An Intro to Immersive Audio for Heritage VR

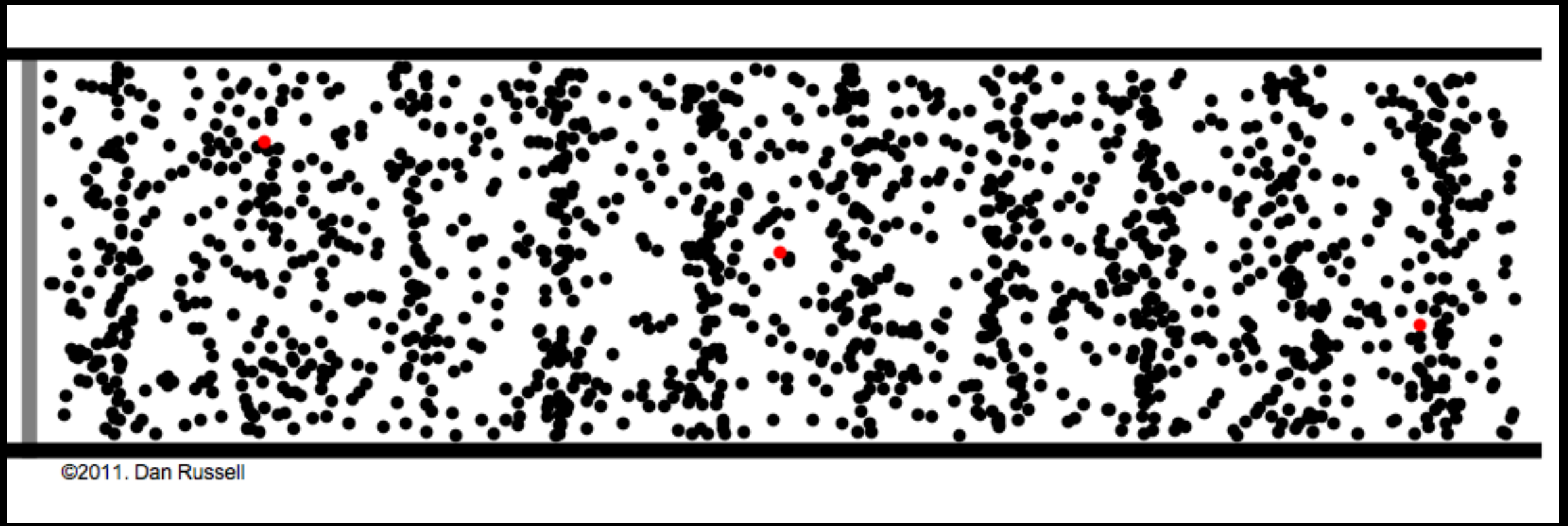
COLLARTS

Lecture overview:

- Basics of audio perception, audio localisation
- Intro to immersive audio
- Commercial systems (stereo, 5.1, 7.1, 10.2, Atmos etc)
- Non-commercial & research systems (Ambisonics, wavefield synthesis)
- HRTFs & head tracking
- Comparisons and implications for listeners and production
- Real world systems development tools & specification
- Uses of immersive data (verisimilitude / reality, data exploration, etc)
- Latest research

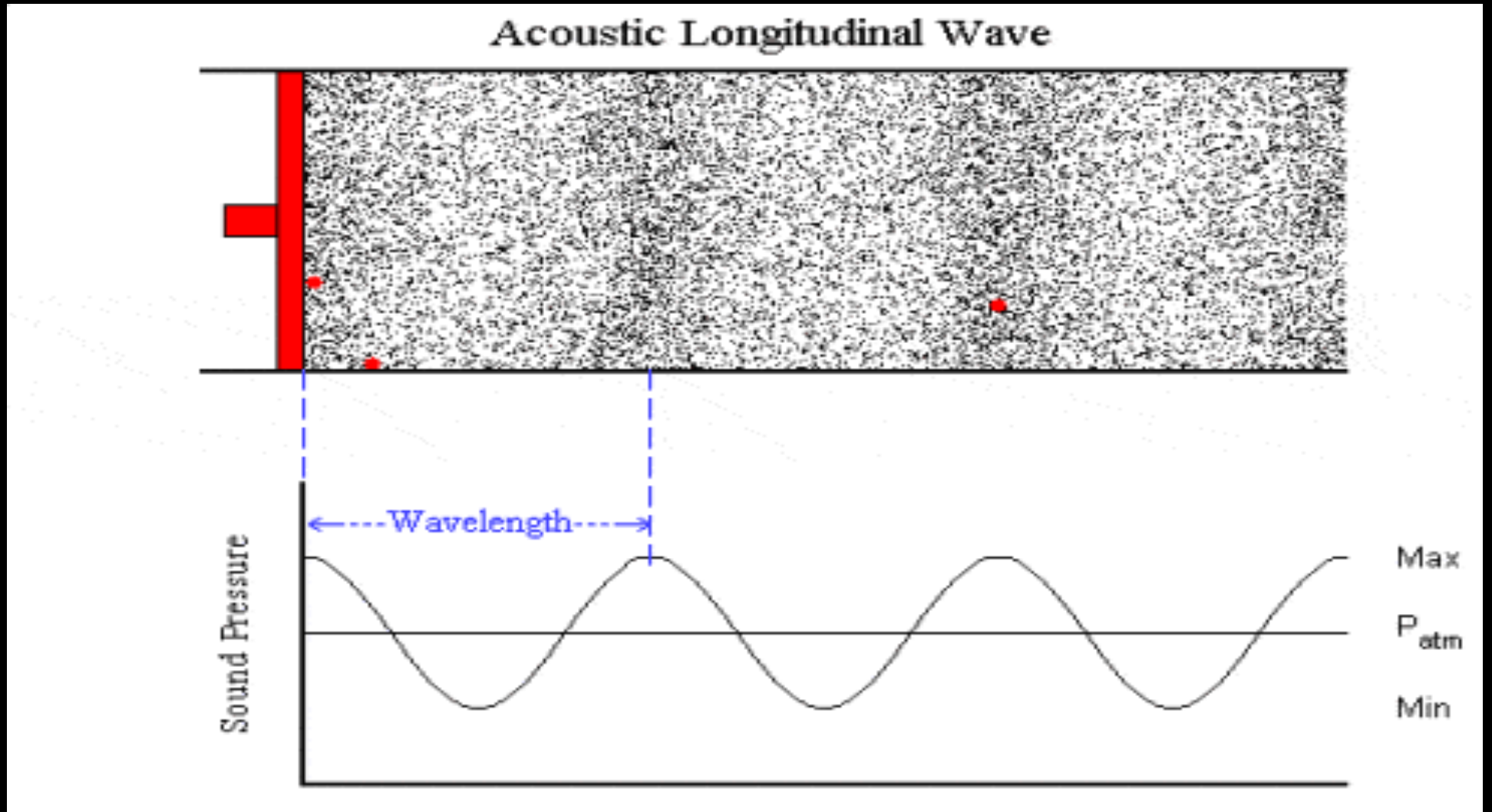
A quick recap – what is sound?:

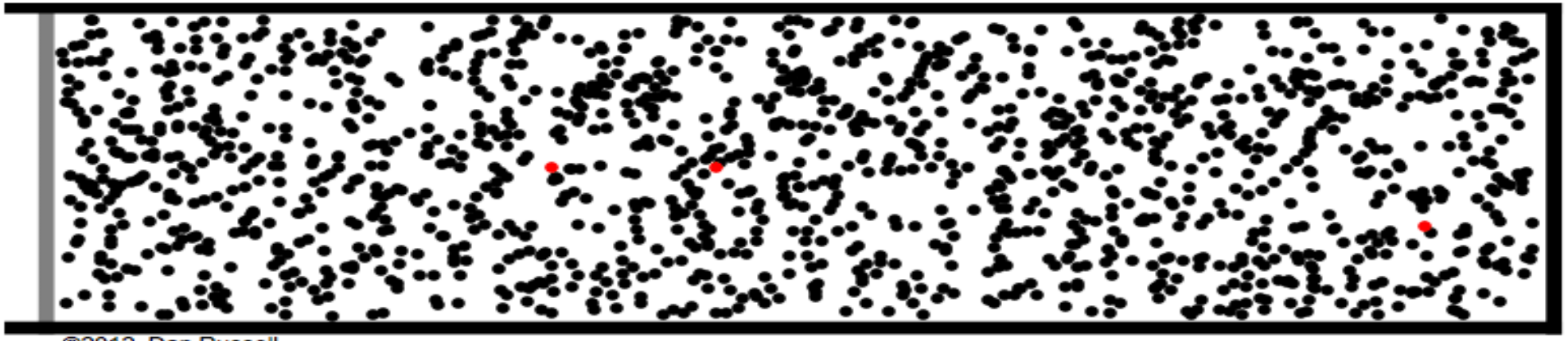
- Sound is pressure waves in a medium, typically air

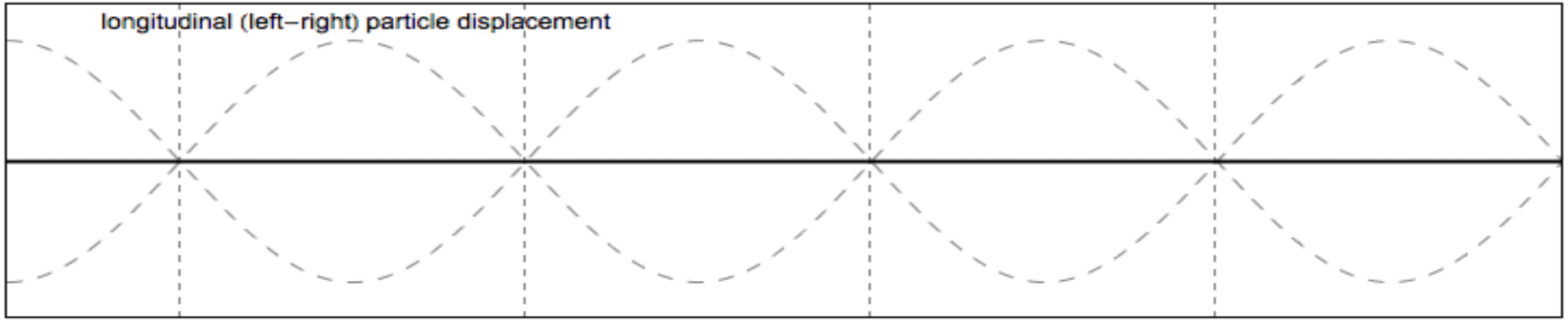


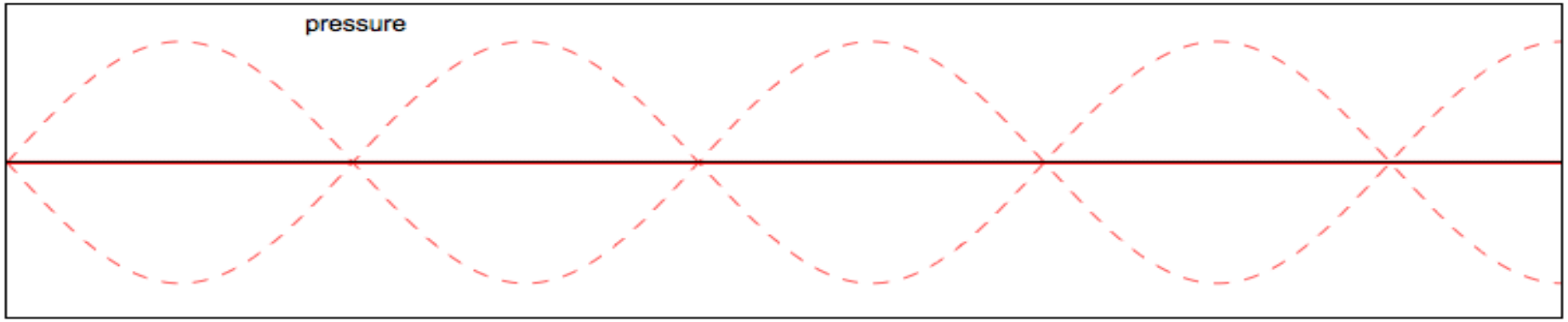
A quick recap – what is sound?:

- Sound is pressure waves in a medium, typically air







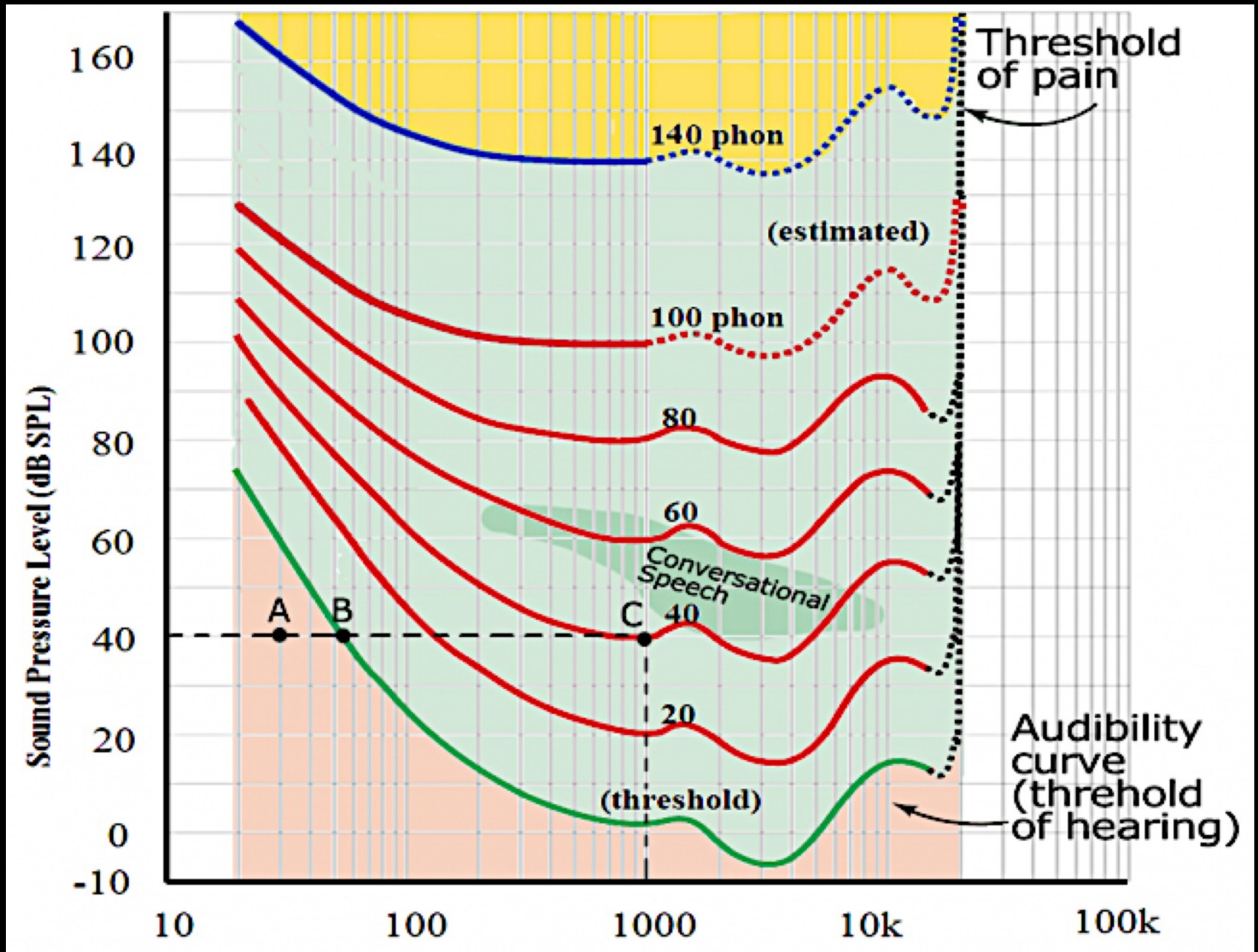


A quick recap – what is sound, how do we perceive it?:

