Spatial Audio for VR: An Overview

What is spatial audio?

- also referred to as 3D audio or 360 audio
- sonic experience where
  - **the audio changes with the movement of the viewer's head**
- produced by stereo speakers, surround-sound speakers, speaker-arrays, or headphones.
- **Spatial music** is music composed to intentionally exploit sound localization
  - in use since prehistoric times in the form of antiphon
  - in use since around 1928 as 'Raumusik' or "space music" from Germany

Spatialization:

- the projection and localization of sound source in a space,
  - Physical
  - simulated
- and its spatial movement in space.

- technically known as spatial domain convolution of sound waves using **head-related transfer functions (HRTF)**.

Creating positional sound

- Amplitude
  - (or more)
- Synchronisation
  - Audio delays
- Frequency
  - Head-Related Transfer Function (HRTF)

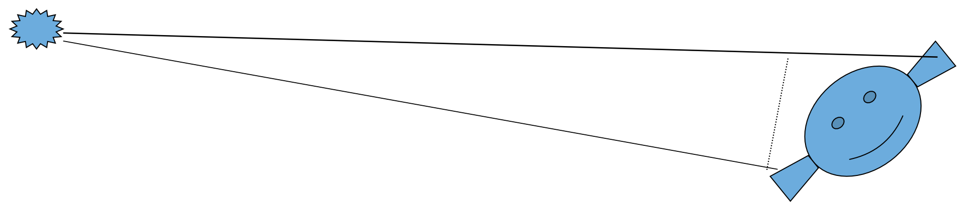
Amplitude

- Generate audio from position sources
- Calculate amplitude from distance
- Include damping factors
  - Air conditions
  - Snow
  - Directional effect of the ears

Synchronisation

- Ears are very precise instruments
- Very good at hearing when something happens after something else
  - Sound travels slowly (c 340 m/sec in air): different distance to each ear
- Use this to help define direction
  - Difference in amplitude gives only very approximate direction information

Speed effect



- 30 centimetres =0.0008 seconds
- Human can hear ≤ 700µS

What is 3D sound?

- Able to position sounds all around a listener.
- Sounds created by loudspeakers/headphones: perceived as coming from arbitrary points in space.
- Conventional stereo systems generally cannot position sounds to side, rear, above, below
- Some commercial products claim 3D capability - e.g stereo multimedia systems marketed as having "3D technology". But usually untrue.

3D positional sound

- Humans have stereo ears
- Two sound pulse impacts
  - One difference in amplitude
  - One difference in time of arrival
- How is it that a human can resolve sound in 3D?
  - Should only be possible in 2D?

Frequency

- Frequency responses of the ears change in different directions
  - Role of pinnae
  - You hear a different frequency filtering in each ear
  - Use that data to work out 3D position information

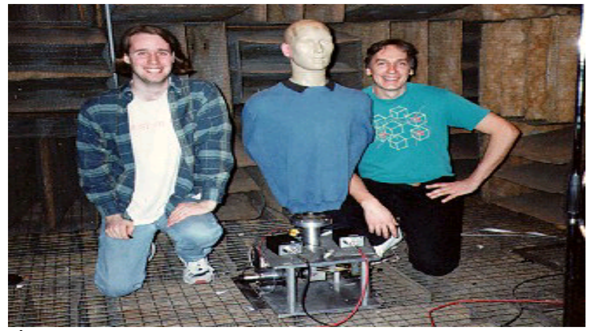
Head-Related Transfer Function

- Unconscious use of time delay, amplitude difference, and tonal information at each ear to determine the location of the sound.
  - Known as *sound localisation cues*.
  - Sound localisation by human listeners has been studied extensively.
- Transformation of sound from a point in space to the ear canal can be measured accurately
  - Head-Related Transfer Functions (HRTFs).
- Measurements are usually made by inserting miniature microphones into ear canals of a human subject or a manikin.

HRTFs

- HRTFs are 3D
  - Depend on ear shape (Pinnae) and resonant qualities of the head!
  - Allows positional sound to be 3D
- Computationally difficult
  - Originally done in special hardware (Convolvotron)
  - Can now be done in real-time using DSP

HRTFs



- First series of HRTF measurement experiments in 1994 by Bill Gardner and Listening Group at **MIT Media Lab**.
- Data from these experiments made available for free on the web.
- Picture shows Gardner and Martin with dummy used for experiment - called a **KEMAR dummy.**
- A measurement signal is played by a loudspeaker and recorded by the microphones in the dummy head.

HRTFs

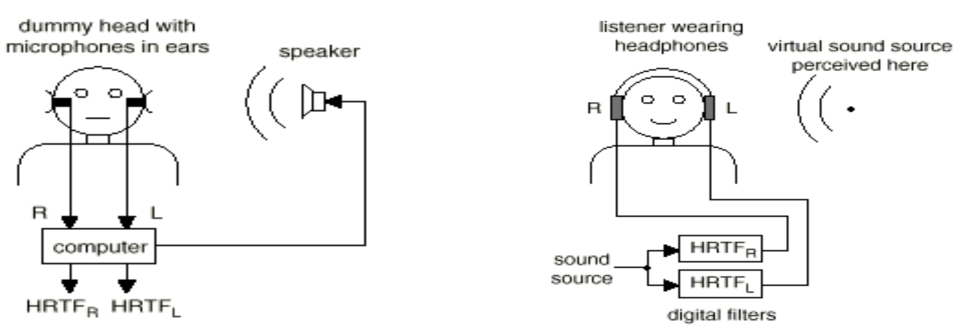
- Recorded signals processed by computer, derives two HRTFs (left and right ears) corresponding to sound source location.
  - HRTF typically consists of several hundred numbers
  - describes time delay, amplitude, and tonal transformation for particular sound source location to left and right ears of the subject.
- Measurement procedure repeated for many locations of sound source relative to head
  - database of hundreds of HRTFs describing sound transformation characteristics of a particular head.

