Aalto University
School of Electrical
Engineering

Amplitude panning, Time-frequency-domain spatial audio

Ville Pulkki

Professor of Acoustics (Associate Professor)
Department of Signal Processing and Acoustics
School of Electrical Engineering
Aalto University, Helsinki, Finland

Virtual acoustics course lecture

February 6, 2019

These slides

- Vector base amplitude panning (VBAP)
- Variants and enhancement of VBAP
- Time-frequency-domain parametric spatial audio
- Directional audio coding

A music student with MSc (Eng) needs extra income (1995)

- Sibelius Academy chamber music hall had lots of loudspeakers on walls and ceiling
- SibA wanted to have a "panning tool" for their loudspeaker system (one month salary for student)
- 1-month joint project btw TKK and SibA

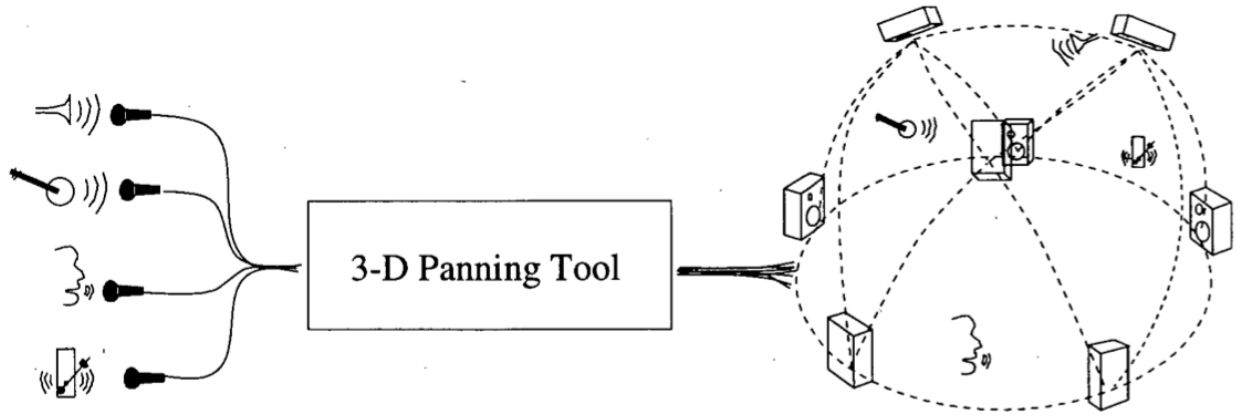
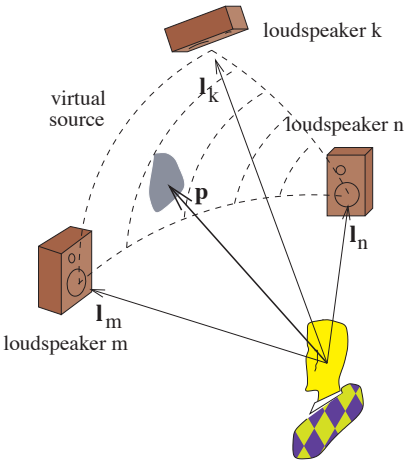


