ELL 788
Computational Perception & Cognition

July – November 2015

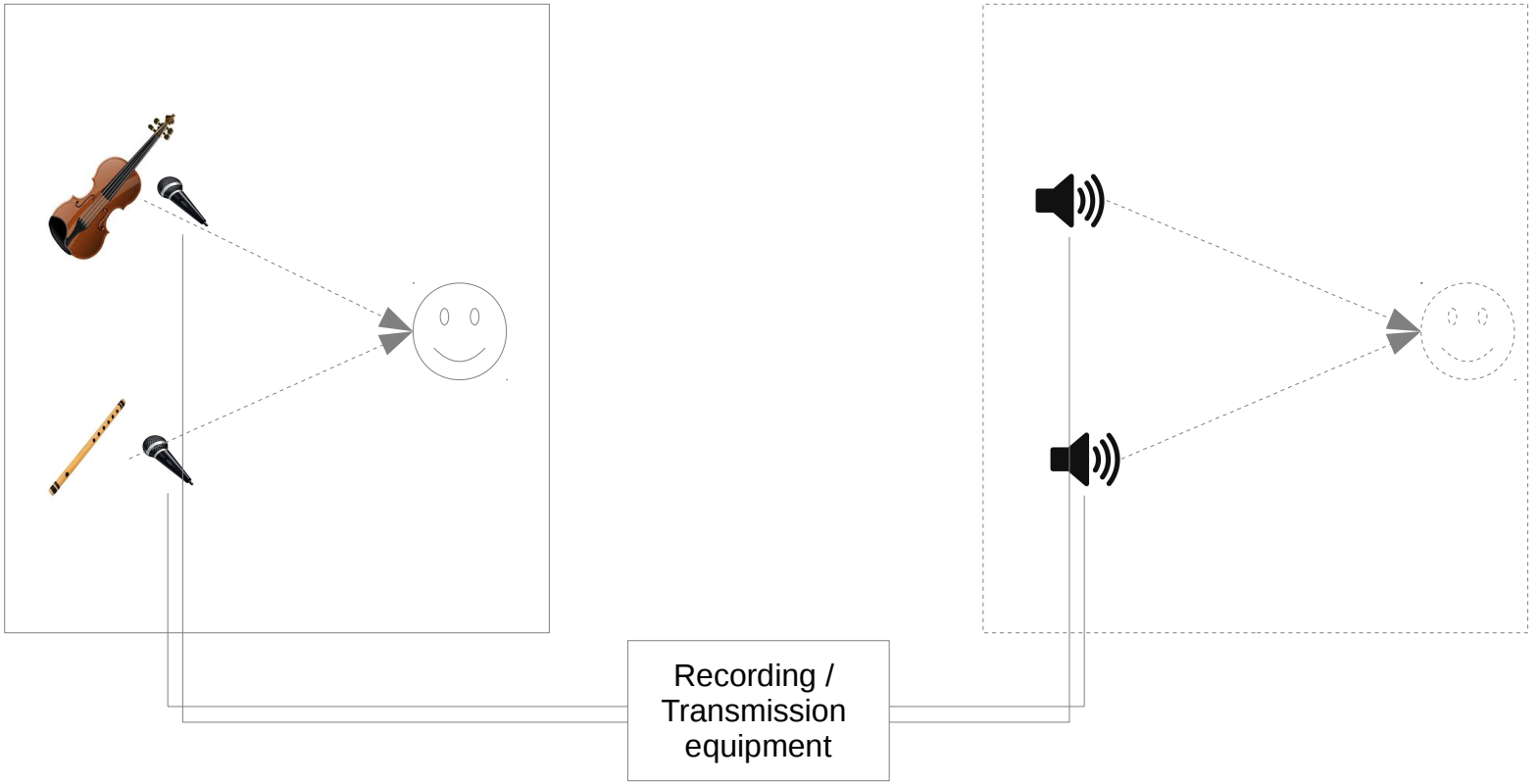
Module 10

Audio Engineering: Spatial Audio

Multichannel sound reconstruction

Accurate physical reproduction of a sound scene

- A microphone for every source
- A speaker for every microphone in the corresponding target location