HRTFs

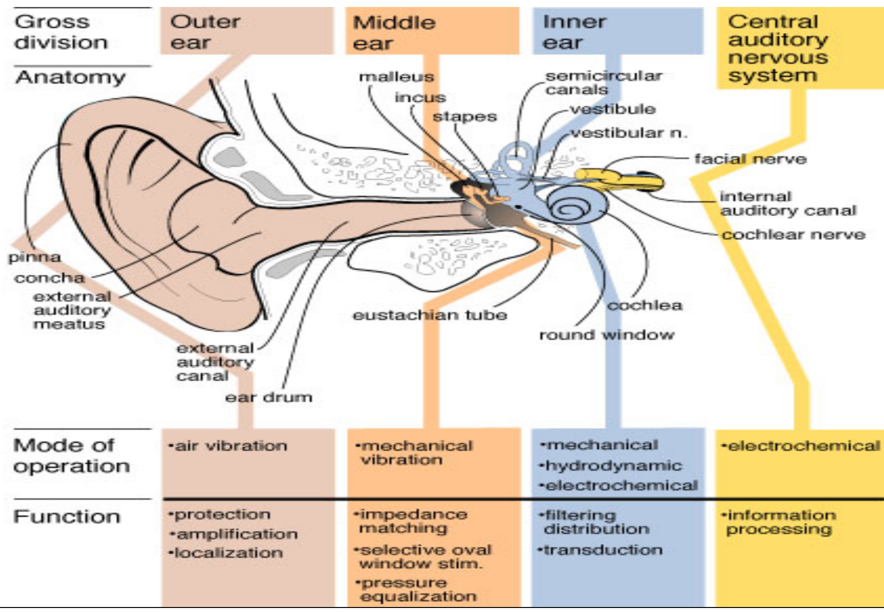
- Mimic process of natural hearing
    - reproducing sound localisation cues at the ears of listener.
- Use pair of measured HRTFs as specification for a pair of digital audio filters.
- Sound signal processed by digital filters and listened to over headphones
    - Reproduces sound localisation cues for each ear
    - listener should perceive sound at the location specified by the HRTFs.
- This process is called *binaural synthesis* (binaural signals are defined as the signals at the ears of a listener).

HRTFs

- The process involved in generating true 3D audio using HRTFs:

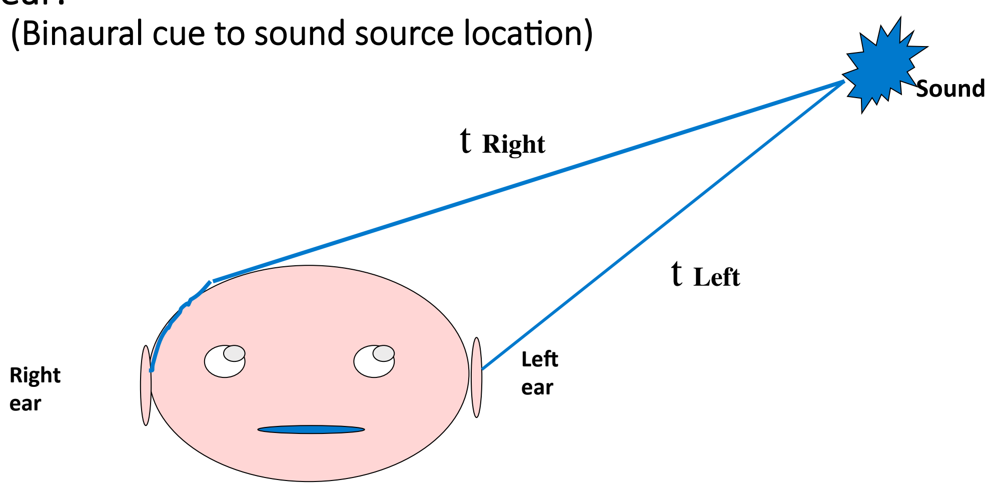


Four major divisions of auditory system – Function



I. Outer ear:

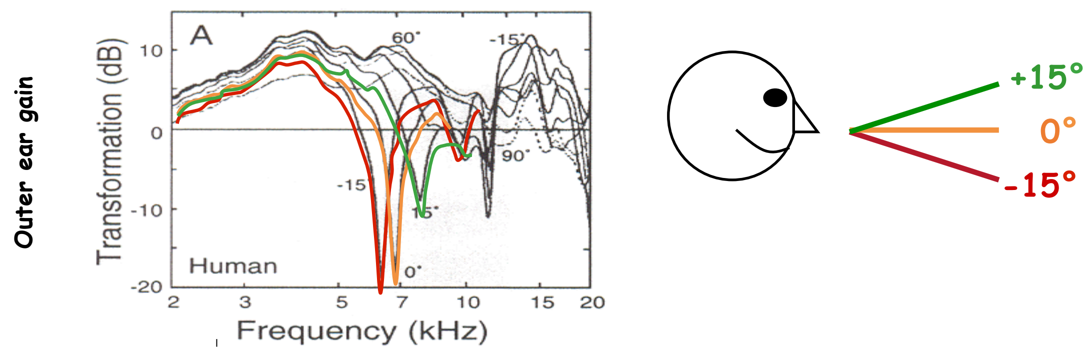
(1) Pinna (Binaural cue to sound source location)



*** Different distances from source to each ear**
=> different arrival times (Interaural time-difference)
and different sound level (interaural level-difference)

I. Outer ear:

(1) Pinna (Spectral cue to sound source location)



The spectral feature of sound is changed depending on the sound elevation
=> **Head Related Transfer Function (HRTF)**

Basics of Object Based Audio

What is an Object Based Audio?

Audio which is generated by

- a Stationary or moving object

OR

- a class of objects that are clubbed together as a collective source of sound

Some Examples of Audio Objects are:

- A Stage artiste
- Chorus
- Cheering Crowd at Cricket ground
- Ocean waves, wind blowing
- Birds chirping
- An Airplane or A Helicopter
- A moving Train
- A bullet fired
- A monologue or a dialogue

Examples of Object-Based Audio

Audio Scene or Sound Field generated by *mixing* (not just adding) audio signals from multiple Objects. Following are a few examples.

