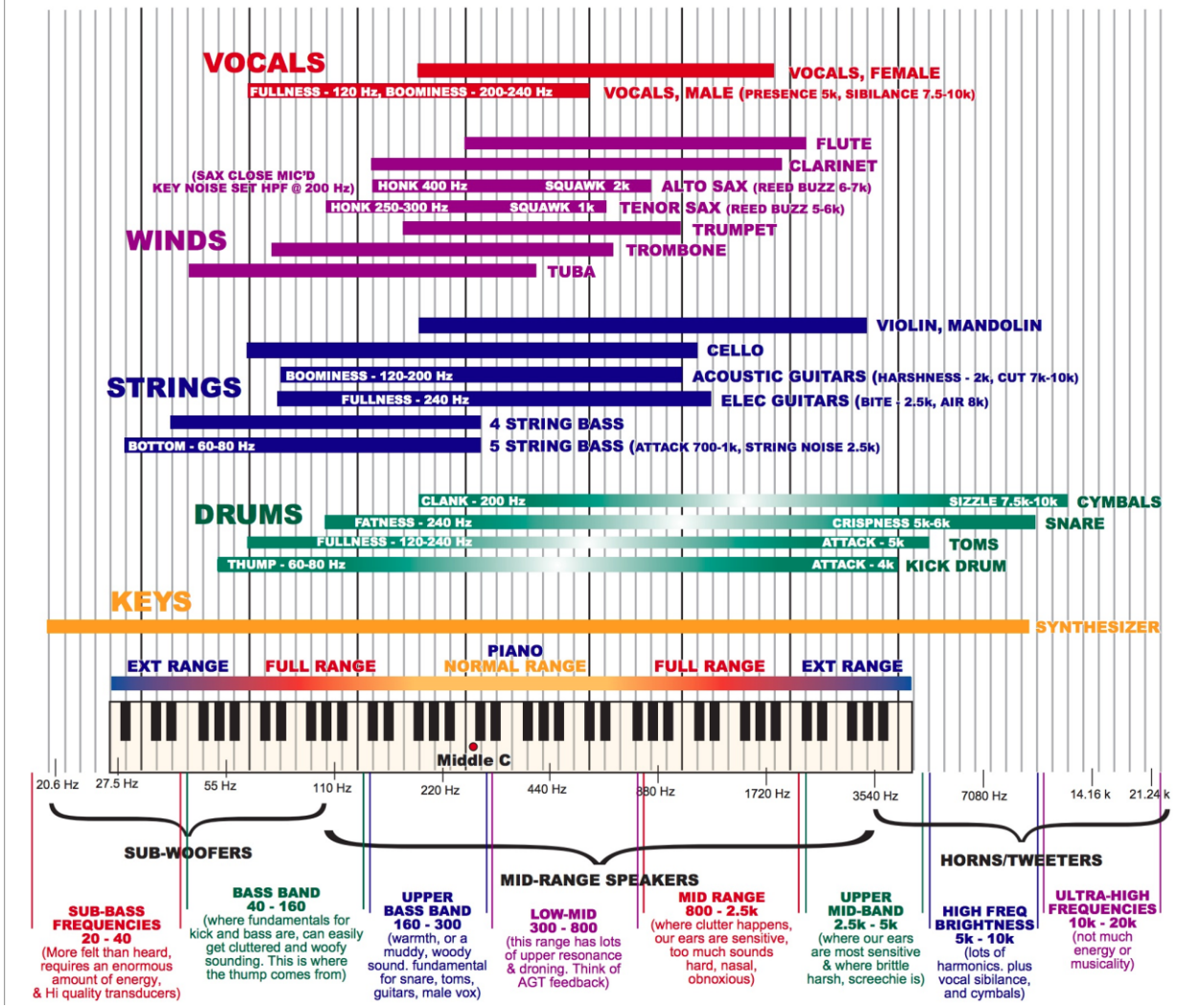
Both sound and light are both propagated by waves, there are many similarities

Audio waves

20 – 25k Hz
very broad spectrum

HF range decrease with age

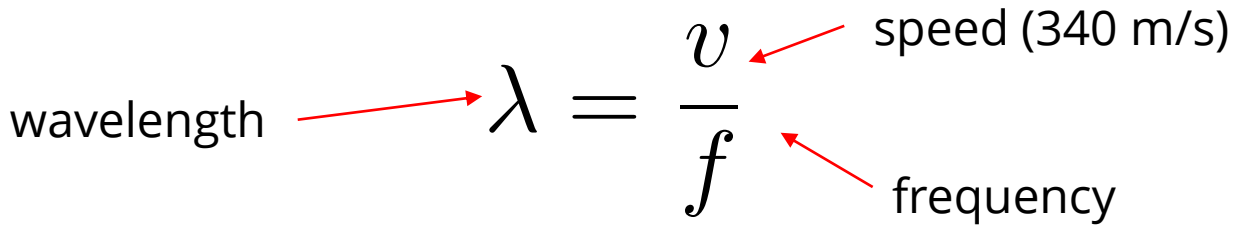
f > 14k Hz cannot be heard by most adult, but do not contain many info



pic from Reddit

Audio- vs Light- Waves

- Light waves can go from 430–770 THz approx
- Audio waves goes from 20 to 25k Hz
  - that's a huge difference – log scale (db)
  - != diffraction / reflection for HF and LF

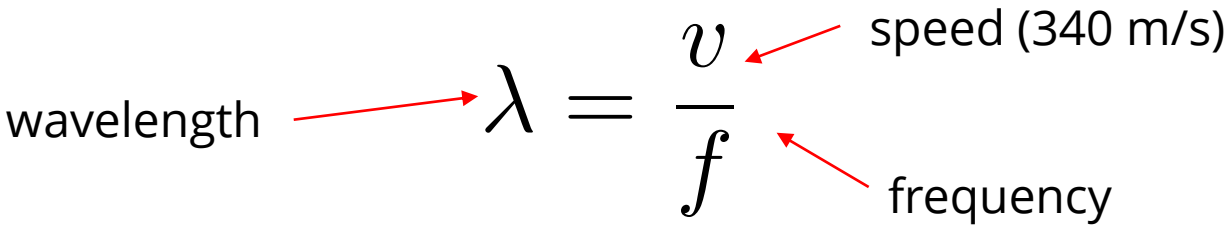


$$f = 20\ \mathtt{Hz} \rightarrow \lambda = 30\ \mathtt{m}$$

$$f = 25000\ \mathtt{Hz} \rightarrow \lambda = 13.6\ \mathtt{mm}$$

Low and High Frequency

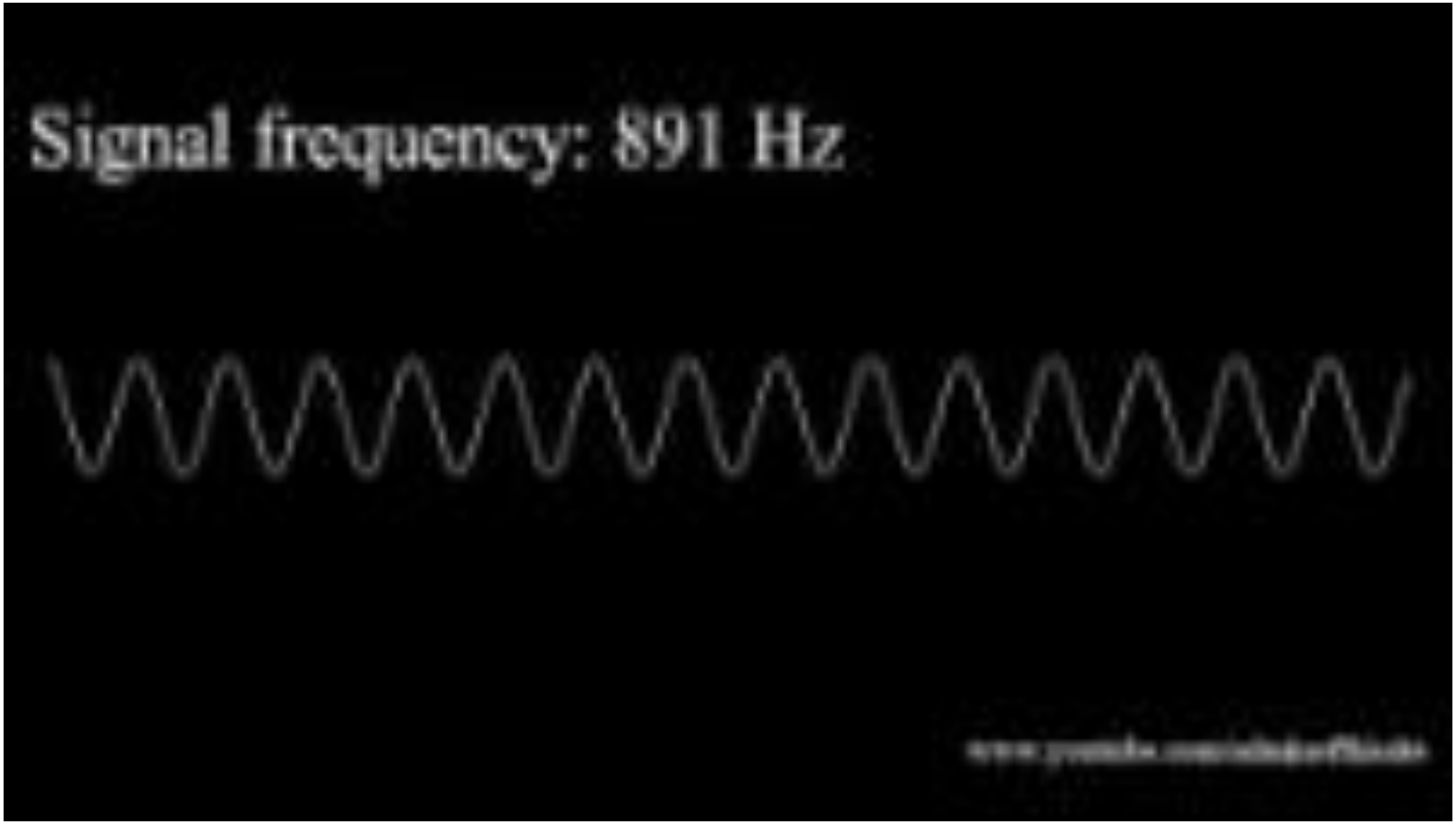
There is a big difference between Low Frequencies **(LF)** and High Frequencies **(HF)** and this affects both how they propagate and how we perceive them.



$$f = 20 \text{ Hz} \rightarrow \lambda = 30 \text{ m}$$

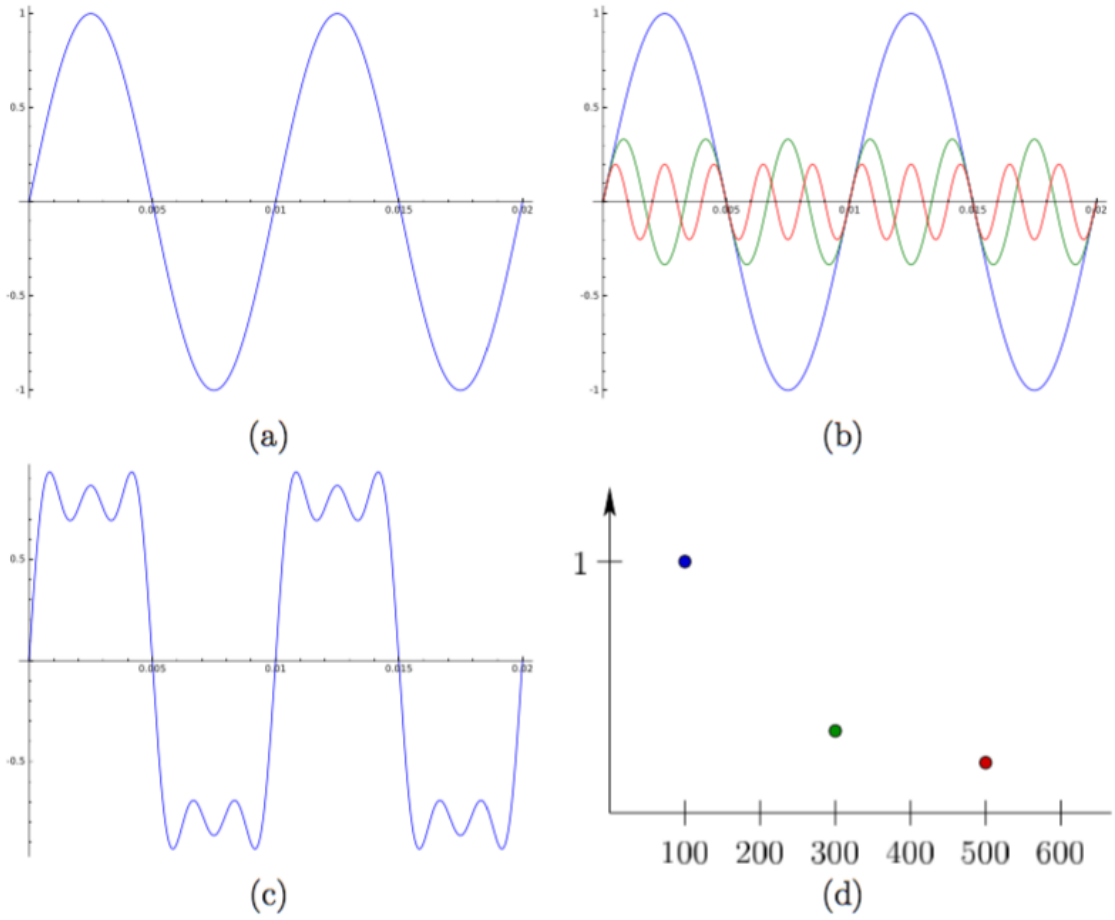
$$f = 25000 \text{ Hz} \rightarrow \lambda = 13.6 \text{ mm}$$

Audio spectrum - video



Audio Waves are Waves…

Spectral decomposition and Fourier analysis are important.