Rely on <u>inaccurate</u> human perceptual system to solve the problem

Sound localization in human auditory system

- Interaural cues
  - Difference in arrival times (ITD)
  - Relative amplitudes -- high-freq sounds (ILD)
  - Assymetric spectral reflection from body parts
  - Ratio of direct signal and reverberations (echoes)
- Mono-aural cues
  - Direction-specific frequency response because of
    - outer ear (pinna) and external sound canal

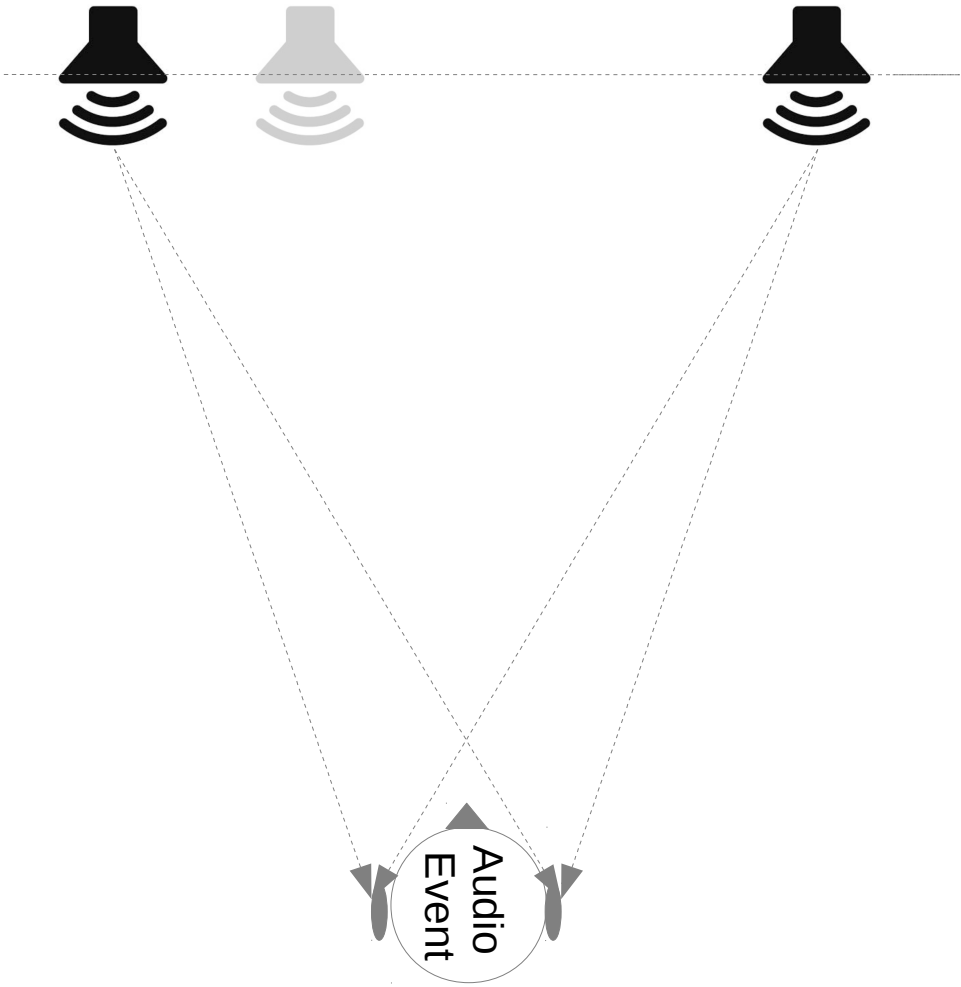
Perception of audio

Sound Event

- Playback from one speaker
  - One audio event
  - Localized at the sound event
  - Pinna, ITD, ISD help in locating source

