Spatial Audio for Immersive Virtual Environments

David Swapp

Department of Computer Science, UCL

<d.swapp@cs.ucl.ac.uk>

Overview

PART 1: Perception of Spatial Audio

- What is spatial audio?
- Why do spatial audio?
- How can spatial audio be simulated?
- Physics of sound
- Psychoacoustics

Overview

PART 2: Synthesis of spatial audio

- Methods for synthesizing spatial audio
- Headphones vs. speaker array
- Hardware & Data requirements
- Environmental acoustic modelling

What Is Spatial Audio?

- “Spatial audio” describes any of a variety of techniques for simulating the 3D soundfield that occurs in a real environment.
- Presentation via headphones or speaker array
- First implementation at 1881 Paris exhibition
- Real environment may contain many audio sources, but the soundfield must be simulated only at 2 ears.
- As well as direct sound waves, reflections & diffractions of these waves from objects in the environment must be accounted for.

Paris Exhibition 1881

- Clément Ader placed microphones at either side of the stage of the Paris Grand Opera.

- Telephone lines connected these to listening rooms 2 miles away at the exhibition.

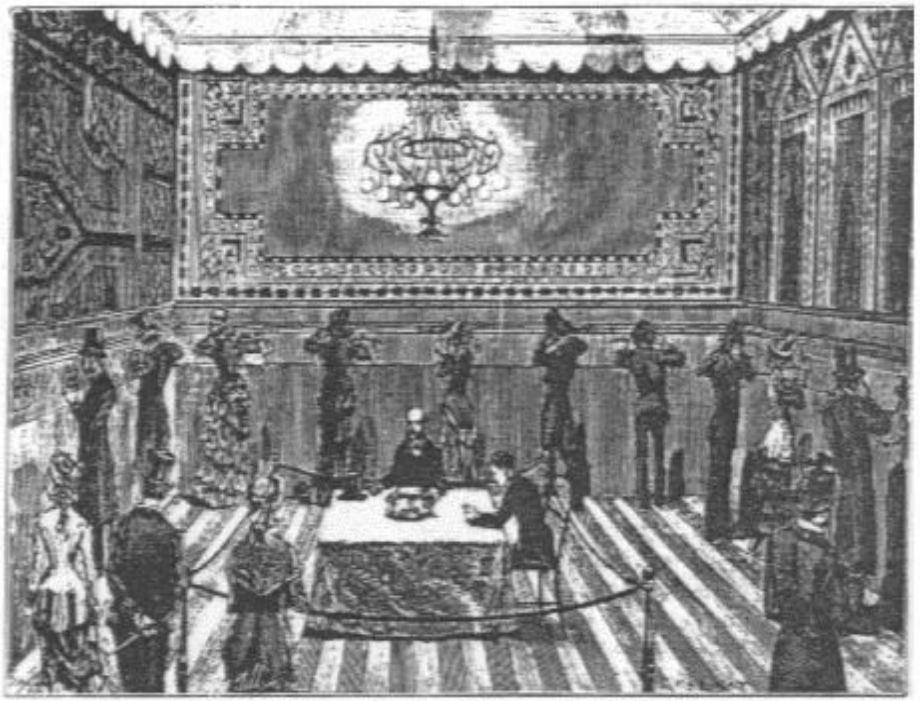


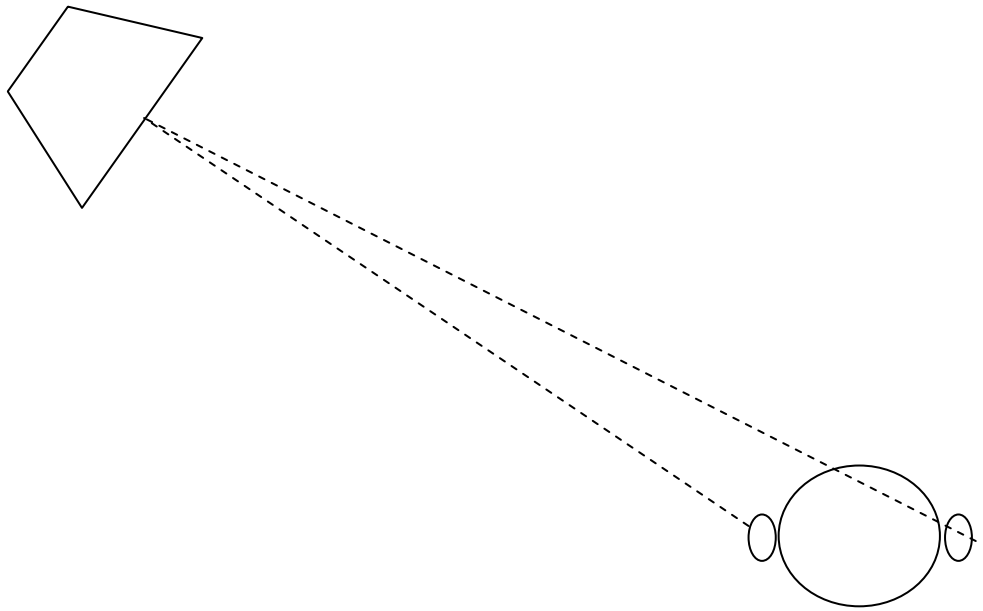
Illustration from "Musical Broadcasting in the 19th Century" by Elliott Sivowitch, Audio, June, 1967, page 21

