Aalto University
School of Electrical Engineering

Acoustics seminar ELEC-E5631
Spatial Audio introduction

Ville Pulkki

Department of Information and Communications Engineering
Aalto University, Finland

March 13, 2024

These slides

- Background
- Devices
- Audio content types
- Applications
- Development of Spatial audio, a history

1 Background

Some terminology

- Acoustics: science of sound

Some terminology

- Acoustics: science of sound
- Spatial: something in 3D space

Some terminology

- Acoustics: science of sound
- Spatial: something in 3D space
- Spatial sound: science of sound with focus on spatial characteristics of sound fields and their perception

Some terminology

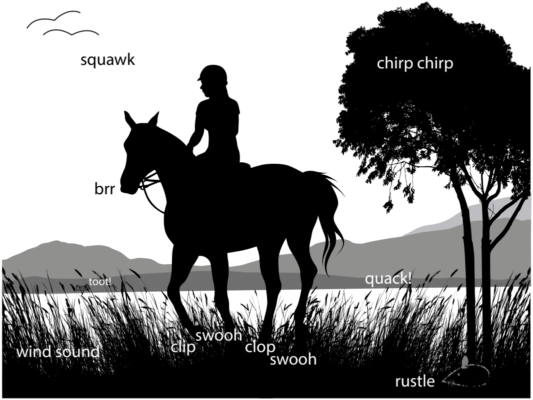
- Acoustics: science of sound
- Spatial: something in 3D space
- Spatial sound: science of sound with focus on spatial characteristics of sound fields and their perception
- Spatial audio: the same, but some electroacoustic transducers (mics, loudspeakers) involved

Some terminology

- Acoustics: science of sound
- Spatial: something in 3D space
- Spatial sound: science of sound with focus on spatial characteristics of sound fields and their perception
- Spatial audio: the same, but some electroacoustic transducers (mics, loudspeakers) involved
- 3D sound: usually means HRTF processing for headphones, can mean also something similar to "spatial audio"

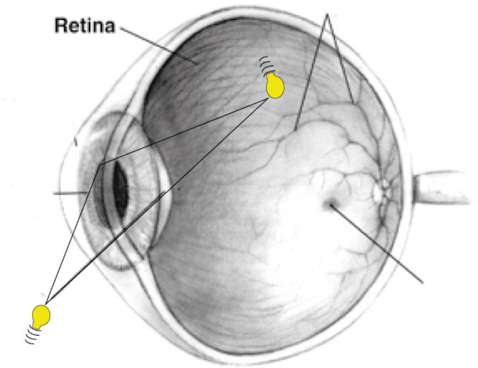
Where and what?

- Localization of sources
- Listening selectively towards different directions



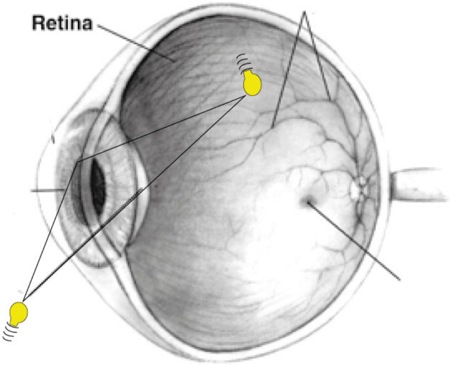
Human eye

Perception



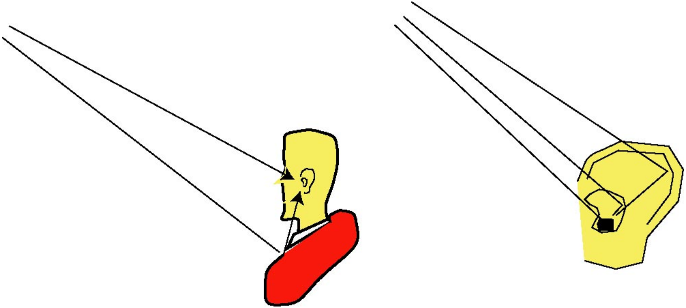
- The cells in the eye are a priori sensitive to direction of light

Human eye



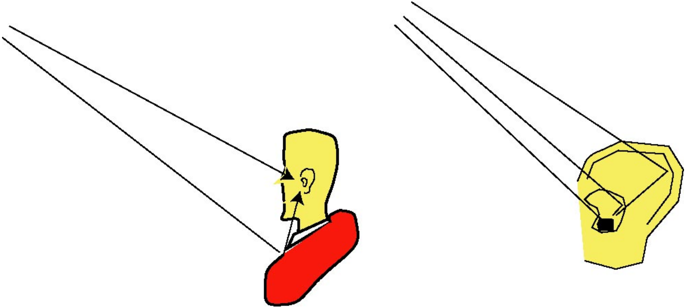
- The cells in the eye are a priori sensitive to direction of light
- Response to quite limited range of wavelengths (380-740nm)

Human spatial hearing



- Response to very large range of wavelengths (2cm–30m)
- Ear canal diameter $<$1cm, sound just bends into the canal