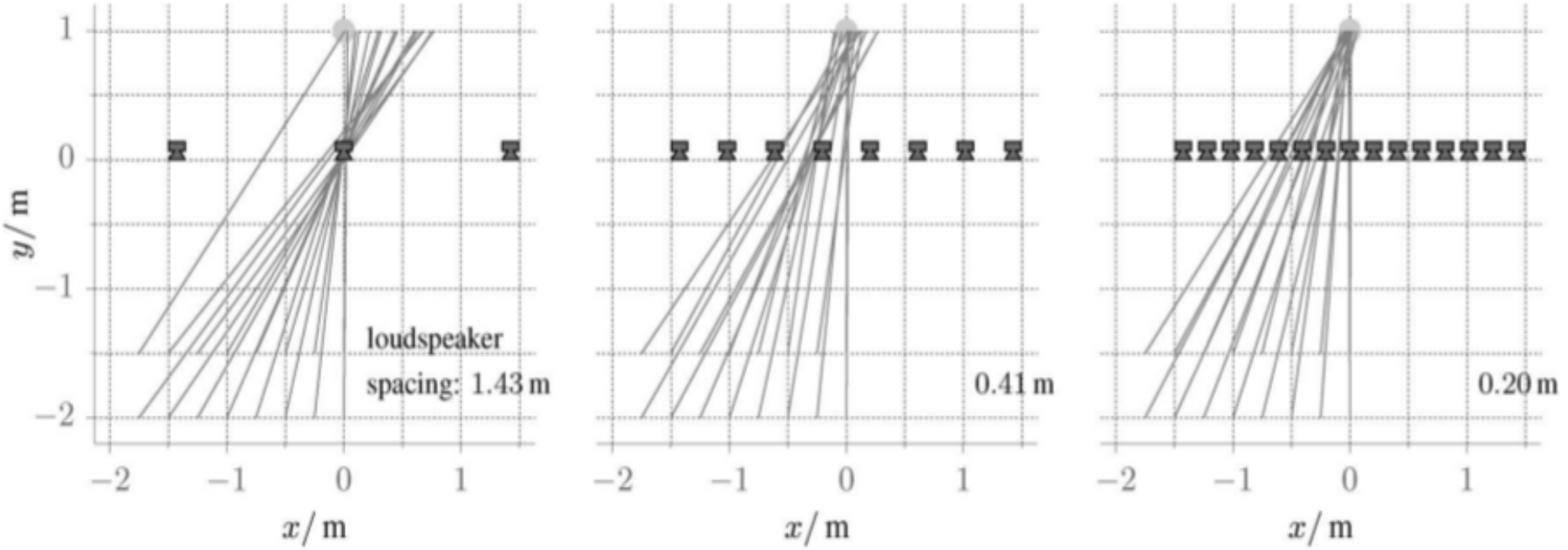
Playback of same sound from two speakers

- No one to one correspondance
- Perceived as single sound localized in between the speakers
  - If listener is (almost) equidistance from the speakers
  - **Within 1 ms** relative delay between the arrival for the two sound events (in two ears)
  - Perception of azimuthal angle depends on relative loudness and delay



Back then (in 1981), two parallel telephone channels were used in order to transmit musical performances from the Paris Opera House to a pair of earphones at the listeners' homes.

Reconstruction over an extended listening area (Perceptual experiment)