- Watching a football match in a stadium with home crowd. (Sports TV Channel)
  - 3 objects : Home Crowd as a ring object, Commentator as a point object, Players-Umpire conversations as another object.
- Participating as a player in a field game. (Computer Games)
  - 4+ objects : Home Crowd as a ring object around you, Commentator as a point object, Player's own voice responses as a point object, Other Players-Umpire as multiple moving objects.
- Being a part of Scuba-diver team searching an underwater treasure. (VR)
- Listening to a conversation between different actors & backgrounds in a movie scene. (cinema)
- Attending a music concert or a simple Jazz performance (concert)

Channel Based Audio vs Object Based Audio

Channel Based Immersive Audio	Object Based Audio
Content Creation	
• Each signal track is associated with a specific speaker feed & setup at listener end. • Content is created for a specific Listener Environment or setup. (mobile, home, or theater)	• Audio Object based signal tracks are independent of speaker-setup. • => Content created is independent Listener Environment or setup. (mobile, home, or theater)
Playback at Listener End	
• At Listener end, the contents (channels) are mapped onto user speaker setup • Need to use Predefined channel-mapping to headphones, stereo speaker, 2.1, 5.1, 11.1 etc.	• At Listener end, the objects are mapped onto user speaker setup • Objects based on positions and movements are mapped on the fly to the speaker-setup.

Channel Based Audio vs Object Based Audio

Channel Based Immersive Audio	Object Based Audio
Content Creation	
• With inputs as the recorded contents or tracks, each Channel track is carefully designed and created at the recording studios. OR at the gaming developer studios for creating good immersive effects.	• Audio Objects can be simply identified encoded as separate tracks. • Associated meta-data should be carefully designed to capture shape, movement, appearance/disappearance of the objects assuming the listener at the center.
Playback at Listener End	
• If the content-target speaker == user speaker setup, then simple-mapping and playback. • Else use some good pre-defined maps and delays for rear speakers to create the content.	• Objects are decoded to create audio signals. • Frame-by-Frame, positions of "active objects" are mapped on to user speakers in form of gains and delays for these objects. Mix and playback.

Channel Based Audio vs Object Based Audio

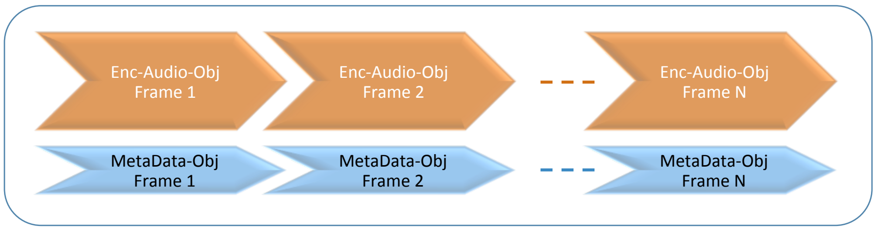
Channel Based Immersive Audio	Object Based Audio
Content Creation	
• Creation is a complex careful process. • Encoding steps and procedure is complex and hence is done by skilled well trained sound designers.	• Creation and encoding object-audio is a relatively simpler process and can be done without much pre-thinking of user-setups & environment. • Audio object meta-data needs to be carefully associated with it.
Playback at Listener End	
• Decoders are Renderers are fairly simple.	• Decoders are simple (as simple as channel based Audio). • However the Renderers are much more complex. • Renderers need to map these objects with its positions to speakers on a frame-by-frame basis.

A summary – basic of Object-Based Audio

Typical Object-Audio Encoded Stream contains 8 to 16 encoded audio object-tracks.

And each audio-object track has two parts

Object Audio Track



Meta-Data:

Stream-level:
- Max Number of Audio Objects present in the scene.

Frame-level:
- Shape of the object, Data related to position, speed of the object, Appearance / Disappearance of object

Compressed PCM Audio-Data:

Standard DD or AAC encoded audio signals associated with a specific object.

Object-Based Audio Stream Decoding & Rendering

- Decoding: The basic audio from object is encoded using standard legacy encoders. Therefore, decoding uses standard mp3, aac, dolby-digital decoding to provide basic audio PCM for the object.
- Renderers: Challenges are in Rendering the decoded object-based audio PCM contents & use object's shape/motion meta-data to create –
  - An immersive audio experience on **headphones**. (VR, Gaming, smartphones, and tablets)
  - An immersive audio experience on our **multi-speaker layouts** at homes or theaters

Object-Based Audio Renderer on Headphones

HRTF Model
(Head Related Transfer Function)

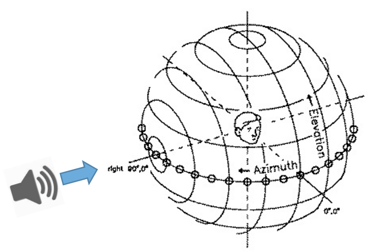
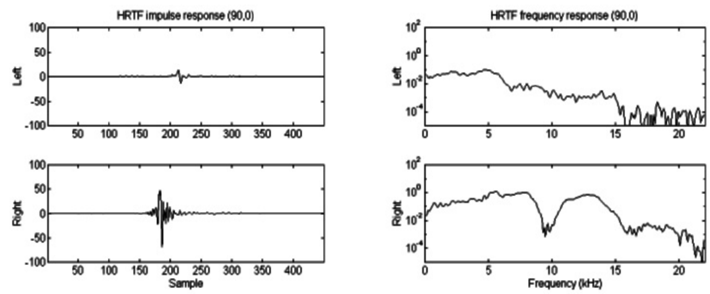


