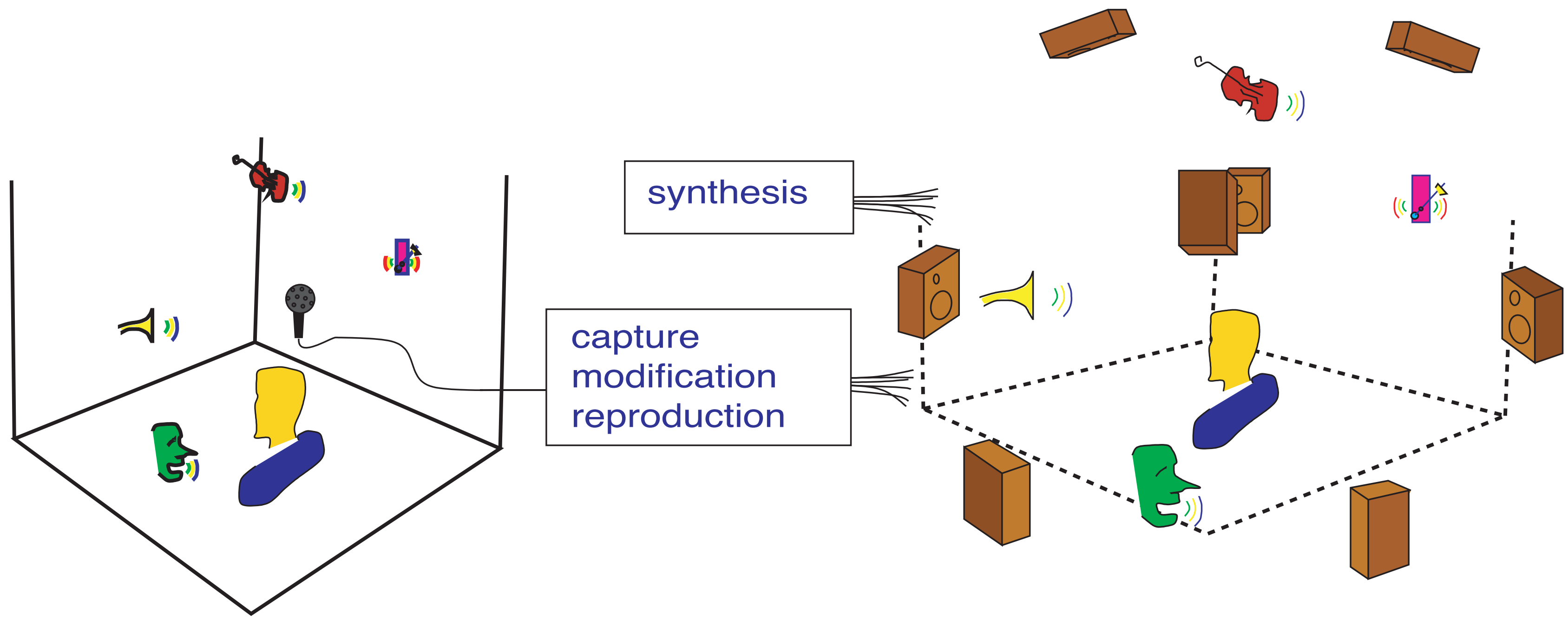
Aalto University
School of Electrical
Engineering

Spatial sound recording and reproduction

Ville Pulkki, Archontis Politis

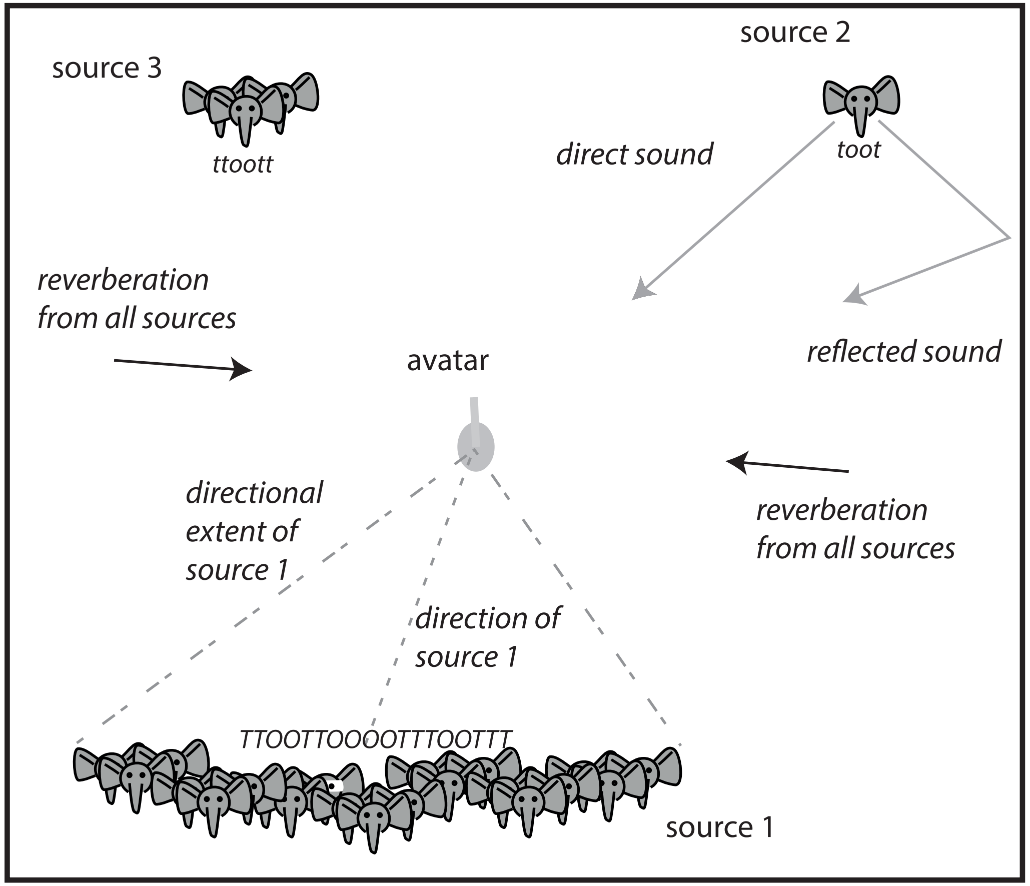
Intro:
Spatial audio technology

Technologies for capturing, transmitting, synthesizing and reproducing spatial properties of a sound scene.



Rendering acoustic virtual reality

A multitude of acoustic paths from each source to the listener



Typical tasks for virtual reality audio engines

Tasks for virtual reality audio engines

- Source directivity modeling
- Direct sound path and distinct reflections modeling
- Render arriving sounds to correct directions, with control of spatial extent
- Generate reverberation and render it to loudspeakers
- Distance rendering

Intro:
Spatial audio technology

Technologies for capturing, transmitting, synthesizing and reproducing spatial properties of a sound scene.

Some use cases:

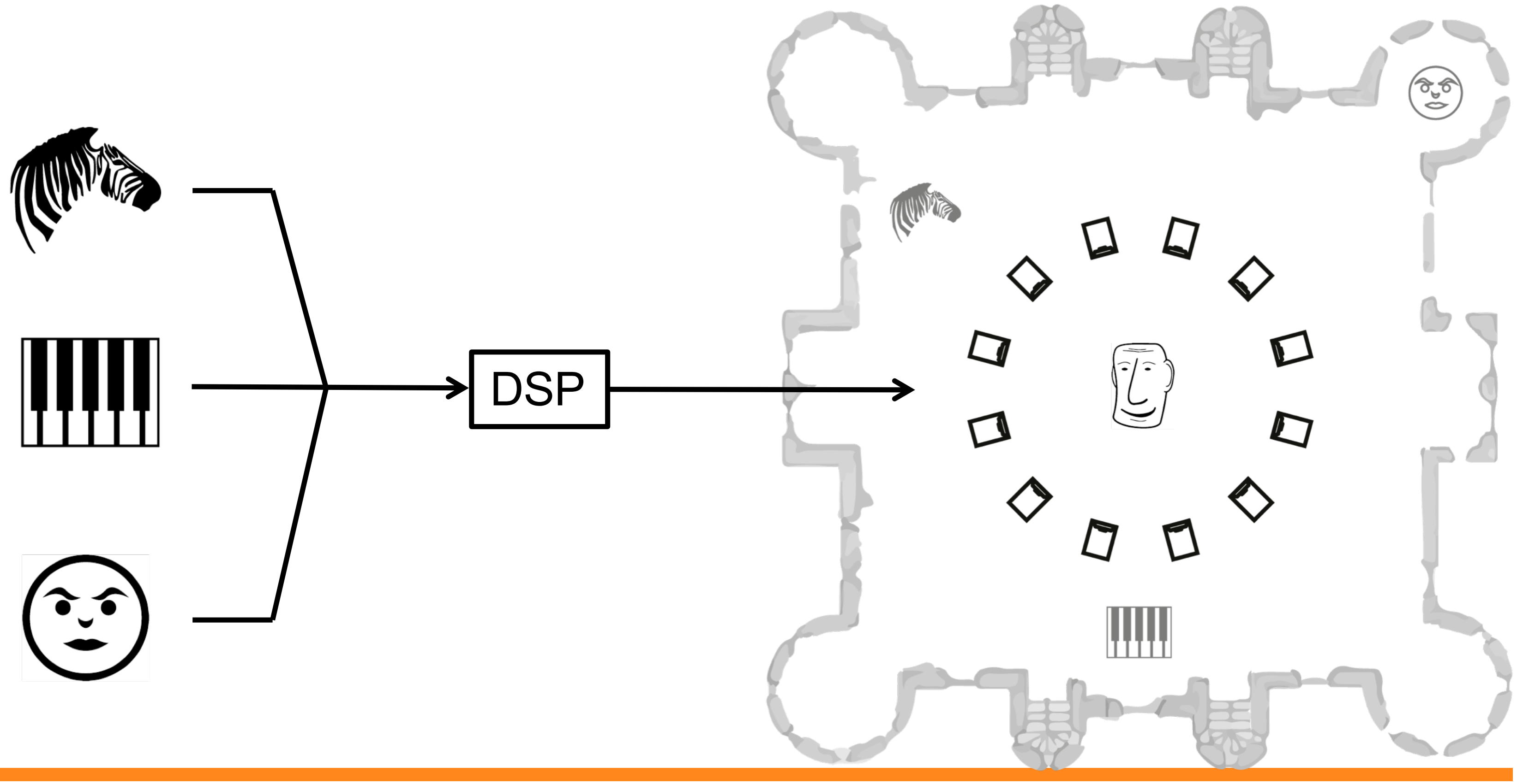
- Immersive audio/music production and reproduction
- Cinema audio
- Virtual and augmented reality
- Telepresence
- Auditory displays

This lecture:
Spatial audio technology

Technologies for capturing, transmitting, synthesizing and reproducing spatial properties of a sound scene.