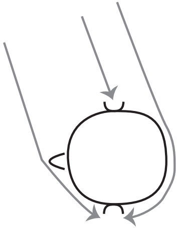
Human spatial hearing

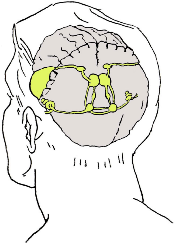


- Response to very large range of wavelengths (2cm–30m)
- Ear canal diameter $<$1cm, sound just bends into the canal
- One ear alone knows quite little of direction

Human spatial hearing

Perception



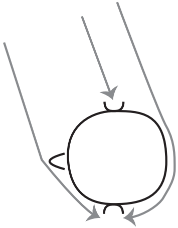


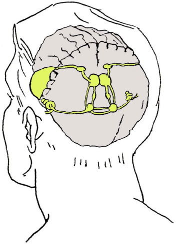
©J. Blauert

- Spatial perception is a result of signal analysis in the brains

Human spatial hearing

Perception



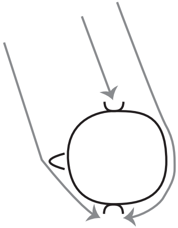


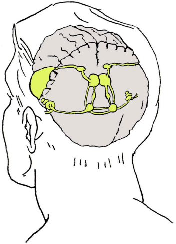
©J. Blauert

- Spatial perception is a result of signal analysis in the brains
- Signal characteristics in one ear / Signal differences between two ears

Human spatial hearing

Perception



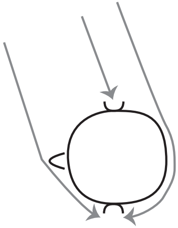


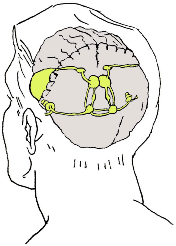
©J. Blauert

- Spatial perception is a result of signal analysis in the brains
- Signal characteristics in one ear / Signal differences between two ears
- Hearing mechanisms estimate the most probable direction for sound

Human spatial hearing

Perception





©J. Blauert

- Spatial perception is a result of signal analysis in the brains
- Signal characteristics in one ear / Signal differences between two ears
- Hearing mechanisms estimate the most probable direction for sound
- Hearing can be fooled easily by audio techniques!

Eye vs ear

Perception

- Eye: lens projects light to retina. Retina has a great number of light-sensitive cells. Retina cells are thus per se sensitive to direction.
- Eye: very high spatial accuracy, limited range of wavelengths (400 - 800nm)

Eye vs ear

- Eye: lens projects light to retina. Retina has a great number of light-sensitive cells. Retina cells are thus per se sensitive to direction.
- Eye: very high spatial accuracy, limited range of wavelengths (400 - 800nm)
- Ear: cochlea sensitive to vast range of wavelengths (2cm - 20m)
- Ear: cochlea has no direct sensitivity to direction of sound

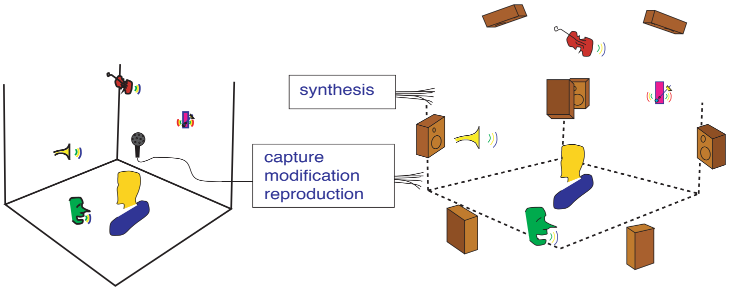
Eye vs ear

- Eye: lens projects light to retina. Retina has a great number of light-sensitive cells. Retina cells are thus per se sensitive to direction.
- Eye: very high spatial accuracy, limited range of wavelengths (400 - 800nm)
- Ear: cochlea sensitive to vast range of wavelengths (2cm - 20m)
- Ear: cochlea has no direct sensitivity to direction of sound
- Different principles of operation $\rightarrow$ audio and video technologies do not share much

2

Spatial sound applications

Spatial audio



- Relay the perception
- Synthesize a desired perception

Spatial sound, tasks for R/D engineers

- Deliver audio with desired spatial characteristics
  - Capture, microphone array processing
  - Reproduction over loudspeakers or headphones
  - Synthesis of spatial characteristics
  - How to store and transmit spatial audio

Spatial sound, tasks for R/D engineers

- Deliver audio with desired spatial characteristics
  - Capture, microphone array processing
  - Reproduction over loudspeakers or headphones
  - Synthesis of spatial characteristics
  - How to store and transmit spatial audio
- Evaluate the quality
  - Subjective testing
  - Instrumental quality evaluation based on computational models of listeners
  - Simplified evaluation

Headphone listening

Applications

- One signal for each ear
- Any audio content, telecom

Headphone listening

- One signal for each ear
- Any audio content, telecom
- All static auditory cues can be controlled, but often results in inside-head perception of sound