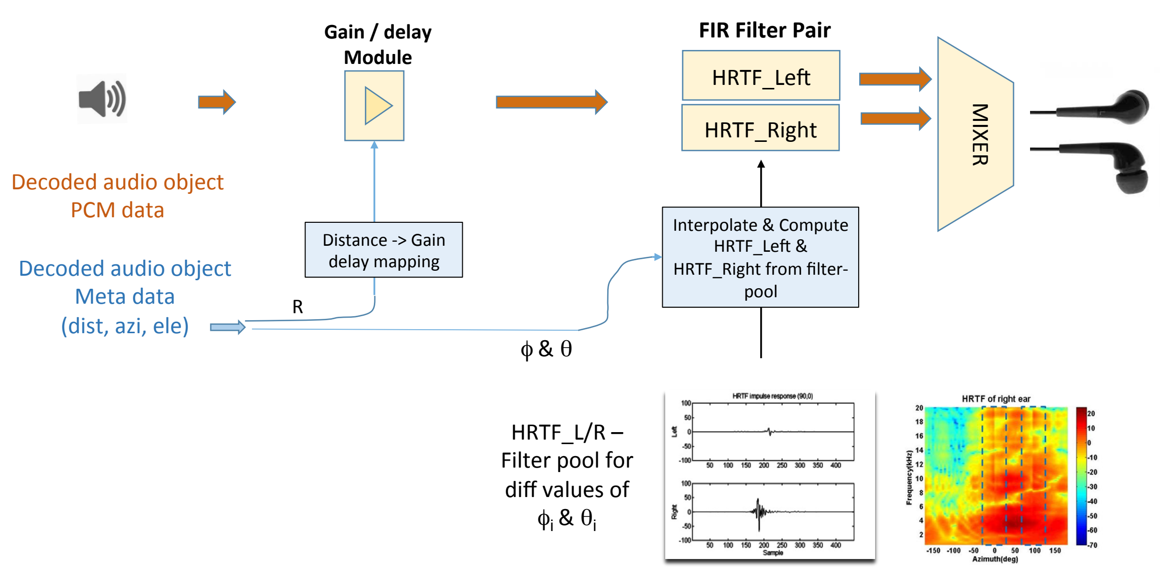
Diagram of spherical coordinate system (Wightman and Kistler, Univ Winsconsin) – 1989 [1]



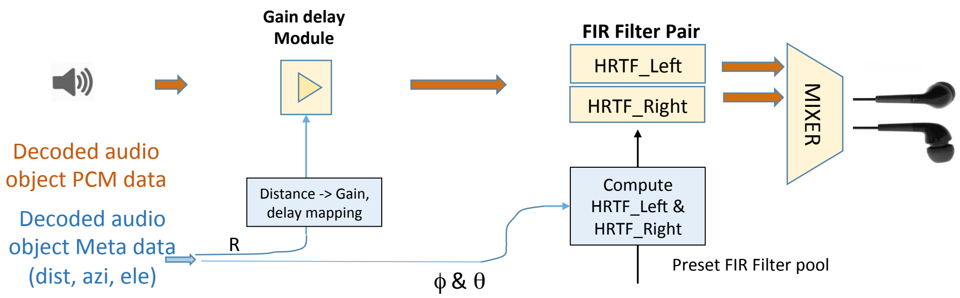
Ear impulse & frequency response for orientation shown in picture on left (90º azimuth, 0º- elevation) from WK SDO set. [2,3]

(Note: 5 to 9 msec duration impulse response width @ 16kHz sample rate)

Object-Based Audio Rendering on Headphones



Object-Based Audio Rendering on Headphones

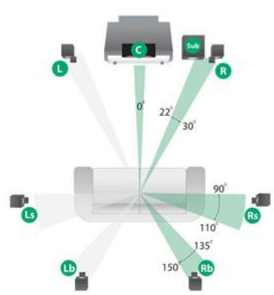


Signal Processing Challenges

- VR / Computer Games related: Head / Joystick movements changes "R, ϕ & θ"
  - ⇒ An additional Head-Tracking or Joystick movements module which feeds user orientation parameters "Ru, ϕu & θu"
  - ⇒ Additional module to Perform 3-D geometry computations to derive final object position parameters from the above 2 sets "R, ϕ & θ"

Object-Based Audio Rendering on Immersive Speaker-Layouts

Examples of Immersive Speaker Layouts

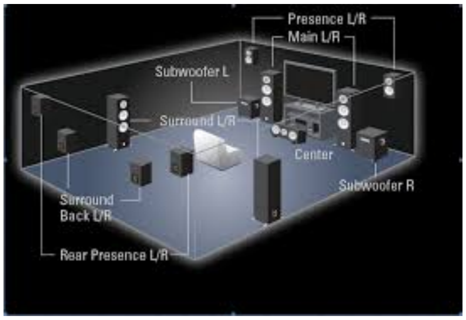


DTS / DD+ 7.1 Speaker Layout

Dolby ATMOS 11.1 Speaker Layouts

Object-Based Audio Renderer on Immersive Speaker-Layouts

Examples of Immersive Speaker Layouts

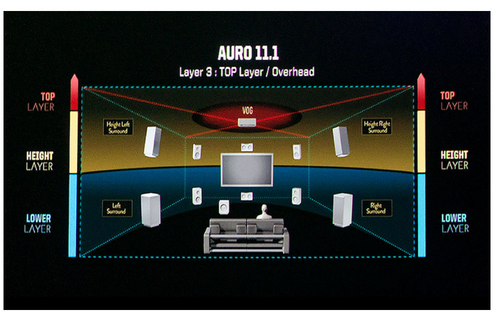


DTS-X 7.2.4 Speaker Layout

DTS Neo:X 11.1 Speaker Layout

Object-Based Audio Renderer on Immersive Speaker-Layouts

Examples of Immersive Speaker Layouts



Auro-3D 11.1 Speaker Layout

Ambisonics 13.1 Speaker Layout

Object-Based Audio Renderer on Immersive Speaker-Layouts

Two Main Techniques.

- VBAP – Vector based amplitude Panning: Mapping object audio to Virtual Speaker Array
- HOA – Higher Order Ambisonics : Creating desired "Sound-Field" at listeners' sitting position

VBAP (Vector Based Amplitude Panning):

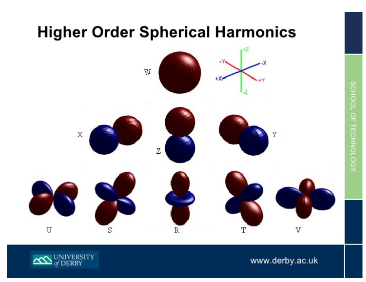
- A large array of "Virtual" Speaker Positions are assumed to surround the listener. Audio-Objects and their motions / positions w.r.t. the listener are mapped on a larger set of "Virtual" Speaker Positions.
- Audio signals for each object is mapped on this virtual speaker positions using VBAP method
- The audio associated with virtual speakers is then mapped to standard user speaker layouts using pre-defined down-mixing matrices & set of delays.