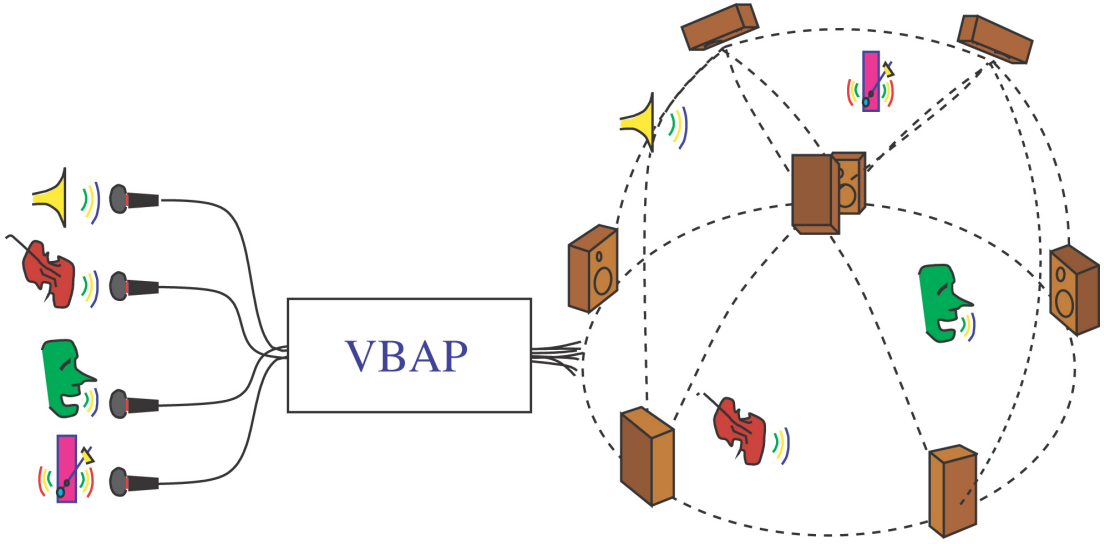
Fig. 9. Possible use of three-dimensional VBAP panning tool. Number of sound sources can vary up to eight; loudspeaker placement is arbitrary; virtual sources may be moving or stationary.