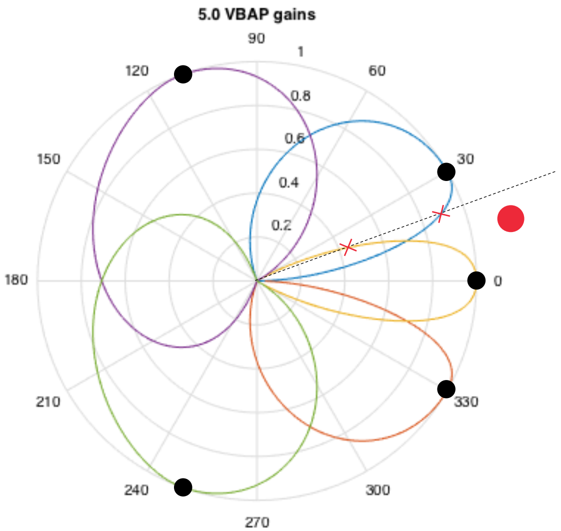
a) Commonly-used tools for synthesis and reproduction of artificial sound scenes
b) Challenges and solutions for recording and reproduction of real sound scenes

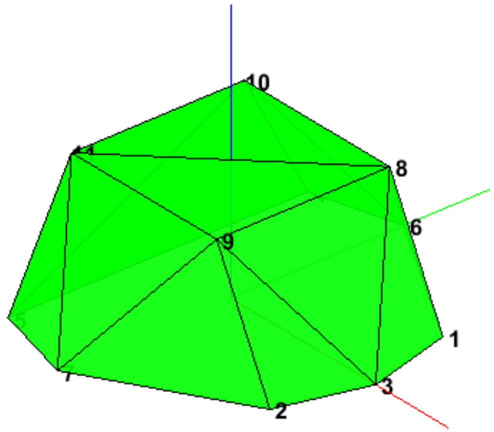
Spatial sound scene synthesis

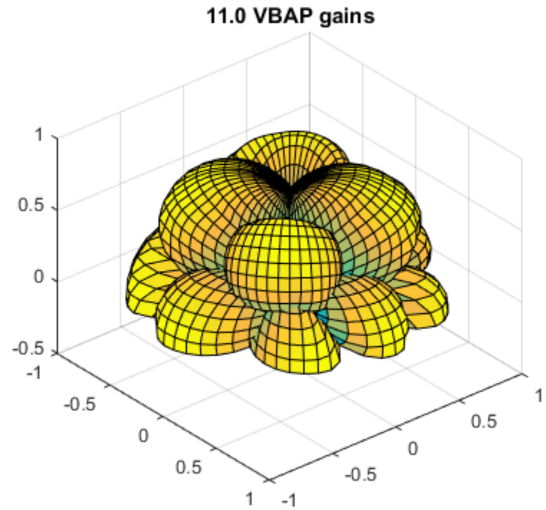


Spatial sound scene synthesis: Render sound to certain direction

Amplitude panning is the standard method for loudspeaker reproduction

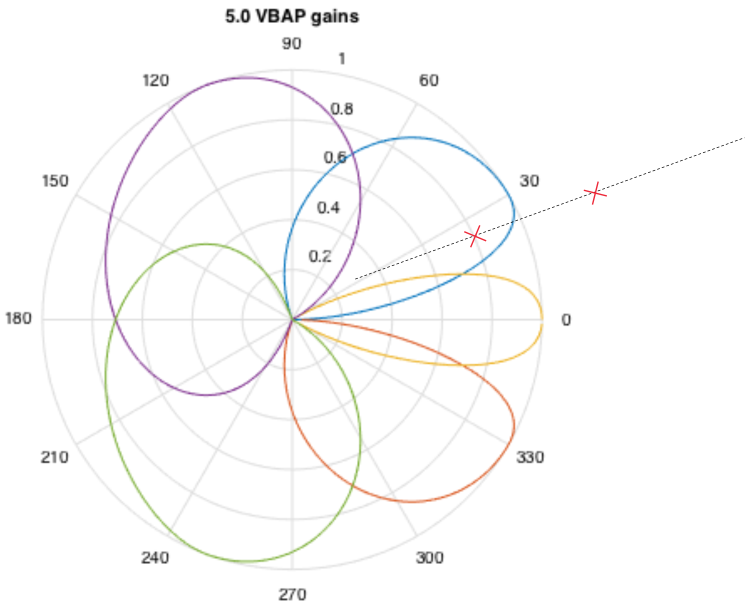






Spatial sound scene synthesis: Render sound to certain direction

Amplitude panning is the standard in loudspeaker reproduction



$$\boxed{\mathbf{y}(\theta, t) = \boldsymbol{g}_L(\theta)\, s(t)}$$

$s(t)$ source signal

$\boldsymbol{g}(\theta) = \begin{bmatrix} g_1(\theta) \\ \dots \\ g_L(\theta) \end{bmatrix}$ panning gains

$\boldsymbol{y}(\theta, t) = \begin{bmatrix} y_1(\theta, t) \\ \dots \\ y_L(\theta, t) \end{bmatrix}$ loudspeaker signals

Aalto University
School of Electrical
Engineering

Spatial sound scene synthesis: Diffusion/decorrelation/spreading

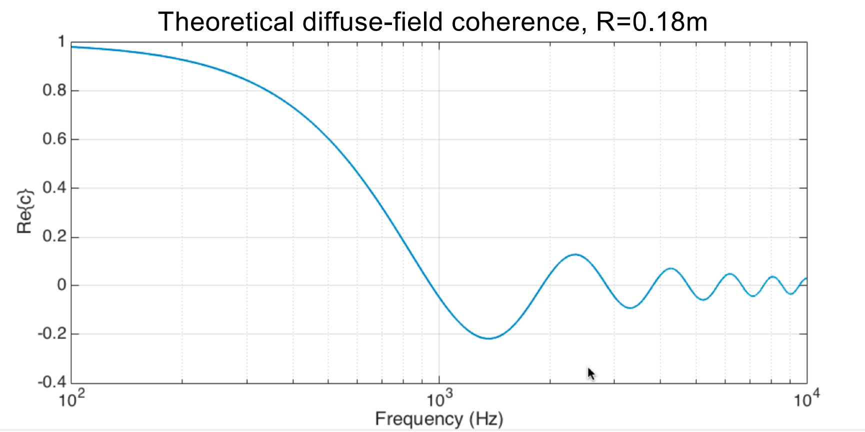
A set of refined audio tools for:

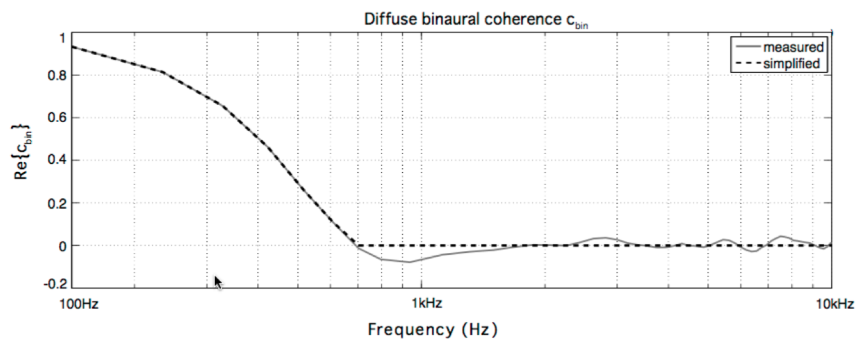
- reverberation and filtering for perceptual distance, room and environmental effects, spatially large sound sources
  - reverberation filters / ASP course!
  - diffusion/decorrelation / today
  - source spreading / today

Spatial sound scene synthesis: Surrounding reverberation effect

A **diffuse** acoustic field contains waves traveling with equal probability from all directions carrying signals that are uncorrelated between them.

Minimum coherence at the ears of the listener!

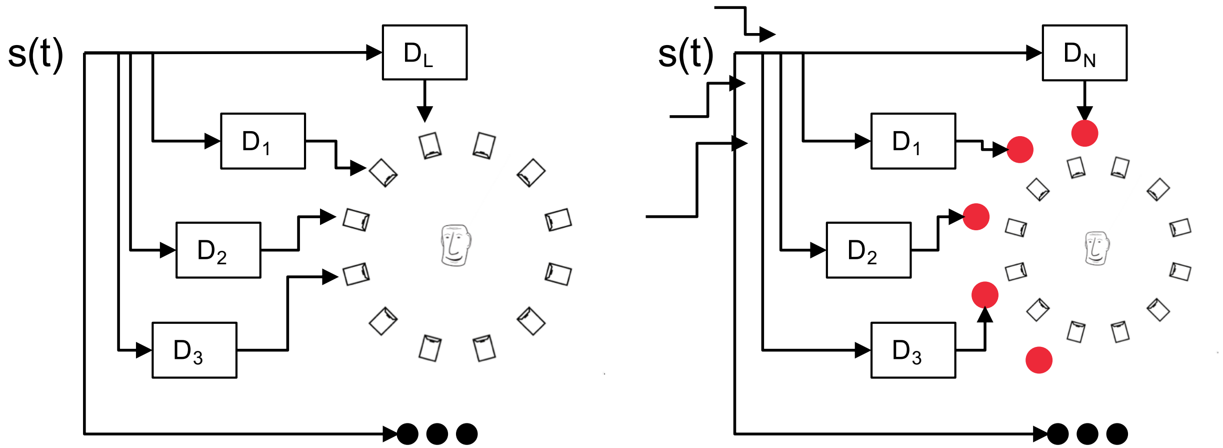




Spatial sound scene synthesis: Surrounding reverberation

In practice:

- Multiple real sources (loudspeakers) or virtual sources serve **decorrelated** versions of the same signal



$D_i[s(t)]$: decorrelation operator

Spatial sound scene synthesis: Diffusion/decorrelation/spreading

In practice:

- Multiple real sources (loudspeakers) or virtual sources serve uncorrelated versions of the same signal
- Some properties of decorrelator operators (or filters)
  - create copies of the input that are as uncorrelated as possible
  - preserve the magnitude response of the input as much as possible
  - affect the temporal structure of the input as little as possible

Spatial sound scene synthesis: Diffusion/decorrelation/spreading

Some ways to do decorrelation:

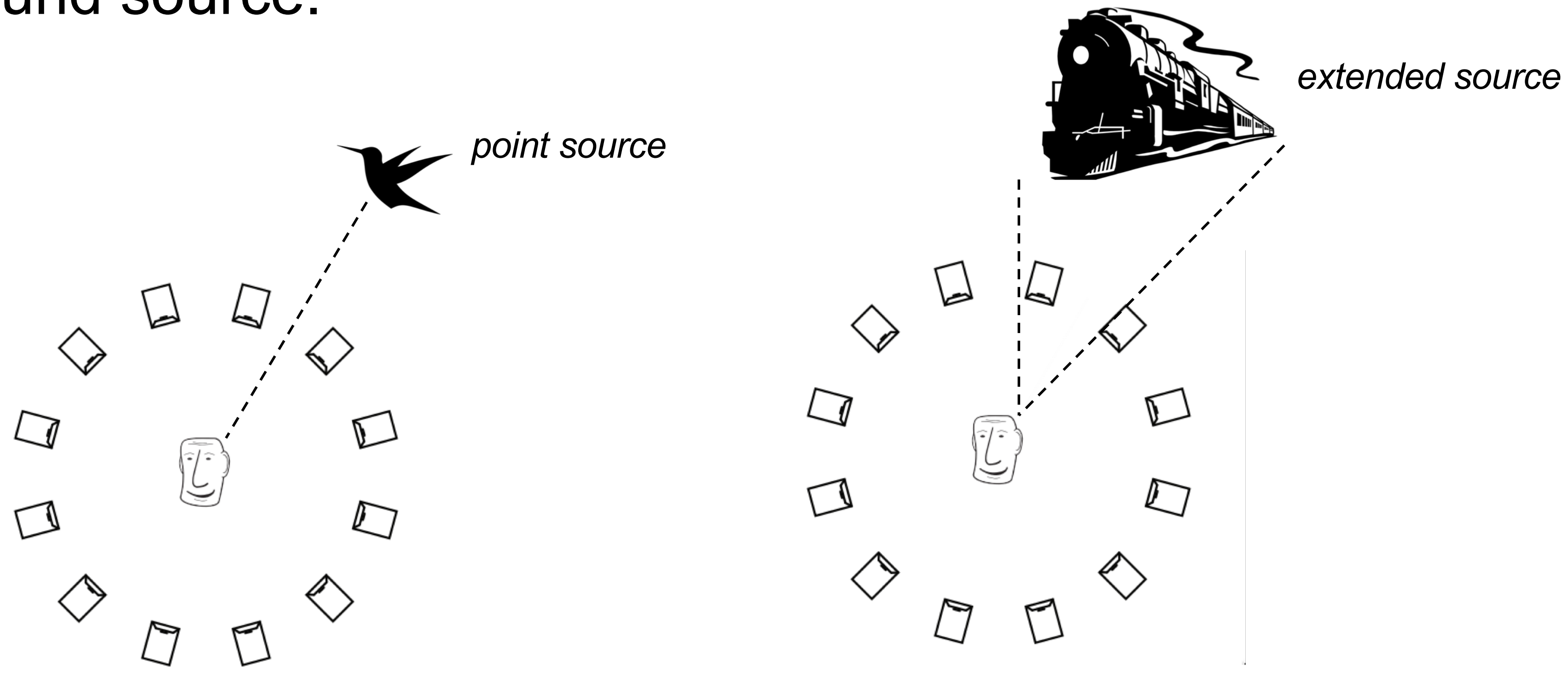
- convolve input with a short noise burst
- randomize the phase response of the input
- apply different delays at different bands of the signal
- pass the signal through multiple cascaded all-pass filters
- time-varying delays
- combinations/variations of the above

Decorrelation problems

- The length of the filter should be minimized to avoid temporal smearing of transients
- The length of the filter should be maximized to obtain an even frequency response
- Trading between spectral and temporal artifacts
- In our work with parametric reproduction, the best way to improve sound quality has been to apply decorrelation as little as possible, but still giving the benefits to audio quality

Spatial sound scene synthesis: Diffusion/decorrelation/spreading

Spreading refers to giving a spatial extent to a virtual sound source.

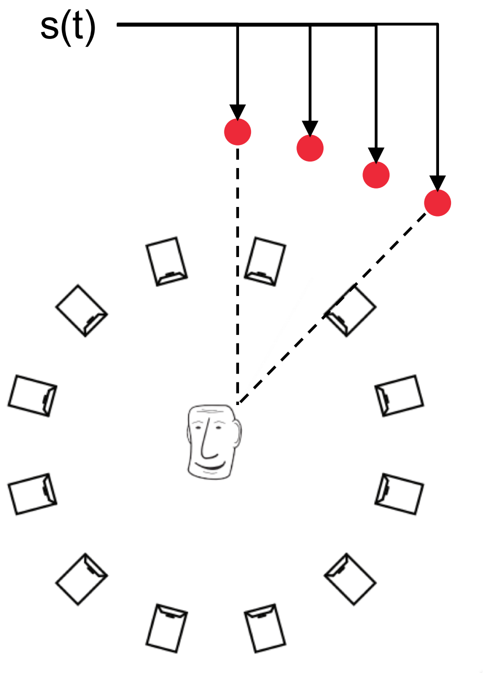


Spatial sound scene synthesis: Diffusion/decorrelation/spreading

Spreading refers to giving a spatial extent to a virtual sound source.

- A naive approach:

  panning

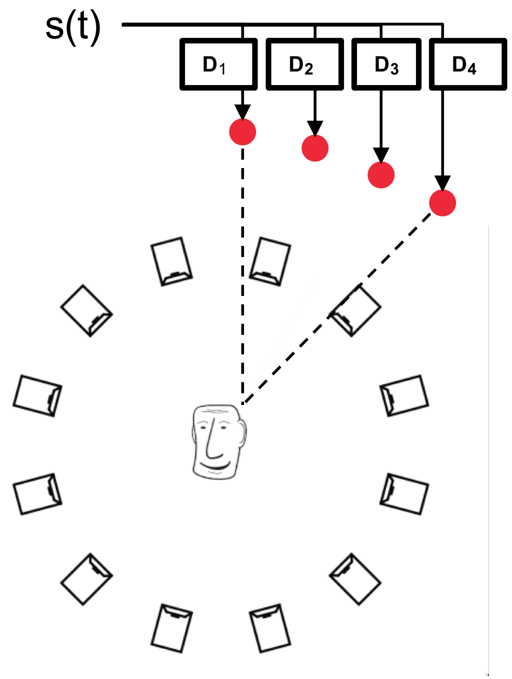


Spatial sound scene synthesis: Diffusion/decorrelation/spreading

Spreading refers to giving a spatial extent to a virtual sound source.

- A better approach:

  decorrelation + panning

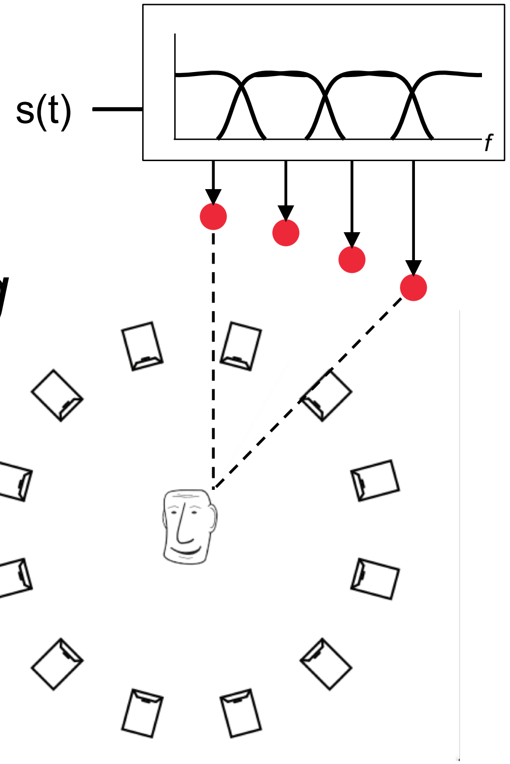


Spatial sound scene synthesis: Diffusion/decorrelation/spreading

Spreading refers to giving a spatial extent to a virtual sound source.

- An alternative approach:

  freq. decomposition + panning

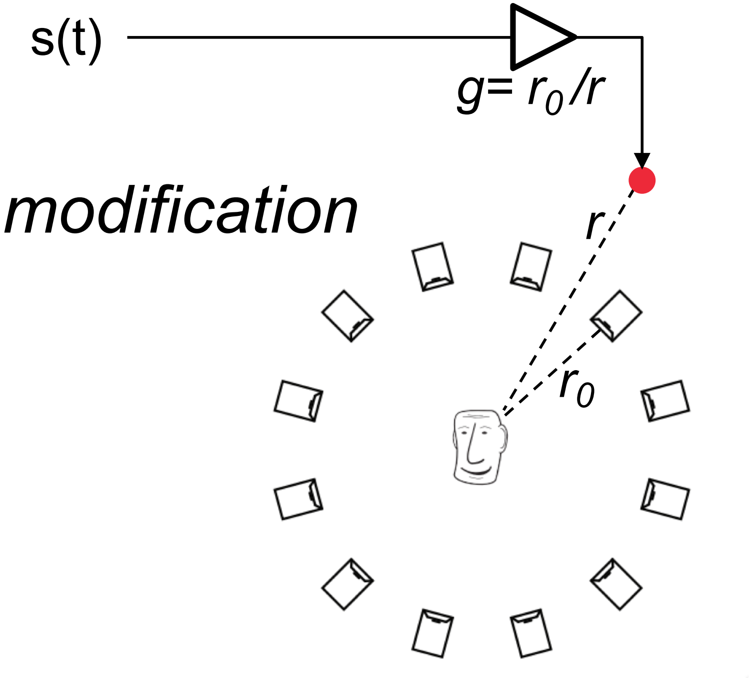


Spatial sound scene synthesis: Diffusion/decorrelation/spreading

Distance perception relies to loudness modifications, spatial extent, and the direct-to-reverberant/diffuse ratio