HOA based object rendering on Immersive Speaker-Layouts

Higher Order Ambisonics **[Gerzon 1970]**

- Creates a sound field generated by audio-object(s) when it gets captured by **directional microphones** located at the listener's position

An Ambisonic Microphone

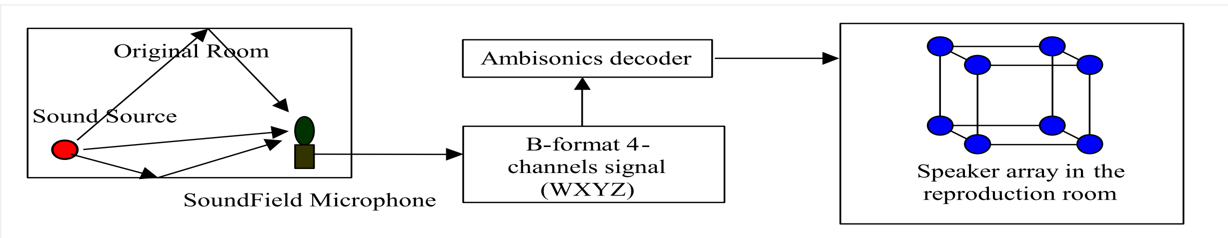


First Order Ambisonics fields

Second Order Ambisonics fields

- HOA channels are encoded and these channels are decoded and then mapped onto any standard "user speaker layouts" from 5.1, or 7.2.4 or 13.1". These mappings are easy and less complex.
- HOA technique makes it easy to modify the sound-field for different user (listener) orientations (required mainly in VR & Computer Gaming)

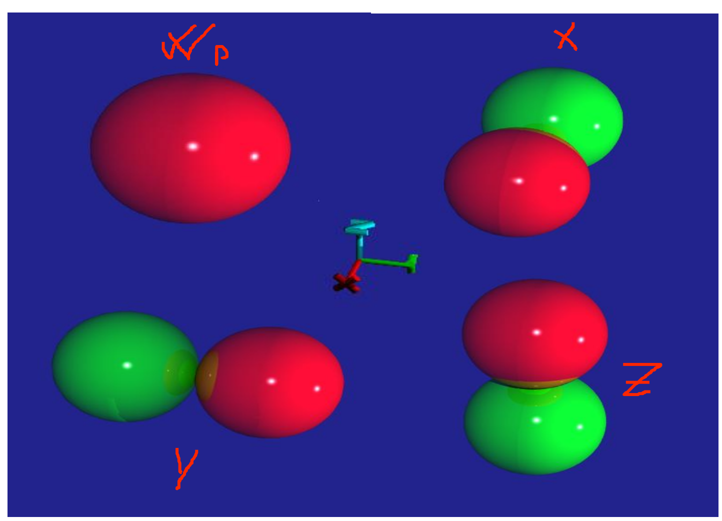
Ambisonics 3D (1st order)



Reproduction occurs over an array of 8-24 loudspeakers, through an Ambisonics decoder

3D extension of the pressure-velocity measurements

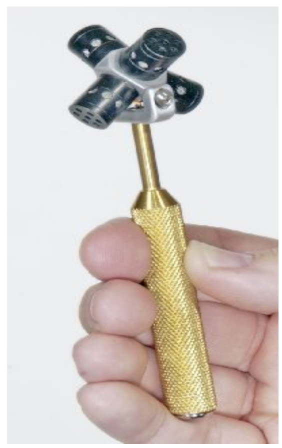
- The Soundfield microphone allows for simultaneous measurements of the omnidirectional pressure and of the three cartesian components of particle velocity (figure-of-8 patterns)



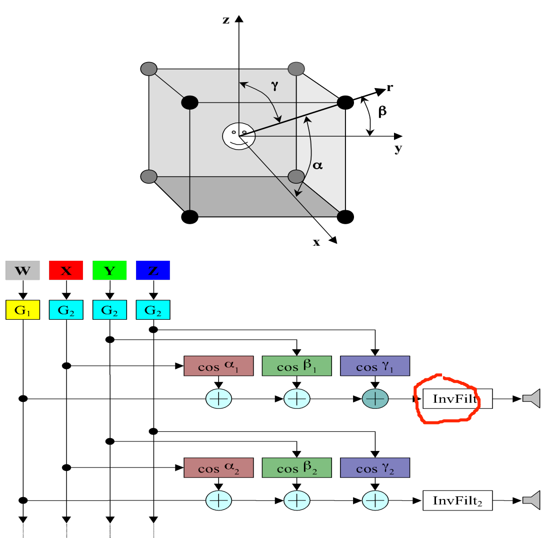
A-format microphone arrays

- Today several alternatives to Soundfield microphones do exists. All of them are providing “raw” signals from the 4 capsules, and the conversion from these signals (A-format) to the standard Ambisonic signals (B-format) is performed digitally by means of software running on the computer





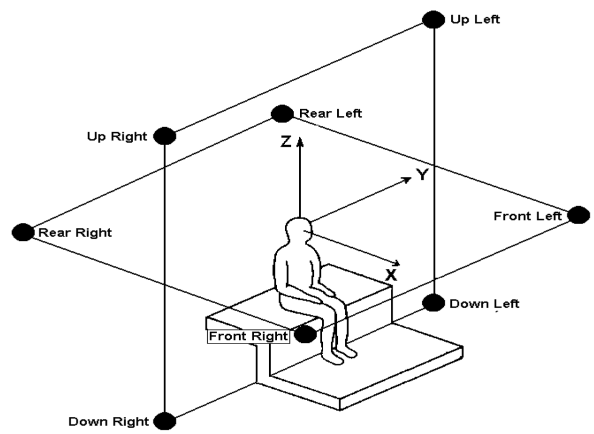

Ambisonics decoding



- Each speaker fed is just obtained as a weighted sum of the 4 B-format signals (WXYZ)
- The weighting factors only depend on the position of each loudspeaker
- It is possible to add a small FIR filter for matching perfectly the gain and phase of all loudspeakers