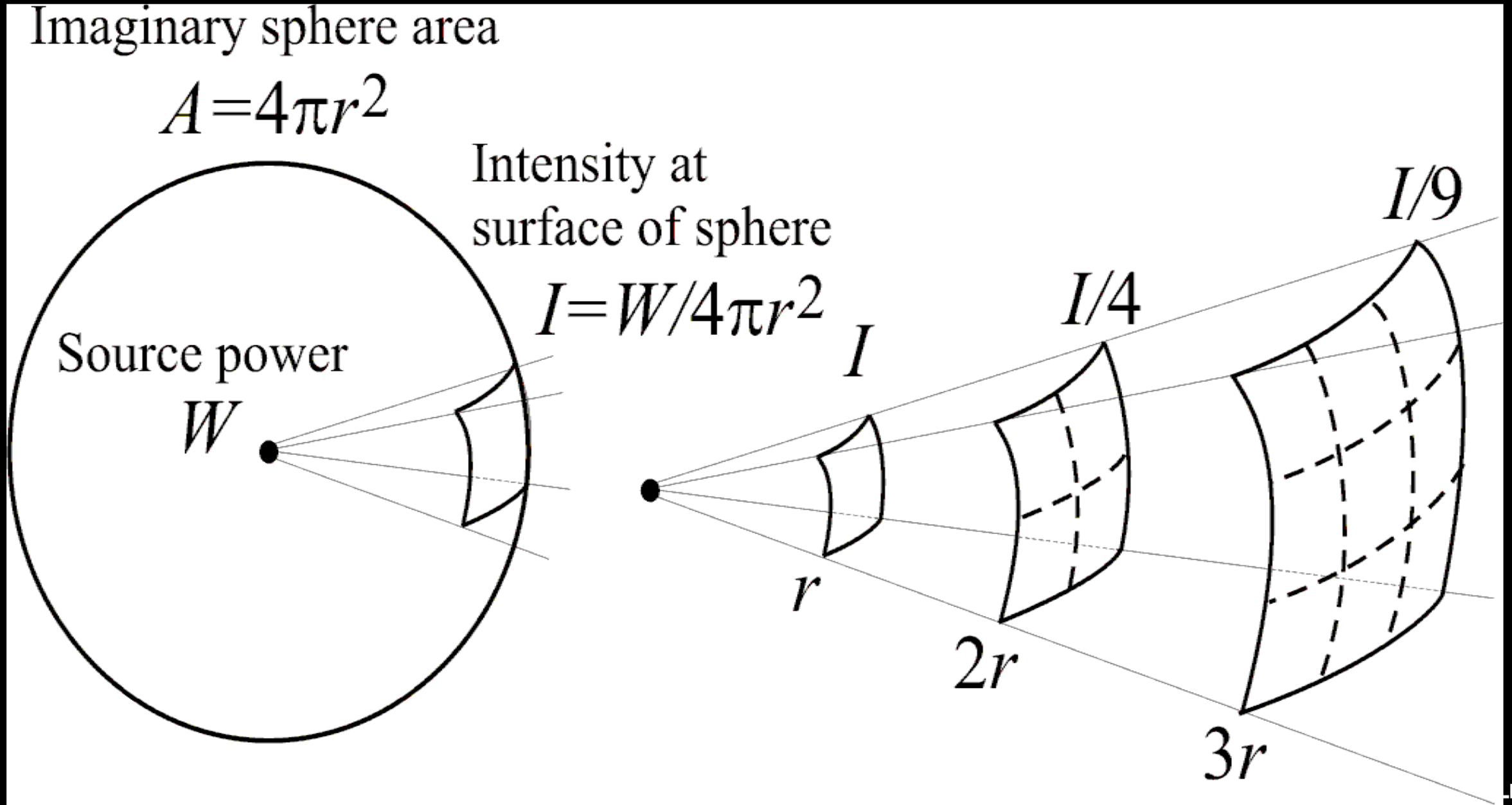
- Sounds have frequency and amplitude.
- Humans respond to audible frequencies from 20Hz to 20KHz.
- We perceive frequency logarithmically.
- Humans perceive a very wide range of amplitudes & frequencies.
- 20µP at 1KHz is the threshold of human hearing, defined as 0dB
- We perceive amplitude logarithmically, measured in deci Bells (dB) to the threshold of hearing. dB = 20 log10 (p2 / p1)
- 1dB is defined as the just noticeable level difference at 1KHz.
- 140dB is the threshold of pain in the ear.
- Exposure to loud sounds causes ***permanent*** hearing loss.

A quick recap – what is sound, how do we perceive it?:

Equal loudness curves

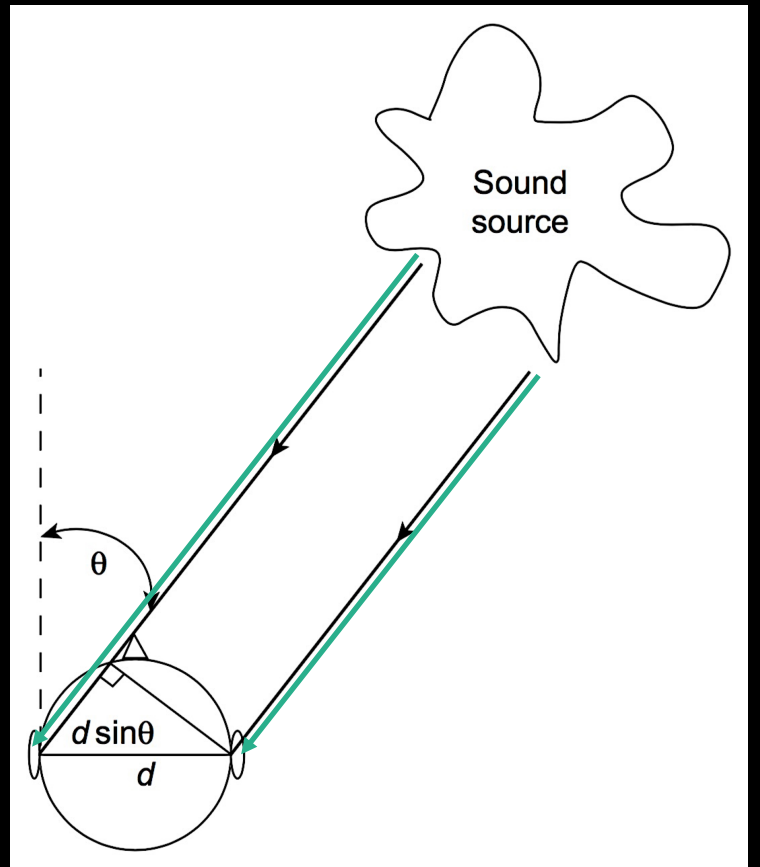


A quick recap – In free space sound propagates spherically:



Basics of audio perception, audio localisation:

- We use 2 ears to localise sounds through 3 mechanisms:
  - Interaural Time Differences (ITD) [LF < 700Hz]
  - Interaural Level Differences (ILD) [HF > 2.5KHz]
  - Pinna effects & head movements

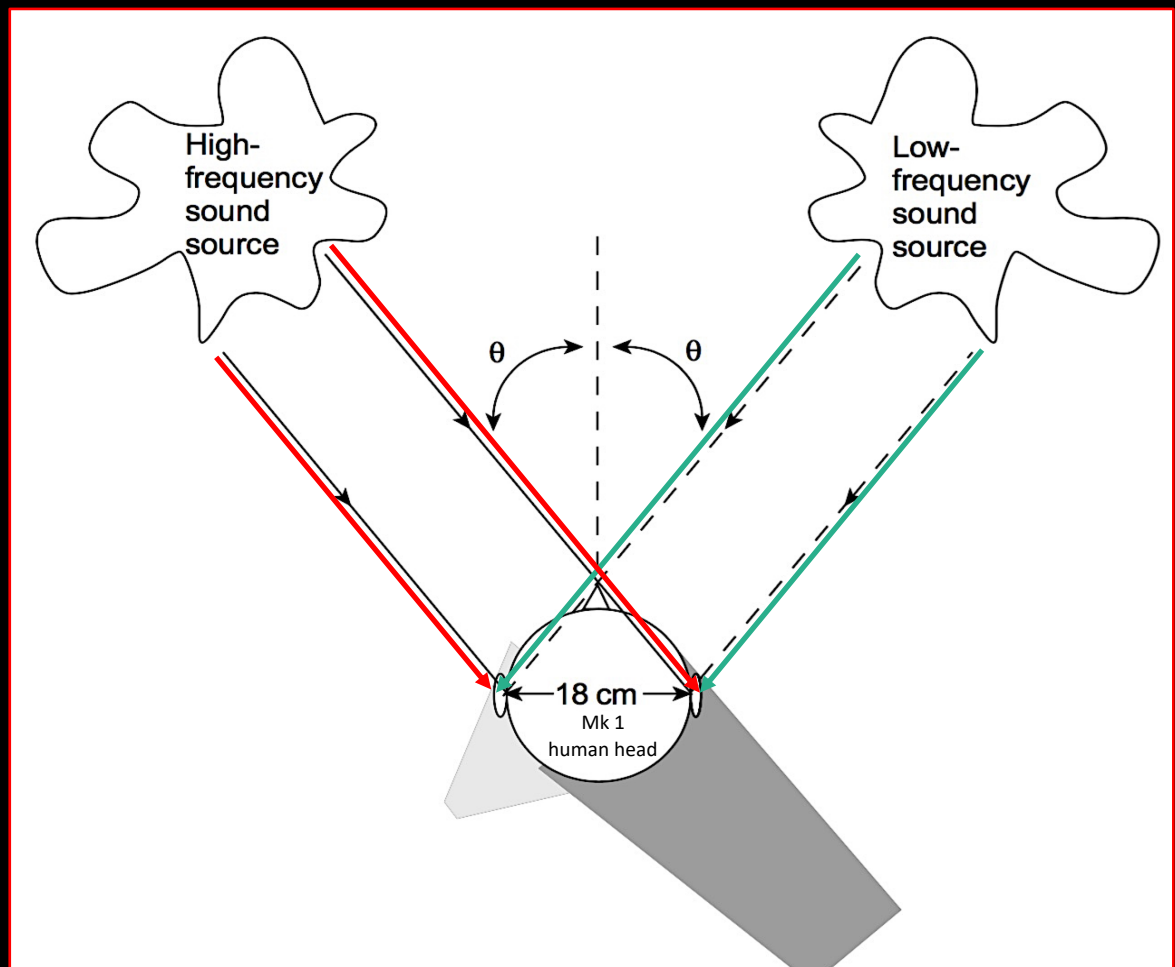


At low frequencies, the sound arrives at the furthest ear later

Basics of audio perception, audio localisation:

- We use 2 ears to localise sounds through 3 mechanisms:
  - Interaural Time Differences (ITD) [LF < 700Hz]
  - Interaural Level Differences (ILD) [HF > 2.5KHz]

At high frequencies, the sound arrives at the furthest ear quieter

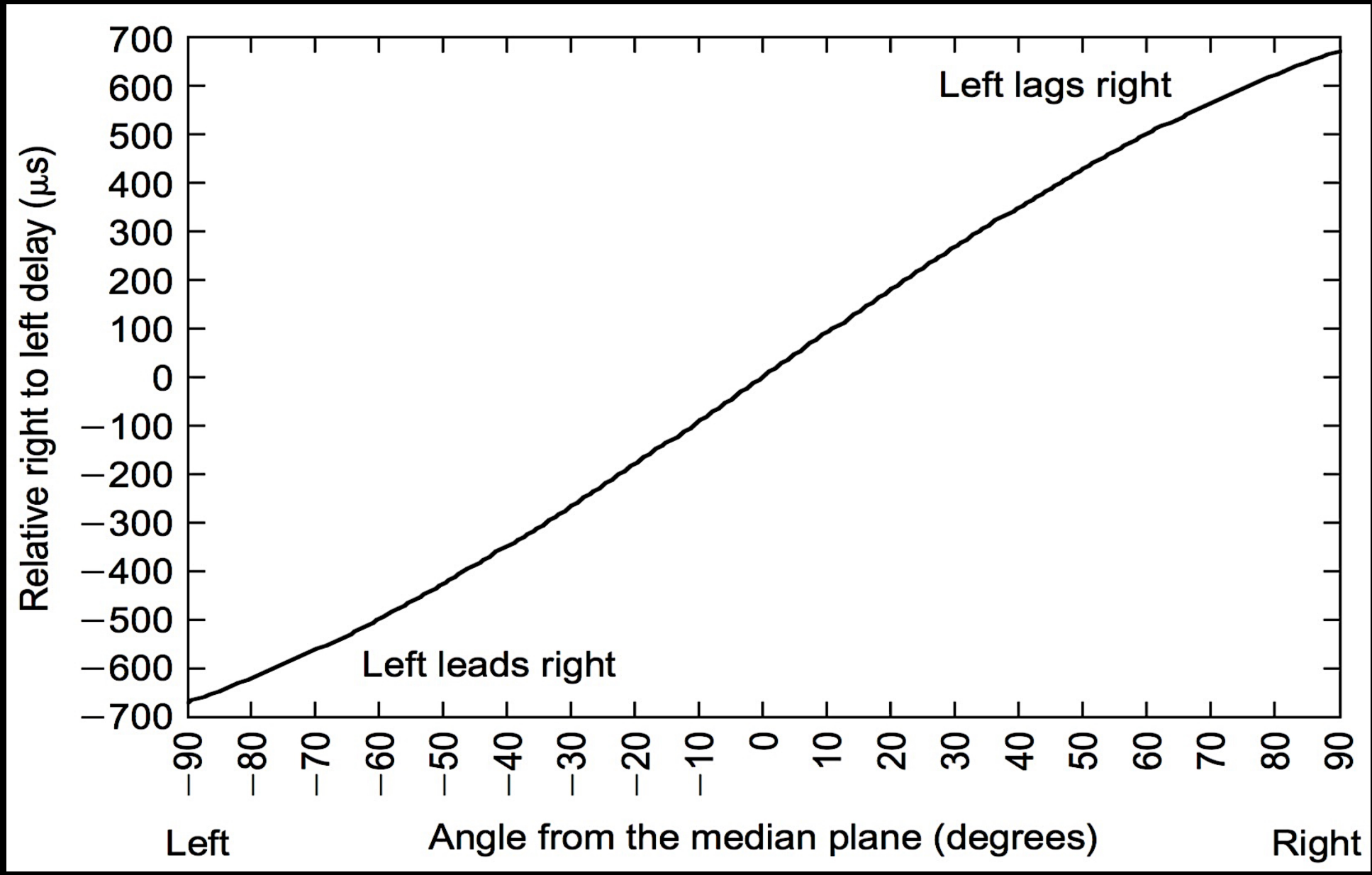


At low frequencies, the sound arrives at the furthest ear later

Basics of audio perception, audio localisation: ITD

- Our ears are separated by about 18cm, so there is a time difference for sounds offset from the median plane.
- When the sound is off to the left the left ear will receive the sound first, and when it is off to the right the right ear will hear it first.
- If the sound is directly in front, or behind, or anywhere on the median plane, the sound will arrive at both ears simultaneously.
- The time difference between the two ears will depend on the difference in the lengths that the two sounds have to travel.