Headphone listening

- One signal for each ear
- Any audio content, telecom
- All static auditory cues can be controlled, but often results in inside-head perception of sound
- Tasks
  - Record existing sound scape with any mic system, and play it back while preserving spatial cues
  - Synthesize a sound source in desired direction (HRTF processing)

Headphones with head tracking

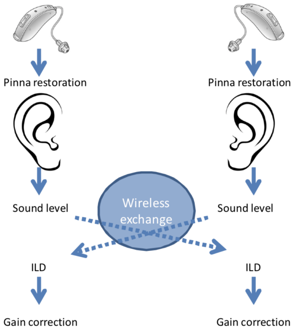
- VR/AR environments, gaming, professional auditory displays VR storytelling (?)
- Binaural rendering is updated with head orientation: externalized sound scene

Headphones with head tracking

- VR/AR environments, gaming, professional auditory displays VR storytelling (?)
- Binaural rendering is updated with head orientation: externalized sound scene
- Tasks
  - Measure head orientation fast and accurately enough
  - Implement the effect of orientation change in audio, fast and accurate enough
  - Evaluation of quality is problematic when the listener moves their head

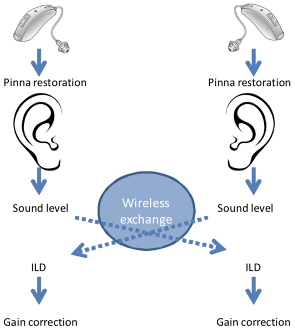
Hearing aids

Applications



- Hear-through devices: microphone + processing + playback

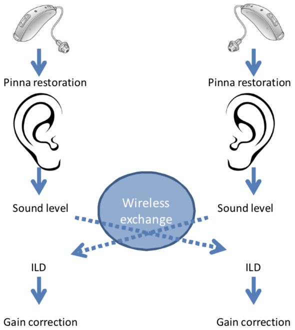
Hearing aids



- Hear-through devices: microphone + processing + playback
- Processing and non-ideal sound capture often create wrong location of sound, also internalized inside listener's head

Hearing aids

Applications



- Hear-through devices: microphone + processing + playback
- Processing and non-ideal sound capture often create wrong location of sound, also internalized inside listener's head
- Lots of R/D interest in industry

Acoustic communication through protective gear

Applications

- Protect the worker from surrounding loud sounds
- They should still
  - hear what is going on in surroundings
  - localize the sources

Audio content production

Applications

- Music / speech / TV / movies / radio

Audio content production

Applications

- Music / speech / TV / movies / radio
- Domestic listening / TV / Car audio / public venues

Audio content production

- Music / speech / TV / movies / radio
- Domestic listening / TV / Car audio / public venues
- Mono / stereo / surround

Audio content production

- Music / speech / TV / movies / radio
- Domestic listening / TV / Car audio / public venues
- Mono / stereo / surround
- In some cases it is important to reproduce sound spatially correct

Audio content production

- Music / speech / TV / movies / radio
- Domestic listening / TV / Car audio / public venues
- Mono / stereo / surround
- In some cases it is important to reproduce sound spatially correct
- Often the directions of virtual sources are not very important, e.g. in music
- However, more loudspeakers will make also the timbral quality better

Cinema sound

Applications

- Synchronized presentation of moving 2D picture and spatial audio

Cinema sound

- Synchronized presentation of moving 2D picture and spatial audio
- Dilemma
  - picture is only in front, and audio covers all directions
  - presenting localized sounds in back can cause viewers to look away from the picture
- Back directions are used for ambient sounds, not grasping the attention

Cinema sound

- Synchronized presentation of moving 2D picture and spatial audio
- Dilemma
  - picture is only in front, and audio covers all directions
  - presenting localized sounds in back can cause viewers to look away from the picture
- Back directions are used for ambient sounds, not grasping the attention
- "VR cinema" (?) would need all directions, though.

User interfaces

Applications

- Aviation
  - Pilots communicate with tower and other airplanes, hear the voices from directions matching with actual directions
  - Warning sounds from the direction of the hazard

User interfaces

Applications

- Aviation
  - Pilots communicate with tower and other airplanes, hear the voices from directions matching with actual directions
  - Warning sounds from the direction of the hazard
- Automotive audio: warning sounds, device usage sounds
- physical sound events / produced with audio system

User interfaces

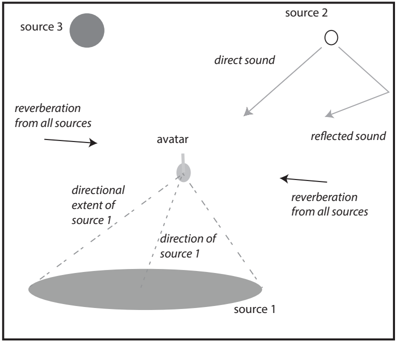
Applications

- Aviation
  - Pilots communicate with tower and other airplanes, hear the voices from directions matching with actual directions
  - Warning sounds from the direction of the hazard
- Automotive audio: warning sounds, device usage sounds
- physical sound events / produced with audio system
- Computer earcons (ear + icon) etc