Ambisonics decoding

Bi-square Ambisonics array



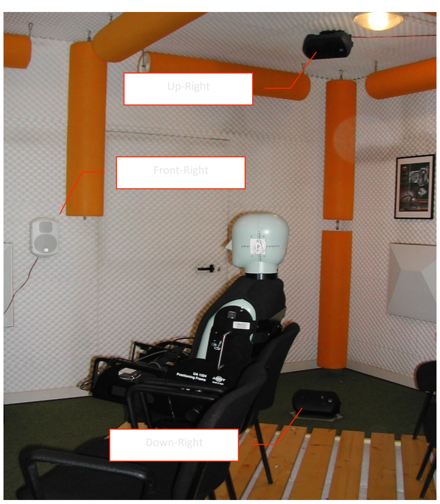
advantages:

- Tridimensional
- Good later sound perception
- Good bass response
- Wide "sweet spot", no colouring outside it

disadvantages:

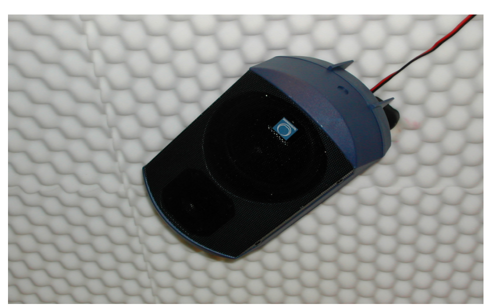
- Not isotropic
- Advanced decoding required (Y gets more weight than X and Z, and this must be compensated for)

Bi-square Ambisonics array



8 Turbosound Impact 50 loudspeakers:

- Light, easily fixed and oriented
- Good frequency response
- Very little distortion



What is Ambisonics?

- Ambisonics is a method for recording, mixing and playing back three-dimensional 360-degree audio. It was invented in the 1970s but was never commercially adopted until recently with the development of the VR industry which requires 360° audio solutions.
- The basic approach of Ambisonics is to treat an audio scene as a full 360-degree sphere of sound coming from different directions around a center point.
- The center point is where the microphone is placed while recording, or where the listener's 'sweet spot' is located while playing back.

Ambisonics B-format

- The most popular Ambisonics format today, widely used in VR and 360 video, is a 4-channel format called **Ambisonics B-format**
- uses as few as four channels to reproduce a complete sphere of sound.

Ambisonics vs. Surround

- Traditional surround technologies are more immersive than simple two-channel stereo, but the principle behind them is the same:
- They all create an audio image by sending audio to a **specific, pre-determined array of speakers**.
- Stereo sends audio to two speakers; 5.1 surround to six; 7.1 to eight; and so on.

- By contrast, Ambisonics does not send audio signals to any particular number of speakers; it is "speaker-agnostic." Instead, **Ambisonics can be decoded to *any* speaker array**. Ambisonic audio represents a full, uninterrupted sphere of sound, without being restricted by the limitations of any specific playback system.

Ambisonics as standard in 360 video and VR:

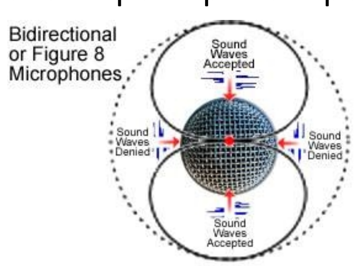
- Traditional surround formats can provide good imaging when static; but as the sound field rotates, the sound tends to 'jump' from one speaker to another.
- Ambisonics can create a smooth, stable and continuous sphere of sound, even when the audio scene rotates (as, for example, when a gamer wearing a VR headset moves her head around). This is because Ambisonics is not pre-limited to any particular speaker array,
- Traditional surround speaker systems are usually 'front-biased': information from the side or rear speakers is not as focused as the sound from the front. By contrast, Ambisonics is designed to spread the sound evenly throughout the three-dimensional sphere.
- Finally, whereas traditional surround systems have various difficulties representing sound beyond the horizontal dimension, Ambisonics is designed to deliver a full sphere complete with ***elevation***, where sounds are easily represented as coming from above and below as well as in front or behind the user.

First-order Ambisonics B-format

- The four channels in first-order B-format are called **W, X, Y and Z**. One simplified and not entirely accurate way to describe these four channels is to say that each represents a different directionality in the 360-degree sphere: center, left-right, front-back, and up-down.
- A more accurate explanation is that each of these four channels represents, in mathematical language, a different **spherical harmonic component** – or, in language more familiar to audio engineers, a **different microphone polar pattern pointing in a specific direction**, with the four being coincident (that is, conjoined at the center point of the sphere).

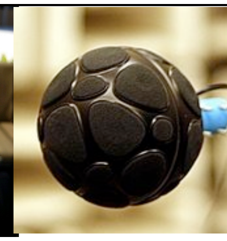
First-order Ambisonics B-format

- **W** is an omni-directional polar pattern, containing all sounds in the sphere, coming from all directions at equal gain and phase.
- **X** is a figure-8 bi-directional polar pattern pointing forward.
- **Y** is a figure-8 bi-directional polar pattern pointing to the left.
- **Z** is a figure-8 bi-directional polar pattern pointing up.