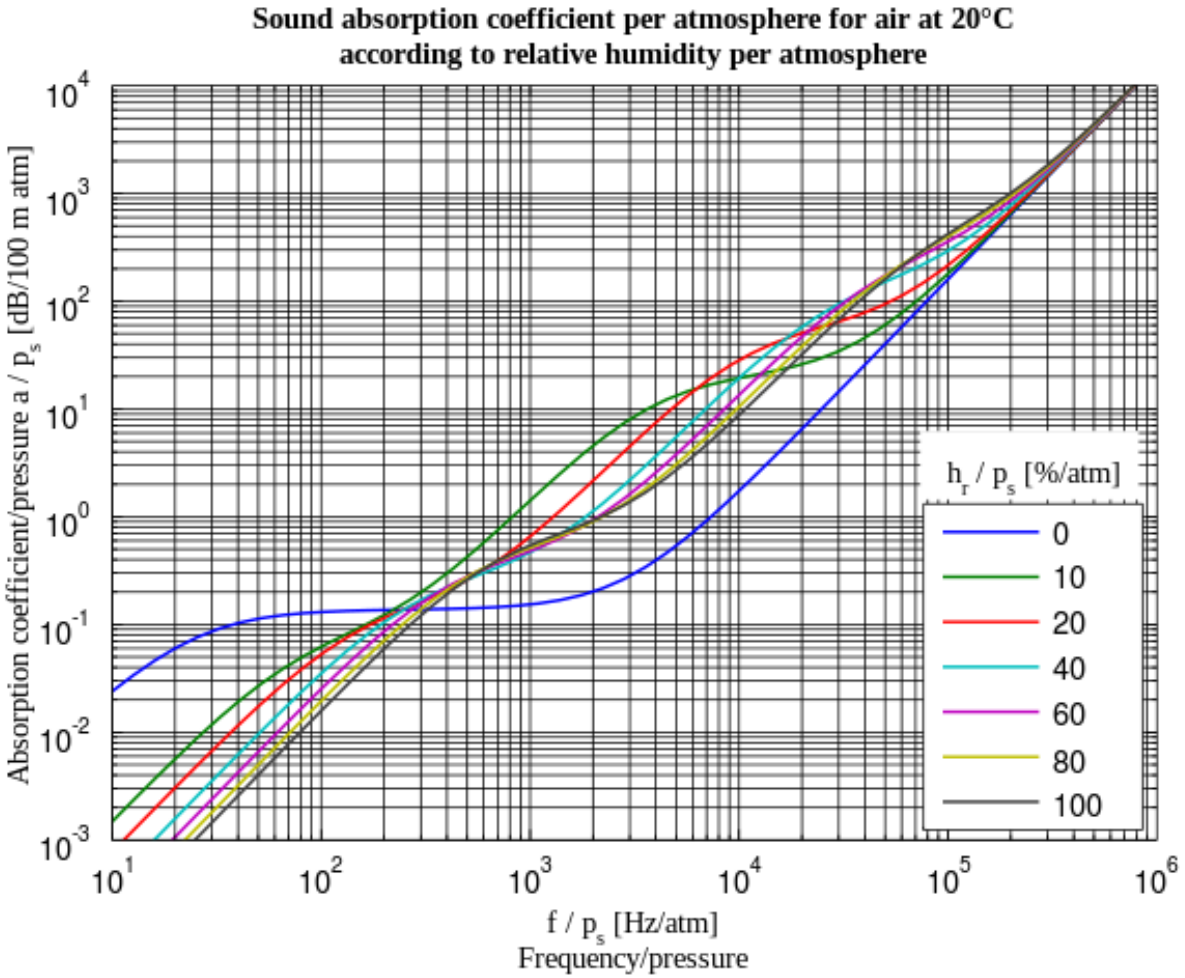


Sound sources

In “our” world audio sources are complex and diffused

In virtual worlds audio sources are single points in the space and this can cause modelling issues

Sound Absorption



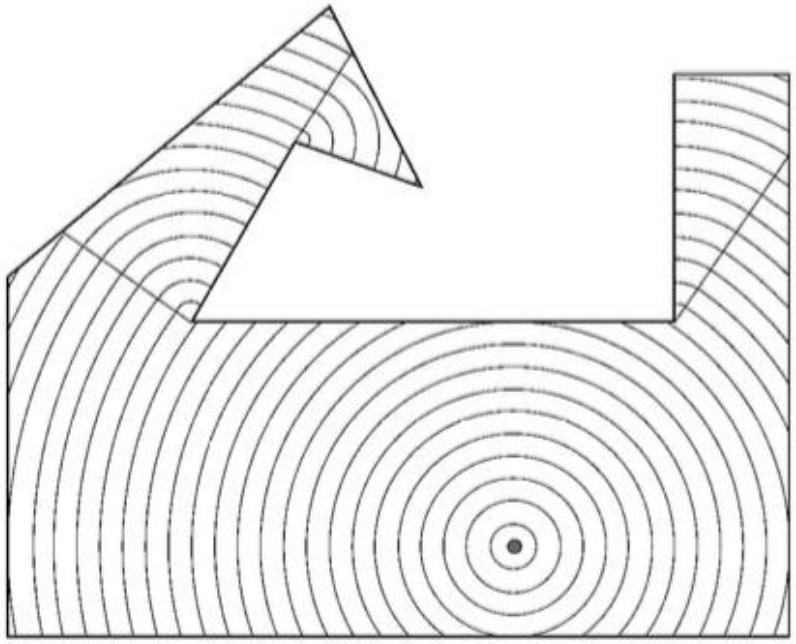
Environmental sound absorption changes a lot from LF to HF.

Consequently, sound changes with distance from the source as different spectral components are subject to different absorption level.

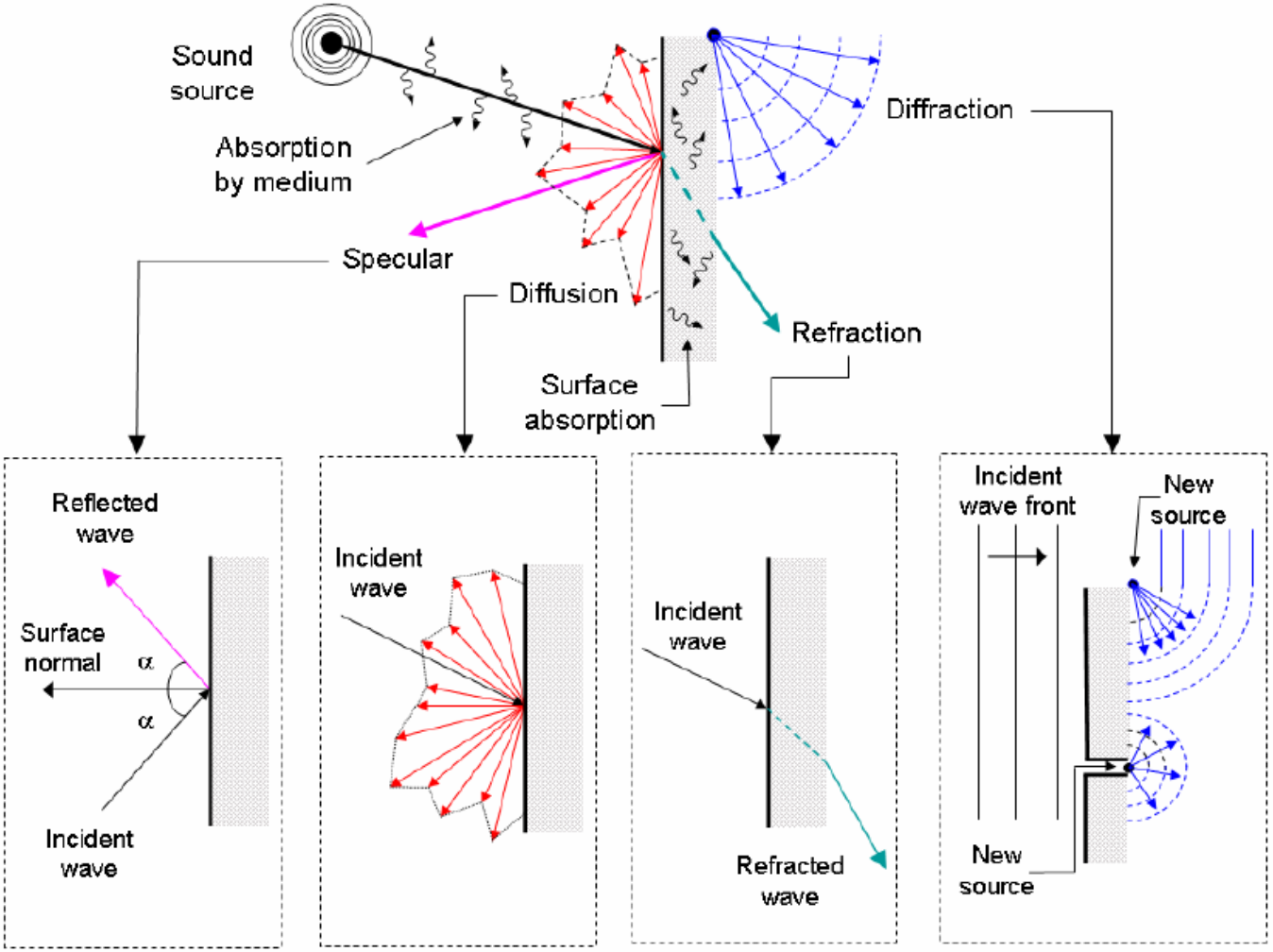
Reflection – Transmission – Diffraction

If an audio wave hits a wall:

- (most) of its energy will bounce – **reflection**
- (some) penetrate– and propagate faster than in air – **refraction**
- (some) trespass the wall and goes on after- **transmission**
- waves bend across corners / obstacles - **diffraction**
- **HF** and **LF** behave very differently