462 J. Audio Eng. Soc., Vol. 45, No. 6, 1997 June

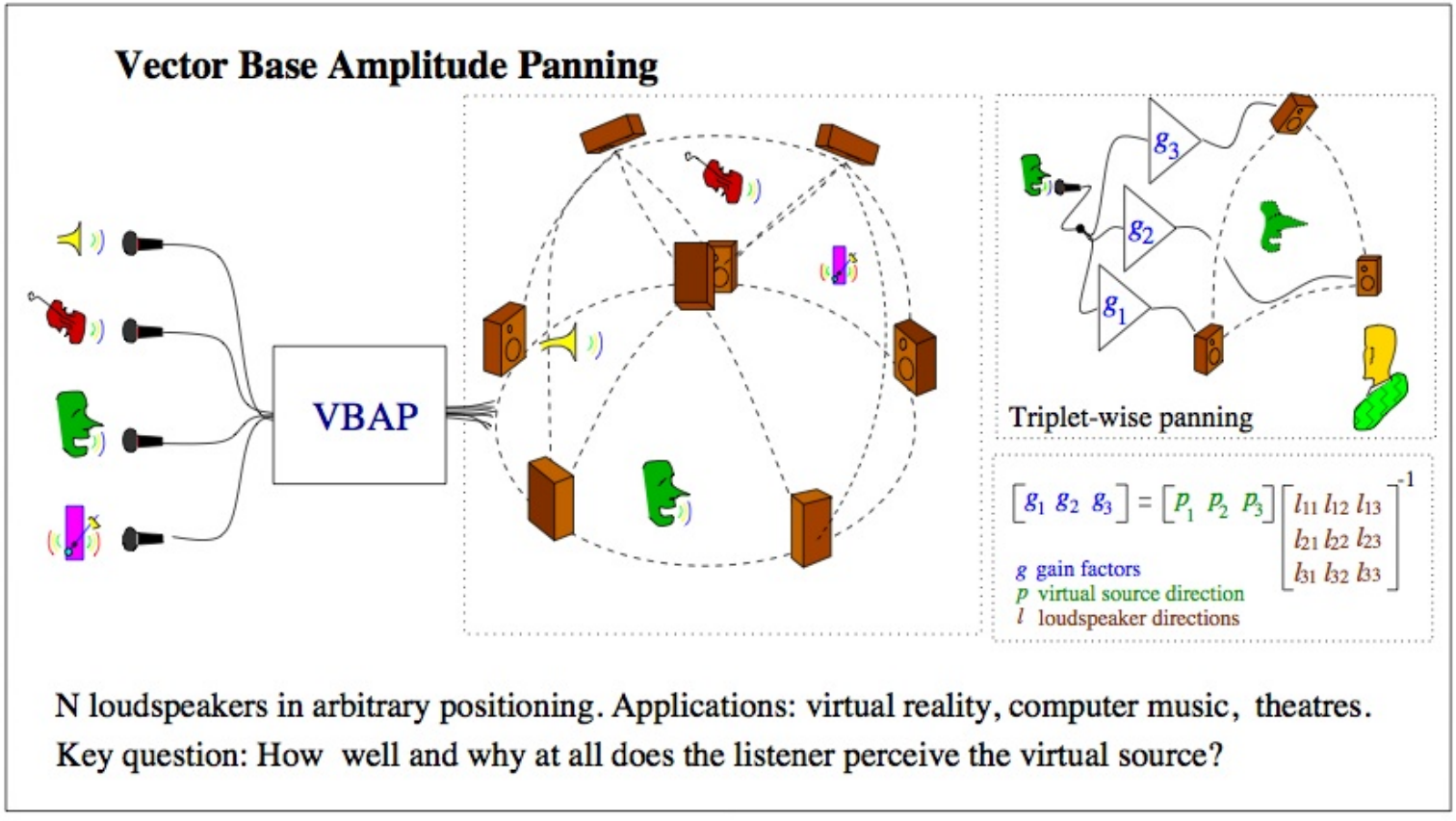
Reformulation of amplitude panning

- Tried to generalize the sine panning law to 3D, no luck
- "Could this be formulated with vector bases?" – "Yes!"
- Vector base amplitude panning (VBAP) was born
- Divide setup into triplets, and compute gain factors for each





Vector base amplitude panning



PhD degree in 2001.

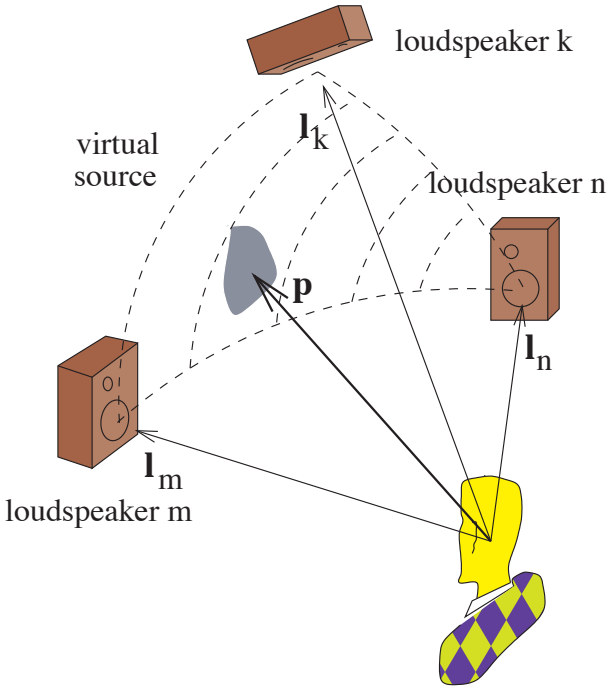
Dissemination of VBAP

- Published VBAP paper in Journal of the Audio Engineering Society (JAES) 1997
- Provided free software implementations of the method
- Article has been cited $>$1200 times in google scholar (2019)
- The first paper of all JAES papers, when ranked with the number of citations (scopus)

Products with "VBAP inside"

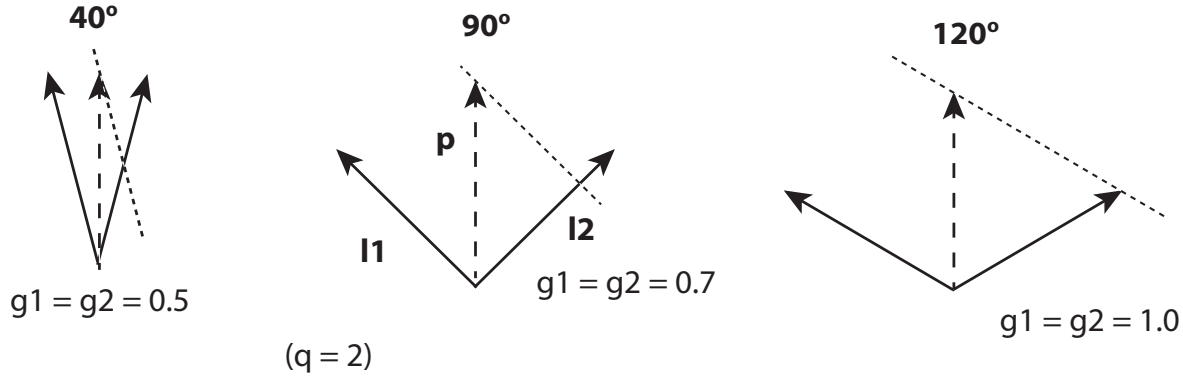
- ITU MPEG-H audio standard (broadcast)
- DTS:X audio format (cinema + blueray)
- Sony Playstation VR (gaming)
- Dedicated audio programming softwares