X, Y, and Z channels:

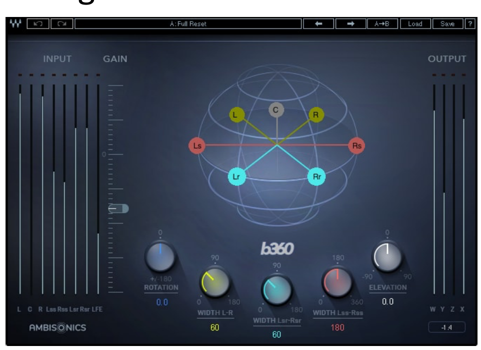




- A figure-8 microphone has a positive side and a negative (inverse phase) side. While the X channel's figure-8 polar pattern points forwards, its negative side points backwards. The resulting audio signal on the X channel contains all the sound that is in the front of the sphere with positive phase, and all the sounds from the back of the sphere with negative phase.
- The same goes for the Y and Z channels: The Y channels pick up the left side of the sphere with positive phase and the right side with negative phase. The Z channel picks up the top side of the sphere with positive phase and the bottom with negative phase. This way, by means of differential gain and phase relations, the four channels combined represent the entire three-dimensional, 360-degree sphere of sound.

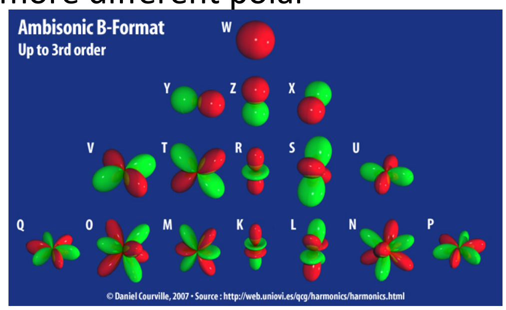
AmbiX vs. FuMa

- two conventions within the Ambisonics B-format standard: **AmbiX** and **FuMa**. They are quite similar, but not interchangeable: they differ by the sequence in which the four channels are arranged, with AmbiX, for example, arranged WYZX instead of WXYZ.

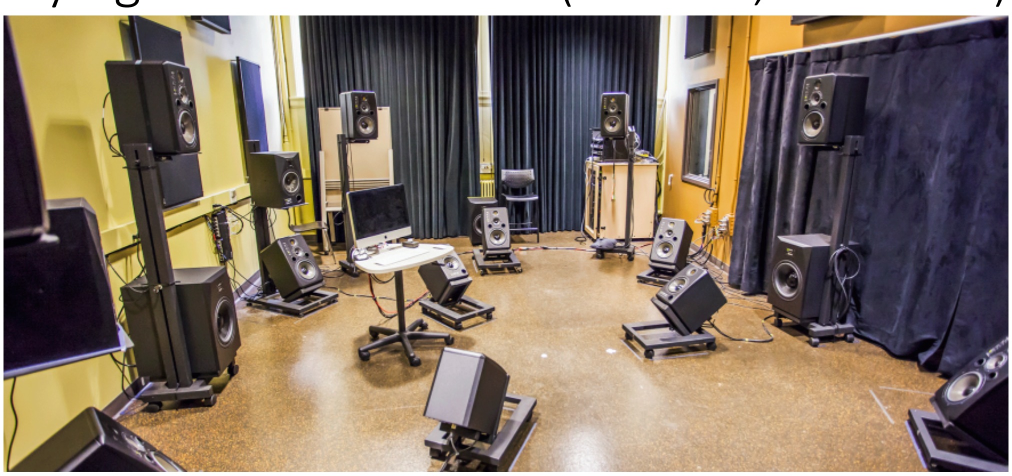


First-order to sixth-order Ambisonics

- The 4-channel format is only a simple, **first-order** form of B-format, which is what most Ambisonics microphones and playback platforms support today.
- Higher-order B-format audio can provide even higher spatial resolutions, with more channels providing more different polar patterns.
- Second-order Ambisonics uses 9 channels,
- Third-order Ambisonics uses 16 channels,
- Sixth-order Ambisonics uses 49 channels.



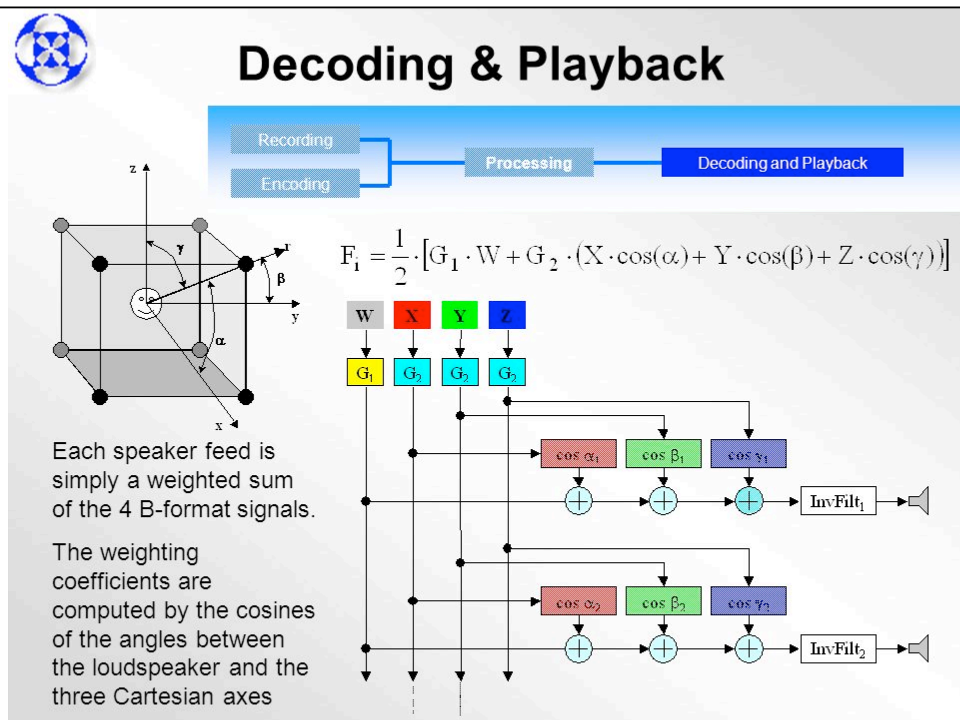
Playing back Ambisonics (DxARTS, Raitt Hall)



Room 117 uses 24 full-range speakers and 4 subwoofers for full height spatial sound reproduction. There is a routing decoding system to care of patching, ambisonic decoding, speaker balancing and room correction, and the crossovers for distributing sound to the subs. You can send a B-format signal to various decoders or address each speaker individually, depending on settings.