Reflection – Transmission – Diffraction

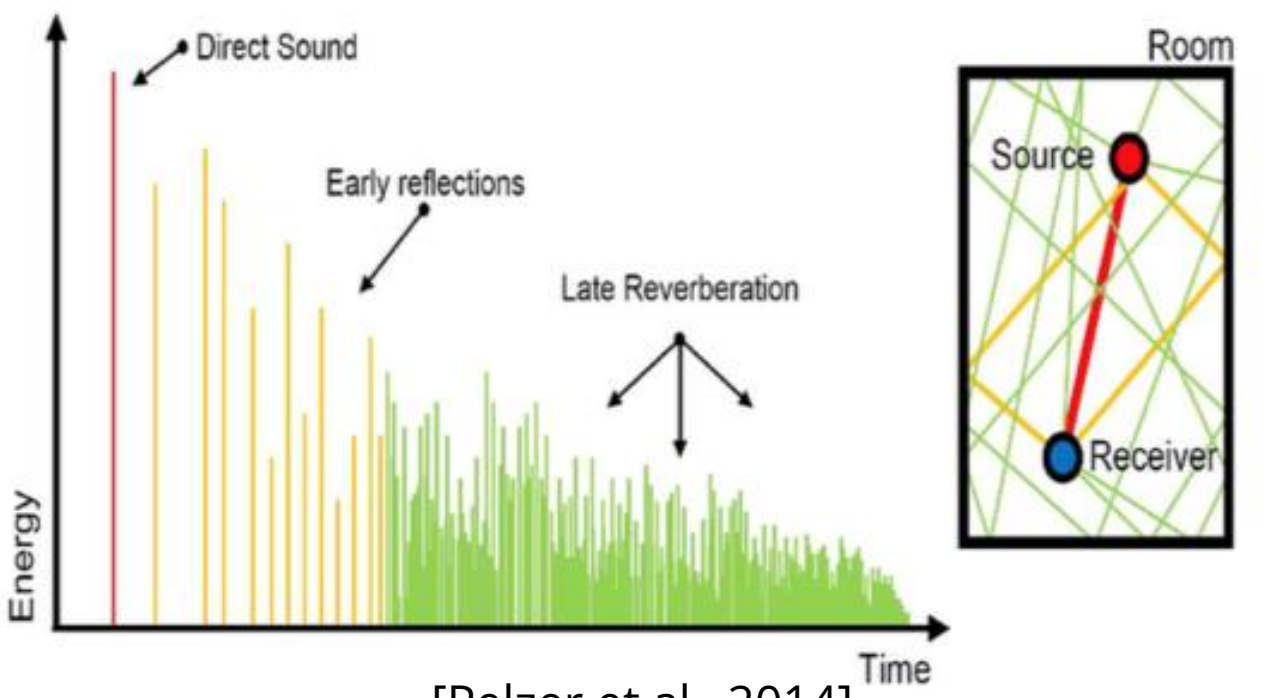


[Kapralos et al, 2008]

Audio Propagation

As a result a single sounds is replicated into many (many) slightly modified copies, all of which are then perceived by the user.

Audio is a simpler information stream than video, but audio propagation is more complex than light propagation



[Pelzer et al., 2014]

Our *sensors*: eyes and ears

Eyes capture at (relatively) low frequency a complex data

- e.g. 1920*1080 (2M) pixel ~ 1k Hz (60 Hz) (fullHD)

Our *sensors*: eyes and ears

Eyes capture at (relatively) low frequency a complex data

- e.g. 1920*1080 (2M) pixel ~ 1k Hz (60 Hz) (fullHD)

Ears capture a 1-dimensional signal (each) at high frequency

- e.g. left + right @ 44100 Hz

Each ear can be seen as a 1-pixel high-frequency high-resolution camera
(but we still have only two-pixel resolution)

Ear physiology is complex

(and difficult to model – from a VR perspective)

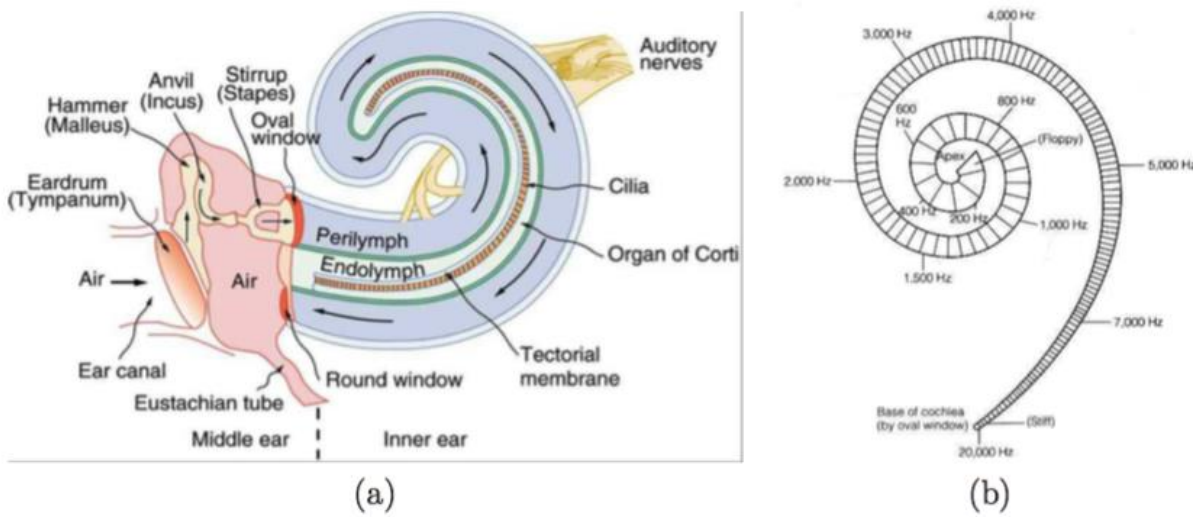
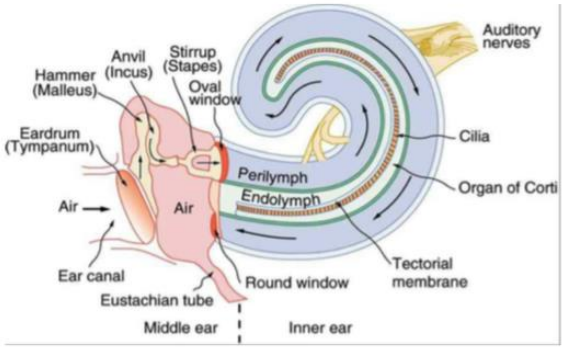


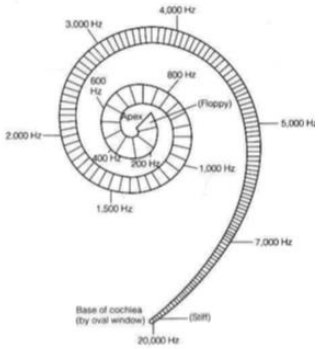
Figure 11.5: The operation of the cochlea: (a) The perilymph transmits waves that are forced by the oval window through a tube that extends the length of the cochlea and back again, to the round window. (b) Because of varying thickness and stiffness, the central spine (basilar membrane) is sensitive to particular frequencies of vibration; this causes the mechanoreceptors, and ultimately auditory perception, to be frequency sensitive.

Ear physiology is complex

(and difficult to model – from a VR perspective)

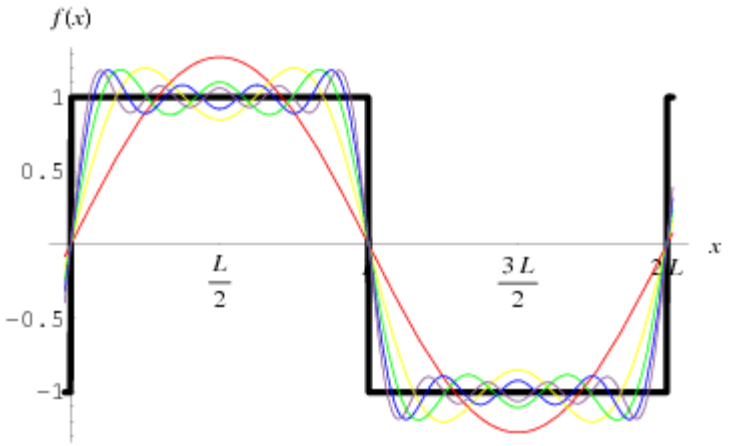


(a)



(b)

Figure 11.5: The operation of the cochlea: (a) The perilymph transmits waves that are forced by the oval window through a tube that extends the length of the cochlea and back again, to the round window. (b) Because of varying thickness and stiffness, the central spine (basilar membrane) is sensitive to particular frequencies of vibration; this causes the mechanoreceptors, and ultimately auditory perception, to be frequency sensitive.



However, what “roughly” ears do is to perform a spectral decomposition and frequency-based analysis of sound waves

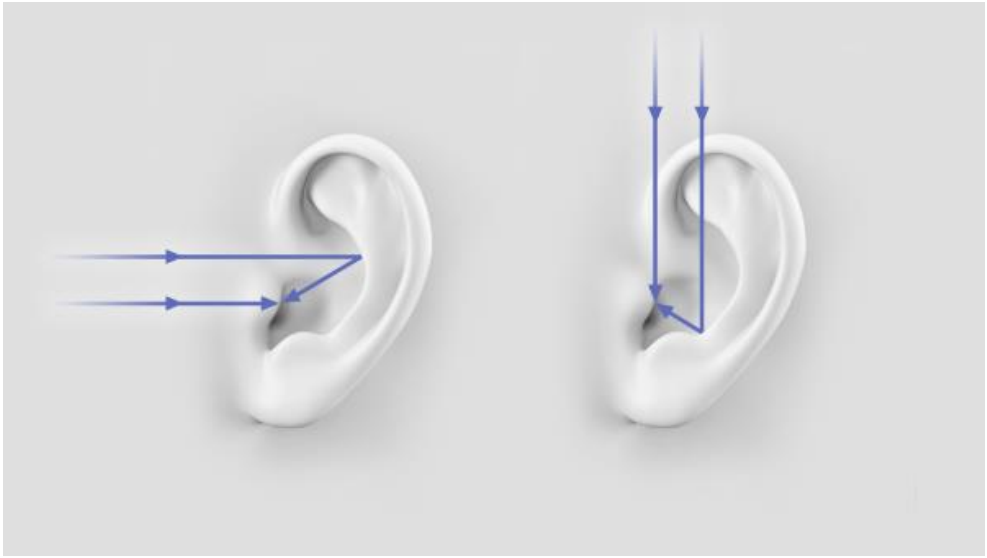
Auditory perception and interpretation

Audio perception involves a lot of “brain processing”, due to our the ear ability and to the different types and phenomena which affects sound waves

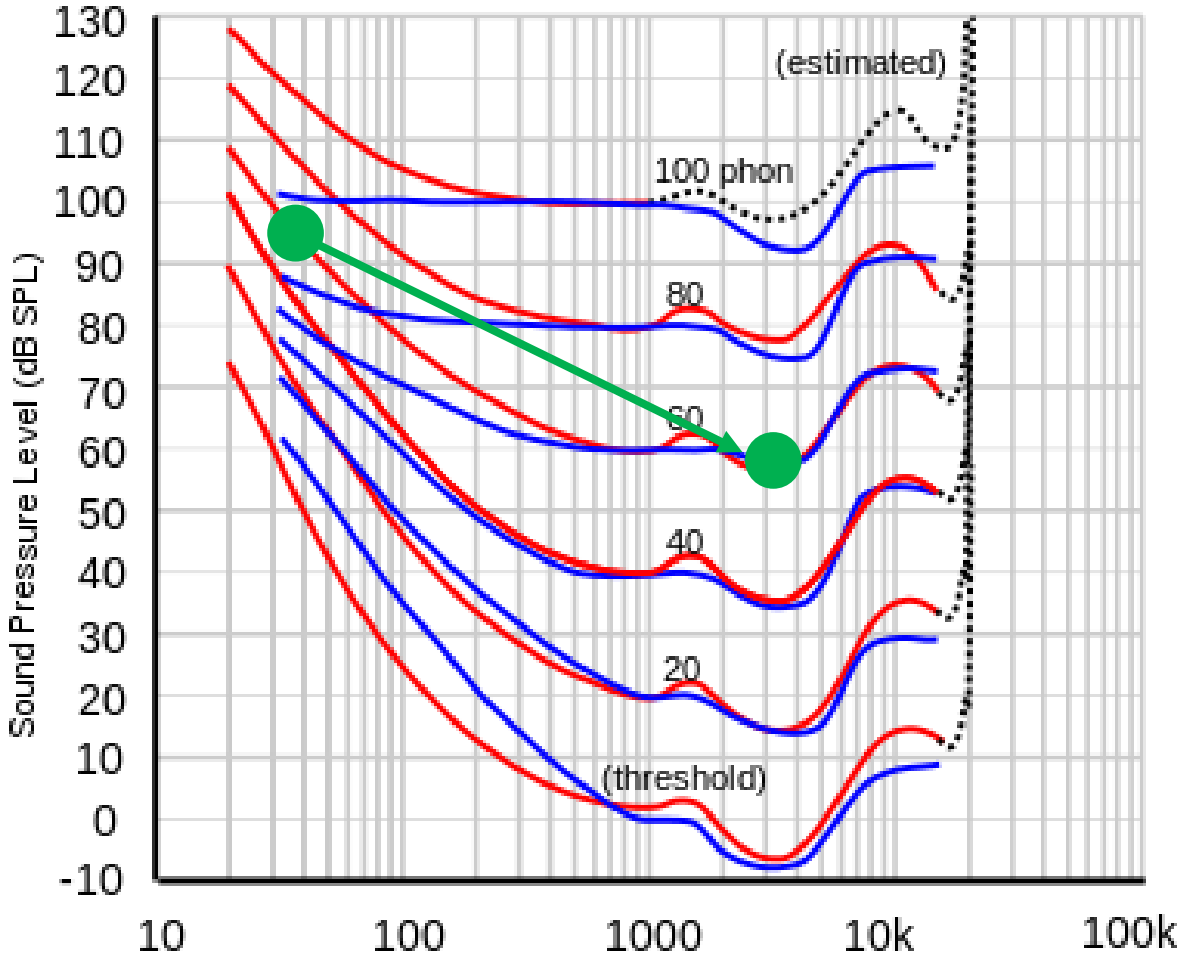
- different wavelengths results into different diffraction and reflections;
- also, HF and LF waves propagates differently and have different energies
- we have 2 ears listening to two versions of the same sound at once
- ears has to cope with adaptation, missing data and assumption.