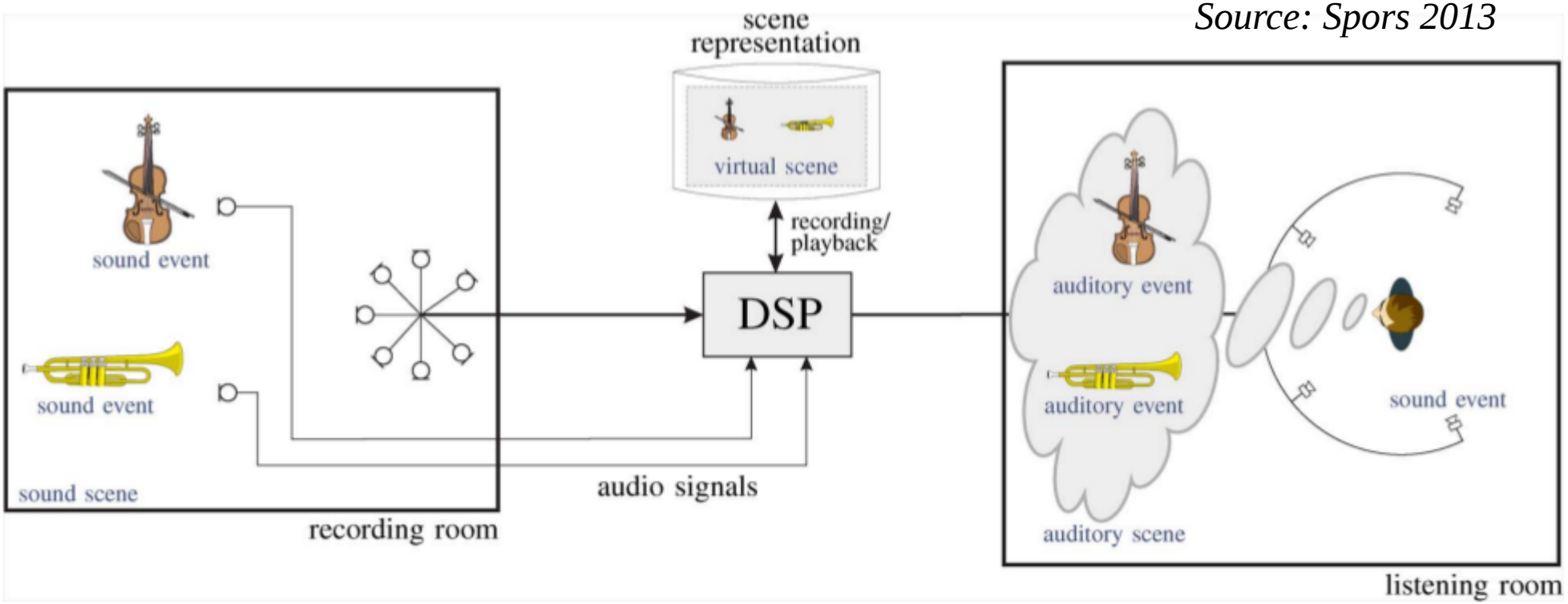
Source: Spors 2013

- Localization using array of speakers
- More convergence as speakers are placed at shorter intervals
  - Depends on frequency (Higher frequency → shorter interval)
- For high freq (~ 20 kHz), the distance between speakers < 1 cm for satisfactory convergence
  - ***Massive multi-channel sound reproduction method***

Virtual sound scene:

Sound scene → Auditory scene

Source: Spors 2013



Virtual sound scene

- Recorded audio and metadata
  - Location, level, ...
- Contains many virtual sources and virtual sound events
- May be related to models of individual sound events
  - Directivity, Diffusedness, Reverberation, Density ...

Quality attributes of an auditory scene

attribute	description
spatial fidelity	degree to which spatial attributes agree with reference
spaciousness	perceived size of environment
width	individual or apparent source width
ensemble width	width of the set of sources present in the scene
envelopment	degree to which the auditory scene is enveloping the listener
depth	sense of perspective in the auditory scene as a whole
distance	distance between listener and auditory event
externalization	degree to which the auditory event is localized in- or outside of the head
localization	measure of how well a spatial location can be attributed to an auditory event
robustness	degree to which the position of an auditory event changes with listener movements
stability	degree to which the location of an auditory event changes over time

Spatial attributes

attribute	description
timbral fidelity	degree to which timbral attributes agree with reference
coloration	timbre-change considered as degradation of auditory event
timbre, color of tone	timbre of the auditory event(s)
volume, richness	perceived "thickness"
brightness	perceived brightness or darkness (dullness)
clarity	absence of distortion, clean sound
distortion, artifacts	noise or other disturbances in auditory event

Timbral attributes

Recording styles

- Explicit scene description
  - Each sound source is independently recorded
  - Audio scene = recorded audio + explicit spatialization parameters
- Implicit scene description
  - A main microphone array record several sound events simultaneously
  - Implicit spatialization parameters
- In practice, a combination of the two is used
  - A microphone array to record several sound events
  - Spot microphones close to specific instruments or instrument groups