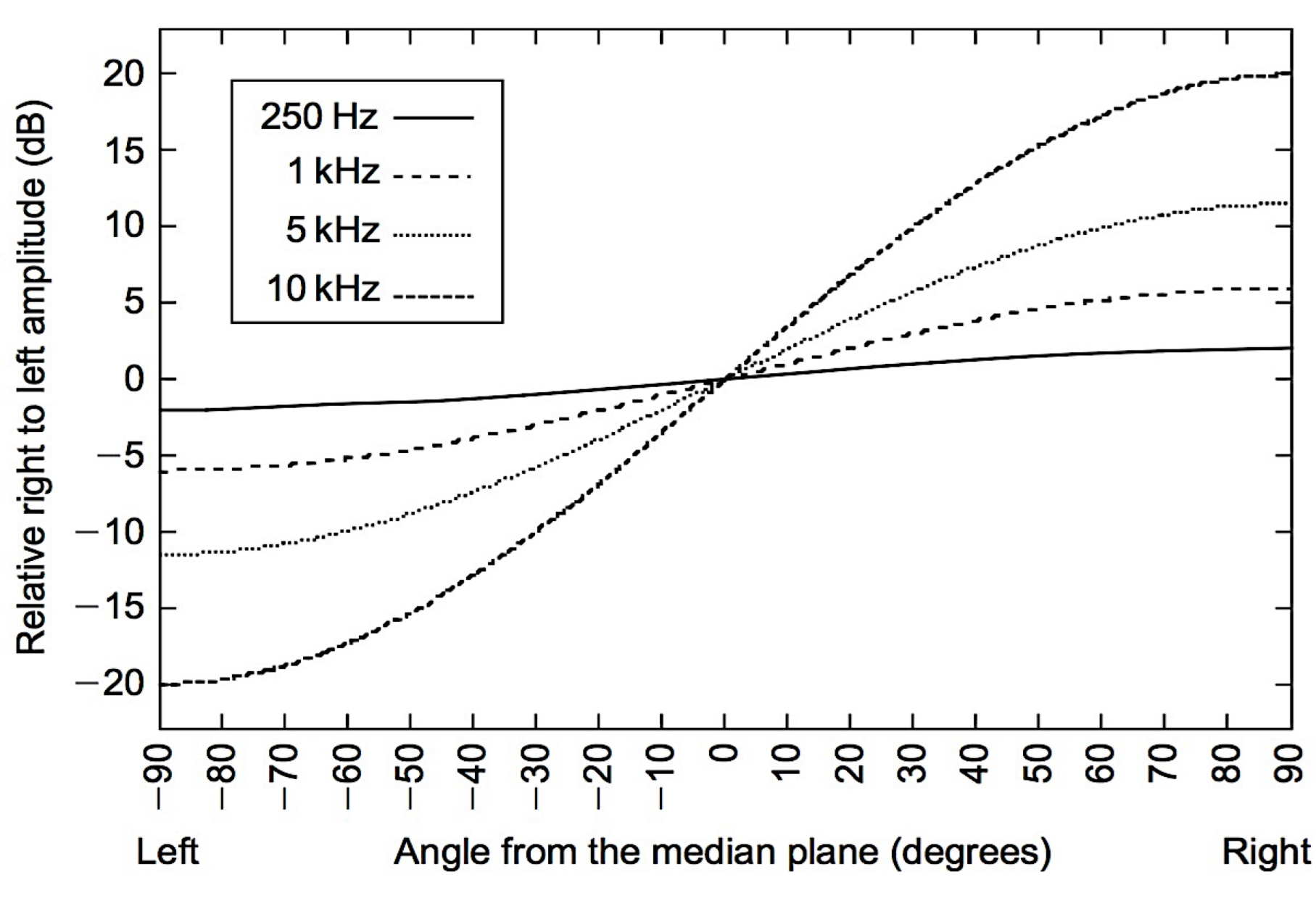
Basics of audio perception, ITD:



Basics of audio perception, audio localisation: IID

- The other main cue that is used to detect the direction of the sound is the different levels of intensity or “loudness” at each ear due to the shading effect of the head.
- The levels at each ear are equal when the sound source is on the median plane but the level at one ear progressively reduces, and increases at the other, as the source moves away from the median plane.
- The level reduces in the ear that is furthest away from the source. *This effect is frequency dependent.*
- The shading effect of the head is harder to calculate. Experiments indicate that the intensity ratio between the two ears varies sinusoidally from 0 dB up to 20 dB, depending on direction & frequency, see below.

Basics of audio perception, ILD:



Basics of audio perception:

Interaural Intensity / Time difference (IID / ITD)

- Interaural intensity difference is a cue for direction at high frequencies
- interaural time difference is a cue for direction at low frequencies.

- Note that the cross-over between the two techniques starts at about 700 Hz and would be complete at about four times this frequency at 2.8 kHz.
- In between these two frequencies the ability of our ears to resolve direction is not as good as at other frequencies.

Basics of audio perception:

Pinnae and Head Movement Effects

The above models of directional hearing do not explain how we can resolve front to back ambiguities or the elevation of the source. These are explained by two other aspects of hearing.

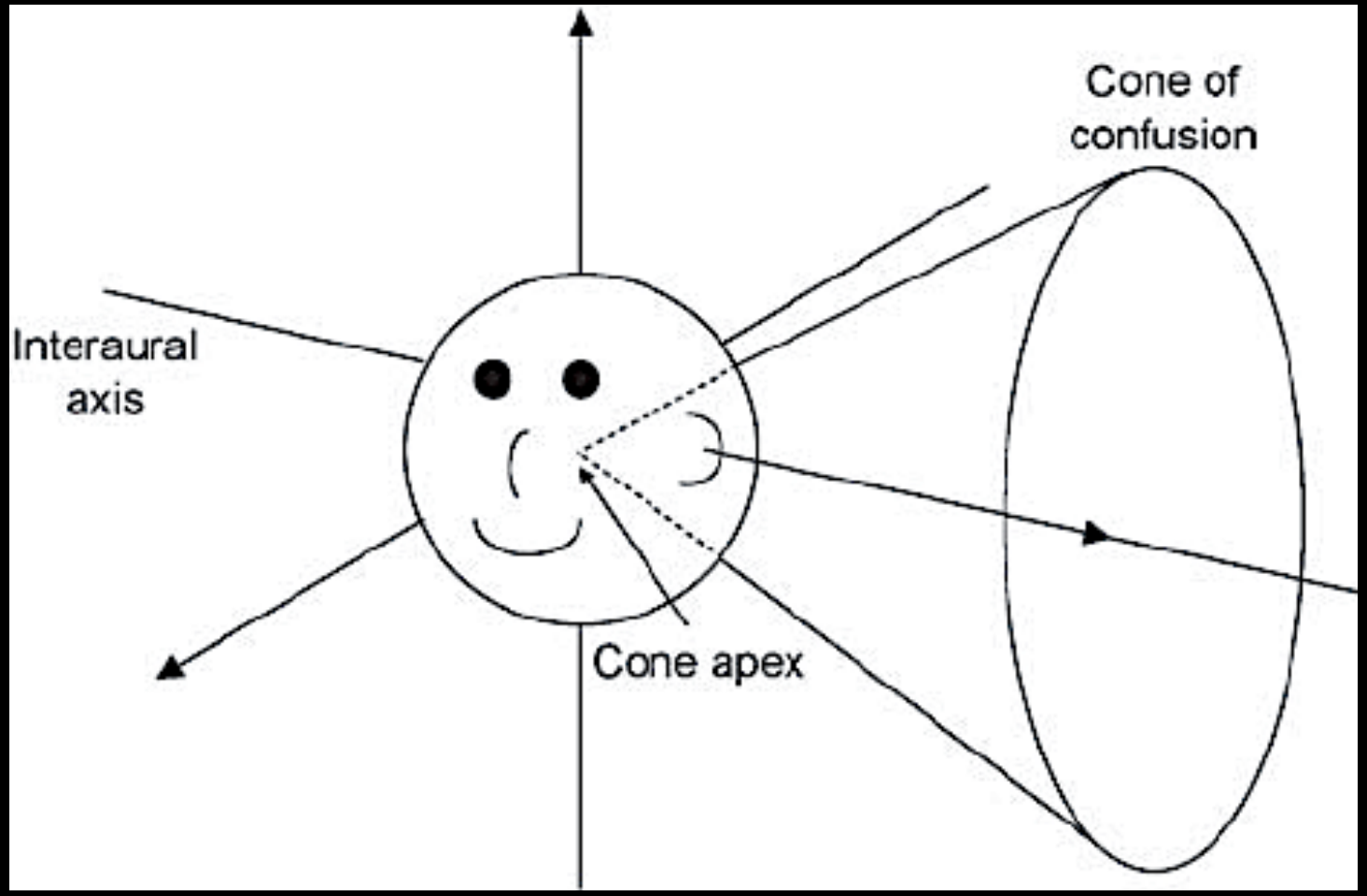
- First is to use the effect of our pinnae on the sounds we receive from different directions to resolve the angle and direction of the sound. The pinnae's set of complex ridges cause reflections (very small but significant) so cause comb filter interference effects on the sound the ear receives that are unique to its direction of arrival, in all three dimensions. We use these cues to resolve ambiguities in direction that are not resolved by the main directional hearing mechanism. The delays are very small and so these effects occur at high audio frequencies, typically above 5 kHz.

- The second, and powerful, means of resolving directional ambiguities is to move our heads.

Basics of audio perception: Cone of confusion

Sound localisation is less precise the further the sound is from the front.

The region of uncertainty is called the “cone of confusion”.



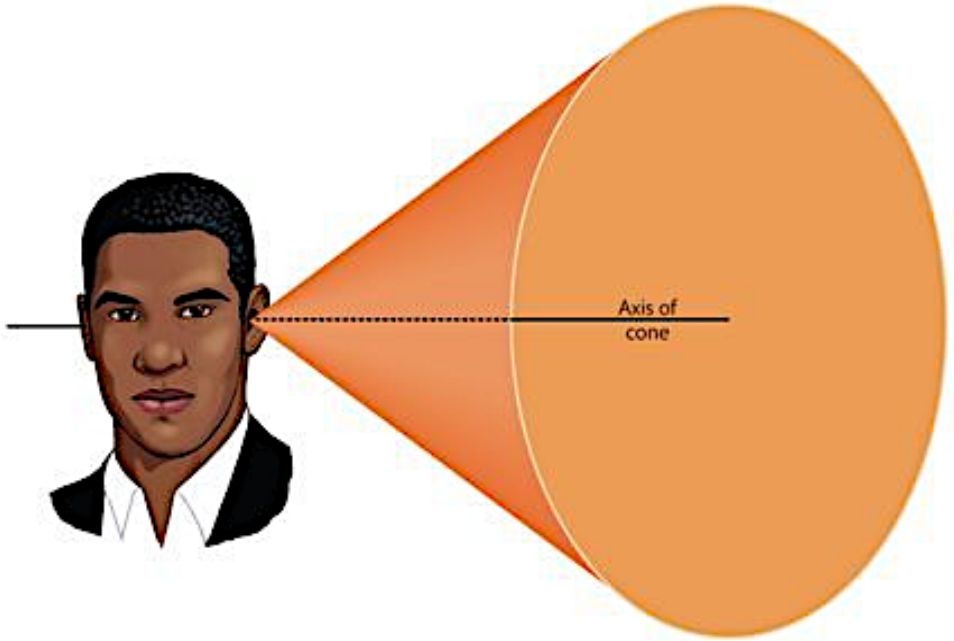
(not the cone of silence)

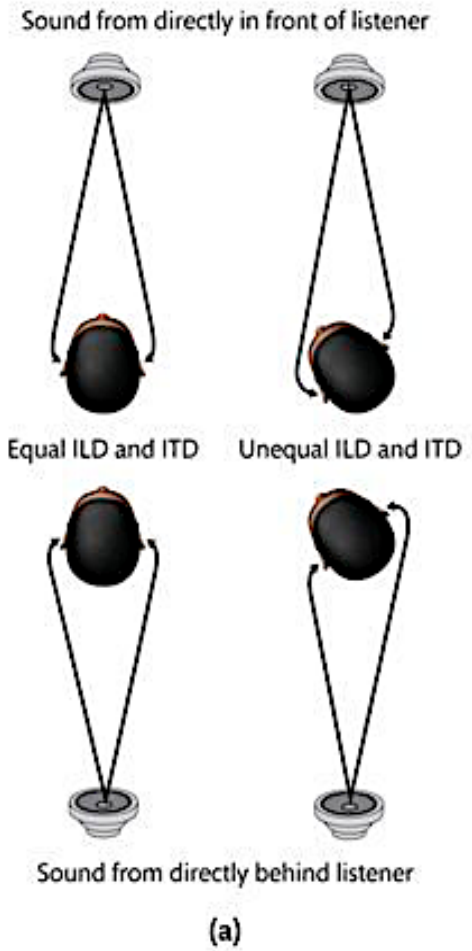
Basics of audio perception: Cone of confusion

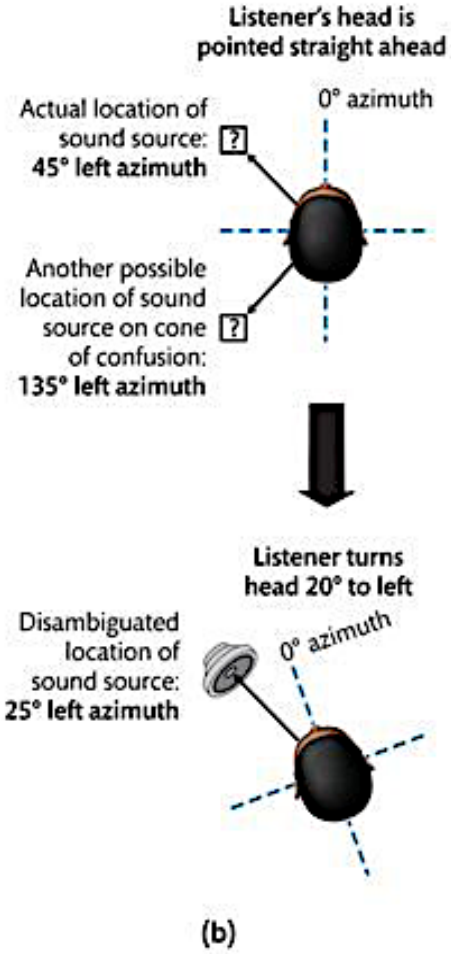
The ambiguity is resolved with head movement.