Auditory perception and interpretation

Auditory perception also involves some “mechanical” components; e.g., our pinna (outer ear) distort the sound in a controlled way that is used by the brain for processing



Equal loudness contour curves



Equal-loudness contours (red) (from ISO 226:2003 revision)
Fletcher–Munson curves shown (blue) for comparison

As LF and HF behaves differently, we also perceive them differently.

A LF require a (far) higher volume to be perceived as loud as an HF one.

Auditory perception + freqs

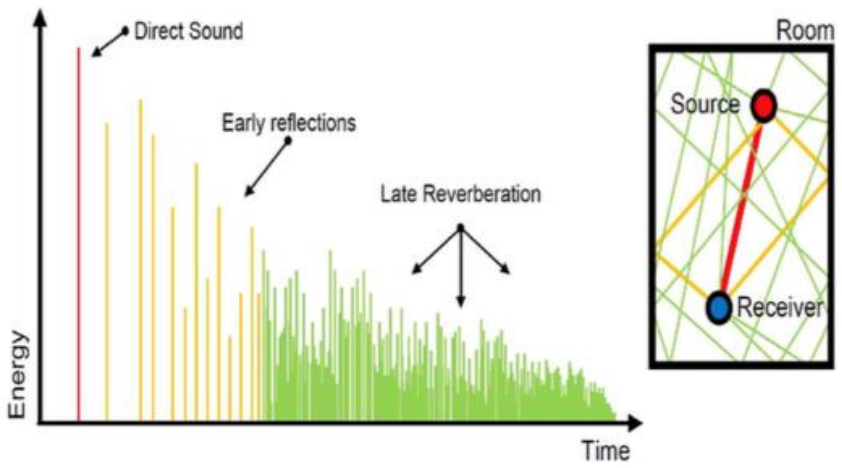
low frequencies

- high wavelength
- require high energy (db) to be perceived
- can go trough objects / walls
- air absorption is low
- can be perceived far from their source

high frequencies

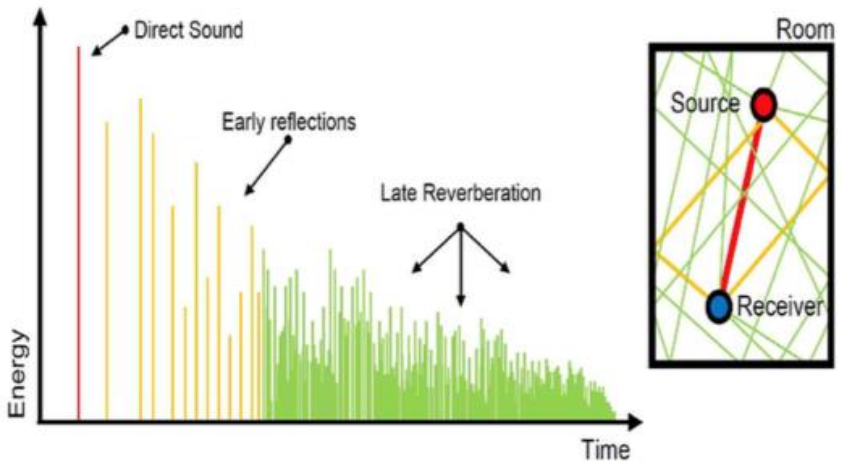
- low wavelength
- require low energy (db) to be perceived
- occlusion / objects change sound
- are absorbed by air
- can be perceived, easily, but close to the source

Auditory Perception and diffusion



- An audio perceived far from its source is subject to different types of distortion - different wavelength
- Also, different versions of the same sound are perceived together, with a ***reverb***, because of reflection
- Small delay for directional sound between two ears

Auditory Perception and diffusion

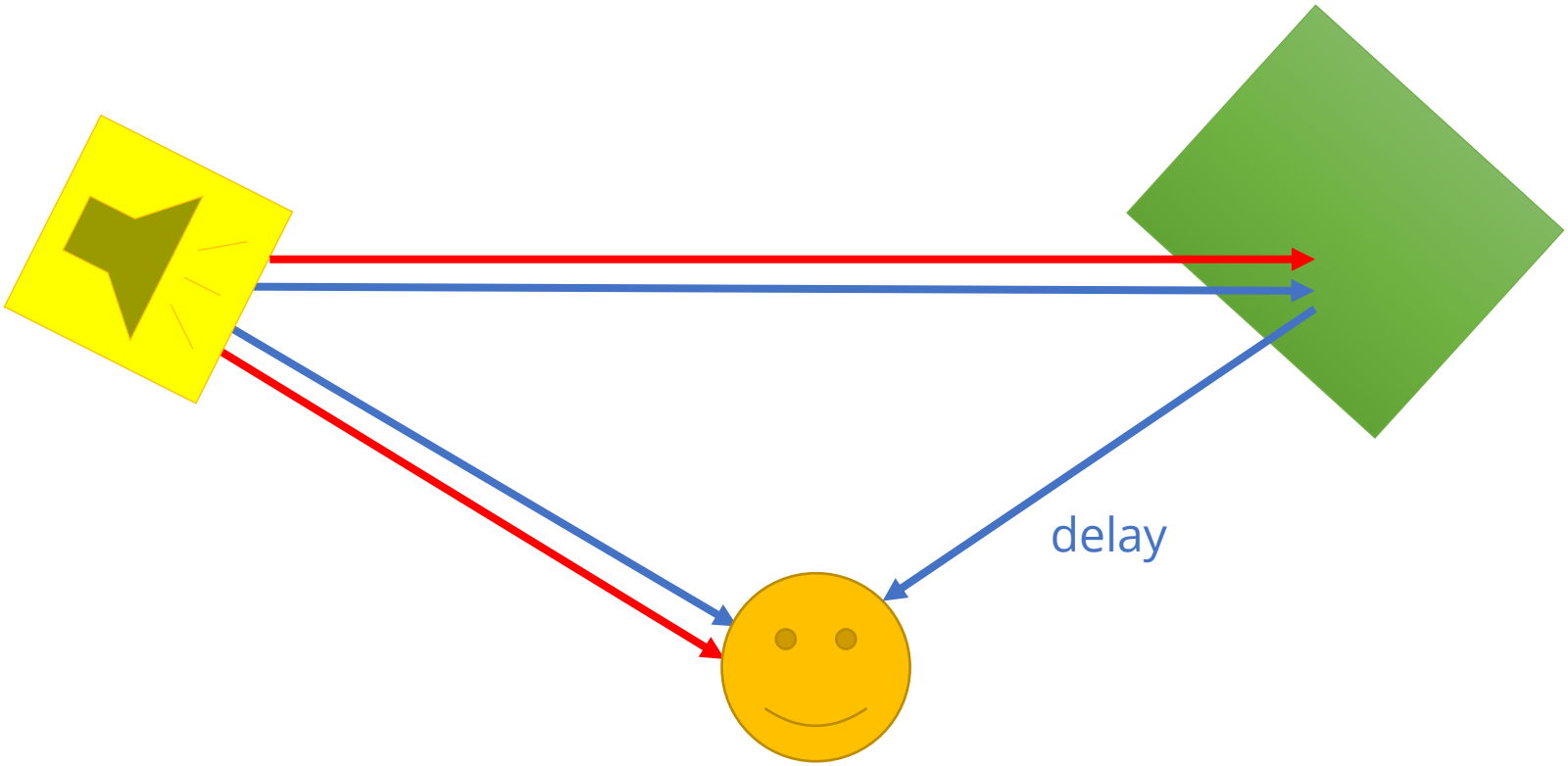


- An audio perceived far from its source is subject to different types of distortion - different wavelength
- Also, different versions of the same sound are perceived together, with a ***reverb***, because of reflection
- Small delay for directional sound between two ears

All of these info (and others) are filtered out by our brain:

- we perceive just one version of the signal …
- … but we use other data to localize the position of the source

Sound perception + propagation



High Frequencies

Low Frequencies

Precedence effect

- If two similar sounds arrive at two different times, only one is perceived
- rather than hearing a jumble, people perceive a single sound.
- based on the first arrival, usually has the largest amplitude.
- echo is perceived if the timing difference is larger than the echo threshold (approx 3 to 61 ms)

Auditory Illusions

The fact that perception involves signal processing that is done by our brain to process its input causes also several auditory illusions

- Precedence effect is one of those
- Glissando Illusion
- Shepard Tone

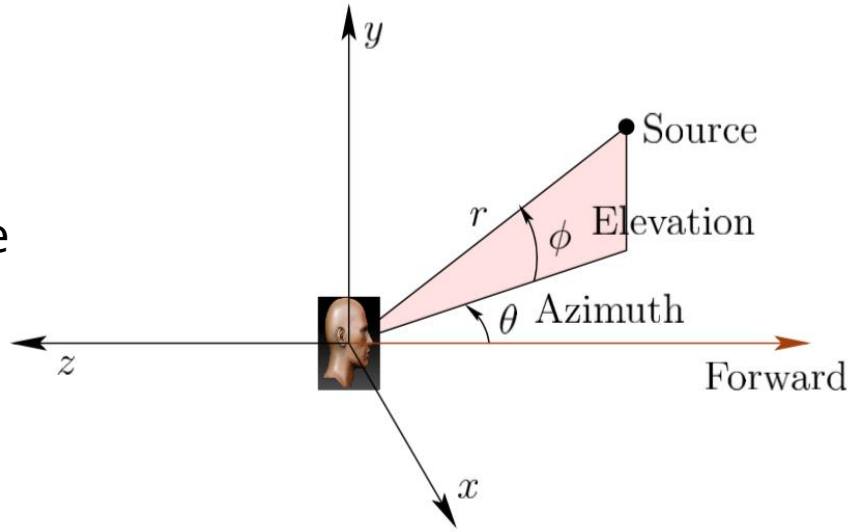
Illusions have been used to give “effects” to games to play with the players

Auditory Illusions

Auditory Perception: Localization

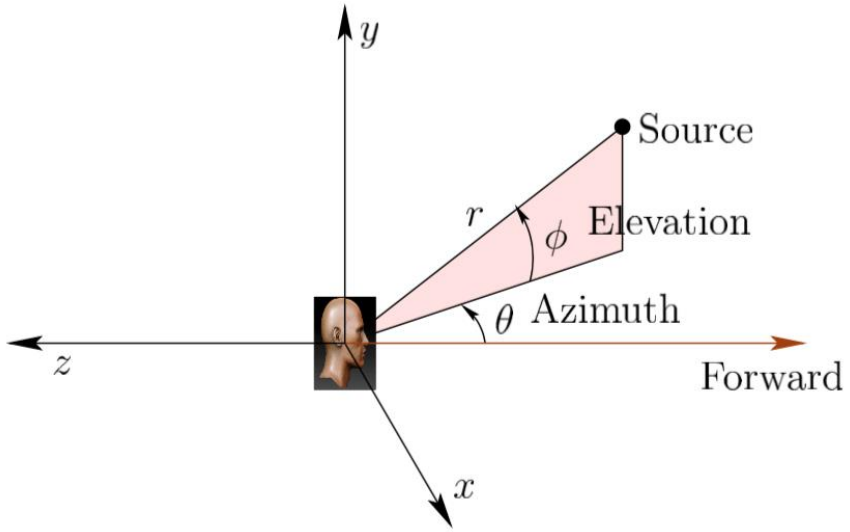
Localization is the estimation of the source of a sound by hearing it.