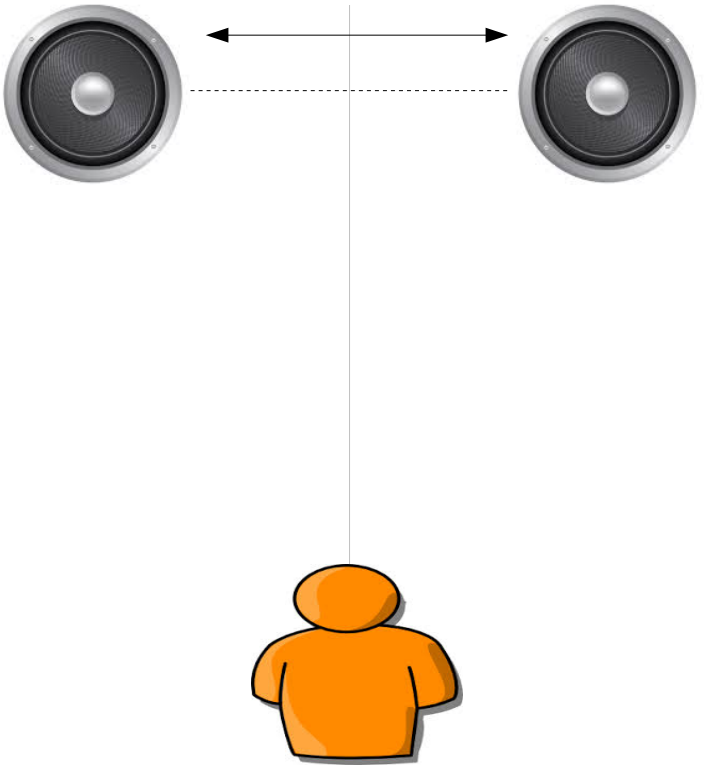
Representation of virtual sound scenes

- Channel-based approach
  - Two channel stereo / 5.1 multi-channel surround
  - Loudspeaker signals that will produce desired audio-scene are stored
- Transform domain based approach
  - Use Orthogonal basis functions to represent the sound scene
  - Reconstruction for specific speaker layout
  - Flexibility in number and layout of speakers
- Object based approach
  - Signals of the virtual sound sources are kept separate
    - Metadata (spatial layout of the virtual sources)
  - Rendering at reproduction site

Channel based approach:
Generating virtual sound scene (Two channel stereo)

- Two speakers, Sound signal s(t)
  - Weights (g_1,g_2) and delays (τ_1,τ_2)

$$d_1(t)=g_1 s(t-\tau_1), d_2(t)=g_2 s(t-\tau_2)$$



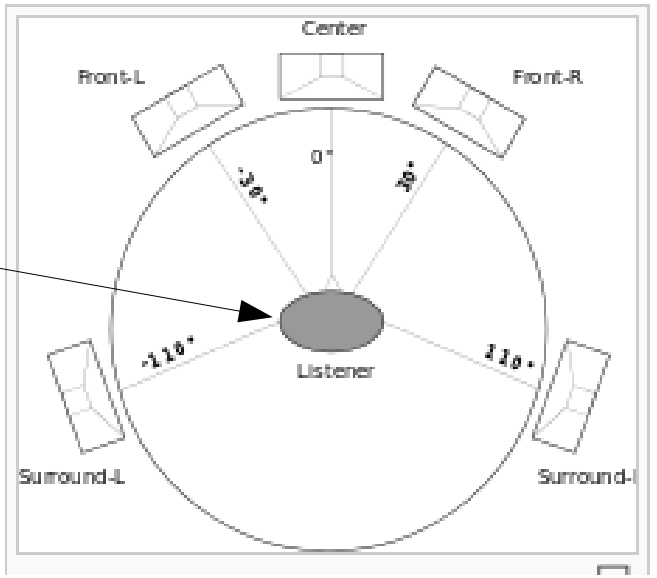
- Adjust relative levels 20 (log g_2 – log g_1) and delay $\tau_2 - \tau_1$ for sound panning
  - Level panning is common – provides good results below 1.5 kHz