Decoding Ambisonics

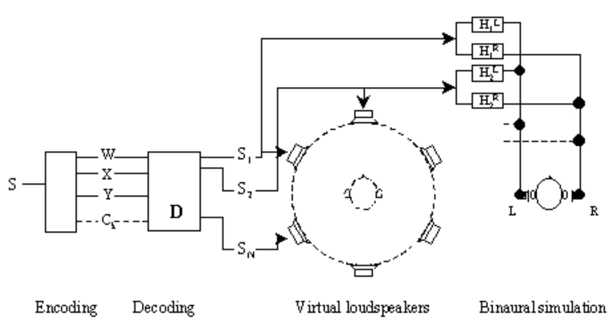
- You can play back Ambisonics on almost any speaker array, recreating the spherical soundfield at the listening spot. But to do that, you need to **decode** the four B-format channels for the specific speaker array.
- All four B-format channels are summed to each speaker feed. Each of the four channels is summed with different gain and phase, depending on the direction of the speaker.
- Some of the sources in the mix are summed in-phase while others are summed out-of-phase at each specific speaker.
- The result is that sources aligned with the direction of the speaker are louder, while those not aligned in the direction of the speaker are lower or cancel out.

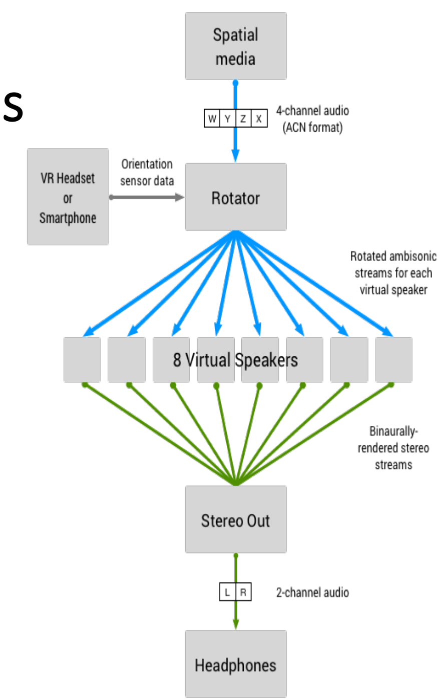


Ambisonics on headphones

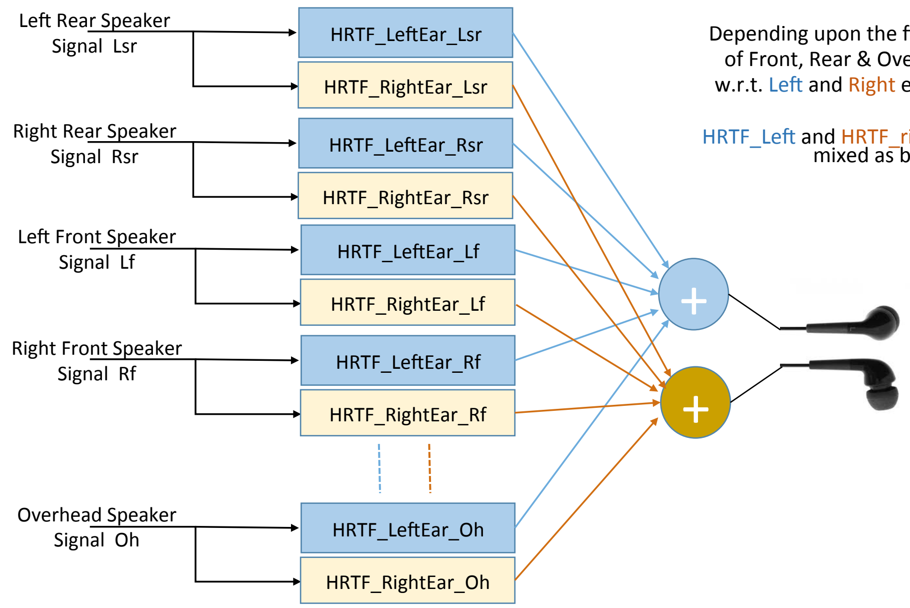
- Spatial sound on headphone is made possible by **binaural audio technologies**. In essence, a binaural processor receives an audio input and a direction in which to position it. the processor adds auditory cues to the signal, so that when played back on headphones it is experienced at the set virtual position.
- The most common way to process Ambisonics for binaural spatial playback on headphones is to decode the Ambisonics channels for a certain speaker array – and send the feeds to a binaural processor which virtually positions them at the direction that the actual speaker would have been.
- The result is that the immersive spherical soundfield is experienced by the listener when monitoring on headphones.

Ambisonics on headphones





BinAural Rendering : Immersive Speakers -> Headphones.



Depending upon the fixed ϕ & θ angles of Front, Rear & Overhead Speakers w.r.t. Left and Right ear of the listener

HRTF_Left and HRTF_right are applied & mixed as below.

Content platforms that support spatial audio

- YouTube
- Facebook 360 Spatial Workstation
- Google VR Resonance Audio
- Samsung VR
- Others

Facebook 360 Spatial Workstation

- The Facebook 360 Spatial Workstation is a software suite for designing spatial audio for 360 video and cinematic VR. It includes plugins for popular audio workstations, a time synchronized 360 video player and utilities to help design and publish spatial audio in a variety of formats. Audio produced with the tools can be experienced on Facebook News Feed on Android and iOS devices, Chrome for desktop and the Samsung Gear VR headset through headphones.

Google VR Resonance Audio

Resonance Audio goes beyond basic 3D spatialization, providing powerful tools for accurately modeling complex sound environments.

- The SDK enables:
  - Sound source directivity customization
  - Near-field effects
  - Sound source spread
  - Geometry-based reverb
  - Occlusions
  - Recording of Ambisonic audio files
- 3rd order Ambisonics formats supported

The end…