Cone of confusion

- A hypothetical cone-shaped surface in auditory space; when two equally distant sound sources are located on a cone of confusion, their locations are confusable because they have highly similar ILD and ITD



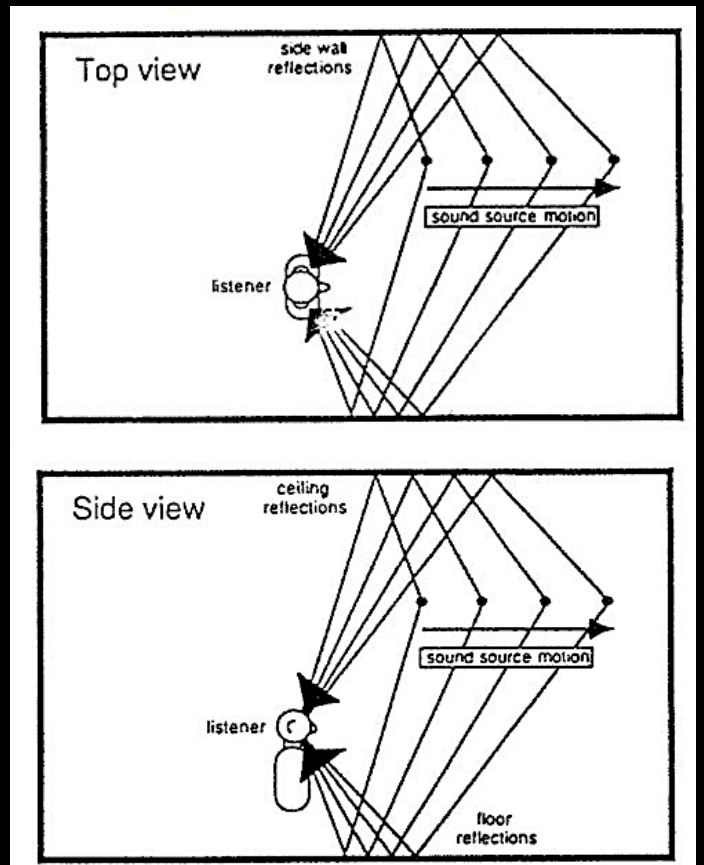


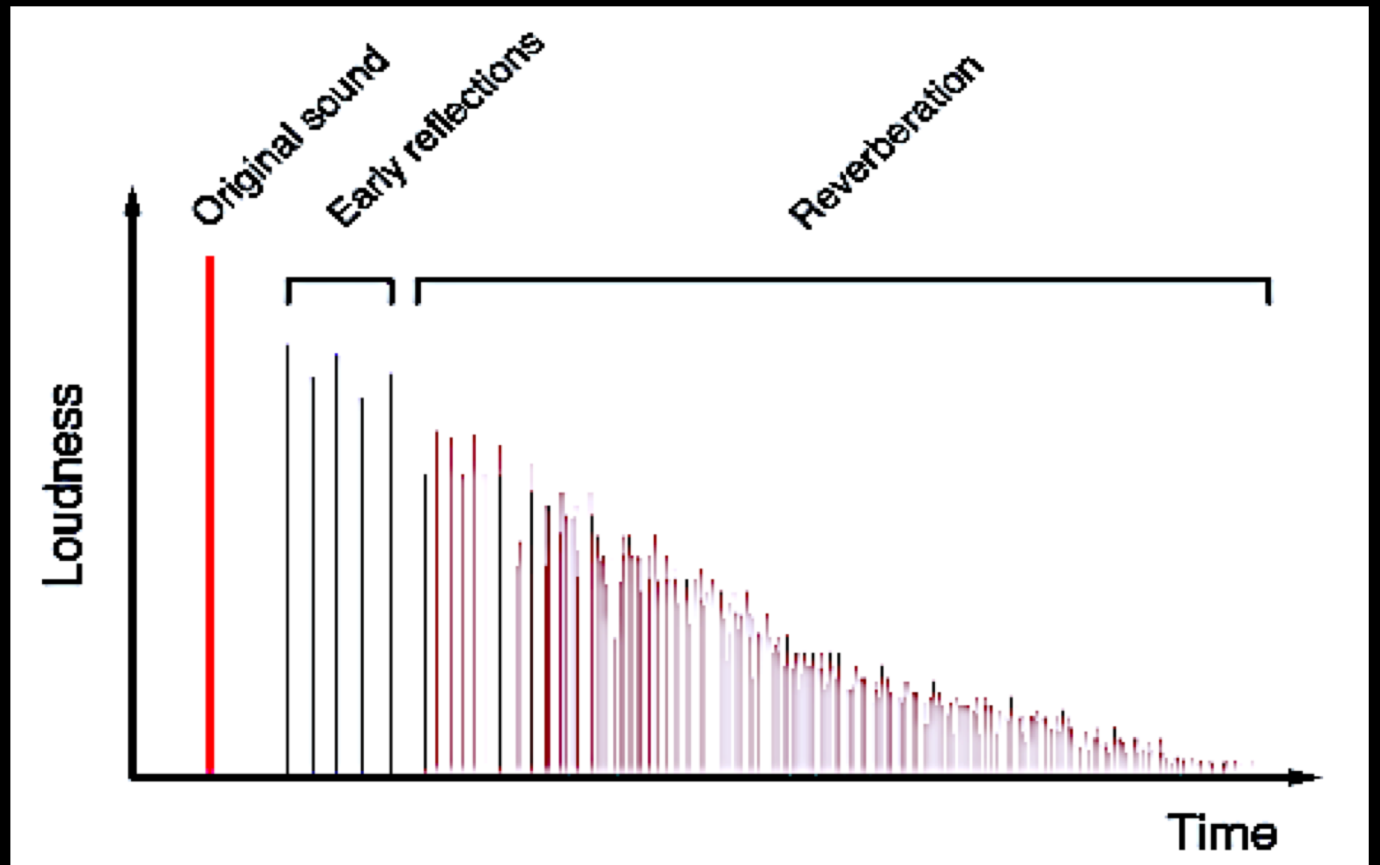


Basics of audio perception: Distance

The perceived distance of a sound is related to several parameters:

- How loud it is.
- The level of the reverberation compared to the original sound.
- The early reflections (first 50-80ms) from the surroundings.





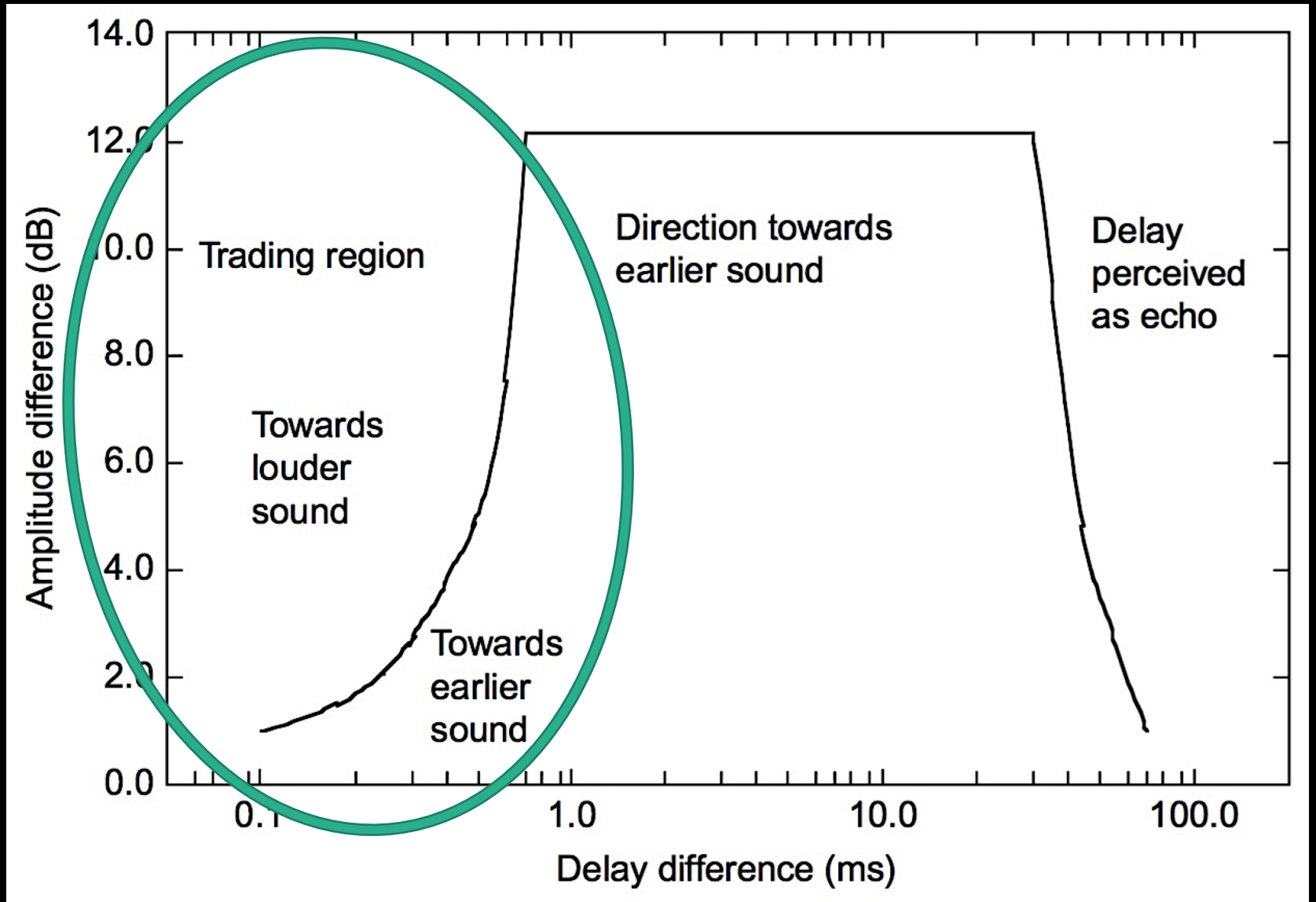
Basics of audio perception:

IID and ITD Trading - important as this is how stereo, 5.1, 7.1 etc. audio works!

- Both intensity and delay cues are used for the perception of sound source direction, they occur in similar areas of the brain and are linked together.
- There is some overlap in the way the cues are interpreted such that intensity can be confused with delay and vice versa in the brain.
- So the effect of one cue, for example delay, could be canceled out by the other, for example intensity.
- This effect does in fact happen and is known as “interaural time difference versus interaural intensity difference trading.” So, within limits, an ITD can be compensated for by an appropriate IID, see below.

Basics of audio perception:

IID and ITD Trading - important as this is how stereo, 5.1, 7.1 etc. audio works!



Basics of audio perception:

IID and ITD Trading

- As expected, time delay versus intensity trading is only effective over the range of delay times which correspond to the maximum interaural time delay of 673µs.
- Beyond this amount of delay, small intensity differences will not alter the perceived direction of the image, the sound will appear to come from the source which arrives first. This effect occurs between 673µs and 30ms.
- However, if the delayed sound's amplitude is more than 12 dB greater than the first arrival then we will perceive the direction of the sound to be towards the delayed sound.
- After 30ms the delayed signal is perceived as an echo and so the listener will be able to differentiate between the delayed and un-delayed sound.

Basics of audio perception:

IID and ITD Trading

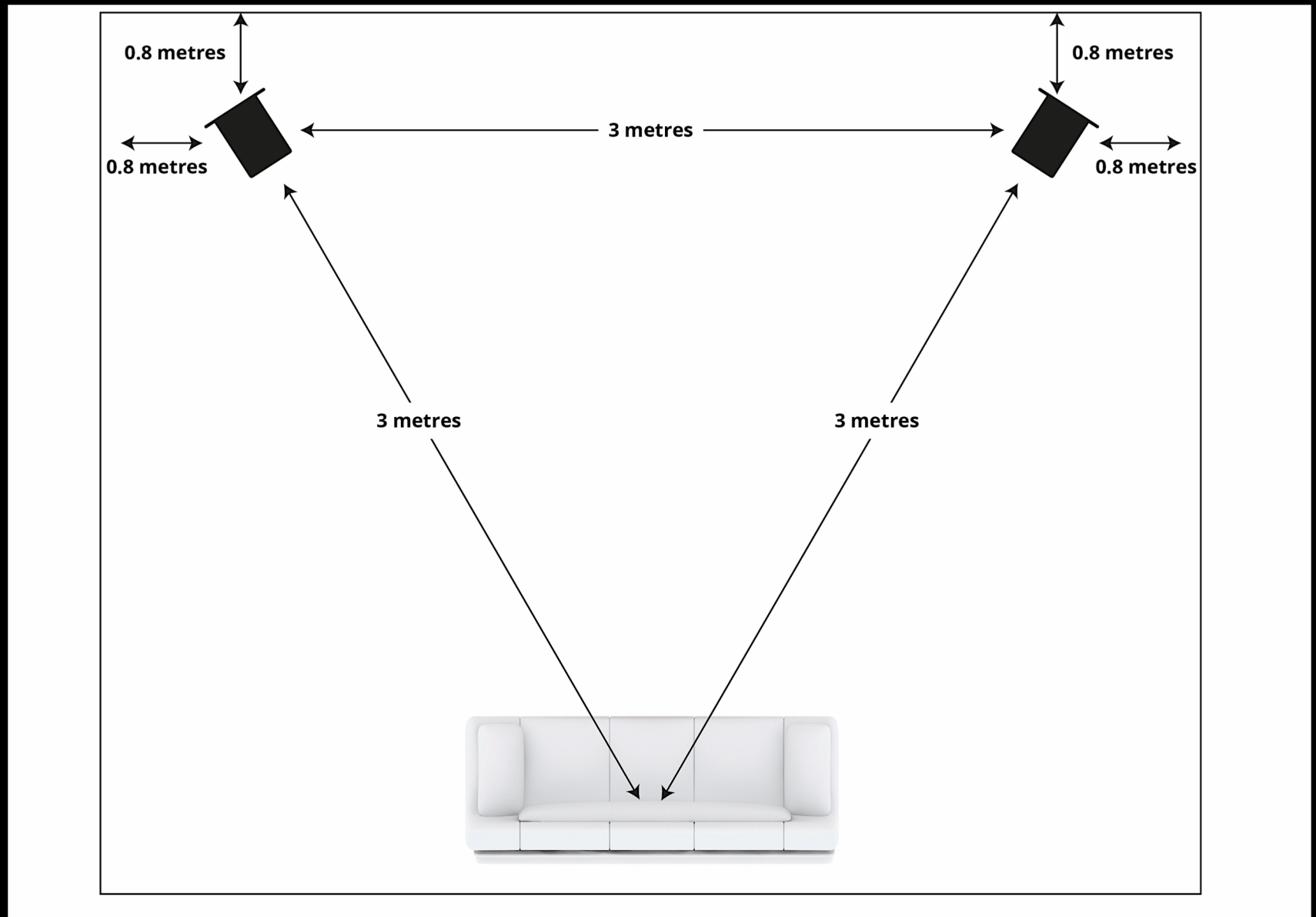
The implications of these results are twofold:

- Firstly, it should be possible to provide directional information purely through either only delay cues or only intensity cues. This is true within the limits previous explained.