It is important in games but is particularly important in VR experience as it what personalizes the audio experience for the user allowing a more immersive and realistic perceptual experience.

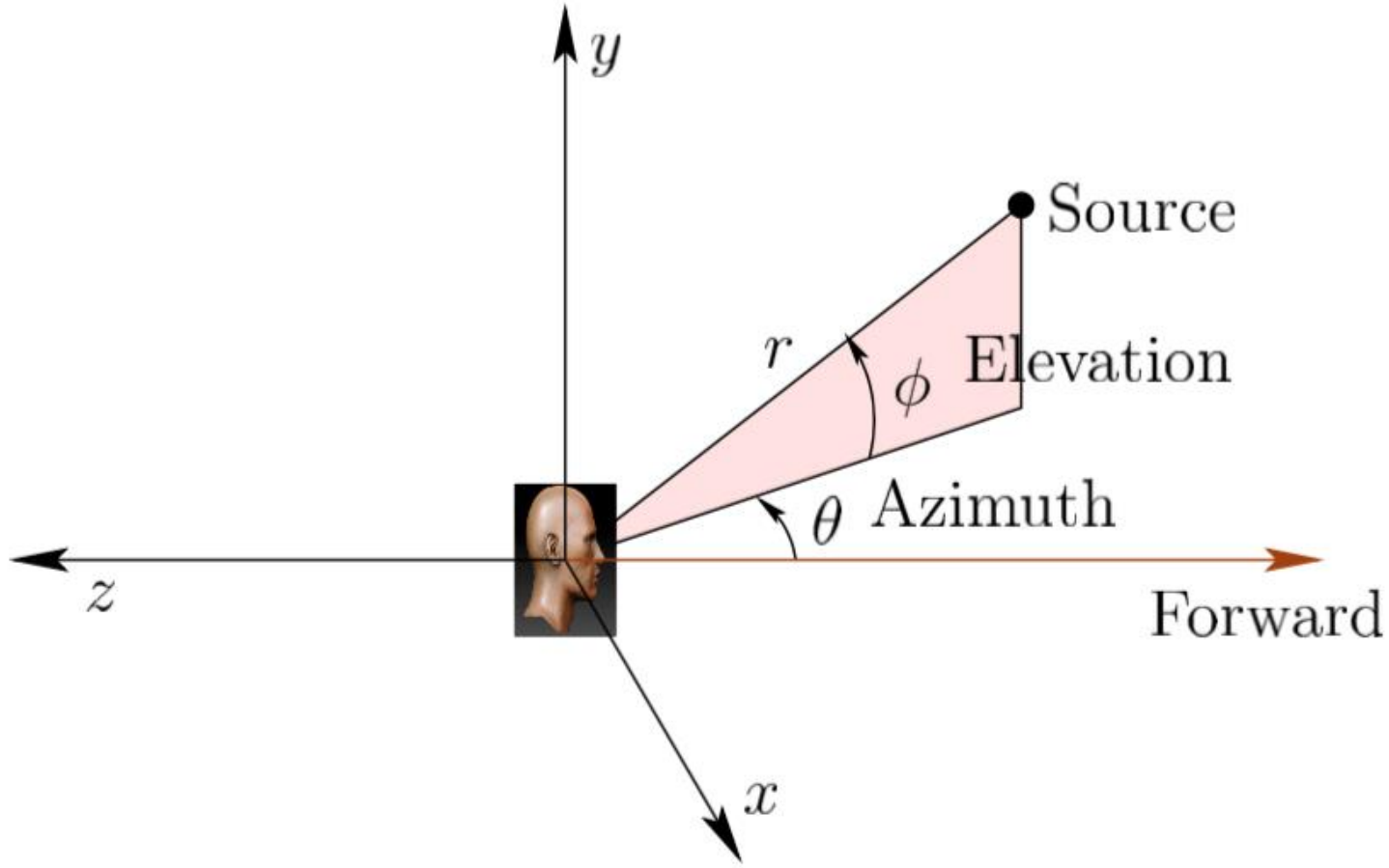


Auditory Perception: Localization

JND (Just Noticeable Differences) in localization (pitch, angle, distance, elevation) are not linear and depend on the several factors (mostly HF and LF)



Measuring distance from the source



minimum audible angle (MAA) depends mostly on frequency and elevation

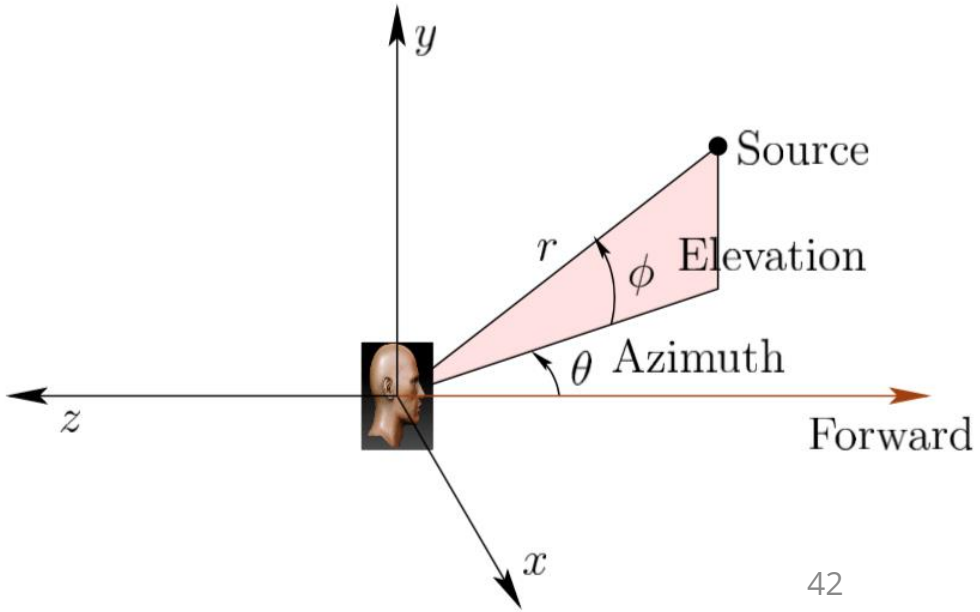
Localization and cues

To localize sound sources we use both:

- Monaural cues rely on sound reaching a single ear;
- Binaural cues based on difference in the same signal when perceived by the two ears jointly.

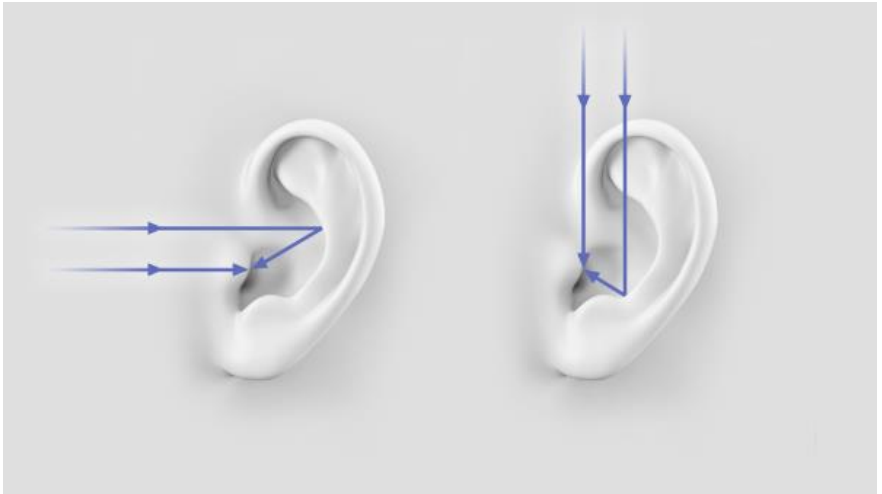
To improve localization we often (and often without noticing) perform movements with our head to improve localization (e.g., tilting moving slightly our head in one of our DOF)

The use of both monaural and binaural cues is similar to what we also to with vision to estimate the distance of an object.



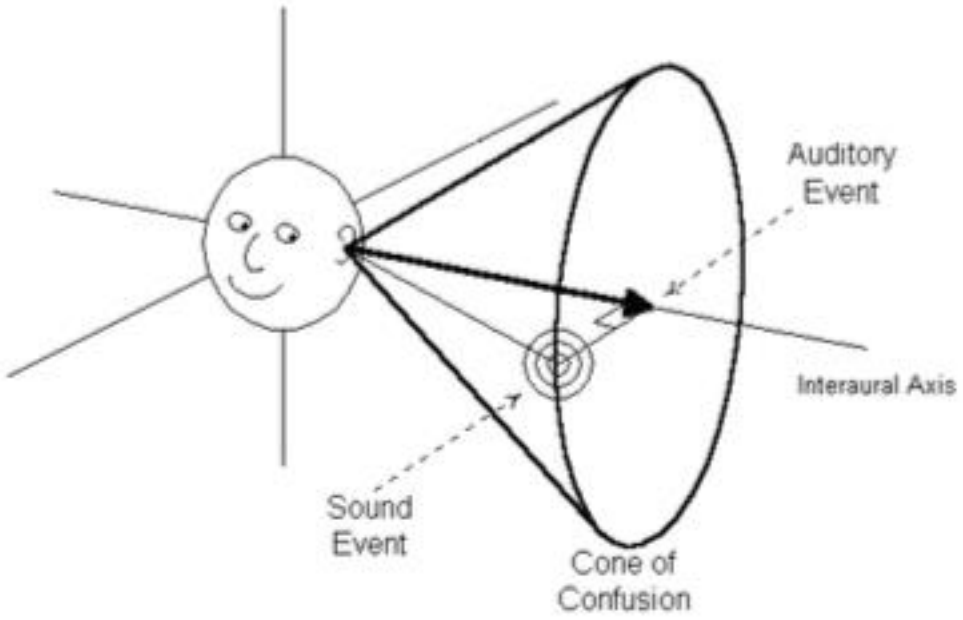
Monaural Cues and localization

- amplitude decrease quadratically w.r.t. distance;
- distant sounds are heavily distorted due to different wavelength attenuation (so, if we know the original sound we can estimate its distance)
- ears are asymmetric; the auricle/pinna distort the sound to identify its source (especially elevation). We do not perceive such distortion, only use it for localization
- reflections and repetition in both ears are particularly important (room / indoor). Indoor we perceive multiple version of the same sound, we filter them and use echoes and reverb for *echolocation*