Why Do Spatial Audio?

- Enhanced spatial awareness in VE
- Sense of presence
- Applications
  1. Architectural
  2. Collaborative
  3. Archaeological/Forensic

How Can Spatial Audio be Simulated (I)

Monaural: Sound comes from direction of loudspeaker



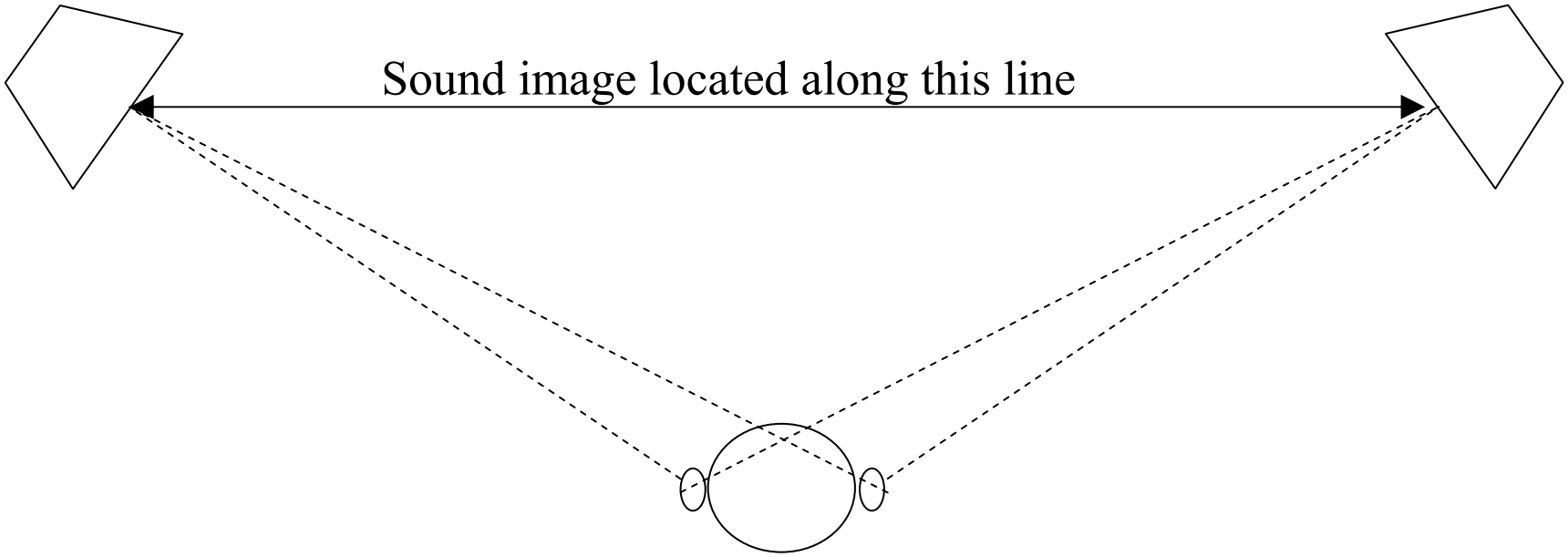
How Can Spatial Audio be Simulated (I)

Monaural:

- Single channel
- Contains no directional information (if single output)
- Distance cues can be simulated
- With two outputs, amplitude panning can produce a moving sound image