- Secondly, once a sound is delayed by greater than about 700µs the ear attends to the sound that arrives first almost irrespective of their relative levels, although clearly if the earlier arriving sound is significantly lower in amplitude, compared with the delayed sound, then the effect will disappear.

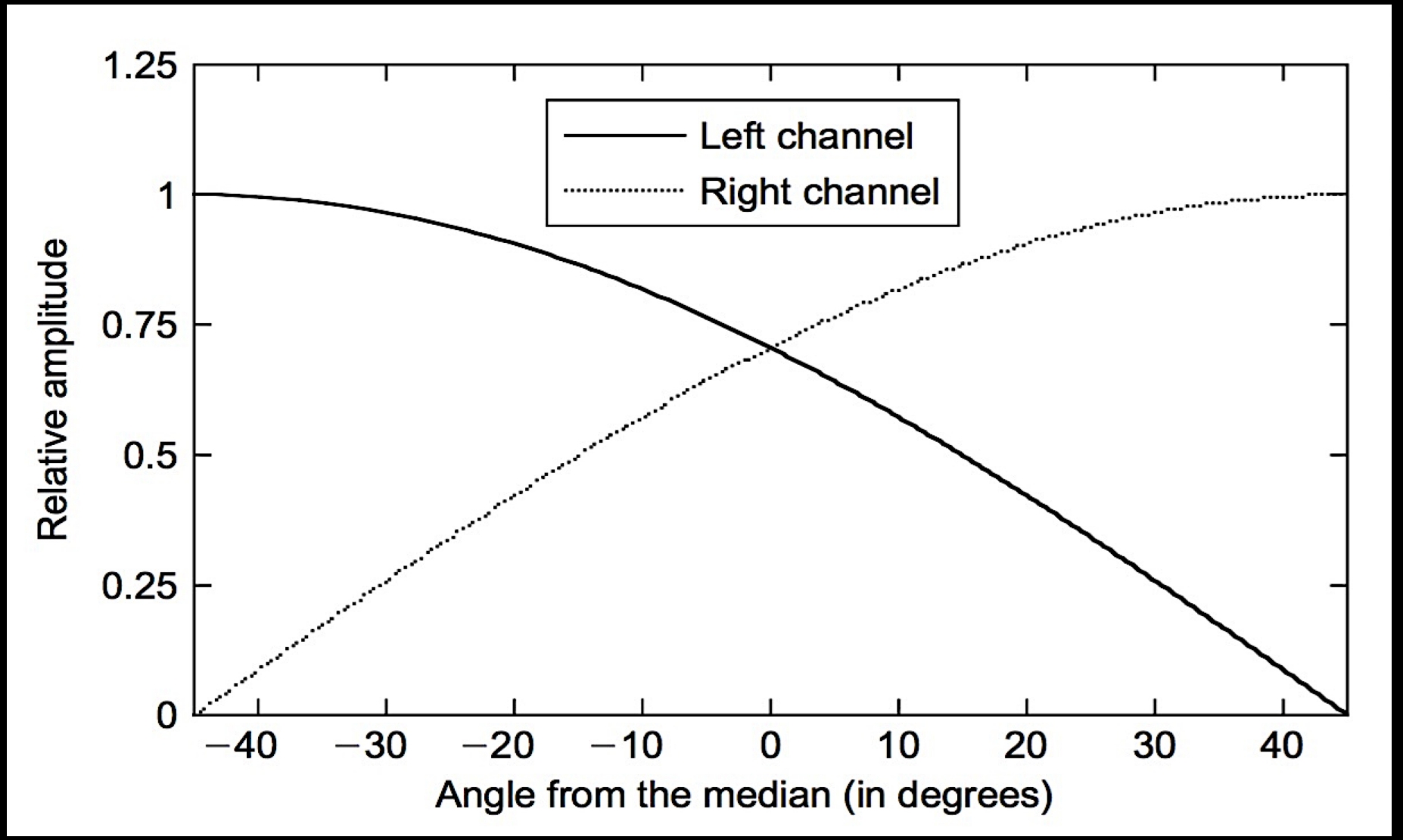
Audio perception - Stereo:

- IIDs are used to position a sound between the two speakers:



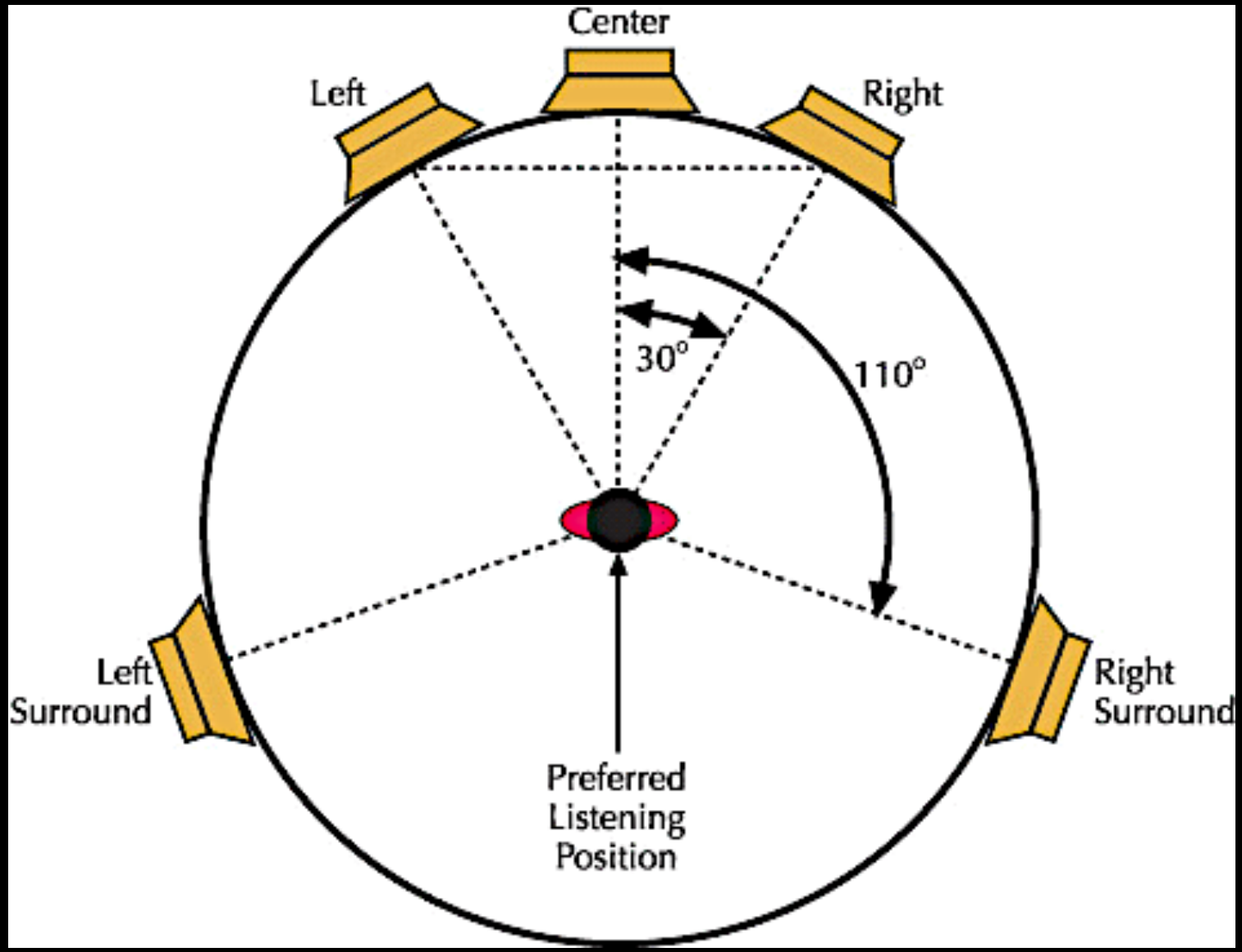
Audio perception - Stereo:

- IIDs are used to position a sound between the two speakers. Here are the panning volume control laws



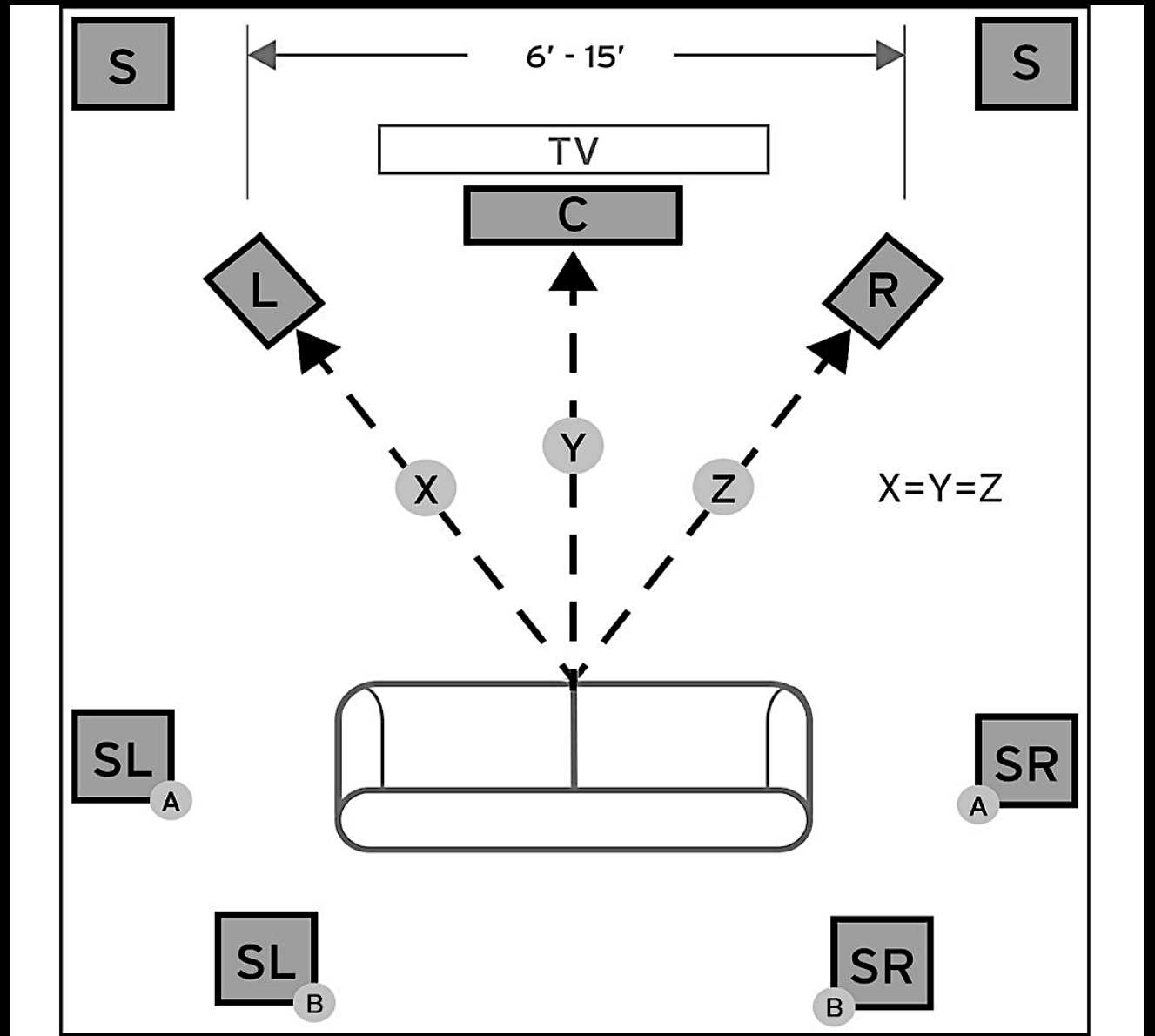
Audio perception – & cinema surround:

- IIDs are used to position a sound. 5.1 cinema setup.



Audio perception – & cinema surround:

- IIDs are used to position a sound. 7.2 cinema setup.



Commercial spatialisation systems:

Stereo, 5.1, 7.1, 10.2, 22.2, n.m… & Atmos, Auro etc.

- Commercial sound spatialisation systems are based on a “sweet spot”, i.e. a small area in the middle of the speakers where the sound localises properly.
- However, the world does not work like that. In the real world we can move anywhere and the sound field is coherent.
- Commercial sound spatialisation systems are convenient because there are tools and workflows that are well documented and understood for the production and playback of such material.
- In production, each “channel” corresponds to a speaker and location.
- The problem is that these systems, while they work “ok” for their intended application (movie and TV viewing), are quite inadequate for high-level VR.