Virtual reality and Game audio engines

Applications

- Perception of physical presence in locations elsewhere in the real world or in imaginary worlds is created for a subject



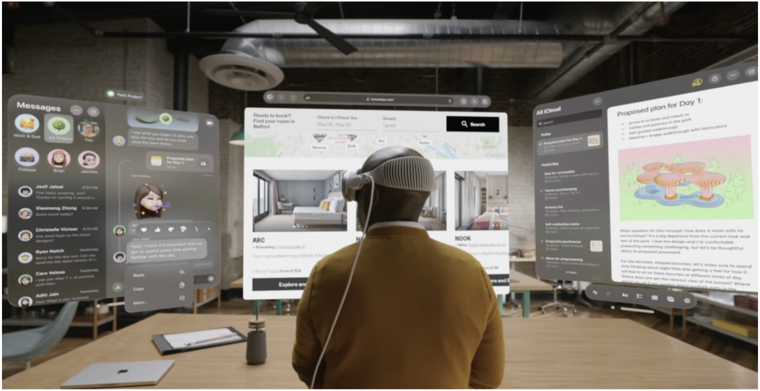
Virtual reality and Game audio engines

Applications

Tasks for audio engine

- Reproduction or synthesis of source signals (from memory or by models of physics)
- Synthesis of source directivity
- Simulation of the direct sound path
- Virtual source positioning
- Spatial extent of virtual sources
- Room effect simulation
- Distance rendering

Augmented reality



- Its like VR, but with hear-through

Augmented reality

Applications

- Somehow the listener should be able to perceive the acoustic environment:

Augmented reality

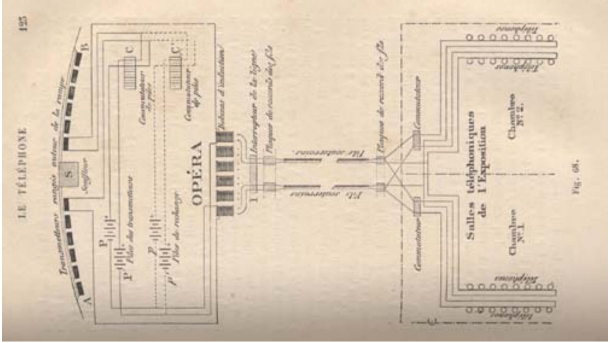
- Somehow the listener should be able to perceive the acoustic environment:
  - headphones are "open", letting the sound through as well as possible, as in the fig
  - headphones are closed, blocking the incoming sound. 2 or more mics are used to capture the sound and bring it to the ears.
- Both approaches have weaknesses, good engineering is needed.

3
Sound reproduction

Telephone

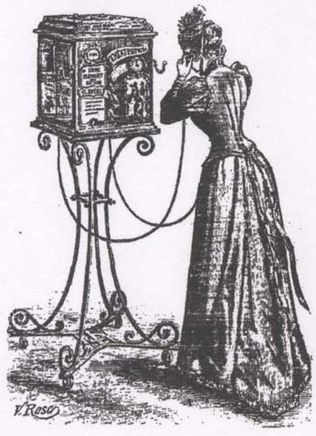
- Monophonic transmission
- Monaural or monophonic listening
  - Sound localized to the position of the loudspeakers
  - Makes sense, a talker is positioned in one spot

First binaural transmission Paris 1881



- Transmission of two telephone lines to listener
- Listen to live concerts in venues with binaural audio

Theatrophone



Théâtrophone.

- Firm active 1890 - 1930
- Listening devices available in several venues, also in homes

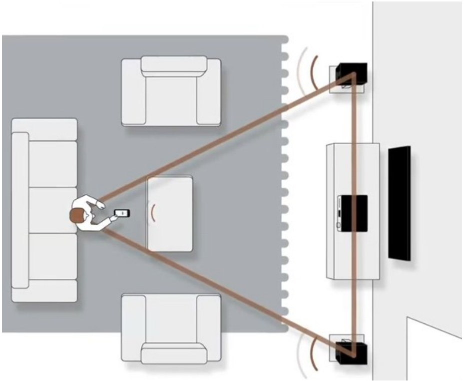
Aircraft detection before radar

- Binaural hearing was cost-efficient and accurate in detecting of direction of incoming aircraft

Gramophone, Mono LP, Mono radio transmission

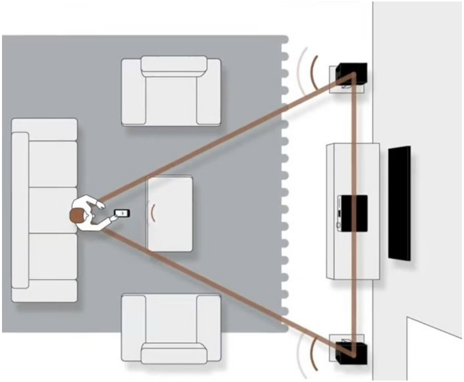
- Single sound channel = monophonic
- Gramophone first versions in 1880's, until about 1960's
- Radio transmissions started in 1920's