VBAP maths



- $\mathbf{p} = g_1\mathbf{l}_1 + g_2\mathbf{l}_2 + g_3\mathbf{l}_3$
- $\mathbf{g} = [p_1\ p_2\ p_3] \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix}^{-1}$
- $\mathbf{g}$ holds the barycentric coordinates of virtual source in vector base
- loudspeaker signals $y_i = g_i x(t)$
- g_i controls the amplitude of signal in each loudspeaker

VBAP maths



- g_i depend on opening angle of the loudspeaker base / not good!
- Length of $\mathbf{g}$ must be normalized to avoid changes in loudness
- $\mathbf{g}_{\mathrm{norm}} = \mathbf{g} / (\sum_i \mathbf{g}_i^{\mathbf{p}})^{\mathbf{1}/\mathbf{q}}$
- Thus: $(\sum_i g_{\mathrm{norm}_i}^q)^{1/q} == 1$
  - $q = 1$ for anechoic cases (also headphones with virtual loudspeakers)
  - $q = 2$ for normal rooms

Matlab code available `https://se.mathworks.com/matlabcentral/fileexchange/53884-vector-base-amplitude-panning-library`

VBAP runtime cycle

init Feed in loudspeaker directions

init 2D: form pairs. 3D form triplets. Compute inverse matrices

run multiply input sound $x(t)$ with $\mathbf{g}$, output to loudspeakers

intrpt1 Start interrupt when virtual source direction changes

intrpt2 Compute gain factors for each LS pair/triplet, select the pair/triplet with positive gains