Sound spatialisation systems for immersive VR:

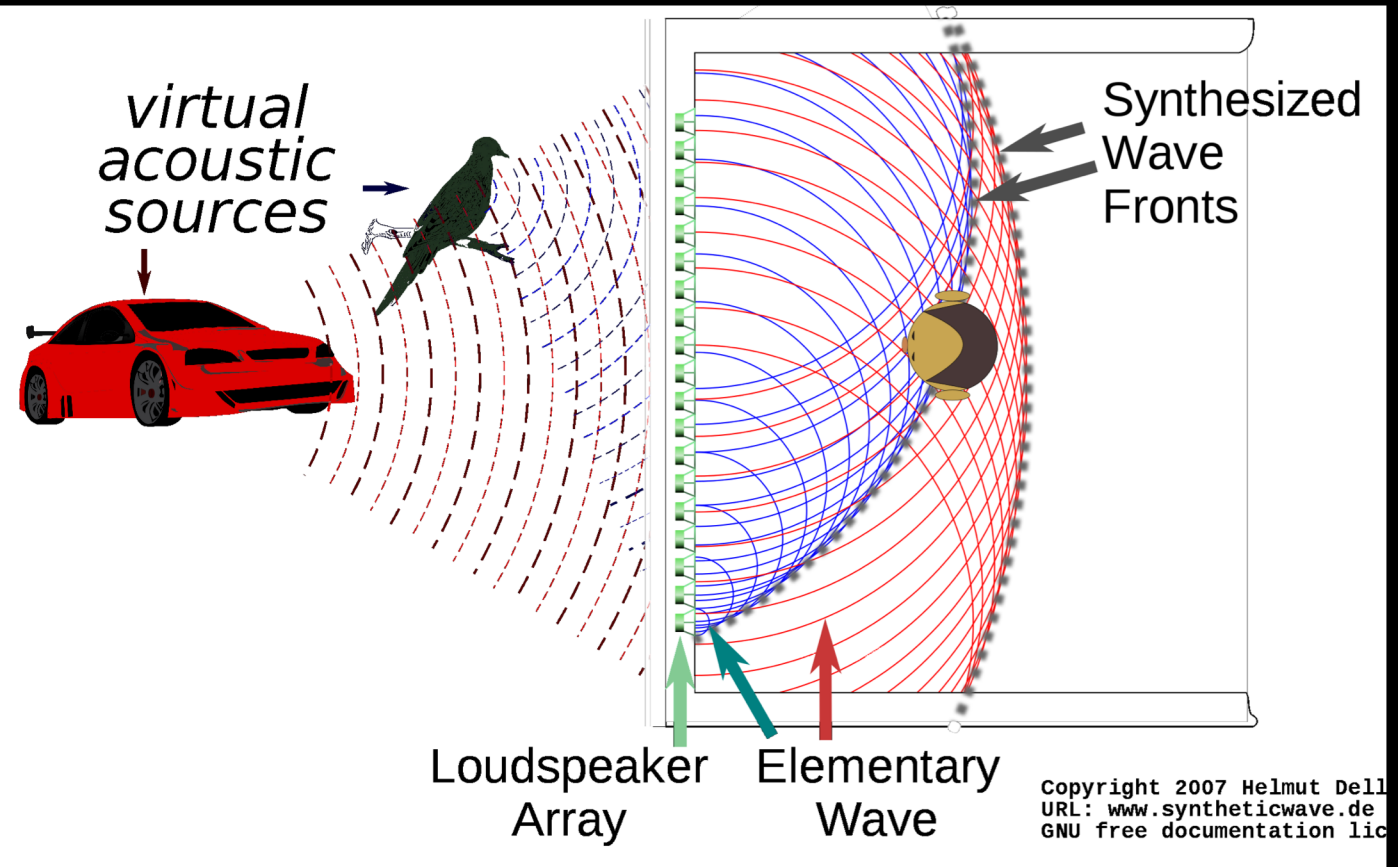
What do we want & what can sound provide?

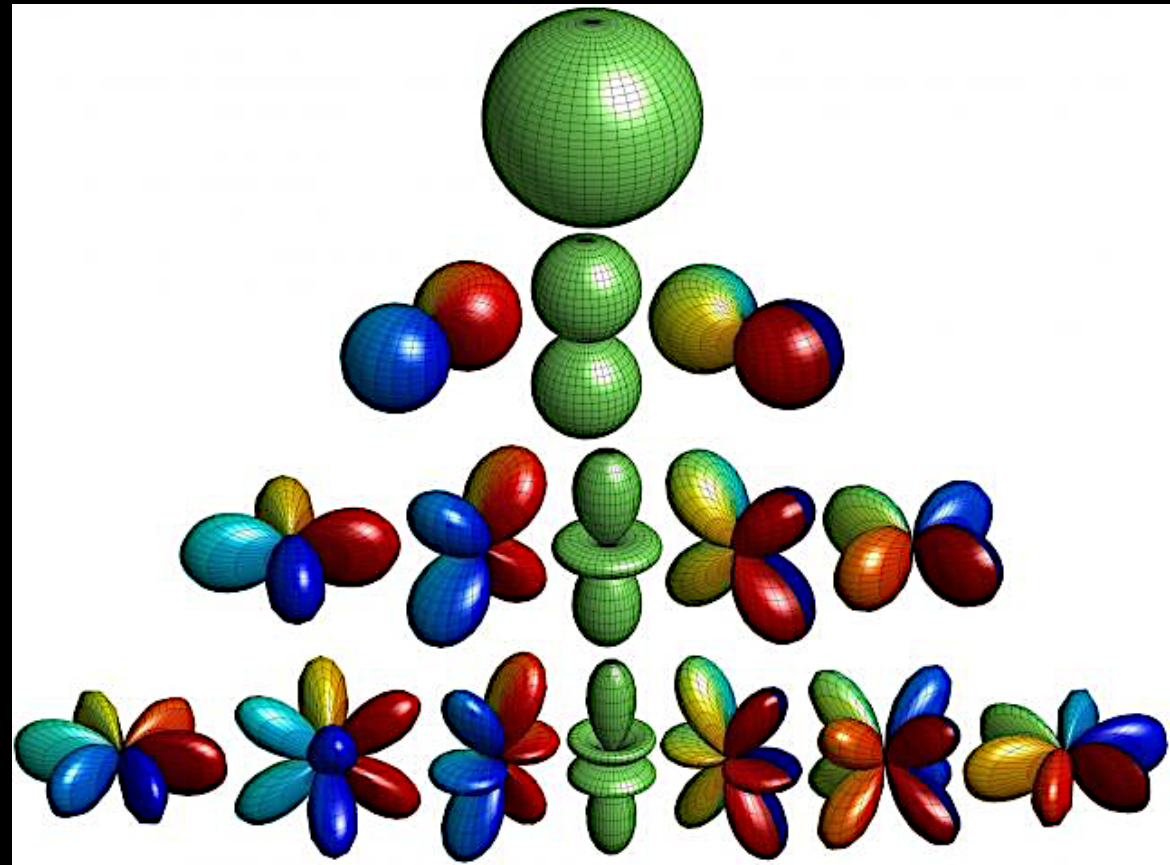
- Sound that is accurate spatially in location and setting
  - This is mostly a technical problem
- Sound without a “sweet spot”
  - Also a technical problem, but potentially bigger
- Sound that provides verisimilitude
  - This is both a technical problem & a production problem
- Sound that provides ***time & spatial continuity***
  - Almost purely a production issue

Alternatives to “sweet spot” sound spatialisation:

There are 2 options-

- Wavefield synthesis & Ambisonics





Wavefield synthesis:

- Wavefield synthesis (WFS, IOSONO & Barco) is a “brute force” approach of recreating the soundfield by solving the Kirchhoff-Helmholtz integral to calculate and render the wave front in real time.
- It is the acoustic equivalent of holography.
- 2D slice – 230 loudspeakers, 230 amplifiers, 230 channel render engine (44Mb/sec at 192kb)

Wavefield synthesis:

- Wavefield synthesis (WFS, IOSONO & Barco) is a “brute force” approach of recreating the soundfield by solving the Kirchhoff-Helmholtz integral to calculate and render the wave front in real time.
- It is the acoustic equivalent of holography.

- 3D version - 2700 loudspeakers, (2700 amplifiers), 832 render channels (160Mb/sec, total throughput 518Mb/sec)

Wavefield synthesis:

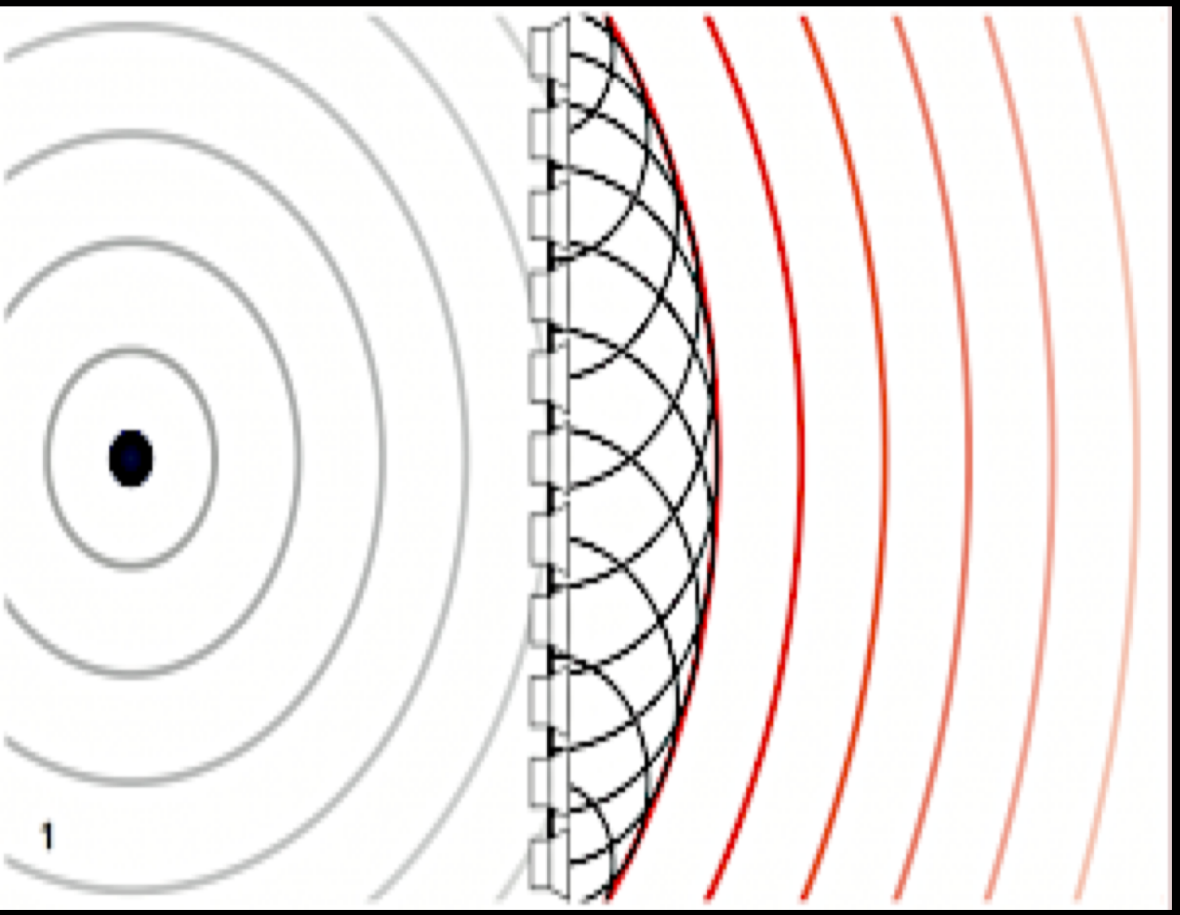
Advantages:

- Excellent localisation
- No “sweet spot”, excellent rendering of space

Disadvantages:

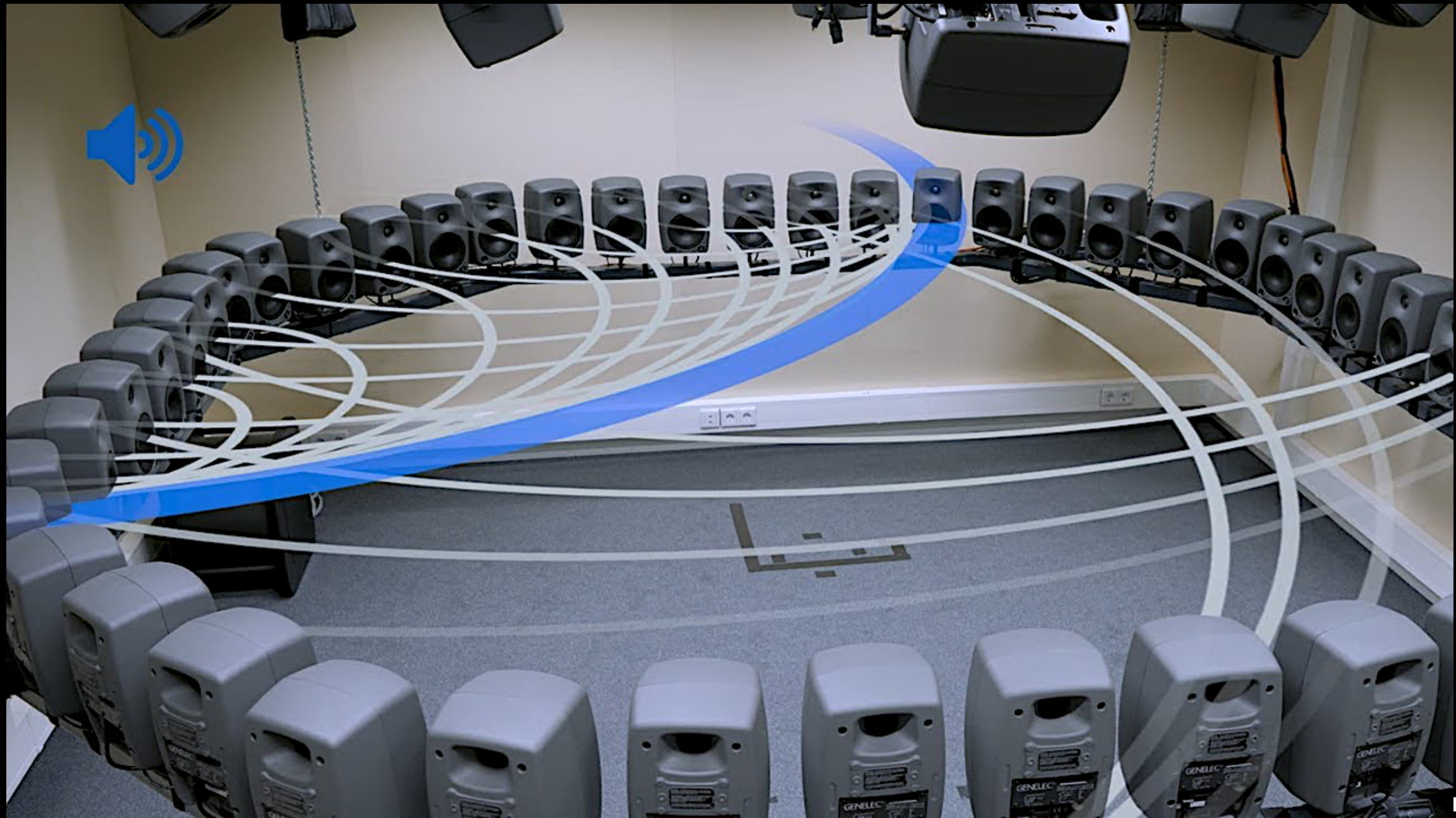
- Resource intensive, 256+ speakers/channels/computers
- Horizontal only (unless we square the number of speakers)
- Does not scale elegantly, or at all (spatial aliases badly)
- Practically no room for a screen
- No recording or capture system, synthesis only
- Needs real time rendering on playback