- A naive approach:

  panning + loudness modification

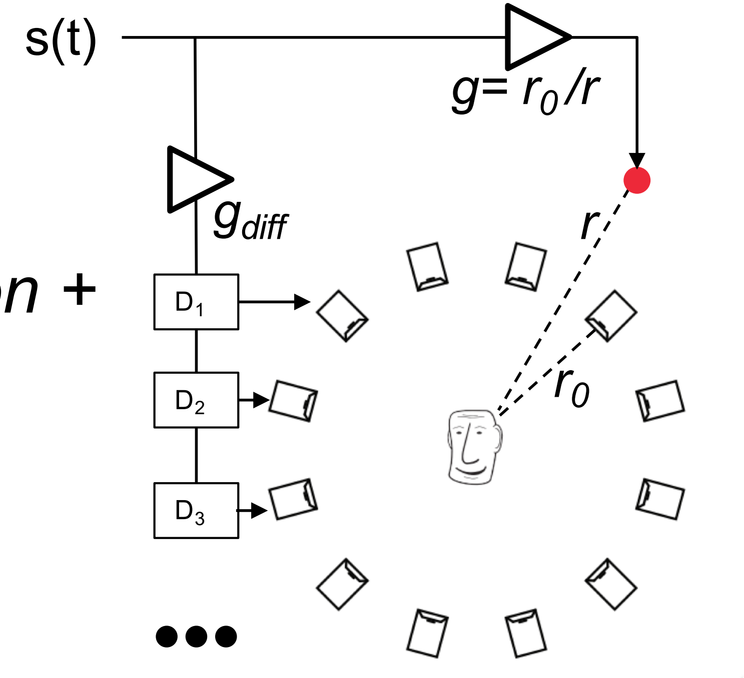


Spatial sound scene synthesis: Diffusion/decorrelation/spreading

Distance perception relies to loudness modifications, spatial extent, and the direct-to-reverberant/diffuse ratio

- A better approach:

  panning +
  loudness modification +
  decorrelation

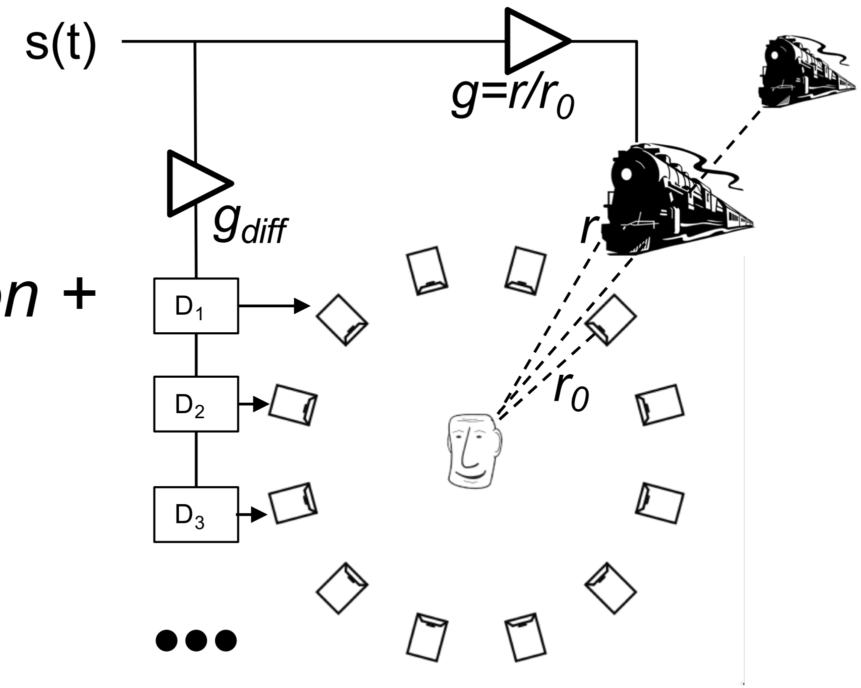


Spatial sound scene synthesis: Diffusion/decorrelation/spreading

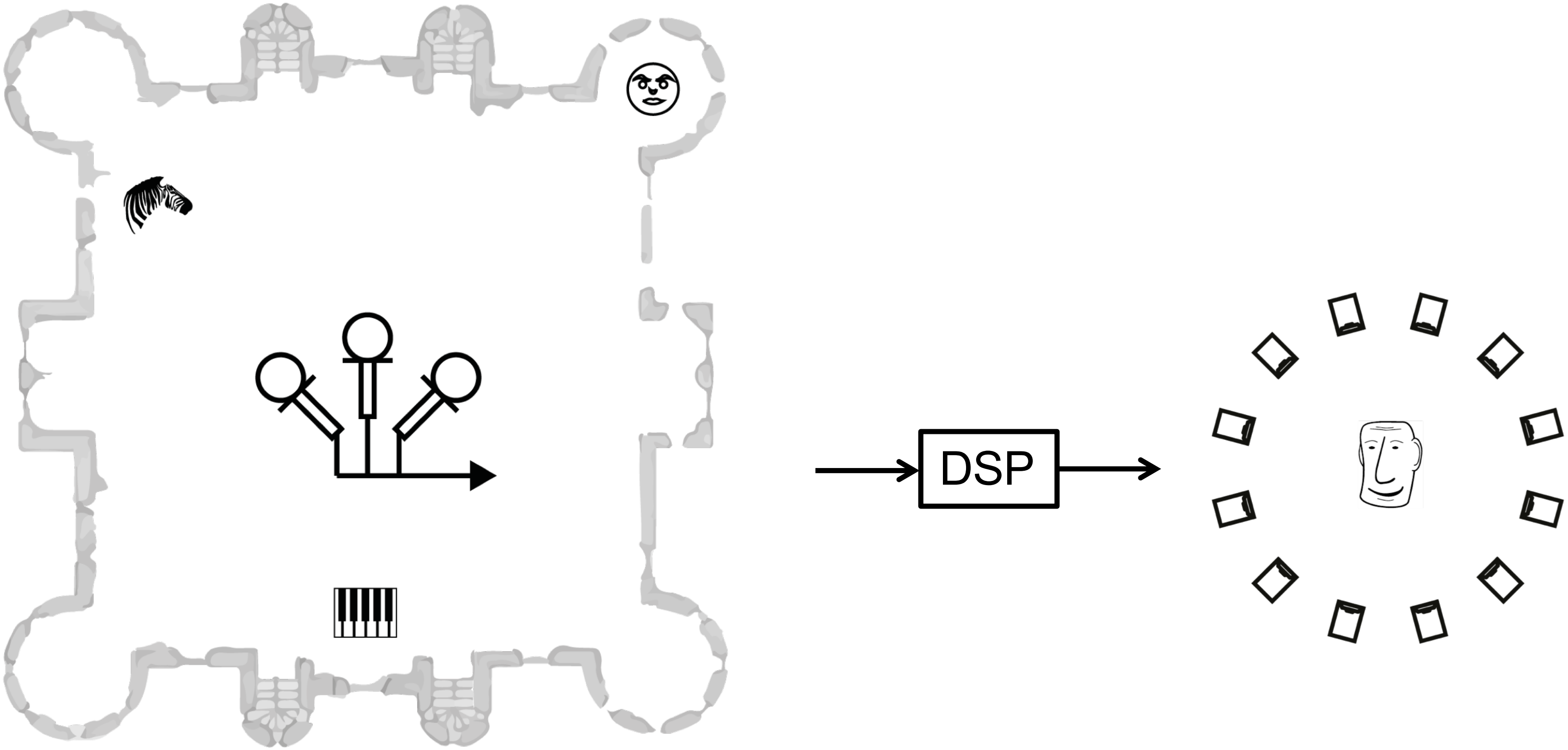
Distance perception relies to loudness modifications, spatial extent, and the direct-to-reverberant/diffuse ratio

- A better approach:

  panning +
  loudness modification +
  decorrelation +
  extent control



Reproduction of real sound scenes



Reproduction of real sound scenes: (Traditional) surround recording

One-to-one channel mapping from recordings to speakers/headphones.

$$\mathbf{y}(t) = \mathbf{I}\,\mathbf{x}(t)$$

$\mathbf{y}$: output signals
$\mathbf{x}$: microphone signals

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Reproduction of real sound scenes: (Traditional) surround recording

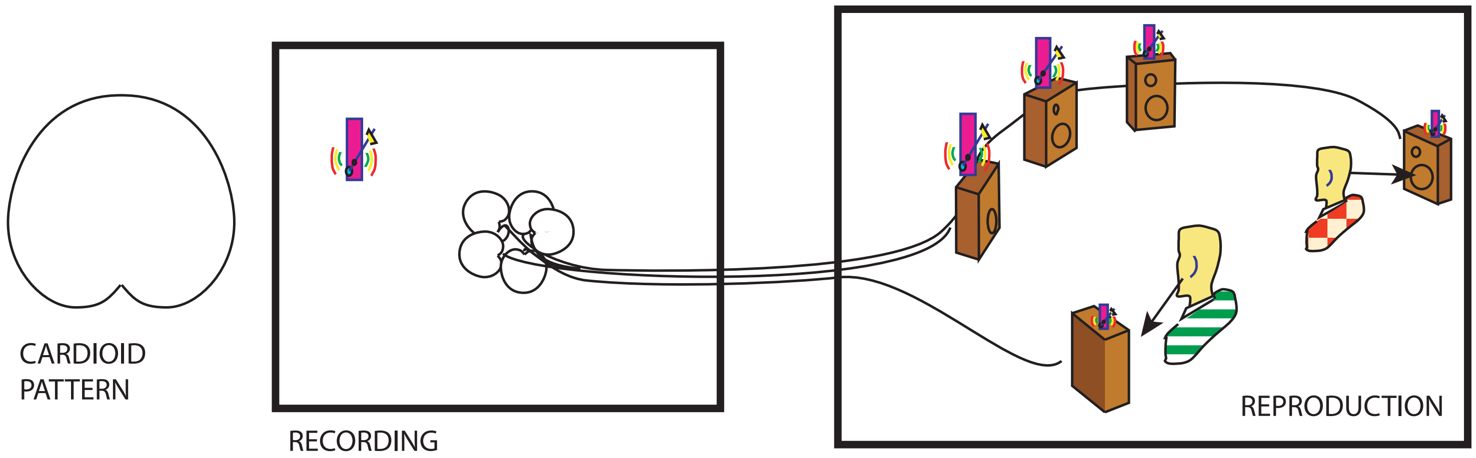
One-to-one channel mapping from recordings to speakers/headphones.

$$\mathbf{y}(t) = \mathbf{I}\,\mathbf{x}(t)$$

y: output signals
x: microphone signals

- The spatial cues delivered by the reproduction system depends solely on the arrangement and properties of the microphone array.