intrpt3 Normalize gains

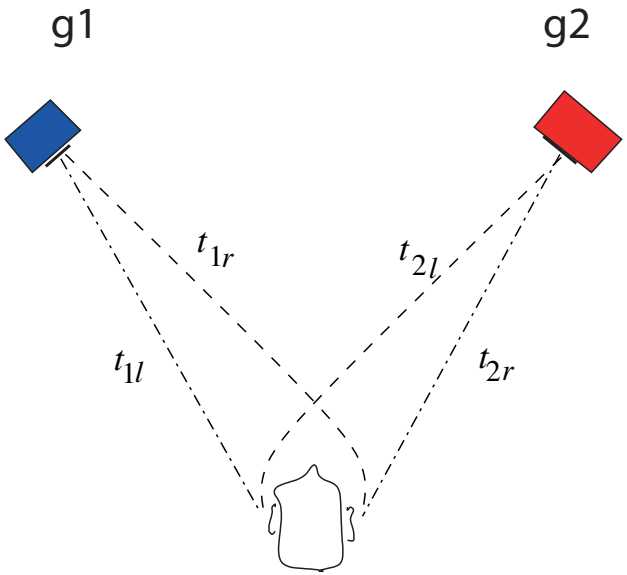
Max demo

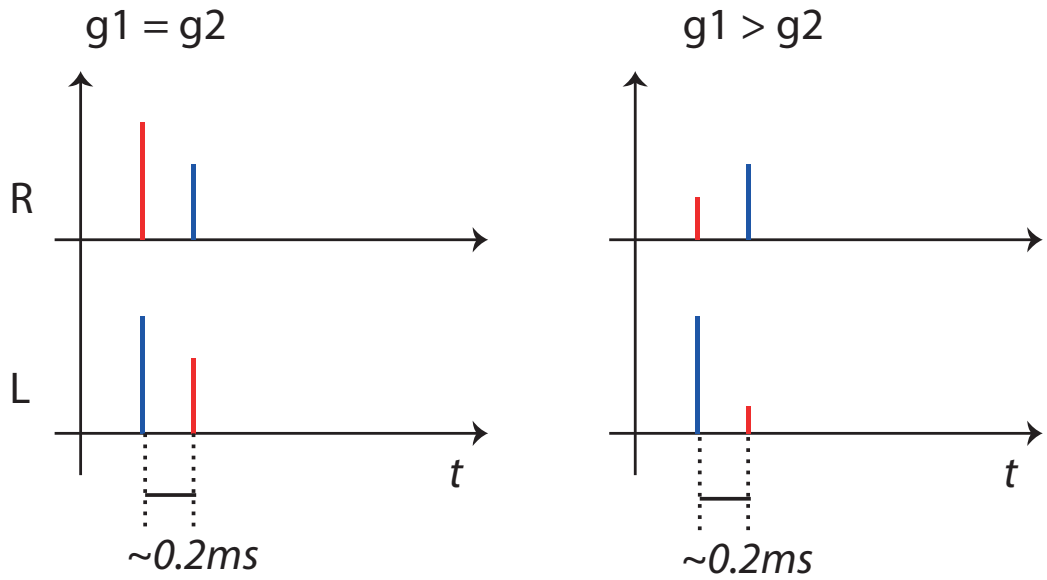
Amplitude panning audio quality

- Direction of amplitude-panned source is perceived relatively accurately in best listening position
- Outside of sweet spot: directional perception is dominated by nearmost loudspeaker
- No prominent coloration issues in normal rooms inside or outside the sweet spot
- Most-used virtual source positioning method: all mixers have "panpot" buttons
- Coloration issues in anehcoic listening

Amplitude panning, mechanism behind formation of perceived direction

Perceiving a virtual source between the loudspeakers does not correspond to actual situation. If you have a red LED in both left and right hands, you see two LEDS, not one in between.
Amplitude panning causes cross-talk and affects both ITD and ILD in complex way