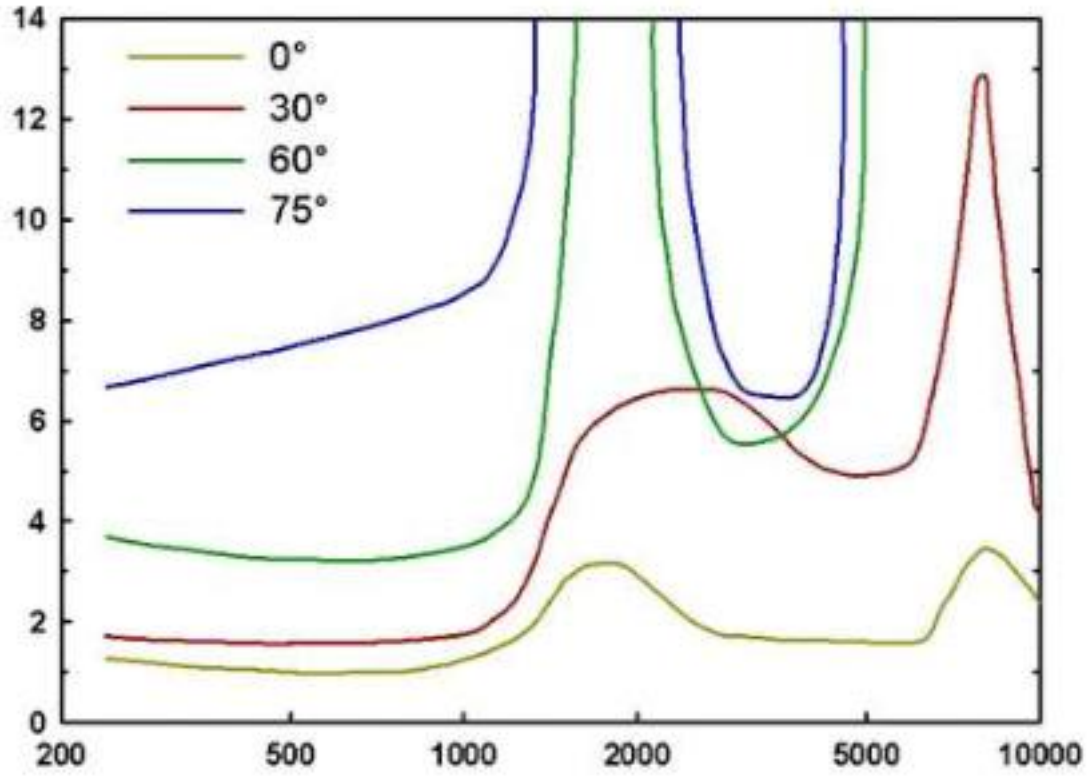
Binaural Cues and localization

- *Interaural level difference (ILD)* – sound magnitude difference – can perceive *acoustic shadows* (attenuation due to occlusion);
- *Interaural time difference (ITD)* – delay perceived between two years (~0.6ms between the ears) when listening the same source
- *Head motion* – Doppler effect - triangulation



We use binaural cues to create a *cone of confusion* where the audio source may be. Slight movements of the head are used to triangulate the sound source and reduce uncertainty

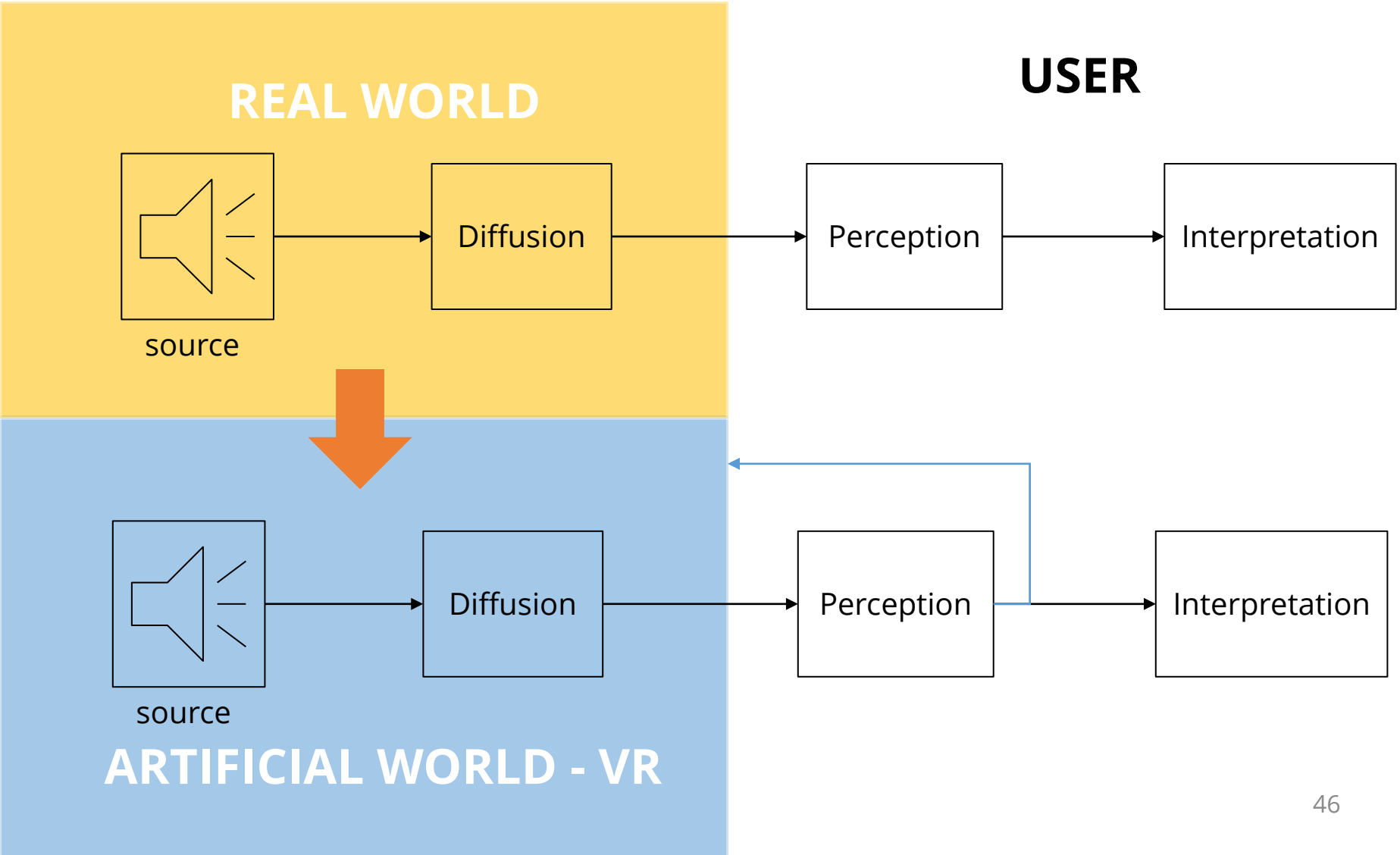
Minimum Audible Angle



Also here, we are better / worse in localizing HF/LF especially when we estimate Azimuth/Elevation

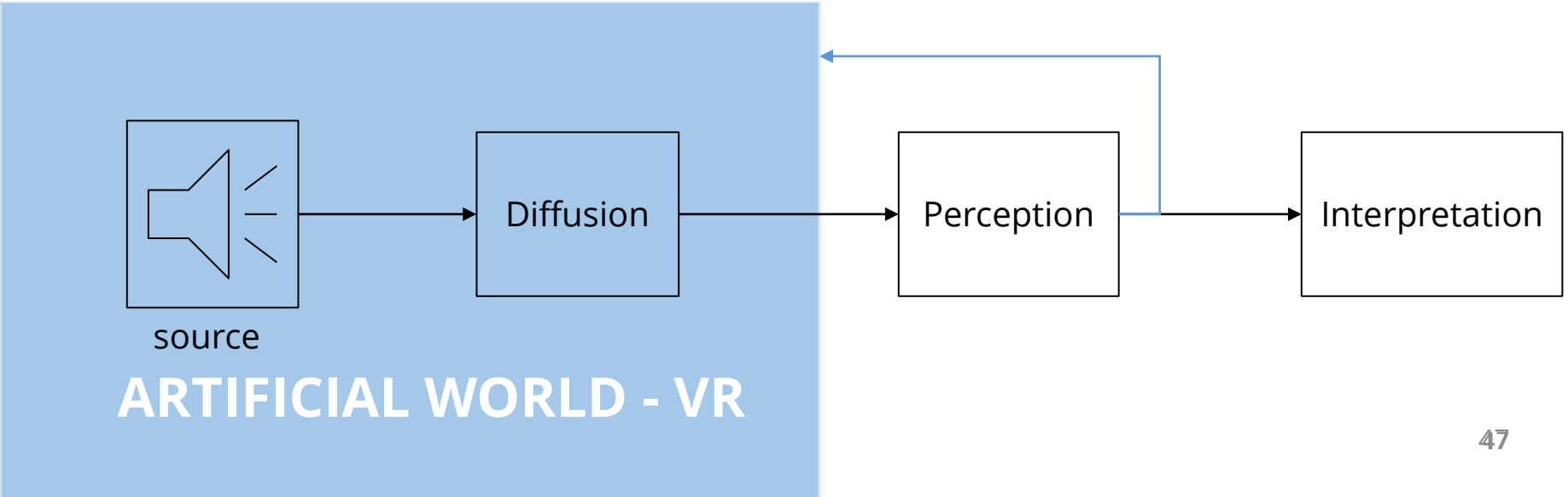
(colors indicate different Azimuth angles)

Audio modelling for VR



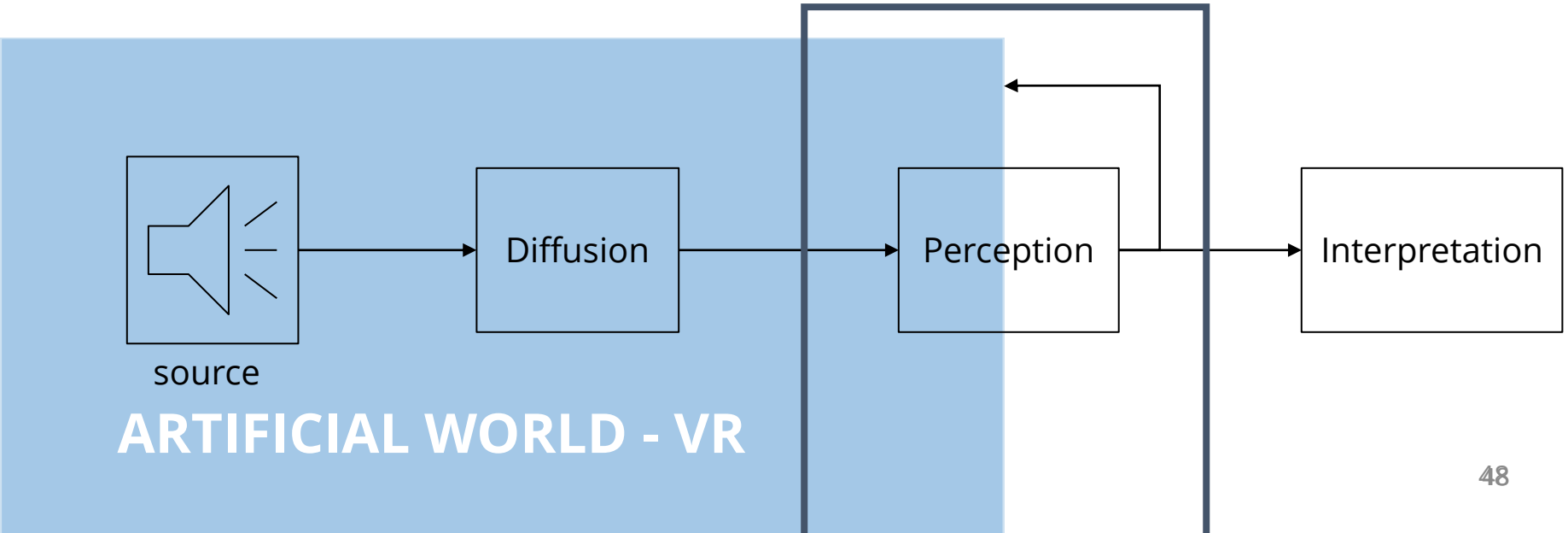
- What environment / world?
- What sounds? What samples?
  - How can we record good sounds for VR?
  - Which samples to use?
  - Where to put those?
- What propagation/diffusion model?

In order to have a proper spatialized audio – one that can be used in a AR or VR setting by a user for localizing the audio source / the movement of the audio source, we have do deal with all these questions.