Stereophony



- Alan Blumlein pioneering work in 1930’s

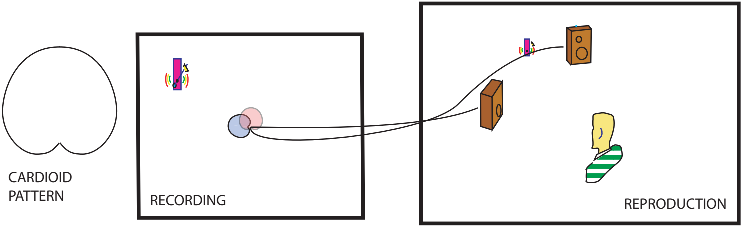
Stereophony



- Alan Blumlein pioneering work in 1930's
- 1960's: Two-channel audio in single vinyl LP groove ($\pm 45°$)
- Separate channel for each loudspeaker

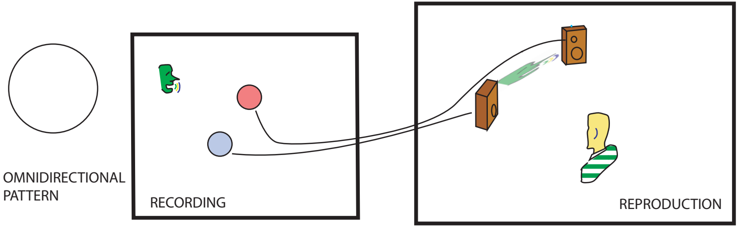
Coincident techniques for stereophony

- Two directive microphones in coincident positioning
- XY (cardioids or similar), Blumlein (Dipoles)
- Virtual sources relatively point-like
- May suppress reverberation



Spaced techniques for stereophony

- Two directive or omnidirectional microphones spaced by 20cm – few meters
- AB technique
- Virtual sources relatively broad, and localization depends on frequency
- Reverberation perceived "airy", "open", not suppressed



Mono → Stereophony: why?

- Major change in audio industry. People had to buy new audio gear. Why was stereo successful?

Mono $\rightarrow$ Stereophony: why?

- Major change in audio industry. People had to buy new audio gear. Why was stereo successful?
- Broader spatial image.

Mono → Stereophony: why?

- Major change in audio industry. People had to buy new audio gear. Why was stereo successful?
- Broader spatial image.
- What about small stereo devices, ghetto blasters?

Mono $\rightarrow$ Stereophony: why?

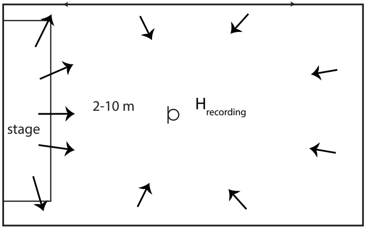
- Major change in audio industry. People had to buy new audio gear. Why was stereo successful?
- Broader spatial image.
- What about small stereo devices, ghetto blasters?
- Why does stereo sound better although the loudspeakers are very close?

Mono → Stereophony: why?

- Major change in audio industry. People had to buy new audio gear. Why was stereo successful?
- Broader spatial image.
- What about small stereo devices, ghetto blasters?
- Why does stereo sound better although the loudspeakers are very close?
- Something in sound color is "more open", "less colored", demo

Monophonic recording of reverberant sound

- Single far-away microphone
  - Captures all sources present
  - Recording room response $H_{\mathrm{recording}}(t)$ from source to microphone
  - Listening room response for ears $H_{\mathrm{listeningL}}(t)$ and $H_{\mathrm{listeningR}}(t)$ (binaural room impulse response)



Monophonic recording

Microphone signal:

$$y(t) = H_{\mathrm{recording}}(t) * x(t)$$

Monophonic recording

Microphone signal:

$$y(t) = H_{\mathrm{recording}}(t) * x(t)$$

Ear canal signals:

$$z_{\mathrm{L}}(t) = H_{\mathrm{listeningL}}(t) * y(t) = H_{\mathrm{listeningL}}(t) * H_{\mathrm{recording}} * x(t)$$
$$z_{\mathrm{R}}(t) = H_{\mathrm{listeningR}}(t) * y(t) = H_{\mathrm{listeningR}}(t) * H_{\mathrm{recording}} * x(t)$$

Monophonic recording

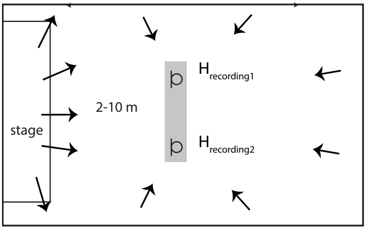
Microphone signal:
$y(t) = H_{\mathrm{recording}}(t) * x(t)$

Ear canal signals:
$z_{\mathrm{L}}(t) = H_{\mathrm{listeningL}}(t) * y(t) = H_{\mathrm{listeningL}}(t) * H_{\mathrm{recording}} * x(t)$
$z_{\mathrm{R}}(t) = H_{\mathrm{listeningR}}(t) * y(t) = H_{\mathrm{listeningR}}(t) * H_{\mathrm{recording}} * x(t)$