How Can Spatial Audio be Simulated (II)

Sound image can appear to move between the 2 speakers



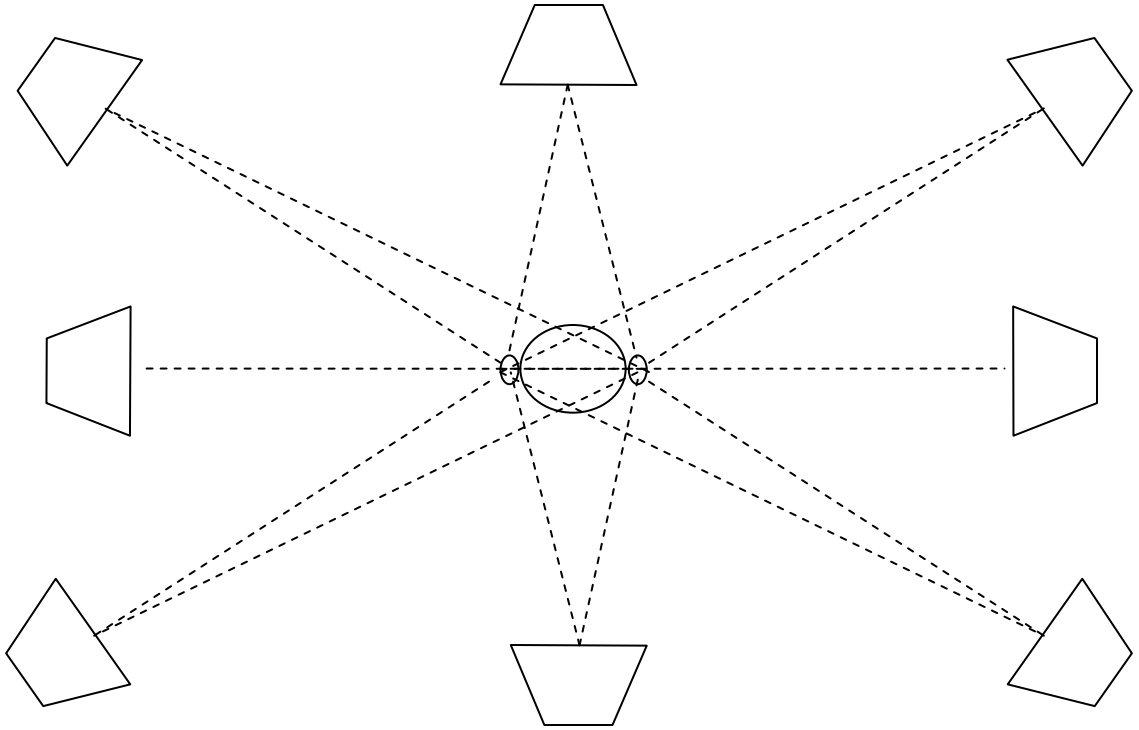
How Can Spatial Audio be Simulated (II)

Stereo:

- 2 distinct audio channels
- Limited directional information in one spatial dimension via amplitude panning
- Amplitude panning can be achieved within the stereo mix (not just at the output)

How Can Spatial Audio be Simulated (III)

Multi-channel audio: surround sound??



How Can Spatial Audio be Simulated (III)

Multi-channel audio:

- Extension of stereo principle to multiple channels and multiple outputs
- Can allow front-back or elevation to be simulated.
- The current state of the art in computer games and movies can (theoretically) simulate up to 14 channels

Limitations of multi-channel audio

- Relies only on manipulation of audio level & environmental effects
- Human audio system also uses other cues so should also try to simulate these:
  - Phase is used for localisation of low-frequency sounds (<1.5kHz) though this diminishes at higher frequencies
  - Level is better for higher frequencies
  - Also need to consider influence of our body on the signal entering our inner ears

Beyond audio panning: reconstructing the soundfield

- Goal is to present sound waves at the listener's ears that would occur in a real environment
- Encoding of phase as well as amplitude information
- Technically difficult to achieve

Physics of sound

- Sound is caused by vibrations of particles in the form of a longitudinal waveform.
- Sound waves travel at ~340ms^{-1} in air
- Audible frequency range of sound is ~20Hz to 20kHz (17m to 17mm wavelength).
- Sound waves reflect specularly if surface is much larger than wavelength.
- Sound waves diffract if surface is approx same size as the wavelength

Physics of sound (2)

- Sound waves refract if they change speed (e.g. moving from cold air to warm air)
- Propagation of sound waves is in many ways analogous to propagation of light waves, but there are crucial differences:
  - Lower speed of sound allows for perceptually significant temporal effects e.g. reverberation and Doppler shifting
  - Coherent nature of sound waves means that interference is more prevalent

Psychoacoustics

Psychoacoustics is the association of measurable audio stimuli (e.g. frequency, power) with the subjective sensations experienced by people (pitch, loudness).