Wavefield synthesis:



Wavefield synthesis:

Wavefield synthesis:

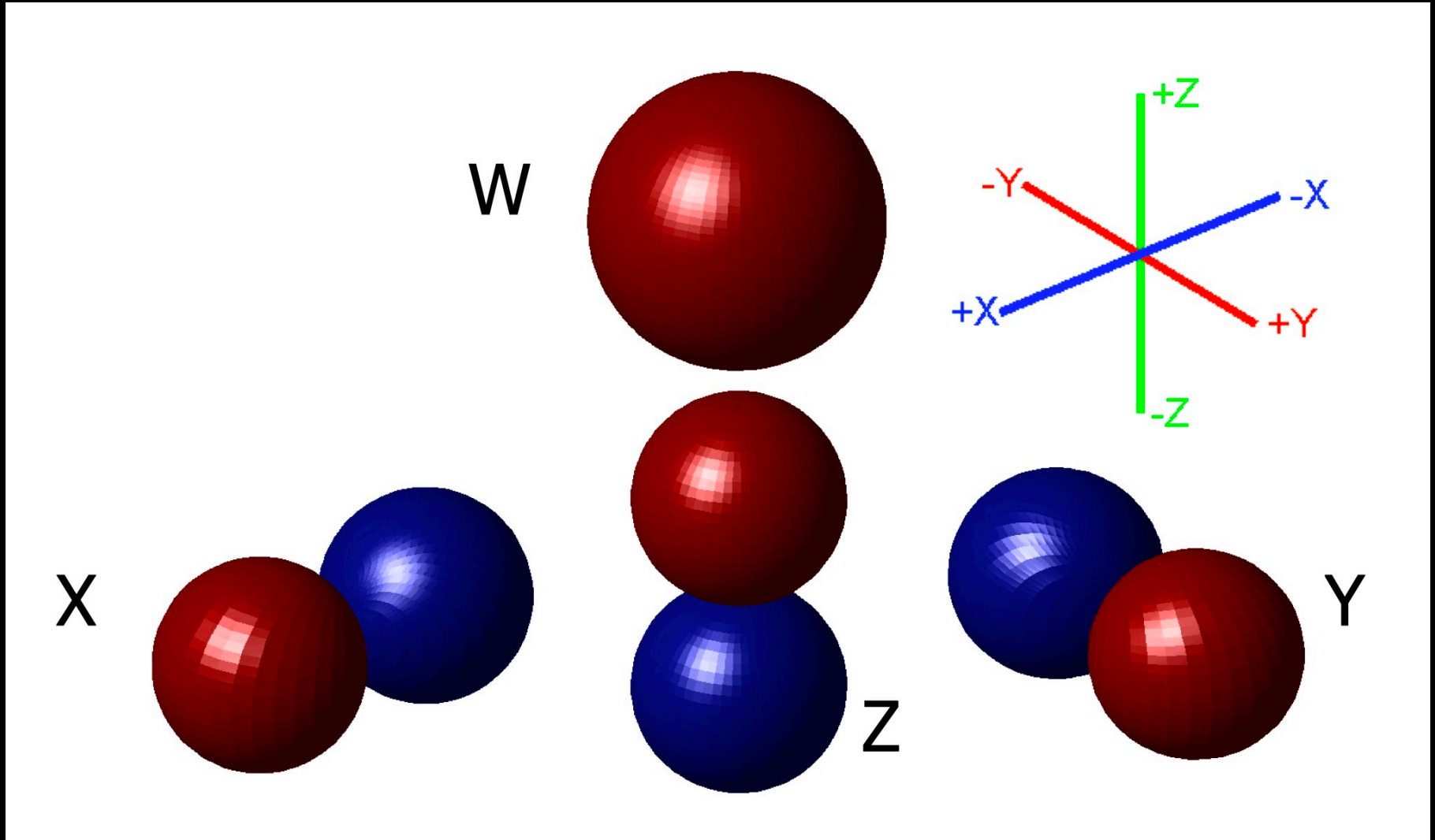


LLARTS

Wavefield synthesis:

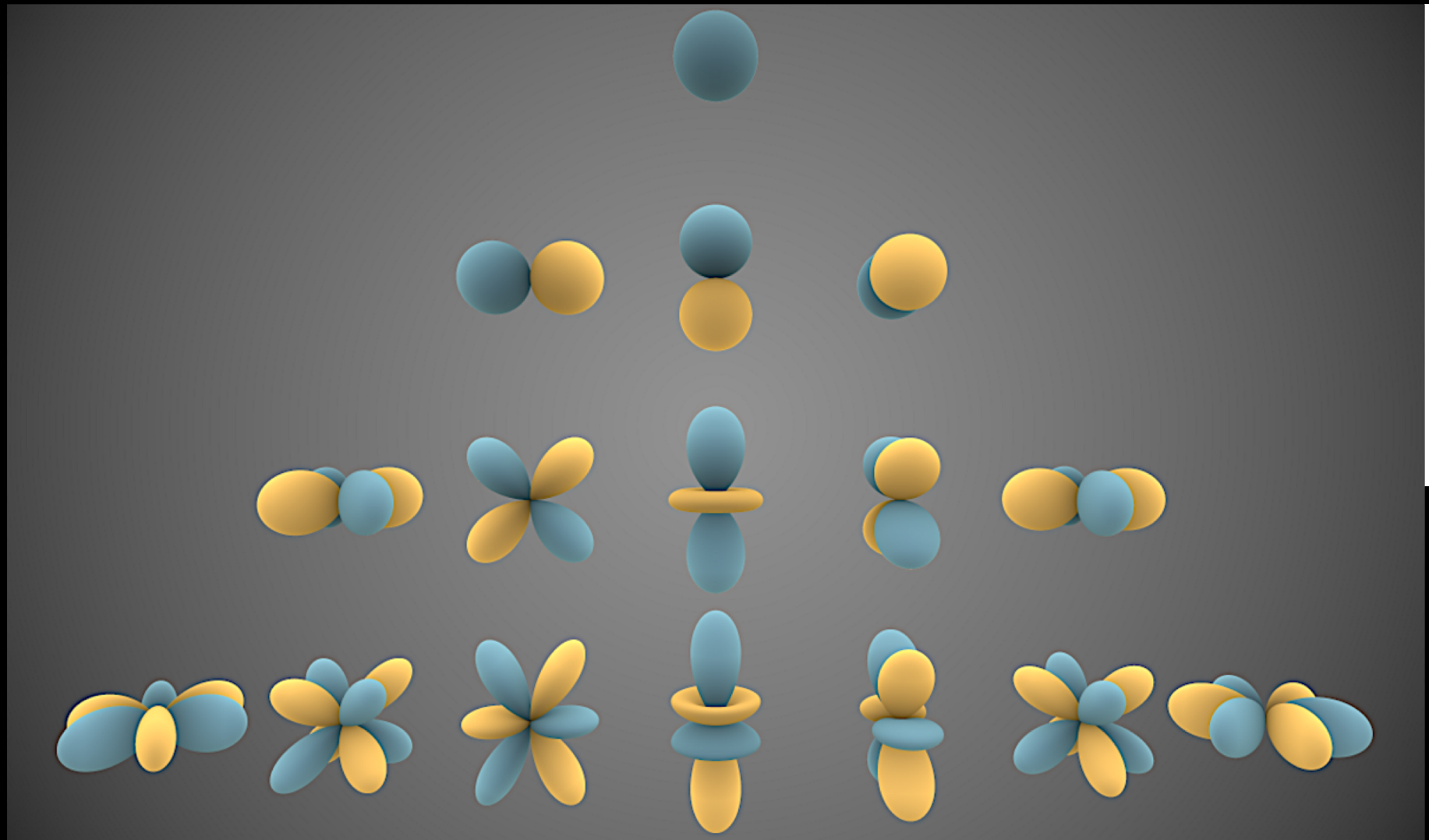
Ambisonics:

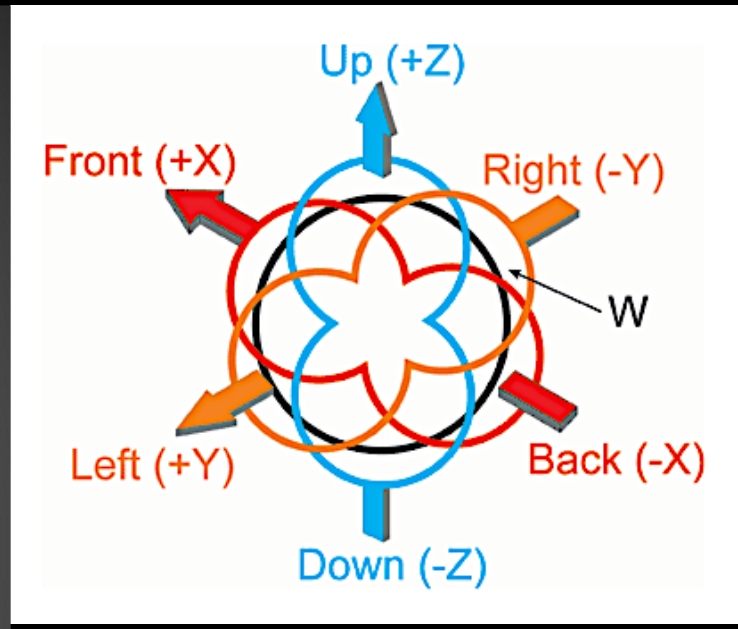
- Ambisonics solves the spatial problem differently, by recreating the spherical harmonics of the original soundfield.



Ambisonics:

- Ambisonics solves the spatial problem differently, by recreating the spherical harmonics of the original soundfield.





Ambisonics:

Advantages:

- Full 3D surround, including height
- Large “sweet spot”, larger with higher orders
- Light on resources (speakers & processing)
- Scalable, 1st order = 4-8 channels, 3rd order =7-17 channels
- Recording & capture system exists (up to 3rd order)

Disadvantages

- Need to go HOA for larger sweet spot, but 3rd order ≃ WFS
- No tools or production (improving)
- Needs decoding on playback

Ambisonics:

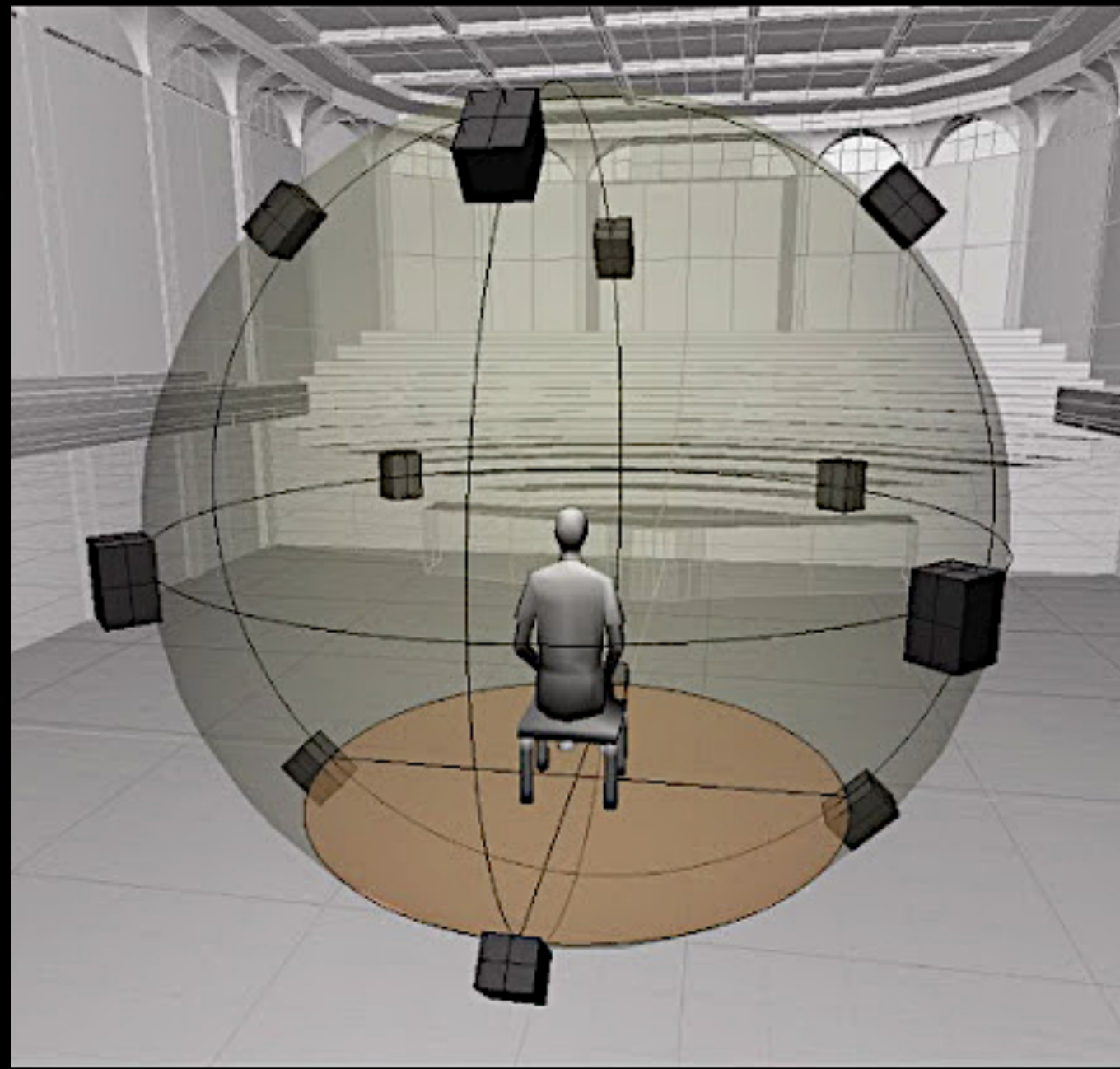
Capture:

Ambisonics:

Playback / production

Ambisonics:

- Files are kept in an intermediate format called “B-Format”.
- This format is hierarchical and extendable, and so could be 4 – 16 or more channels depending on desired spatial acuity and accuracy.
- B-Format files are decoded for the playback on any speaker array.