Both ear canal signals are filtered by $H_{\mathrm{recording}}$: this is not natural condition, room response is always different to the ears in reality
Sound will be a bit "muffled"

Why stereo recordings produce better timbral quality

- Two microphones in recording room with responses
  - $H_{\text{recording1}}(t)$ and $H_{\text{recording2}}(t)$
- Two loudspeakers and two ears in listening room $\rightarrow$ four responses
  - Loudspeaker 1 to left ear $H_{1\text{listeningL}}(t)$
  - Loudspeaker 2 to left ear $H_{2\text{listeningL}}(t)$
  - Loudspeaker 1 to right ear $H_{1\text{listeningR}}(t)$
  - Loudspeaker 2 to right ear $H_{2\text{listeningR}}(t)$



Why stereo recordings produce better timbral quality

Two microphone signals in recording room:

$y_1(t) = H_{\mathrm{recording1}}(t) * x(t)$

$y_2(t) = H_{\mathrm{recording2}}(t) * x(t)$

Why stereo recordings produce better timbral quality

Two microphone signals in recording room:

$$y_1(t) = H_{\mathrm{recording1}}(t) * x(t)$$
$$y_2(t) = H_{\mathrm{recording2}}(t) * x(t)$$

Ear canal signals in listening room:

$$z_{\mathrm{L}}(t) = H_{\mathrm{1listeningL}}(t) * y_1(t) + H_{\mathrm{2listeningL}}(t) * y_2(t)$$
$$z_{\mathrm{L}}(t) = H_{\mathrm{1listeningR}}(t) * y_1(t) + H_{\mathrm{2listeningR}}(t) * y_1(t)$$

Why stereo recordings produce better timbral quality

Two microphone signals in recording room:

$y_1(t) = H_{\mathrm{recording1}}(t) * x(t)$
$y_2(t) = H_{\mathrm{recording2}}(t) * x(t)$

Ear canal signals in listening room:

$z_{\mathrm{L}}(t) = H_{\mathrm{1listeningL}}(t) * y_1(t) + H_{\mathrm{2listeningL}}(t) * y_2(t)$
$z_{\mathrm{L}}(t) = H_{\mathrm{1listeningR}}(t) * y_1(t) + H_{\mathrm{2listeningR}}(t) * y_1(t)$

Ear canal signals do not share the same frequency response, flatter tot recording room sound is less prominent

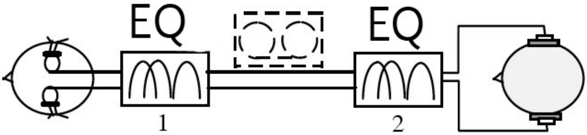
Binaural techniques

- Ear canal signals are the main input to hearing
- Why not replicate only them?
- Recording/reproduction/synthesis of ear canal signals
- Challenges: dynamic cues (head movements), tactile perception

Binaural recording, headphone playback

- careful microphone and headphone equalization
- binaural cues and auditory spectrum reproduced as were in recording
- in some cases this is appealing solution

Applications: personalized recording, academic use, noise measurements, augmented reality audio

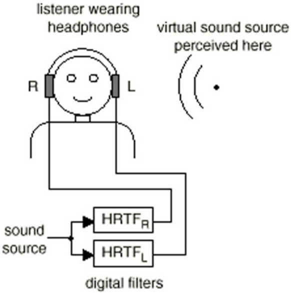


Binaural recording

Challenges

- headphone equalization is problematic
- listener head movements does change binaural cues inside-head localization
- front-to-back confusions
- vision conflicts with audition
- works best only with recordings made with your own head

Binaural synthesis, headphones

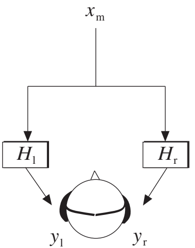


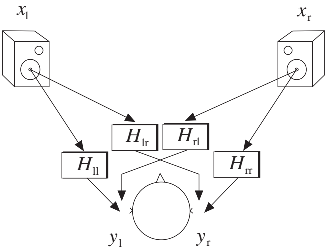
©Bill Gardner

- convolve monophonic sound tracks with measured [individual] HRTFs
- auditory objects can be positioned in 3D virtual space
- inside-head localization, front-back confusions
- need of individual HRTFs
- head tracking may be used to resolve this
- virtual reality, gaming, aviation
- playback of surround audio content over multiple virtual loudspeakers

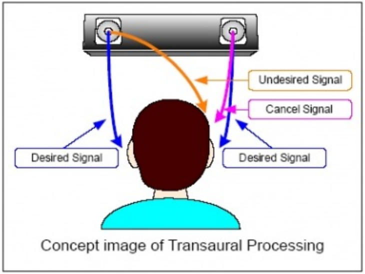
Binaural recording, loudspeaker playback

- Left loudspeaker sound signal reaches also right ear, and vice versa
- "Cross-talk" is a problem
- Could cross-talk be avoided?