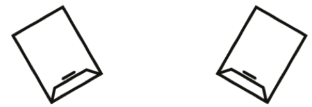
Reproduction of real sound scenes: (Traditional) surround recording



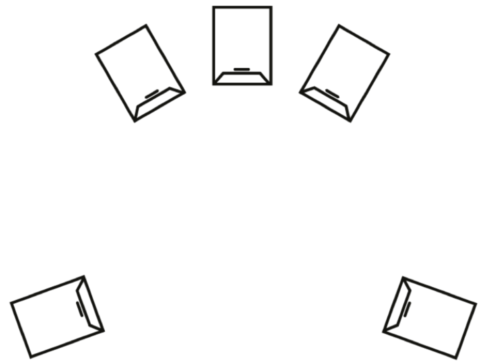
- Loudspeaker signals partially coherent
- Comb filter effects
- Perceived directions are smeared and vague

Reproduction of real sound scenes: (Traditional) surround recording

XY recording setup



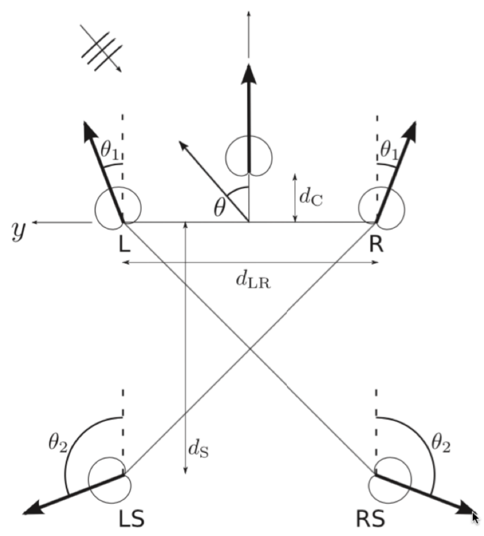
INA-5 recording setup



Reproduction of real sound scenes: (Traditional) surround recording

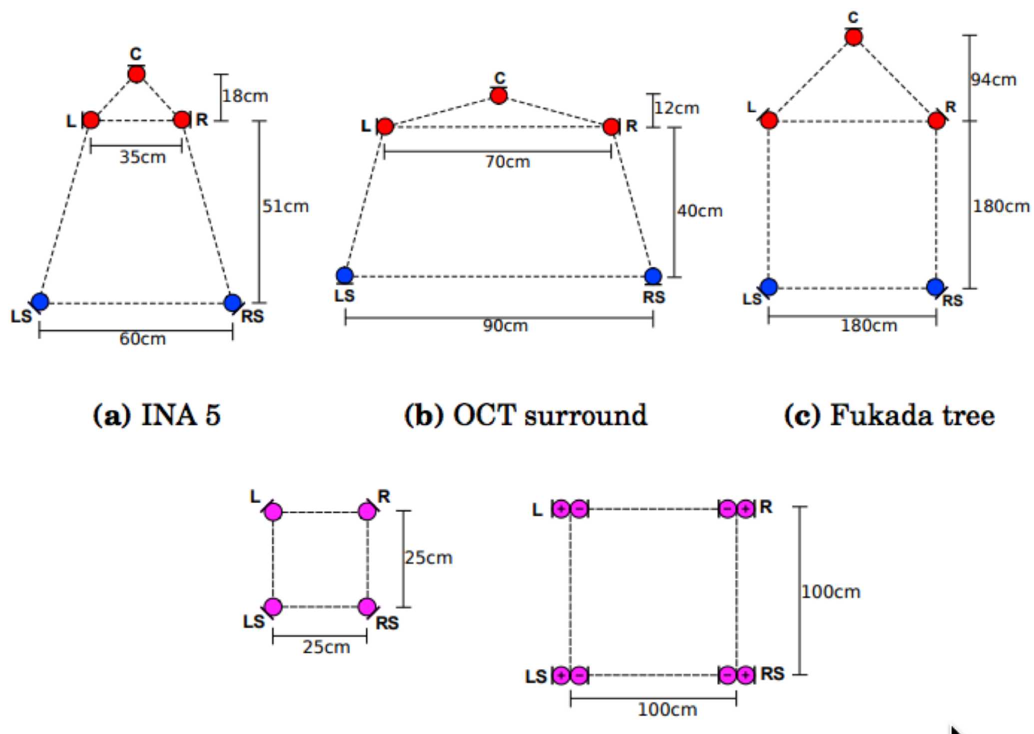
(Spatial) Performance determined by a combination of inter-channel level differences and inter-channel time differences

- Larger distances make signals less coherent, mitigating issues
- Too large delays may cause perceivable echoes, and they also finally smear spatial effects



Reproduction of real sound scenes: (Traditional) surround recording

Some examples:

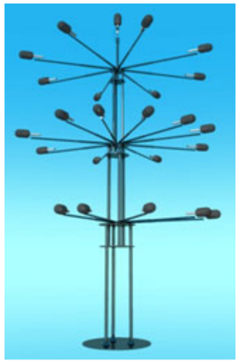


(a) INA 5 (b) OCT surround (c) Fukada tree

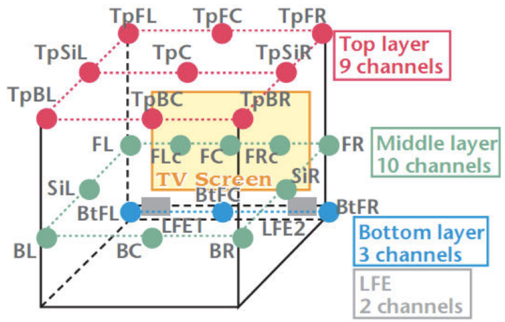
(d) IRT cross (e) Hamasaki square

Reproduction of real sound scenes: (Traditional) surround recording

- Well-defined for stereophony, complex for surround, not a scalable approach for arbitrary setups!
  - inflexible/not portable
  - redundant (hardware-wise)
  - impractical



NHK 22.2 recording setup



Reproduction of real sound scenes: Modern approaches

Record the sound scene:

a) **efficiently**, with a compact recording device
and a practical number of microphones

and reproduce the content:

b) **flexibly**, to any target reproduction setup

Reproduction of real sound scenes: Modern approaches

Two families of approaches:

A. **Non-parametric approaches**: rely only on the properties of the microphone array and the playback setup.

B. **Parametric methods**: rely on properties of the array and the reproduction setup, but also on the recorded signals themselves.

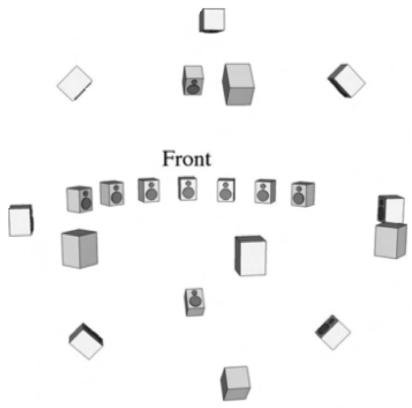
Reproduction of real sound scenes: Non-parametric methods

Rely solely on the properties of the microphone array and the reproduction setup:

M *microphones*

microphone array properties

playback system properties



L *output channels*

$$\mathbf{y}_L(f) = \mathbf{M}_{\mathrm{rep}}(f)\mathbf{x}_M(f)$$

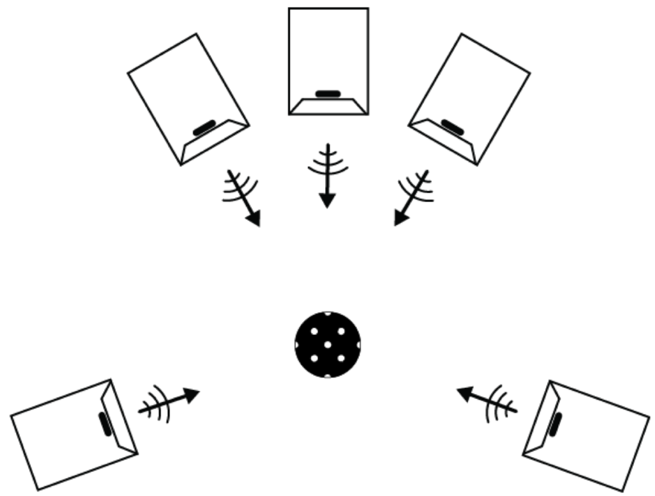
Reproduction of real sound scenes: Non-parametric methods

MIMO inversion/ pressure matching approaches:

$$\mathbf{x}_M(f) = \mathbf{C}_{\mathrm{rep}}(f)\mathbf{y}_L(f)$$

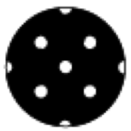
$\mathbf{C}_{\mathrm{rep}}(f)$ measured/modeled response matrix
$\mathbf{C}^{+}_{\mathrm{rep}}(f)$ regularized pseudo-inverse
$\mathbf{M}_{\mathrm{rep}} = \mathbf{C}^{+}_{\mathrm{rep}}(f)$

$$\mathbf{y}_L(f) = \mathbf{M}_{\mathrm{rep}}(f)\mathbf{x}_M(f)$$



Reproduction of real sound scenes: Non-parametric methods

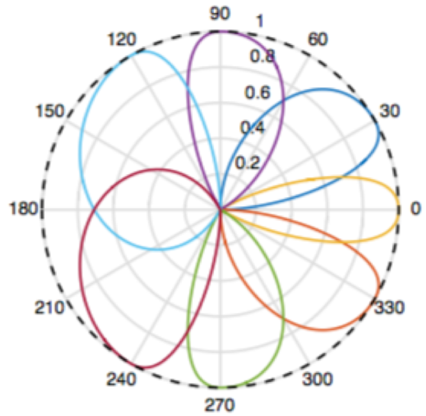
Direct beamforming approaches:

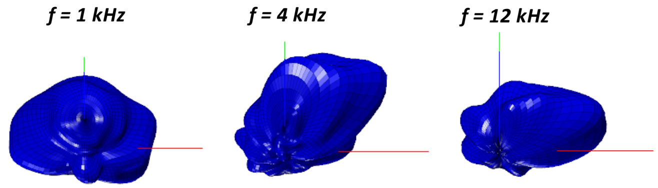


microphone array properties

$\mathbf{W}(\mathrm{f})$ beamforming weight matrix
for best fit to panning functions/HRTFs

$\mathbf{M}_{\mathrm{rep}}(f) = \mathbf{W}^{\mathrm{H}}(f)$

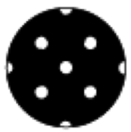




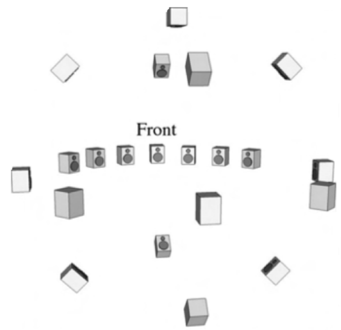
$$\mathbf{y}_L(f) = \mathbf{M}_{\mathrm{rep}}(f)\mathbf{x}_M(f)$$