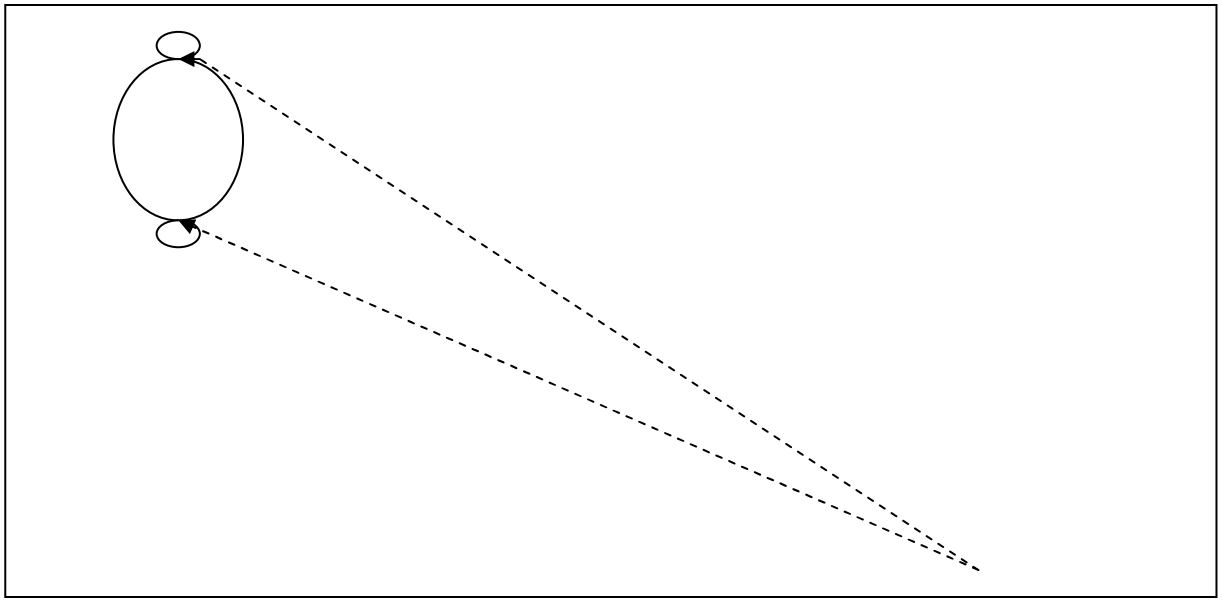
Ranges of sensitivity

Frequency range approx 20Hz to 20kHz (~10 octaves)

Amplitude range hard to determine, but order of magnitude for ratio of loudest audible sound (at pain threshold) to quietest audible sound at 4kHz is one million.

Sound source location

Direction and distance estimates are made on account of differences in the sounds that reach our left and right ears.



Source direction estimation

(1) Interaural time difference:

- Sound arrives at a fractionally different time (less than 1ms) at the left and right ears.
- Human audio system is sensitive to the phase difference between the left and right channels
- Only works for phase differences less than a whole wavelength thus works better for detecting direction of low frequencies.

Source direction estimation (2)