- Use an array of speakers for better spanning
  - Global panning: Use all available speakers to reproduce a sound source
  - Local panning: Use pair or triplets of speakers to reproduce a sound source

Review: channel based approach

- Trivial decoding: suitable for commodity music systems
- Infexible: assumes specific speaker layout
- Users experience coloration, increased impression of width
- Degraded localization -- Existence of a “sweet spot”

Sweet spot



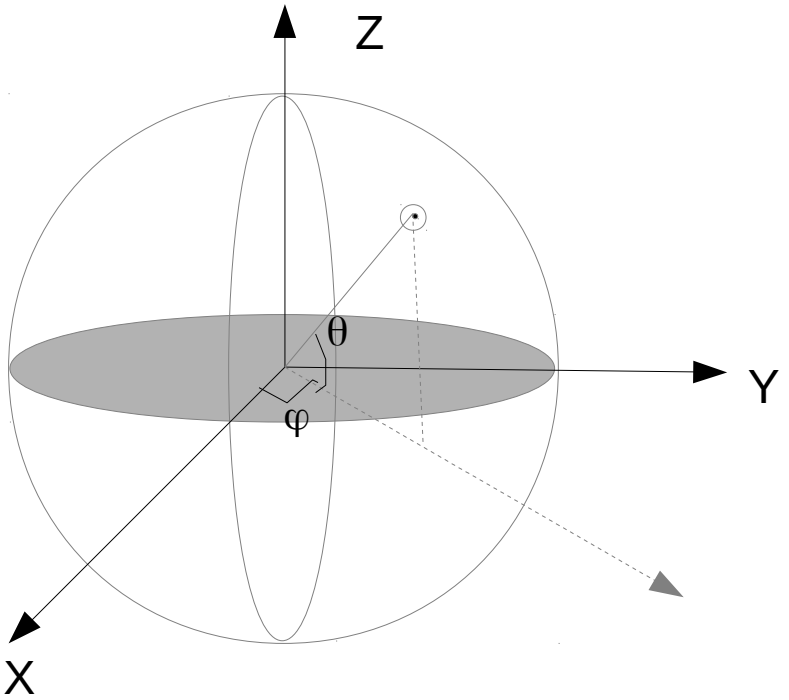
Commonly used technique

Transform based approach
(First order ambisonic)

- Assume that a set of sound sources are located on the surface of an unit sphere
- The sound sources can be represented $\{s_i(t), \varphi_i, \theta_i\}, i = 1..k$

Elevation

Asymuth



B-format encoding
(4-channel encoding)

$$W = \frac{1}{\sqrt{2}} \frac{1}{k} \sum_{i=1}^{k} s_i(t)$$ omnidirectional information

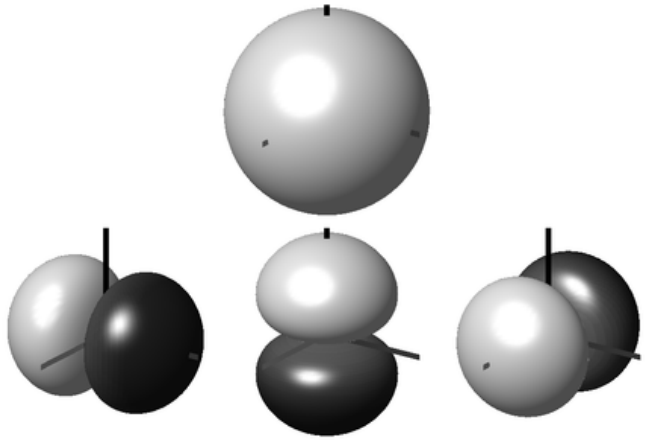
$$X = \frac{1}{k} \sum_{i=1}^{k} s_i(t) . \cos\varphi_i . \cos\theta_i$$ x-directional information (front-back)

$$Y = \frac{1}{k} \sum_{i=1}^{k} s_i(t) . \sin\varphi_i . \cos\theta_i$$ y-directional information (left-right)

$$Y = \frac{1}{k} \sum_{i=1}^{k} s_i(t) . \sin\theta_i$$ z-directional information (elevation)