Reproduction of real sound scenes: Non-parametric methods

Ambisonics:



encoding | decoding



microphone array properties

spatial Fourier transform

$$\mathbf{b}_N(f) = \mathbf{M}_{\mathrm{enc}}(f)\mathbf{x}_M(f)$$

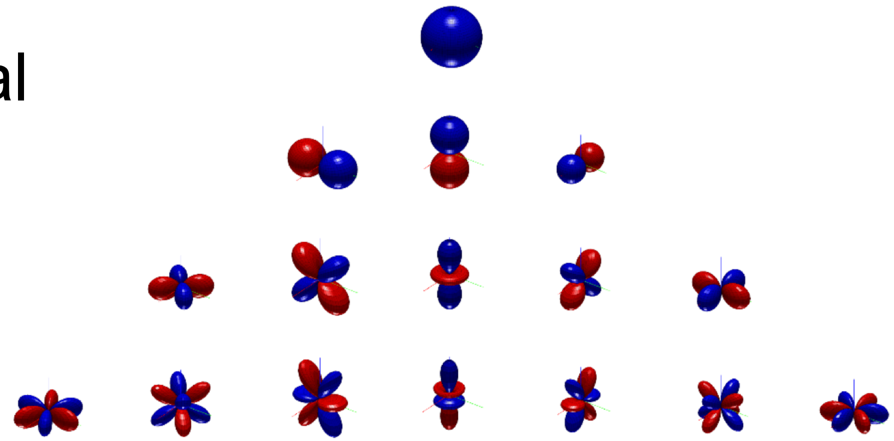
playback system properties

spatial Fourier transform

$$\mathbf{y}_L(f) = \mathbf{M}_{\mathrm{dec}}\mathbf{b}_N(f)$$

Reproduction of real sound scenes: Ambisonics

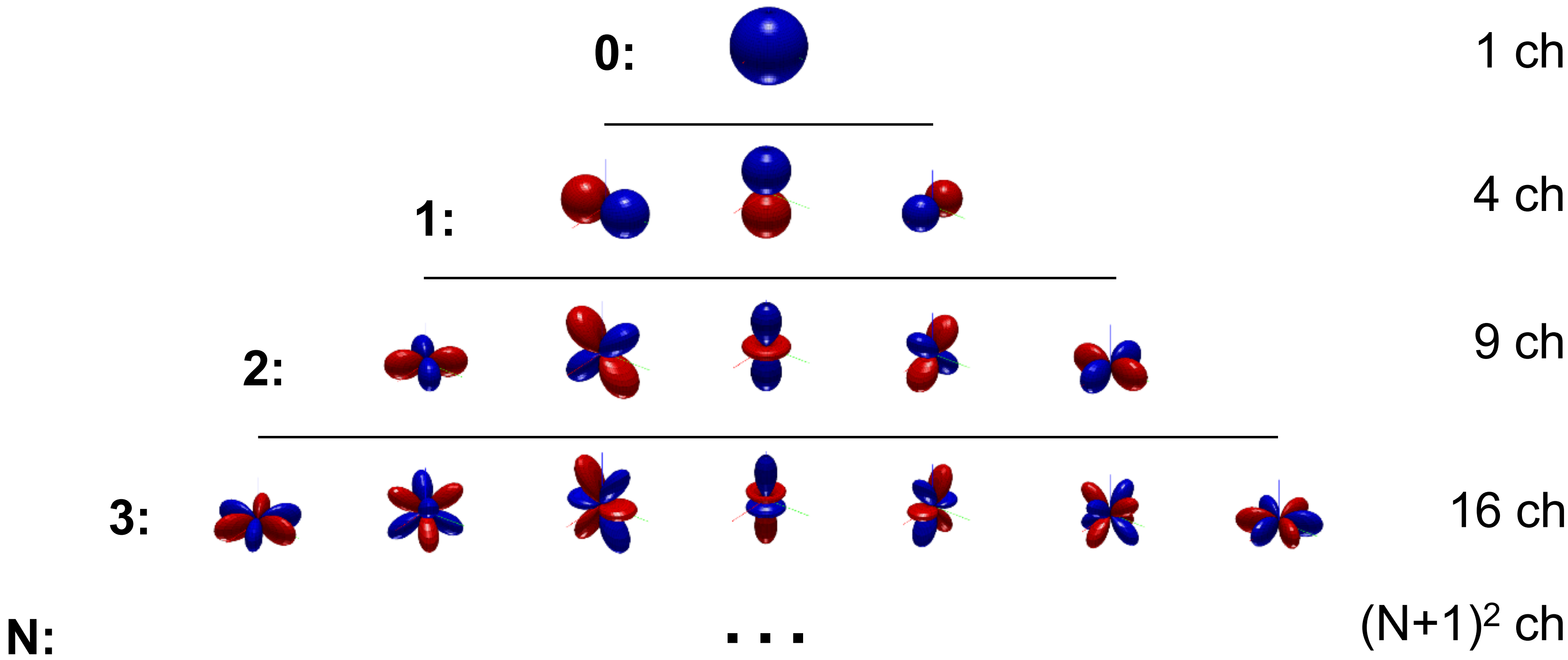
Encoding realizes the spatial transform from the microphone signals to the sound field coefficients.



Decoding redistributes the sound scene directionally to the target setup (physically or perceptually optimized).

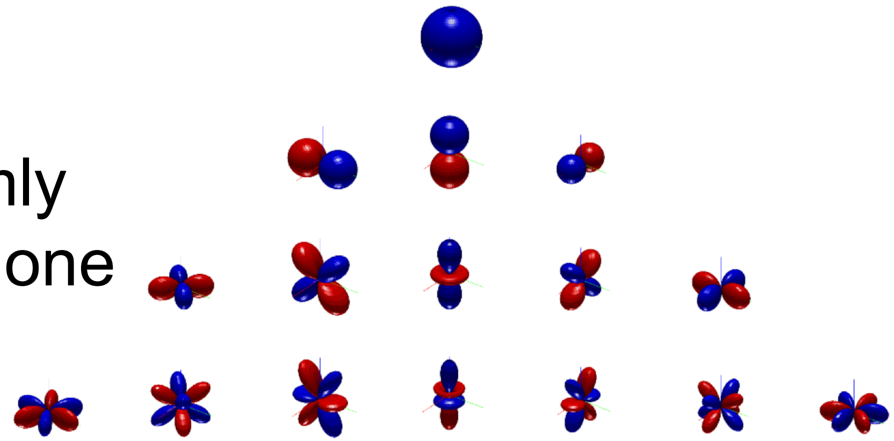
Reproduction of real sound scenes: Ambisonics

Spatial resolution:



Reproduction of real sound scenes: Ambisonics

Ambisonic recording:



Commonly done with uniformly distributed spherical microphone arrays.

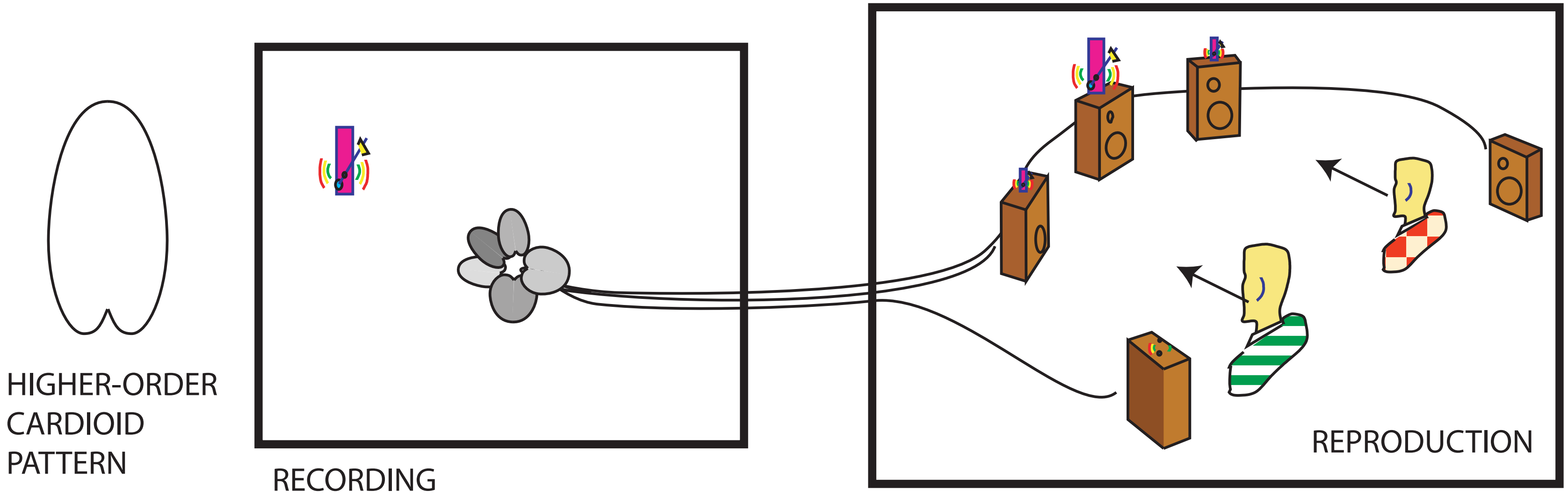
Soundfield SPS200 & ST350

Mh Acoustics Eigenmike

VisiSonics 5/64 AV Camera

Reproduction of real sound scenes: Ambisonics idea in practice

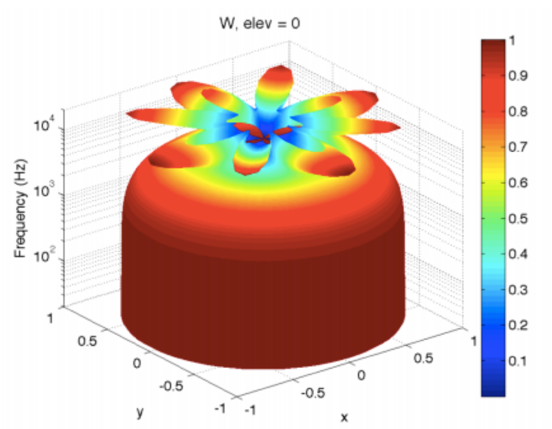
- The system performs beam forming in practice
- Quality depends on the accuracy of beams