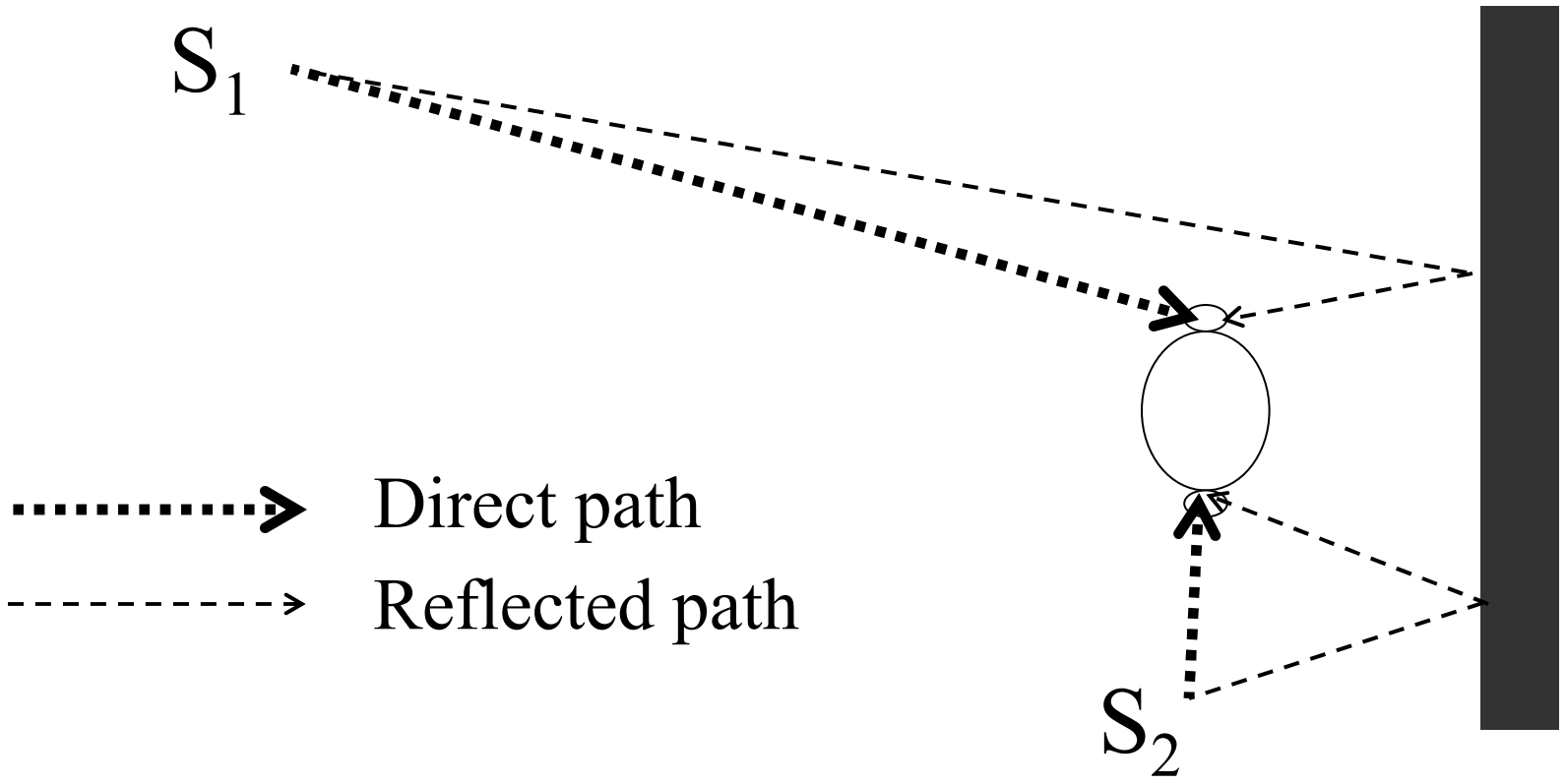
(2) Interaural intensity difference:

- Sound is scattered when it strikes the head and upper torso – an effect known as “shading”.
- Intensity of sound at ear further from the source is consequently reduced.
- Shading is more pronounced for high frequencies since these are scattered more.

Source distance estimation

- Problem is the differentiation of a loud sound far away and a quiet sound close by.
- We can detect the differential attenuation across the frequency range as sound sources move further away from us.
- Difference in the amount of reflected sound: more reflected sound implies that sound source is further away.

Distance estimation using ratio of reflected sound to direct sound



Head-related transfer function (HRTF)

- Filtering effect of our bodies on sounds en route to the ears
- Can be modeled mathematically to reproduce the same effect via headphones .
- Mainly composed of:
  1) Filtering by the outer ear flap (pinna) affects the propagation of different (especially high) frequencies. The precise nature is determined by the ear shape, thus is unique to each individual.
  2) The upper torso reflects frequencies (especially mid-range) to produce very short time-delayed echoes. The length of this time delay varies with the elevation of the sound source.