- What environment / world?
- What sounds? What samples?
  - How can we record good sounds for VR?
  - Which samples to use?
  - Where to put those?
- What propagation/diffusion model?
- **What perception model?**

A key component of spatialized audio is to model also some characteristics of perception (how our ears perceive sound) so to control this step by exploiting the fact that earphones bypass the outer pinna / auricle



Auditory rendering

Virtual audio should be consistent with visual cues + past auditory experiences.
To do so we transform the signal according to the user movements

- Techniques: spectral decomposition – signal processing
  - Frequency domain
  - Fourier analysis
- *Filters* – transform and distort the signal
  - Sampling rate = 2*max frequency ~ 40 kHz = 44100 Hz

E.g.; Linear filters

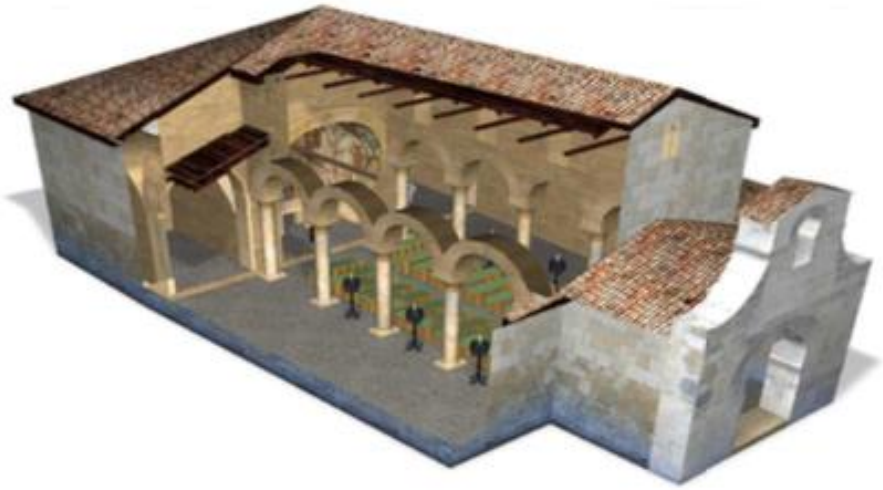
$$y[k] = c_0 x[k] + c_1 x[k-1] + c_2 x[k-2] + \ldots + c_n x[k-n]$$

$$y[k] = \frac{1}{2}x[k] + \frac{1}{4}x[k-1] + \frac{1}{8}x[k-2] + \frac{1}{16}x[k-4]$$

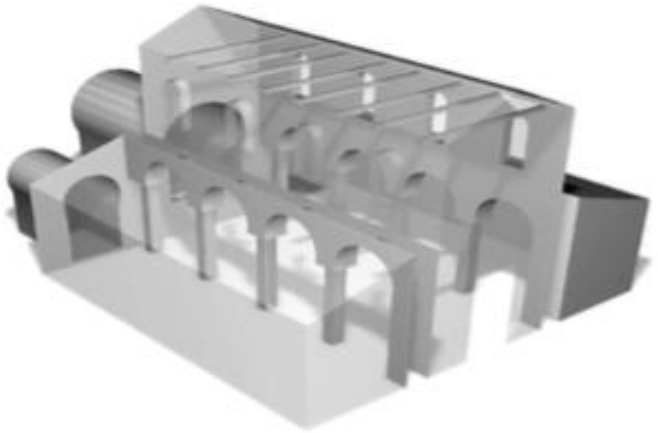
Example: exponential smoothing

Acoustic modelling

Room model for audio rendering can be much more simpler than those used for visual rendering



(a)



(b)

small objects are invisible to sound – spatial resolution of 0.5 m

We can focus only on large elements / materials

Acoustic modelling

Room model for audio rendering can be much more simpler than those used for visual rendering

However:

- different shapes reflect sound wave differently
- Sound waves propagates differently in materials
- Smaller objects / corrugated objects (e.g. bricks) can results in scattering

Audio propagation is difficult to simulate – for hi-fi performance, where user is static, the “best” solution is binaural recordings

Neumann KU1000

Acoustic modelling

Room model for audio rendering can be much more simpler than those used for visual rendering

Audio propagation is difficult– for hi-fi performance the “best” solution is binaural recordings.

However in binaural recordings the head is “fixed” – useful for replicate how the sound is perceived by a passive listener.

For VR a “360” audio is more useful – ambisonic sounds.

E.g. YouTube 360 supports First Order Ambisonic Sounds (L/R/F/B)

Ambisonic Sounds

Ambisonics sounds are full-sphere recordings, trying to record the sound at 360°.

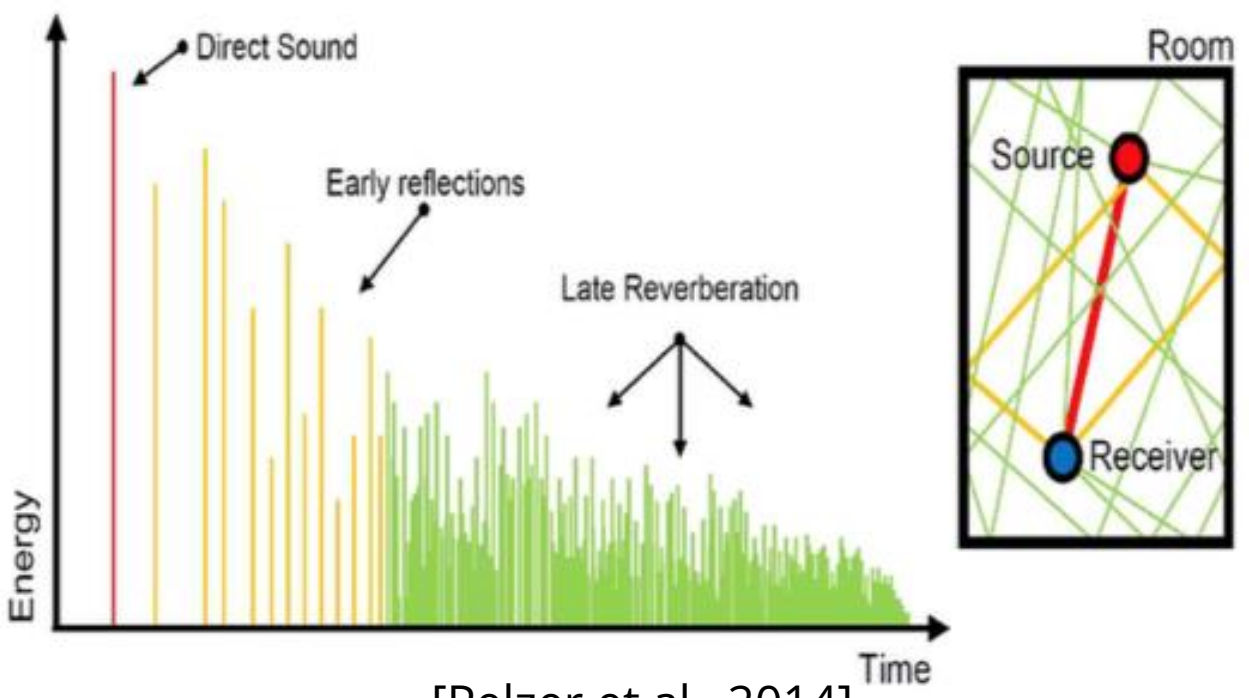
Note that a 360° recording is very different from what we perceive as we “listen” to the sound *after* our brain processed it, while ambisonics capture the raw signal.

When using ambisonics is the artificial world/engine that processes the sound so that the users perceives it correctly

Acoustic modeling

Two components:

1. how sound propagates in the environment
2. how it's perceived



[Pelzer et al., 2014]

Acoustic modeling

Two components:

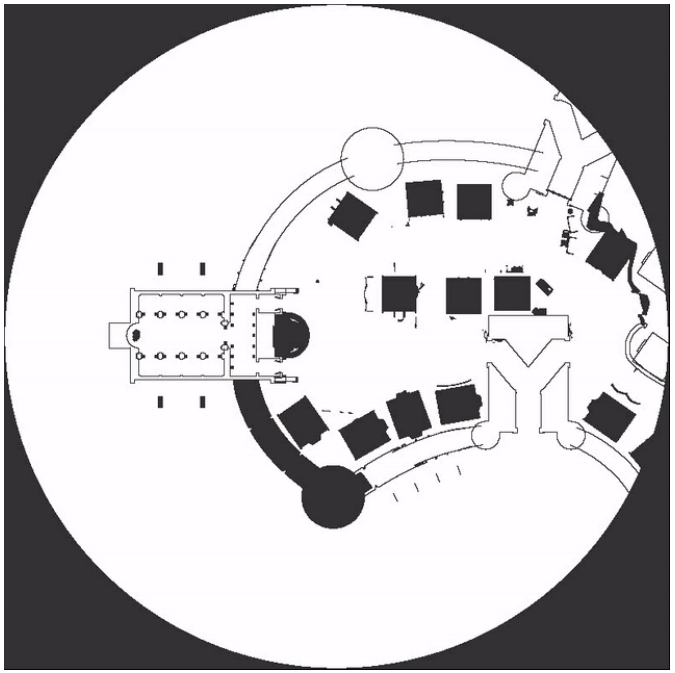
1. how sound propagates in the environment
2. how it's perceived

We can compute wave propagation:
Helmholtz wave equation: closed form solution often do not exist, computationally expensive;

Sound propagation

Microsoft Project Acoustic:

- Compute a propagation model for a given environment using a cloud based solution on Azure
- Pre-computed model is used to process real-time wave diffusion



https://www.microsoft.com/en-us/research/project/project-triton/

Head-Related Transfer Function (**HRTF**)

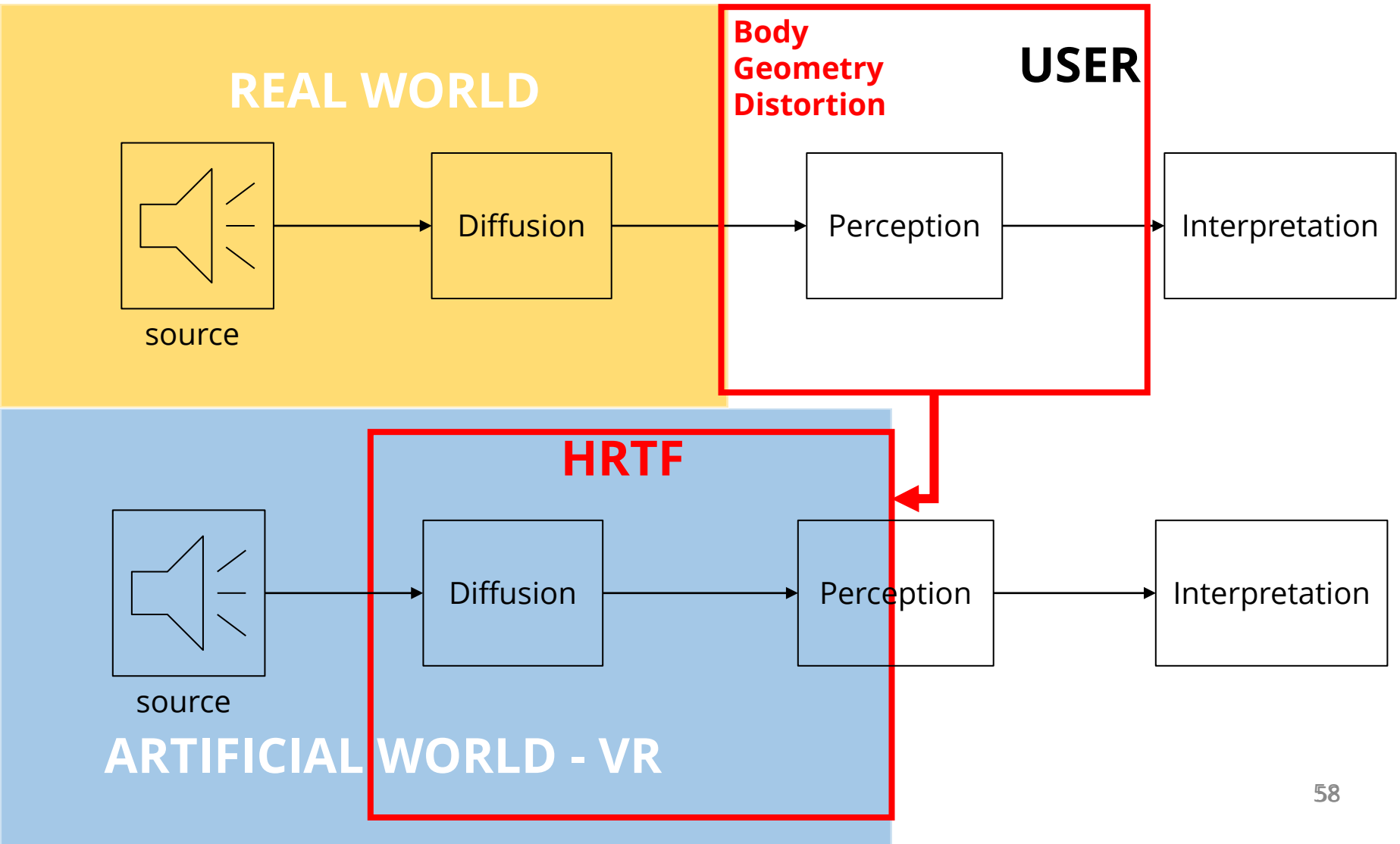
Two components:

1. how sound propagates in the environment
2. how it's perceived

Percepiton involves modelling distortion due to our own head and ears.
This is done using
Head-Related Transfer Functions (**HRTF**) are:

- Linear filter that distort sound by simulating how the sound is perceived source-head
- Model are approximate (HRTF "should" be tailored on each user measuring inner-ear components)

Head-Related Transfer Function (**HRTF**)



Head-Related Transfer Function

HRTFs provide influenced filtering of the sound applied to both ears:

- micro-delay between ears
- Modeling the directional filtering that ear-flaps, the head itself and the shoulders contribute to.