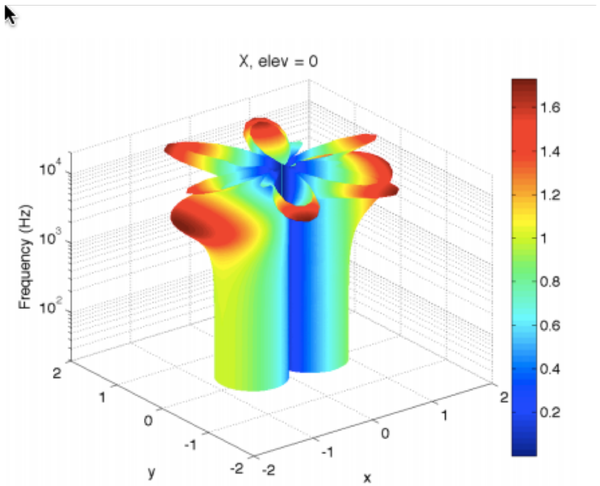


Reproduction of real sound scenes: Ambisonics

Physical limitations prohibit perfect capture of ambisonic signals (especially at high-orders/resolutions)

- Spatial aliasing
- Low-frequency noise



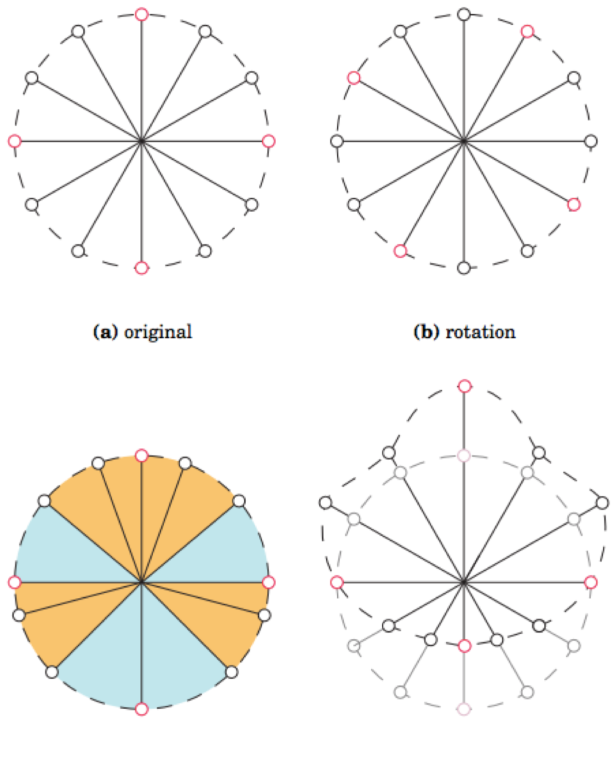


Reproduction of real sound scenes: Ambisonics

Spatial manipulations of the sound scene such as

- **rotations**,
- **directional warping**,
- **directional smoothing**,
- **directional loudness modifications**

and others, conveniently expressed in the spatio-spectral domain.



(a) original (b) rotation

(c) warping (d) amplitude manipulation

$$\mathbf{b}'_N(f) = \mathbf{T}\mathbf{b}_N(f)$$

Reproduction of real sound scenes: Non-parametric methods

Pros:

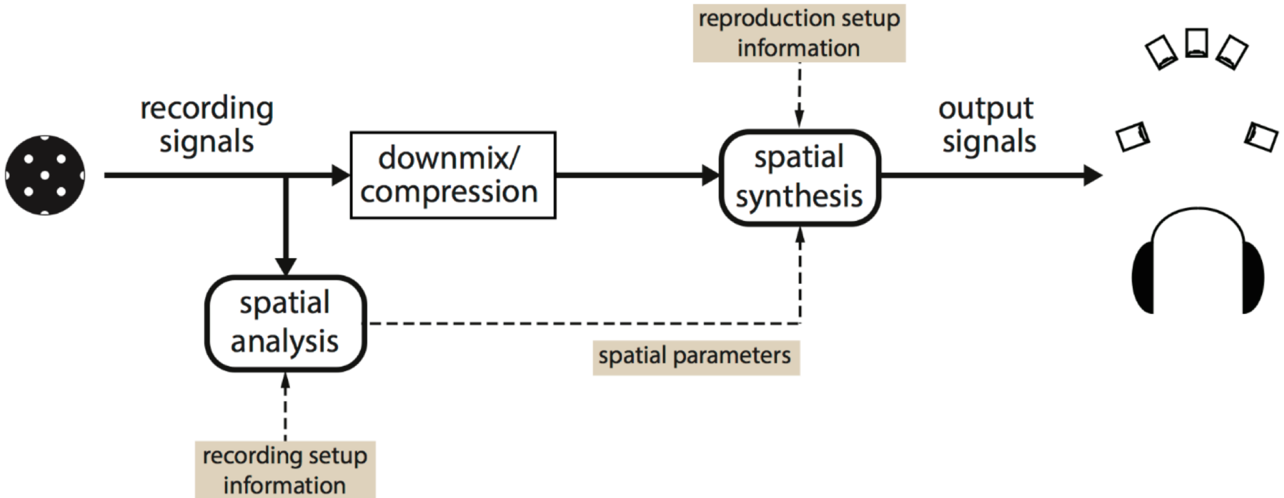
- No time-variant processing, high single-channel quality.
- Efficient, just a single static matrix of filters or gains

Cons:

- Resolution completely determined by geometry and number of microphones, can be too low to deliver the appropriate perceptual cues with compact arrays of a few microphones

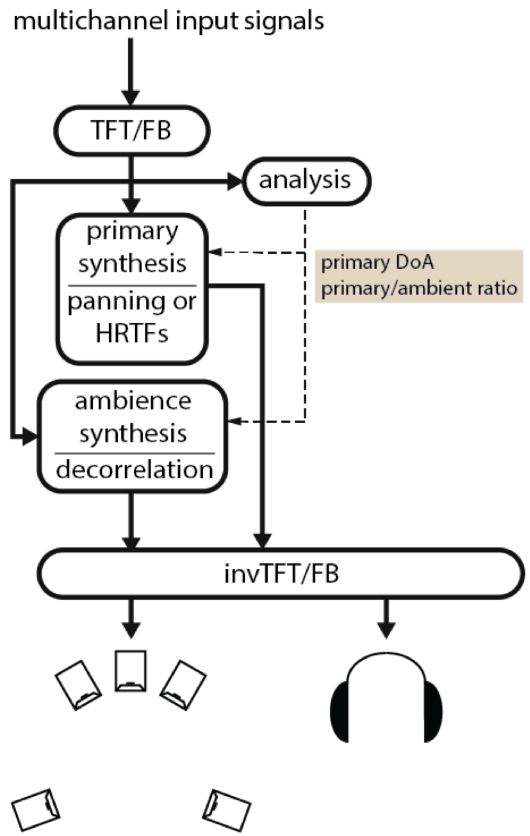
Reproduction of real sound scenes: Parametric methods

Apart from the microphone array and reproduction setup properties, the spatial relations between the recorded signals are used.



Parametric methods: State-of-the-art

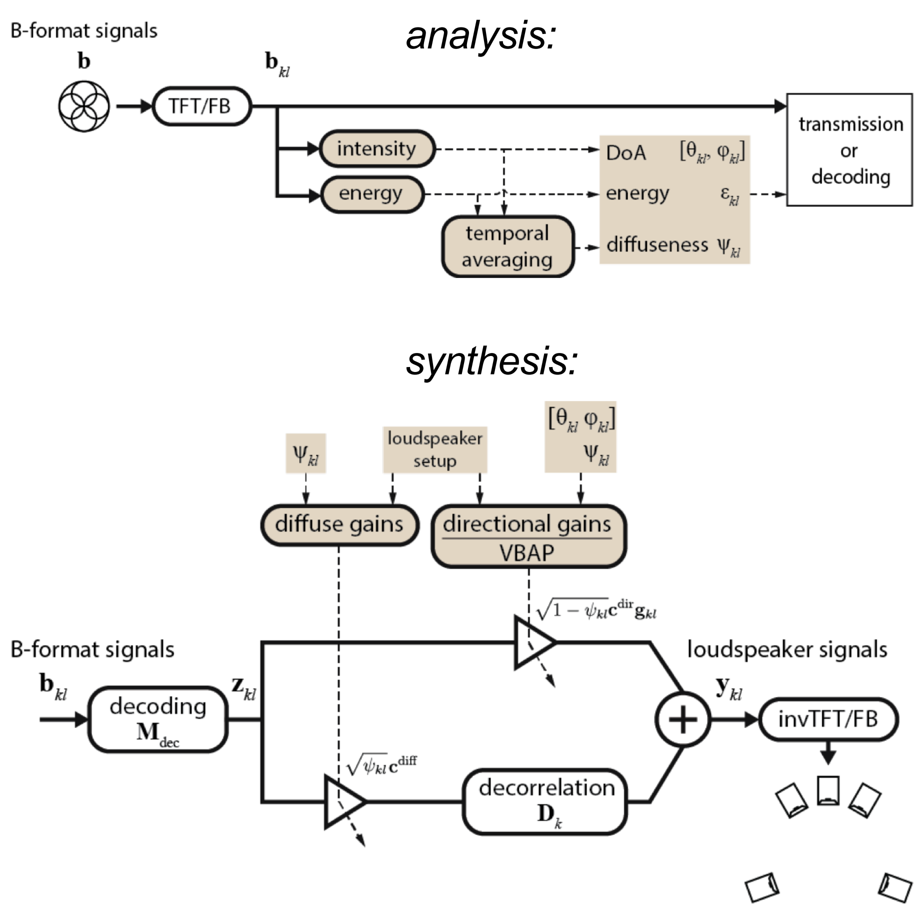
SAC / up-mixing approach:



- Perceptually effective
- Tuned to sound reproduction with quality in mind
- Simple model
- Tuned to channel-based content (not recordings)

Parametric methods: State-of-the-art

Directional Audio Coding:



- Perceptually effective
- Global sound scene parameters
- Mainly for spatial recordings
- Works also for channel-based formats

Reproduction of real sound scenes: Parametric methods

Pros:

- Super-resolution (when estimation is correct)
- Intuitive parameterization of the sound scene
- Flexible rendering

Cons:

- Time-variant spectral processing (requires care)
- Estimation errors should be handled gracefully
- More computationally demanding than non-parametric

Analysis & synthesis of room acoustics

Capture the spatial response of a room or acoustic environment of interest, and reproduce it perceptually.