Properties of HRTF

- Monaural cue (though we have 2 distinct HRTFs)
- Modifies both frequency spectrum and timing of incoming signal .
- Varies with direction of incoming signal
- Affected by changes in clothing, hairstyle etc
- Can be measured by placing small mics at entrance to ear canal and taking very many measurements

Environmental effects

- **Absorption by air:** greater for humid or polluted environments
- **Collision:** material and shape/size of object affects amount of reflection, absorption and diffusion
- **Reflection:** solid surfaces will reflect more than soft furnishings which will absorb energy from the sound waves.
- **Diffusion:** If object is small or collision is near edge: collisions within a 1/4 wavelength from an edge are considered to diffuse (scatter) rather than reflect.

Reverberation

- A sound source will be reflected, absorbed and otherwise attenuated until it loses its energy (generally considered to be a drop of 60dB).
- Reverberation time of a room can be measured as the length of time it takes for a sound signal to drop away to this level.
- Large empty rooms (thus fewer reflections) with solid surfaces (more reflection, less absorption) will have long reverberation times and small rooms with soft surfaces will have short reverberation times.

Part 2 Overview

Synthesis of spatial audio

- Methods for synthesising spatial audio
- Hardware & Data requirements
- Headphones vs speaker array
- Environmental acoustic modelling

Binaural Simulation

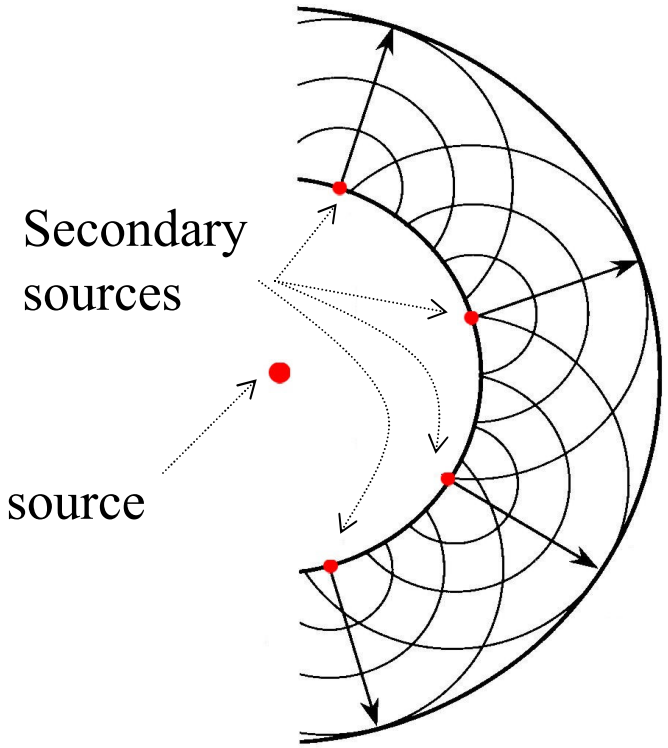
- 2 output channels
- Originally conceived as means of recording acoustics for evaluation of auditoria
- Takes account of listener's physical characteristics via HRTF
- Highly effective though laborious to compute
- Generic HRTFs often used

Soundfield Reconstruction without Headphones

- Many output channels: at least 4 for horizontal simulation, 8 for 3D
- Must account for location of listener relative to the speaker array
- 2 predominant methodologies:
  1. Wavefield Synthesis
  2. Ambisonics

Wavefield Synthesis

- Based on Huygens' Principle
- Soundfield of environment is approximated by array of loudspeakers acting as secondary sources