Binaural recording, cross-talk cancelled playback



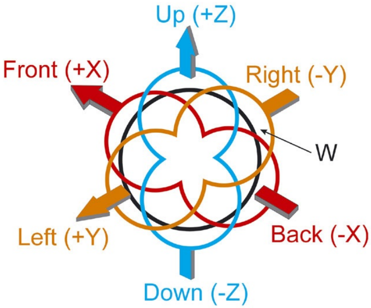
©JVC

- head has to be placed with about 1cm accuracy
- reflections should not exist
- applicable in some special cases
- back-to-front confusions

B-format recording

- B-format microphones
- Omni + 3 dipoles on Cartesian axis
- Steerable first-order microphone
- Cardioid or hypercardioid for each loudspeaker

B-format recording



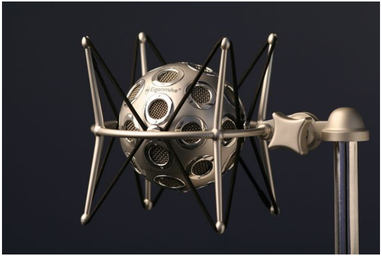
www.soundfield.com

First-order Ambisonics

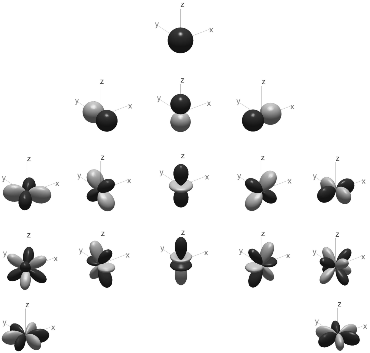
[Gerzon 70's]

- A signal for each loudspeaker is decoded from B-format
- Loudspeaker channels are relatively coherent
- Coloring
- OK quality in best listening position, and in good listening room
- Nearmost loudspeaker dominates outside best listening position

Higher-order microphones



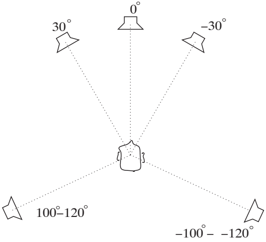
http://www.mhacoustics.com

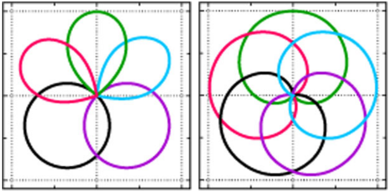


Higher-order microphones

- Requires tens of microphones
- Decoded signals have desired directional patterns only at a limited frequency band
- Serious noise problems at low frequencies in decoded spherical harmonics
- Serious problems at frequencies above spatial aliasing frequency

Microphone techniques for multichannel

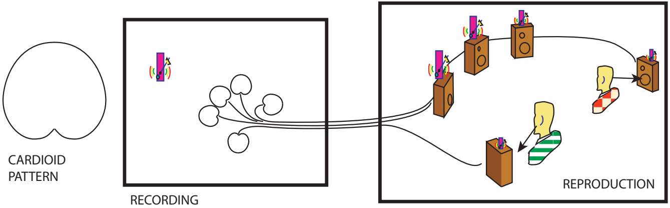




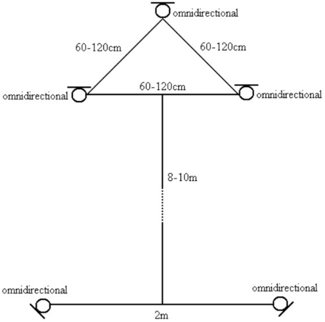
- Center: Ideal microphone patterns for 5.1 loudspeaker setup
- Right: First-order directional patterns
- Too broad patterns cause loudspeaker signals to be coherent
- Comb-filter effects, "muffled" sound, stereo image blurred

Spaced microphone techniques for multichannel

- A set of [usually first-order] directive microphones in some layout
- Large enough spacing to avoid too high coherence btw loudspeaker channels
- Directional patterns provide some kind of reproduction of source directions
- Trade-offs, no generic solution

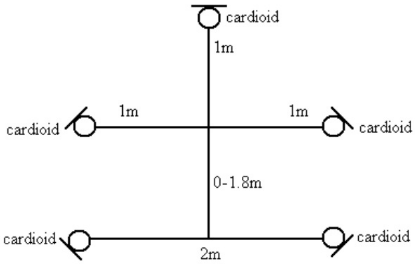


Spaced microphone arrays for multichannel



Decca tree

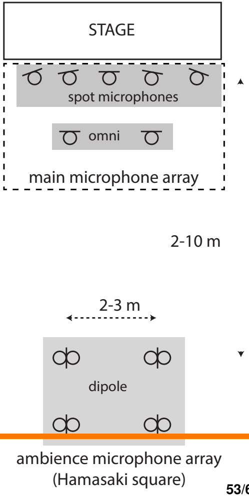
Spaced microphone arrays for multichannel