Some use cases:

- Psychoacoustics of perception in rooms
- Architectural auralization
- Real acoustics-informed reverberation design
- Room acoustics enhancement
- Augmented reality

Analysis & synthesis of room acoustics

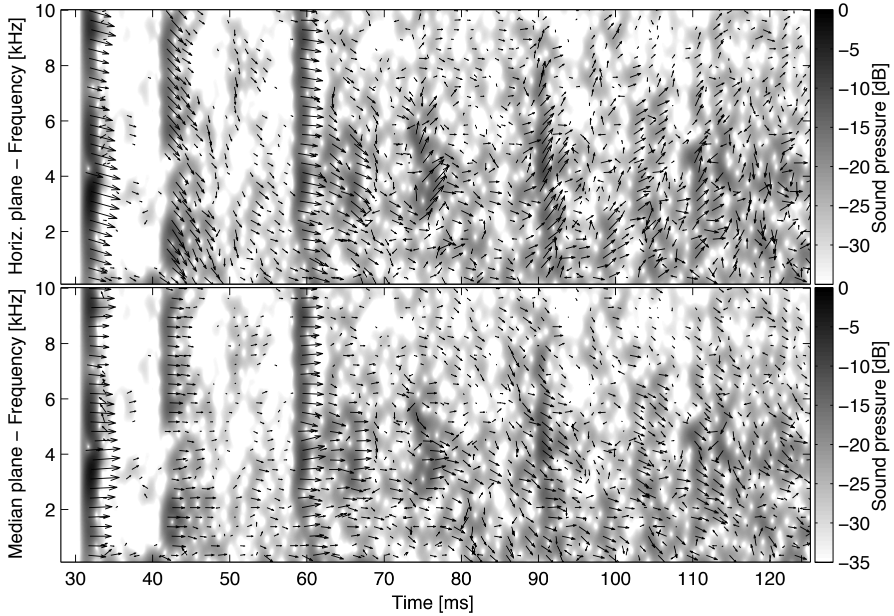
Capture the spatial response of a room or acoustic environment of interest, and reproduce it perceptually.

- non-parametric (Ambisonics)
- parametric (SIRR, SDM)

Analysis & synthesis of room acoustics

Spatial Impulse Response Rendering (SIRR) method:

Same operating principle as DirAC, applied to RIRs.



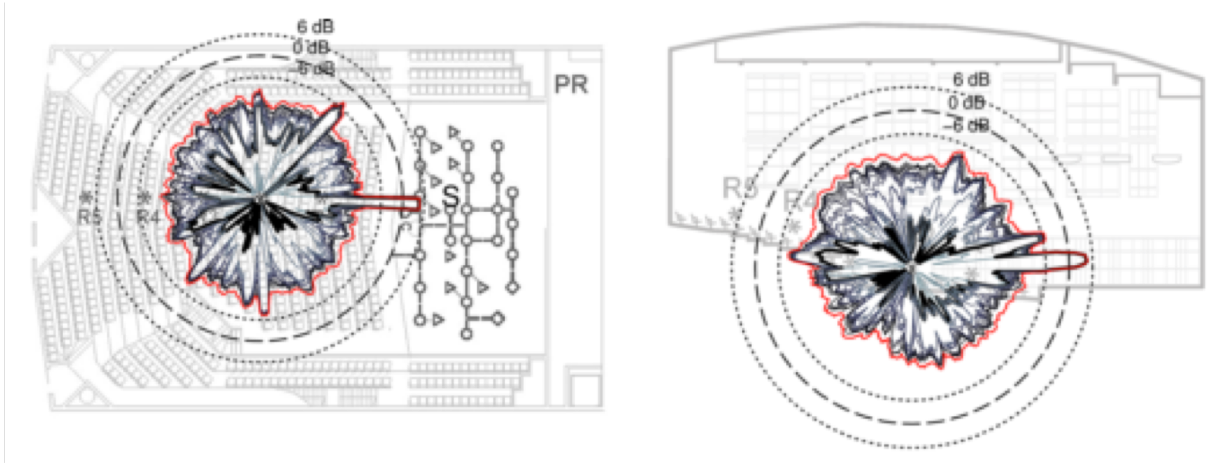
(from *Merimaa & Pulkki JAES 2005*)

Analysis & synthesis of room acoustics

Spatial Decomposition Method (SDM)

(*Tervo et al. JAES 2013*)

Single plane-wave broadband temporal directional analysis.



(from *Pätynen, Tervo & Lokki JASA 2013*)

Parametric methods: Beyond the state-of-the-art

State-of-the-art parametric methods work effectively for the majority of sound scenarios without being too demanding – good balance between resources/performance.

Why try to do more?

- Transparency
- Lossless sound-scene compression
- Higher-level objectification of the sound scene

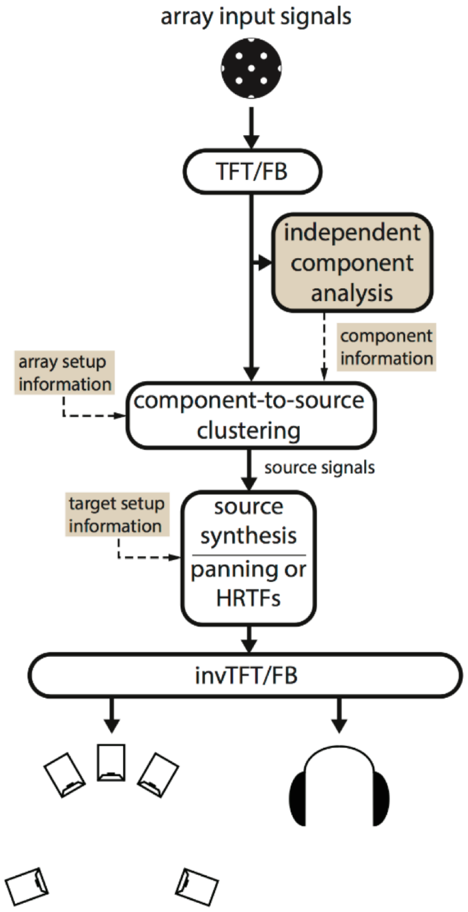
How:

- More powerful/general model

Reproduction of real sound scenes: Parametric methods

Source separation approach:

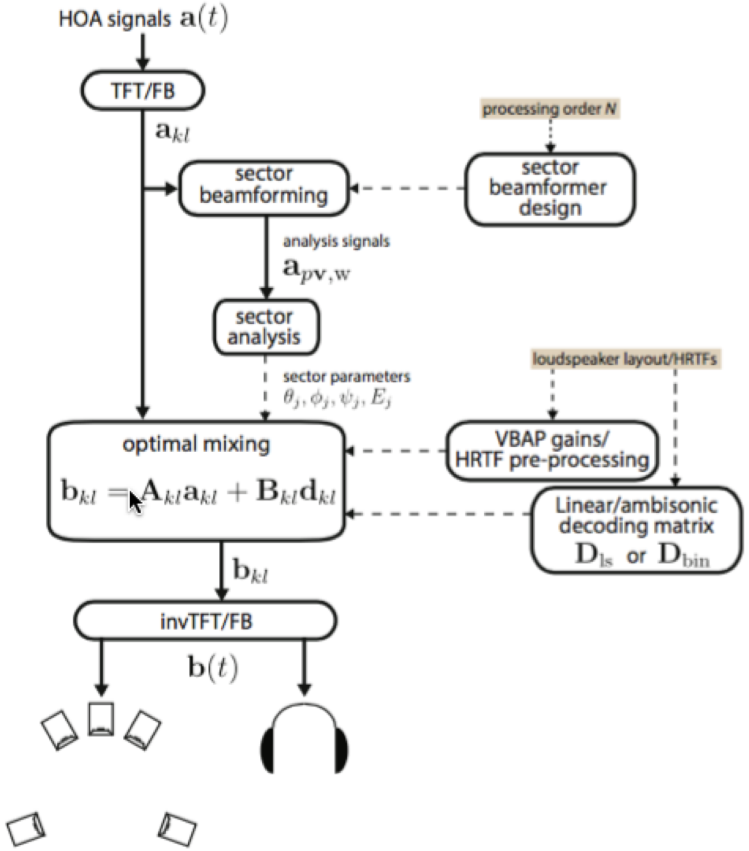
(*e.g.* Nikunen, IEEE TSALP 2014)



- Able to extract multiple directional components
- Perceptually justified
- Usually tuned to separation rather than reproduction
- Slow (for real-time)
- Hard to make robust

Reproduction of real sound scenes: Parametric methods

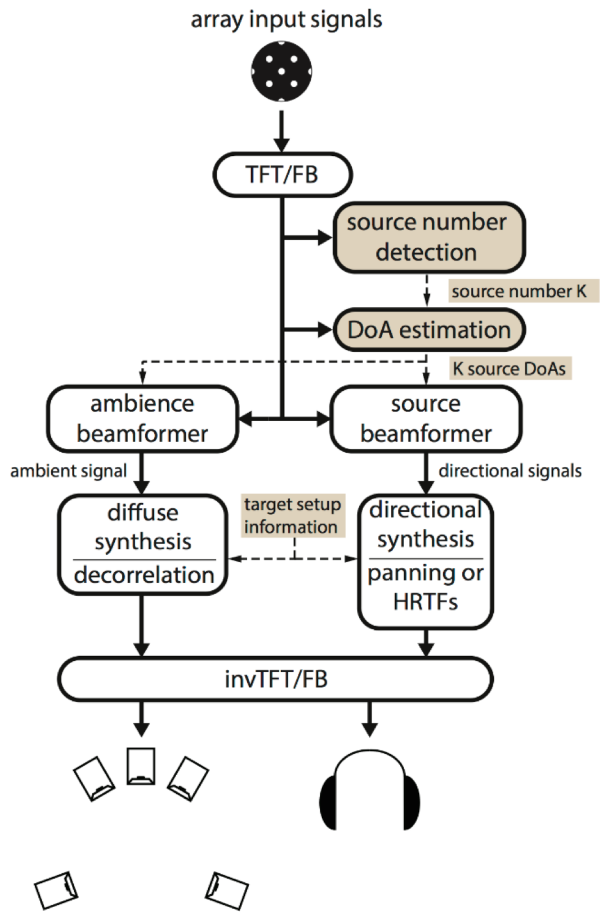
Higher-order DirAC
(Politis et. al. IEEE JSTSP 2015)



- Able to extract multiple directional components
- Perceptually driven
- Robust
- Hard to interpret intuitively parameters

Reproduction of real sound scenes: Parametric methods

Multiwave beamforming approach:
(e.g. COMPASS, Politis & Tervo 2018)



- Able to extract multiple directional components
- Acoustically driven
- Tools usually tuned to suppression or enhancement rather than reproduction
- More complex to make robust