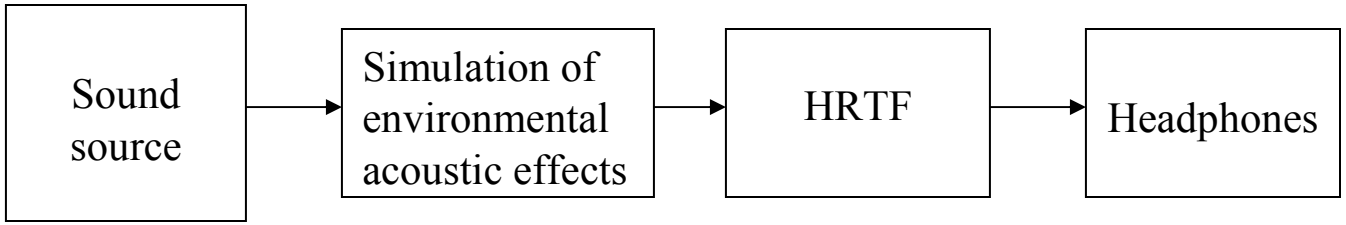


Ambisonics

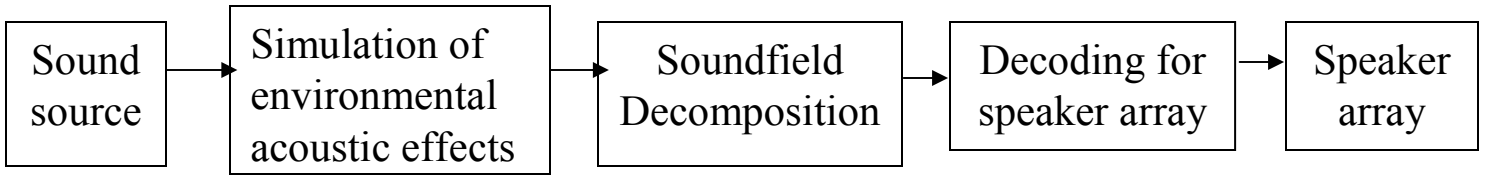
- Normally encodes soundfield in 4 channels (B-format signal):
  - Audio pressure W
  - Front-back X
  - Left-right Y
  - Up-down Z
- Can be easily operated upon by spatial transformations (e.g. rotations)
- Encoding and decoding are separable: signal can be decoded for any arbitrary speaker array.

Headphones vs Speaker Array

Headphone (binaural) presentation



Speaker array presentation



Headphones vs Speaker Array

- Headphones avoid the need to allow for the location of the listener relative to the speaker array
- Speaker array avoids the need for computation of HRTF
- Acoustic isolation useful in some circumstances, and not in others
- Headphones can be an encumbrance
- Speaker array better for loud, low frequency sound

Output hardware

- Minimal requirement is for a binaural display (e.g. via headphones), but display may also be via a speaker array (e.g. in cave-like VR setup).
- Major challenge is the processing of audio signals such that they arrive at either headphone or speaker outputs filtered appropriately to simulate audio sources located in the 3D virtual environment

Data requirements

- Sampling rate of audio source:

  To represent a signal of frequency ***f*** , it must be sampled at a rate of at least ***2f*** (Nyquist theorem); otherwise the signal will sound distorted (aliasing).

- Dynamic range of audio source:

  Dynamic range is nonlinear and can be adequately represented by as few as 256 gradations (which conveniently fits into 8 bits of data) although high quality digital audio will be represented by 16, 20 or 24 bits

- ## Sound source location

  In most simulations, we want the sources of audio to be able to move (e.g. a car driving past, people talking and walking, insects flying). Thus the real-time simulation must continually update the location and orientation (needed for directional sound sources) of each sound source.

- ## Listener location

  This is needed for the same reasons as above, since it is the location of the listener *relative* to the sound sources that is important. In addition to this, the location of the listener relative to the speaker array must be accounted for, since this will affect the sound that reaches the listener's ears. Currently this latter consideration is difficult to achieve in real-time.

HRTF modelling

- Laborious to compute - measurements must be taken of sounds over a range of frequencies and spatial positions
- Listener's HRTF characteristics are subject to change over time
- Effect is very compelling
- Generic simplifications are often used, with trade-off of reduced effectiveness

Environmental acoustic modeling

Environmental acoustic modeling of a room or building is analogous to building a graphical model.

- First the geometry of the environment must be described in terms of the 3D co-ordinates of all significant surfaces (floor, ceiling, walls etc).

- The characteristics of the surface materials that cover the geometry must also be described (analogous to the colour/texture properties of the graphical model). These are described in terms of **absorption** and **diffusion** co-efficients across a range of frequencies.

Simulation of environmental effects

A model of the acoustic properties of an environment can be used as the basis for computing a simulation of the propagation of sound through that environment. There are various ways of achieving this:

- Finite element methods
- Image source methods
- Ray tracing
- Beam tracing

Simulation of environmental effects (2)

Once the propagation paths have been computed, the effect of each reflection, diffraction etc must be accounted for. Physical effects to be considered are:

- Distance attenuation
- Atmospheric scattering
- Diffraction from surfaces of appropriate dimension
- Doppler shifting

Various simplifying assumptions are used e.g. point sources, perfect specularity, Lambertian surfaces