Ambisonic microphone

First order ambisonic signals can be captured with one omnidirectional microphone and three directional (figure of 8) microphones



Decoding B-Format

- Assume L speakers spread over a unit sphere.
  - Locations: (φ_j, θ_j), $\quad j = 1 \,..\, L$
  - $L \geq N$ (number of ambisonic channels)

- Signal feeding j-th speaker (decoding by projection)

$$p_j = \frac{1}{L}\left[W \,.\, \frac{1}{\sqrt{2}} + X \cos\varphi_j \,.\, \cos\theta_j + Y \sin\varphi_j \,.\, \cos\theta_j + Z \,.\, \sin\theta_j \right]$$

- Often Z (elevation) is ignored ... spreakers are arranged on a circle
- More number of speakers $\rightarrow$ better localization
- Layout – as regular as possible

Head movement:

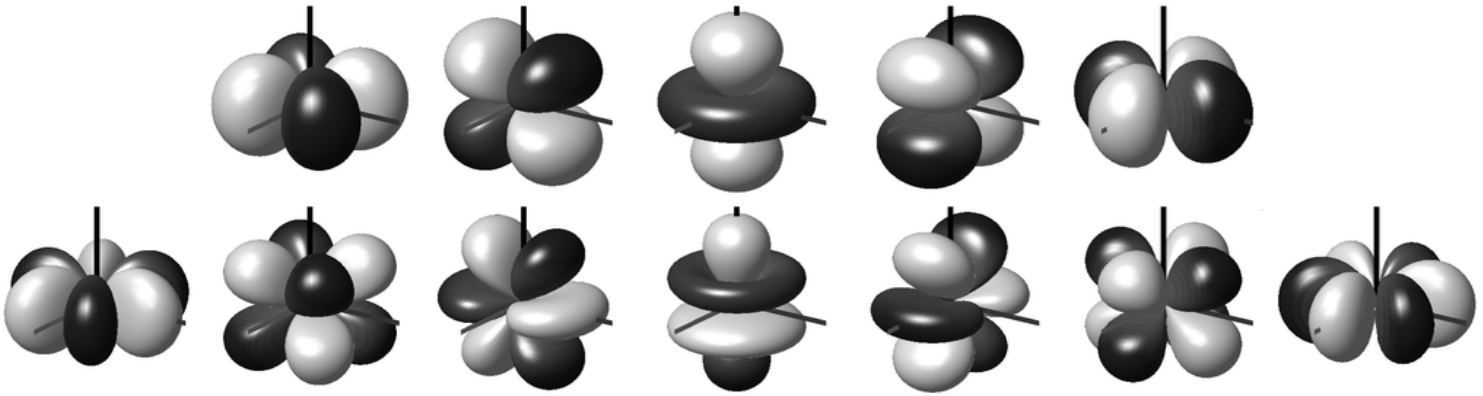
- Simple rotation, tilt and tumble operations can take care of head movement
- Head tracking device

Higher order ambisoncics

- Sweet spot can be extended with higher order ambisonics
- Additional channels need to be introduced
  - 4 + 5 channels for order 2 ambisonics
  - 4 + 5 + 7 channels for order 3 ambisonics ...

2nd order

3rd order



Basic assumptions

1. A soundfield can be regarded as a superposition of plane waves (speakers are sufficiently far from each other)
2. A plane wave can be represented as an infinite series – that can be approximated by spherical harmonic functions for the sweet spot (near the center of the sphere)

Object based encoding

- Encoding each sound source separately for large number of sources is difficult
- Use channel-based / ambisonic coding for a group of sound sources
  - Choir
  - A set of musical instruments
- Save the meta-data (spatial layout)
- Possible to personalize, e.g.
  - Drop some sources
  - Change spatial layout

References

- Spors, et al. Spatial Sound With Loudspeakers and Its Perception: ... Proc. IEEE, Sept 2013
- Hollerweger. An Introduction to Higher Order Ambisonic