Fukada tree

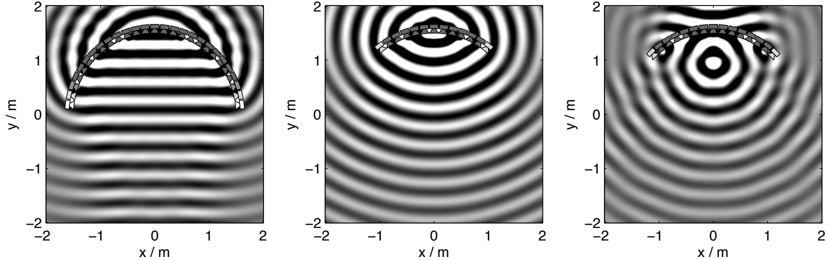
Spot microphone recording

- Multiple sources, e.g., an orchestra on stage
- A "spot" microphone near each source, optimally capturing only single source signal
- Spot microphones are mixed together
- Often far-away "ambience" signals are also recorded with far-away microphones, and mixed with spot microphone signals



Wave field synthesis

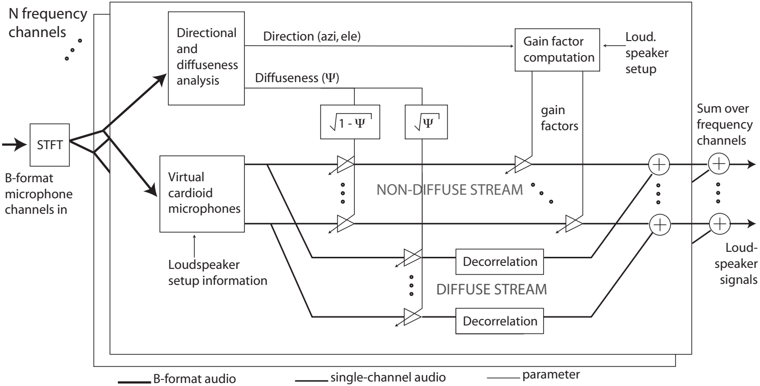
- Try to control the complete wave field
- Helmholtz-Kirchhoff integral
- Can position virtual sources also closer than the loudspeakers are



Wave field synthesis

- Hundreds of loudspeakers needed for 2D loudspeaker setups
- Hundreds of thousands of loudspeakers would be needed for 3D setups
- Not practical as recording technique, possible as virtual source positioning technique
- Spatial aliasing occurs typically near 1kHz, depending on spacing between loudspeakers
- Applications: large venues and installations
- Sound field control, silent and loud zones, noise suppression

Parametric time-frequency-domain reproduction



Adaptive use of assumptions of spatial hearing to mitigate shortcomings in Ambisonics

Spatial audio formats 1

- Channel-based (5.1, 7.1, 7.2.4, 22.2 etc)
  - Each loudspeaker channel transmitted as separate audio signal
  - Lossy coding of each channel separately
  - Bit saving not necessarily huge
  - Potential defects in spatial properties
  - Not easy to transcode btw loudspeaker layouts, or to binaural listening

Spatial audio formats 1

- Channel-based (5.1, 7.1, 7.2.4, 22.2 etc)
  - Each loudspeaker channel transmitted as separate audio signal
  - Lossy coding of each channel separately
  - Bit saving not necessarily huge
  - Potential defects in spatial properties
  - Not easy to transcode btw loudspeaker layouts, or to binaural listening
- Object-based
  - Each audio event as separate audio signal
  - Spatial information (direction, distance, reverberation, etc) as metadata
  - Leads to very large bit streams
  - Dolby Atmos, DTS-X

Spatial audio formats 2

- Spherical harmonics -based (Ambisonics)
  - The sound stream consists of signals whose directional patterns follow spherical harmonics
  - A decent sound quality for 3D obtained with abt 5th order, 36 channels
  - Individual coding of spherical harmonic components will cause spatial artifacts
  - Not straightforward to record higher-order Ambisonics
  - Quite high demand of bit rate

Spatial audio formats 2

- Spherical harmonics -based (Ambisonics)
  - The sound stream consists of signals whose directional patterns follow spherical harmonics
  - A decent sound quality for 3D obtained with abt 5th order, 36 channels
  - Individual coding of spherical harmonic components will cause spatial artifacts
  - Not straightforward to record higher-order Ambisonics
  - Quite high demand of bit rate
- "All-inclusive" codecs
  - Objects, channels, spherical harmonics
  - Different lossy coding strategies
  - MPEG-H, 3GPP IVAS

Spatial audio formats 3

- Time-frequency-domain coding of channel-based spatial audio formats
  - MPEG Surround
  - A stereo mixdown is sent
  - Instantaneous frequency-specific differences btw channels in level and time are the metadata
  - Huge savings in e.g. coding of 5.1 audio format

Spatial audio formats 3

- Time-frequency-domain coding of channel-based spatial audio formats
  - MPEG Surround
  - A stereo mixdown is sent
  - Instantaneous frequency-specific differences btw channels in level and time are the metadata
  - Huge savings in e.g. coding of 5.1 audio format
- Time-frequency-domain coding of Ambisonics
  - DirAC, HO-DirAC
  - Instantaneous spatial parameters in TF-domain (direction, diffuseness) are the metadata
  - N channels with cardioid patterns are sent as audio channels, each of which as individual metadata
  - Already 6 channels + metadata provides very good reproduction of original sound scene