Adding HRTF filtering already immensely improves the sensation of direction over a conventional panning solution.

Direct HRTF is limited as it only is concerned with the direct path of audio and not how it is transmitted in space (e.g., occlusion) – and sound reflection (similar to global illumination problem in graphics).

Unity provides an interface for Spatial Audio using HRTF
https://docs.unity3d.com/Manual/AudioSpatializerSDK.html

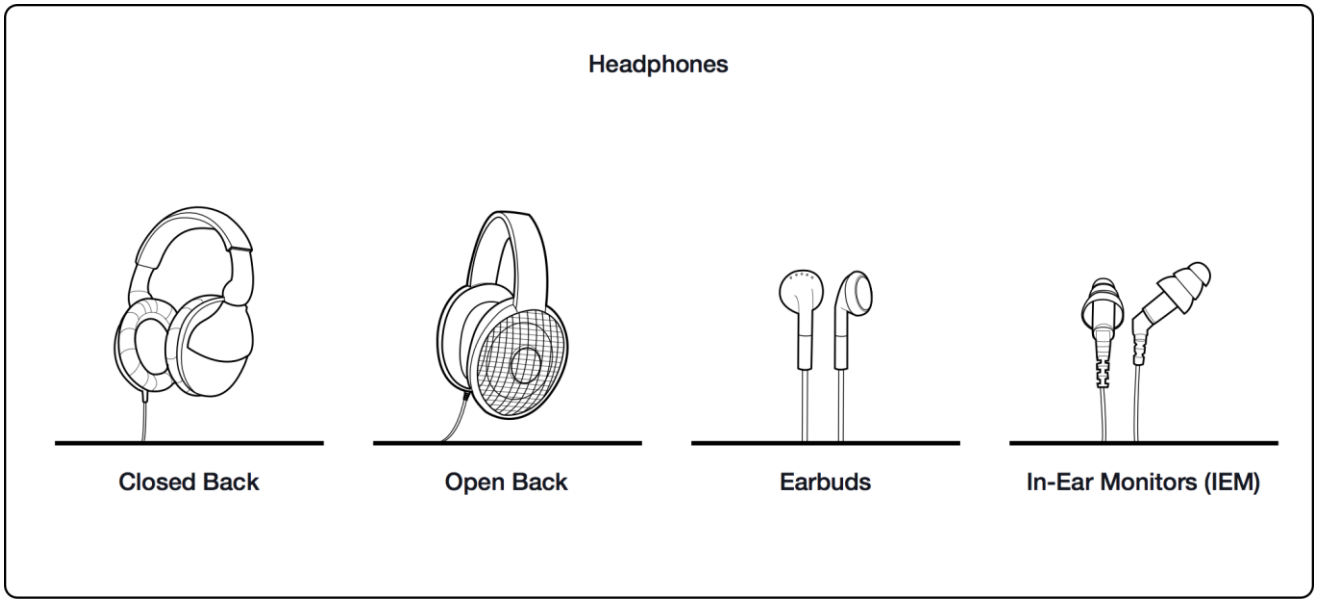
Head-Related Transfer Function

- HRTF are a highly-parametrized transfer function which model the head-related distortion *for each user*
- HRTF computes how audio is perceived from inside the user head
- Using headpones you can bypass the distortion of your own ears and use the simulated one instead

However, the estimation of such parameters is complex

- Solution: *«borrow»* the head of an average listener using an average HRTF of someone else

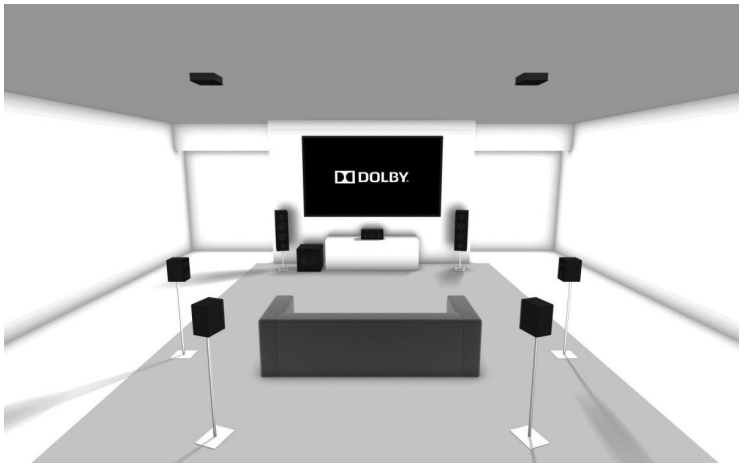
Audio devices for VR



Audio devices for VR

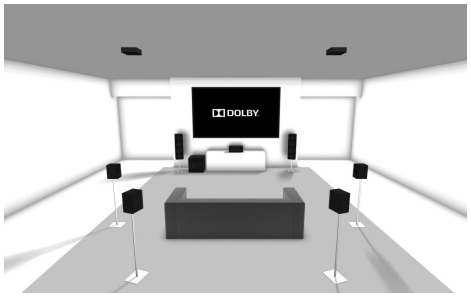
Several type of output devices:

- Stereo Speakers
- Surround Systems
- Headphones



Bluetooth introduce latency (up to 0.5s) which can be annoying for VR, so it should be avoided

External Speakers



External speaker are world-fixed, so not the best solution for VR

- Imprecise imaging due to panning over large portions of the listening area.
- No elevation cues. Sounds only appear in a 360 degree circle around the listener.
- Assumption of immobile listener; in particular, no head tracking.
- (real) Room effects such as reverberation and reflections impact the reproduced sound.
- Poor isolation means that outside sounds can intrude on the VR experience.

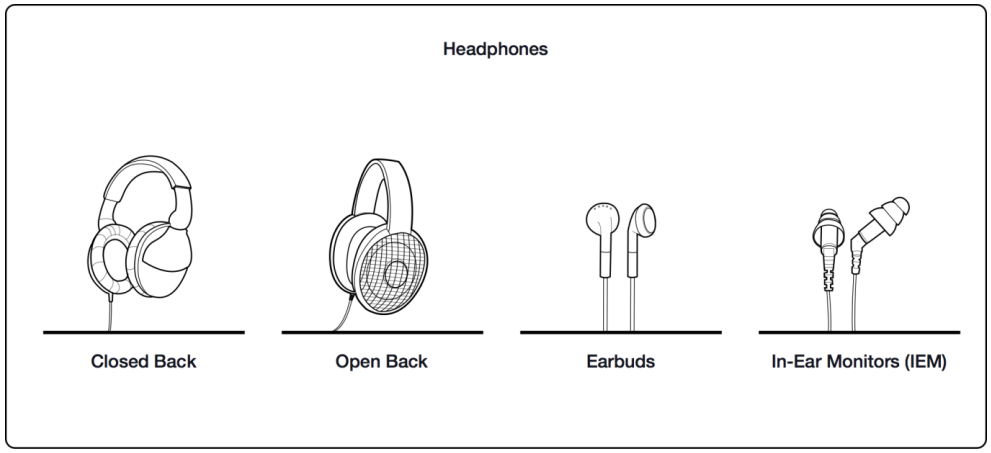
But:

- LF could be “felt” and perceived using subwoofers

Headphones

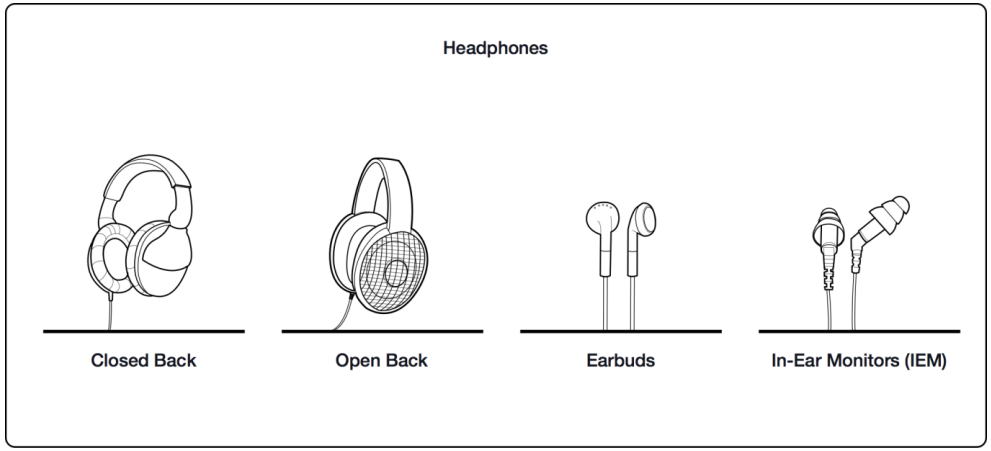
Headphones are the best option for VR:

- Move with user head
- Acoustic isolation
- Don’t suffer from the “doubling down” of HRTF effects, i.e. sounds being modified from the simulated HRTF, and again by the listener’s actual body geometry



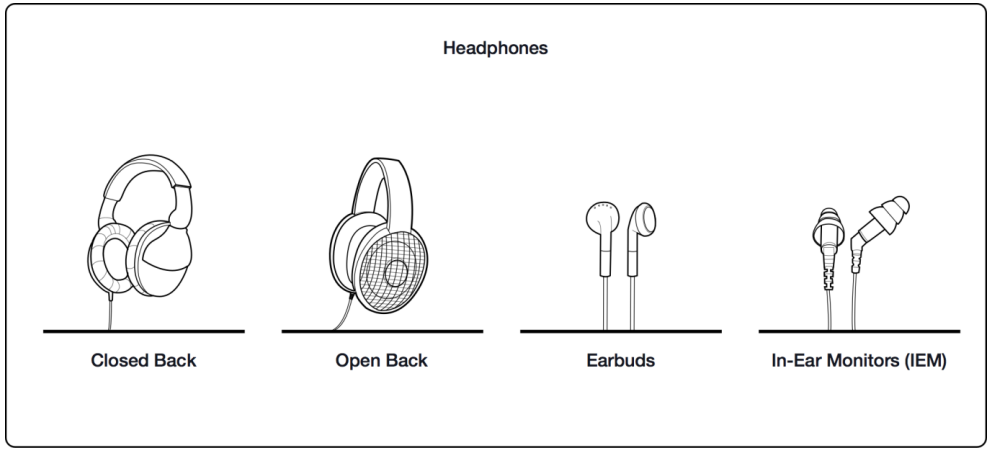
On-ear Headphones

- On-ear headphones reproduce both LF and HF
- LF are less “felt” than subwoofers, which can be used in combination with them
- Provide isolation (not optimal though)
- Outer-ear distortion and HRTF double-down effect slightly present



In-ear Headphones

- In-ear headphones / earbuds poorly reproduces LF
- Better isolation
- Outer-ear distortion and HRTF entirely removed



Try it yourself

https://cdn.rawgit.com/resonance-audio/resonance-audio-web-sdk/master/examples/birds.html

References

Steven La Valle, Virtual Reality, Cambridge Press

http://lavalle.pl/vr/

Book Available (free) online on the page of the author