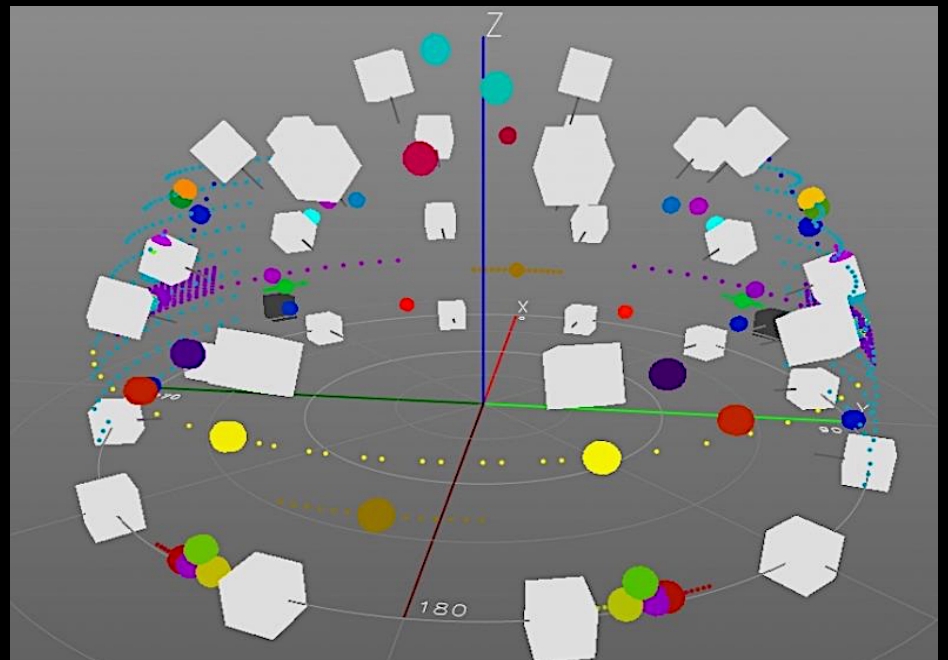
Ambisonics:

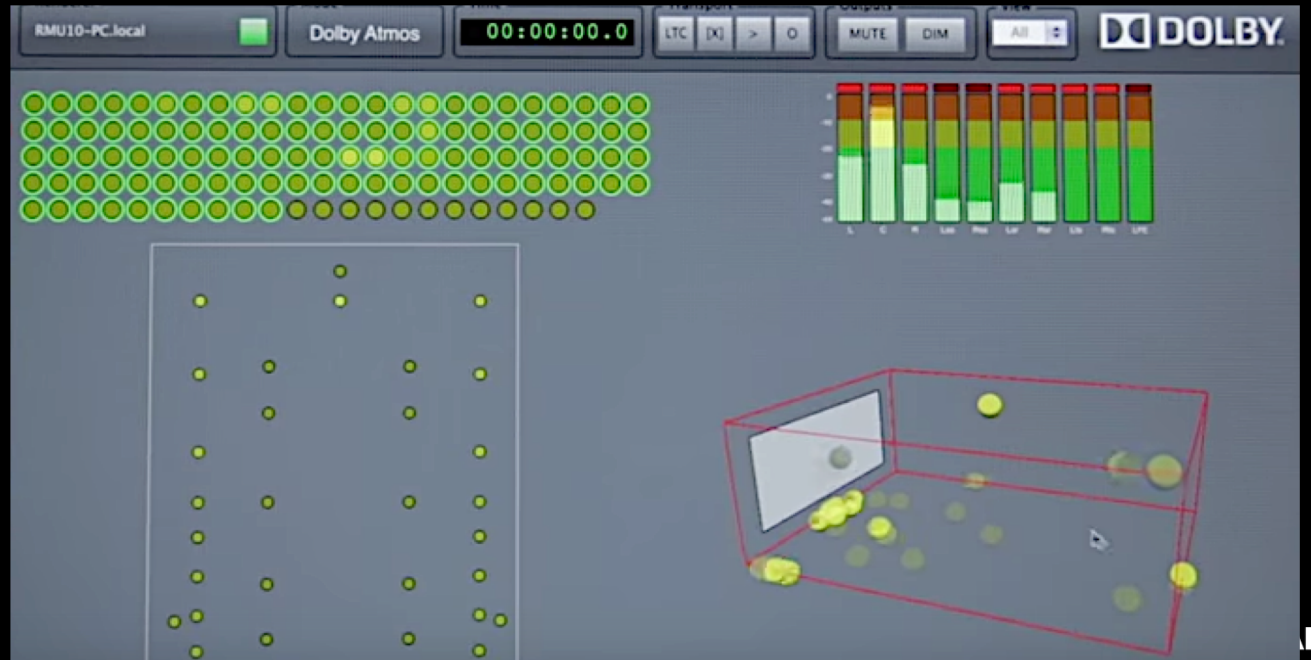
The sound can be rendered over ***any*** practical arrangement of loudspeakers or headphones exhibiting (with some caveats):

- Same relative volume per source
- Same source positions
- Same spatial impression

Ambisonics & Wavefield synthesis:

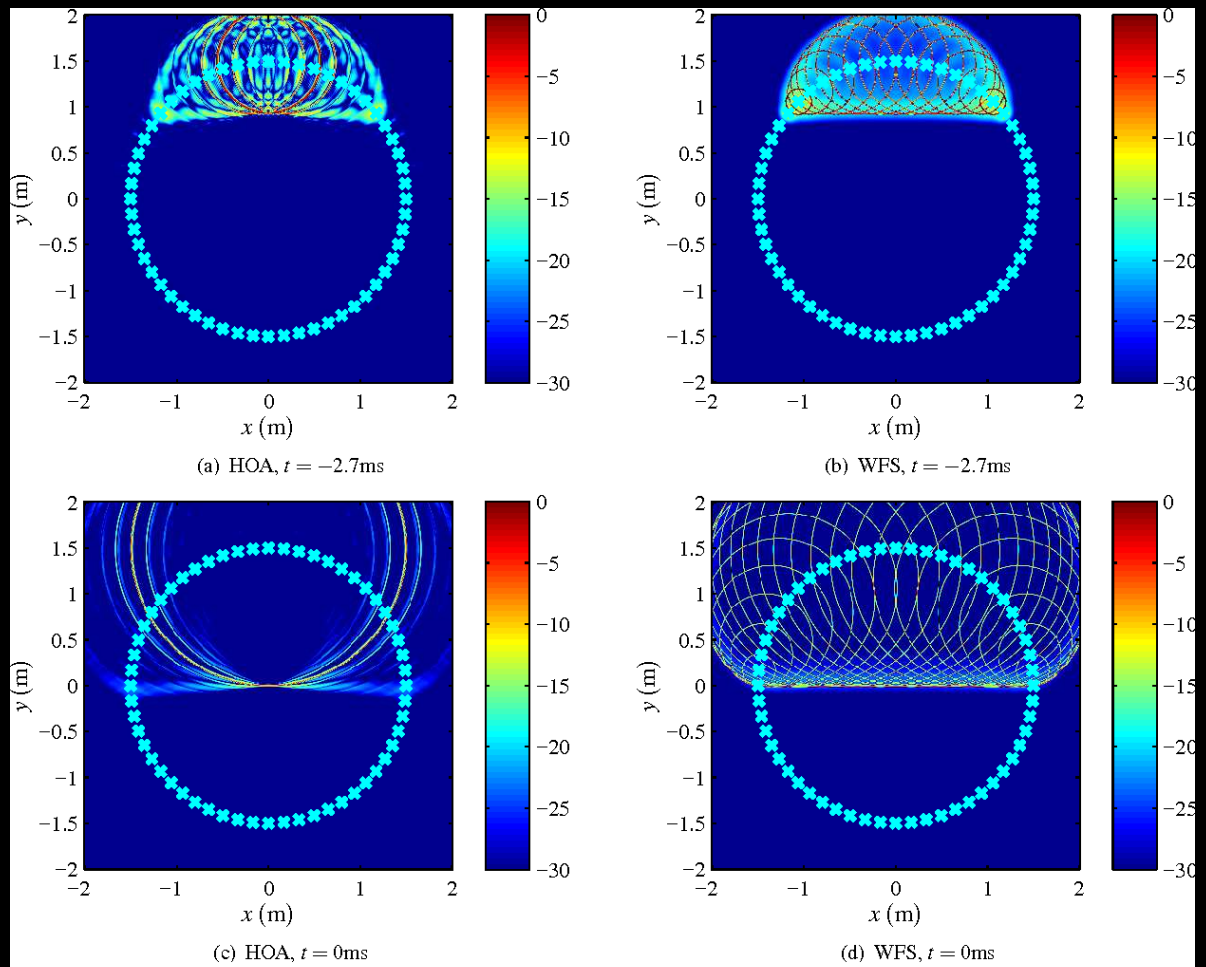
- Both are “object based” audio systems instead of channel based.
- In production, sound objects are placed in space and the system renders the object in that space. Ambisonics adjusts to the playback system.
- Commercial systems are channel based so the playback system is fixed and production makes the speaker feeds. Atmos now has an object renderer, but it is not as sophisticated as WFS or Ambisonics





Ambisonics & Wavefield synthesis:

- Both can render similarly sized listening areas with similar spatial accuracy.



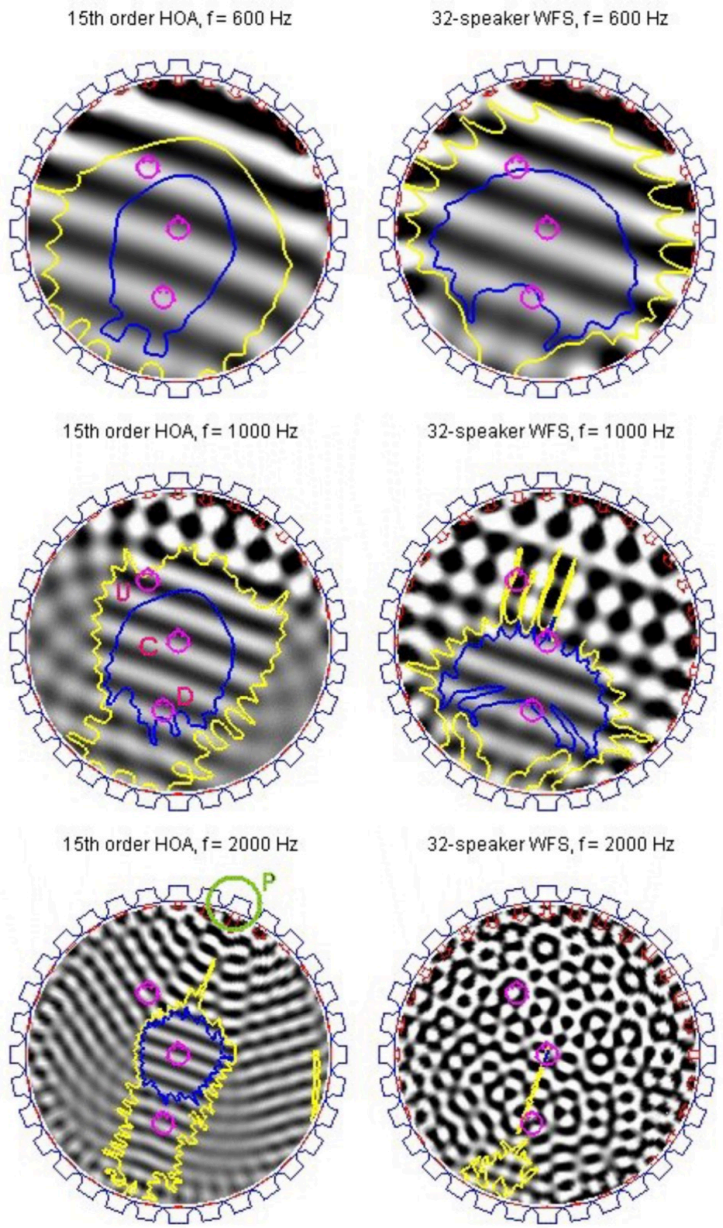


Figure 17 Reconstruction of monochromatic plane waves with HOA and WFS. Blue/dark and yellow/bright contours enclose well-reconstructed areas with error tolerance of resp. 20% and 50%.

What about headphones?

- Headphone based spatialisation systems use HRTF processing
- Head Related Transfer Functions allow full 3D sound over headphones, by simulating the ITD, IID & pinna effect cues for sounds from any direction
- Needs head-tracking for reality
- Has a capture system (dummy head)
- Most spatialisation is synthesised, either static speaker feeds, or sound objects
- HRTF rendering is well understood and implemented now, head-tracking is getting cheaper

What about headphones?

- Headphones may be the best option for some systems

Comparison & implications for listeners

Listeners:

- 5.1 – 10.2 systems limited with a “front”, limited immersion, limited spatial precision, but very common.
- WFS has most accurate sound localization & largest listening area, but horizontal only, no screen space, & resource intensive (crazy).
- Ambisonics can approach WFS for localisation accuracy & listening area, is full 3D, can accommodate a screen.
- HRTF headphone systems can be very good for personal listening & games, may not work in VR but could be considered.

Comparison & implications for production

Production:

- 5.1 – 10.2 systems have numerous production tools & resources.
- WFS is resource intensive (crazy) & practically no tools (research only). Programmable through MaxMSP, PureData, Supercollider, Python etc.
- Ambisonics has some tools (large increase over last 10 years). Programmable through MaxMSP, PureData, Supercollider, Python etc .
- HRTF headphone systems are well supported now through software libraries and tools. See Simon Goodwin's game audio work for an excellent example.

Real world systems development, tools & specifications

- Split development of playback system from production of the audio material, they are separate tasks requiring unique skills
- An interactive playback system needs an interactive audio software system such as MaxMSP, PureData, Supercollider, Python or C(++)
- ***The playback system determines the production requirements***
- Production typically uses 24-bit, 96KHz resolution for production, but ***delivery*** need only be 16-bit, 48KHz, halving bandwidth requirements.

Uses of immersive audio

- Verisimilitude for cultural heritage works, allowing greater suspension of disbelief
- Temporal continuity & cues for participants, creates a narrative because it is time based
- Data exploration through sonification (hearing features of data spatially), if the sound system is good enough

Latest research

- Dolby Atmos uses sonic objects (so that's becoming more common), will allow Ambisonics & WFS to replace it
- Barco Auro uses some elements of WFS, Atmos competitor
- Personalised HRTFs from your phone
- Mixed order Ambisonics (3rd order horizontal, 1st order vertical)
- HOA becoming cheaper & more common, tools also
- AI & ML tools to assist production, e.g. Izotope
- Wavefront curvature perception, capture & rendering (closer to reality, no sweet spot) tighter curvature = closer sound

Latest research

Move away from channel-derived eqi-distant (surround):

- Speaker (j) =Σ(Stream n, {x,y,z}), then scalar spherical to Vector format and point of use render
- Wave field everywhere defined: Reconstructed from n-sources (capture points) Σ{SPL, v}x.y,z continuous space
- Solutions to Helmholtz wave equation, and Huygens Fresnel superposition, will become possible with wavefront rendering
- Vector voltage distribution (E.g. B-Format)

Thank you!

Questions / discussion…