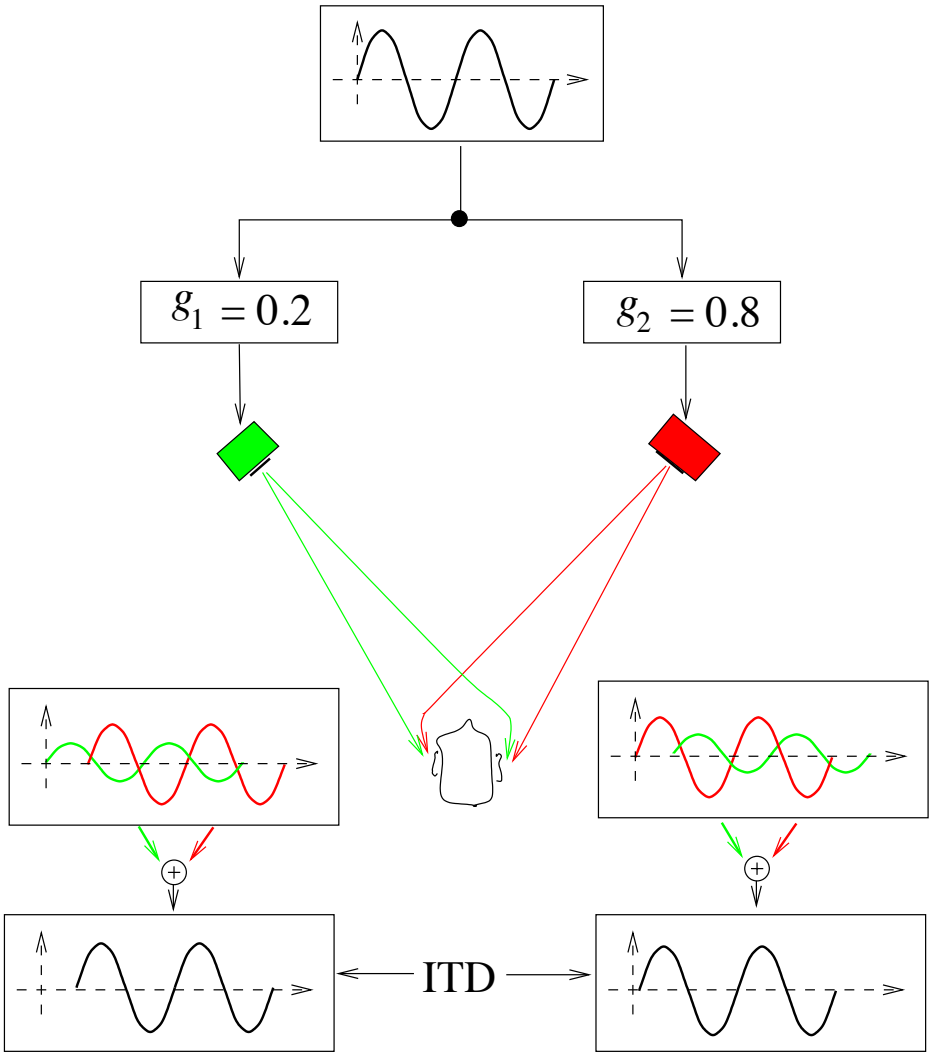




ampltude-panned impulse responses at ear canals

Summing localization at low frequencies:

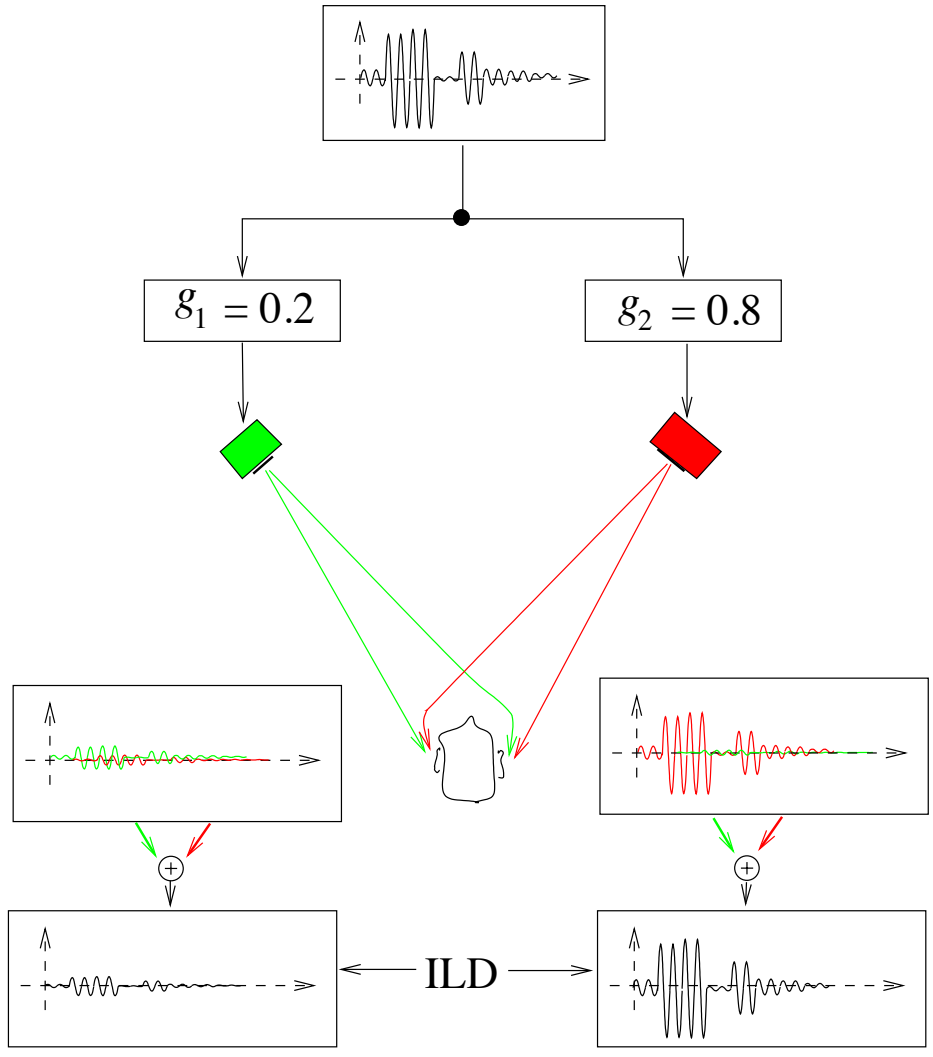
Amplitude panning



ILD = 0 (?)

Summing localization at high frequencies:

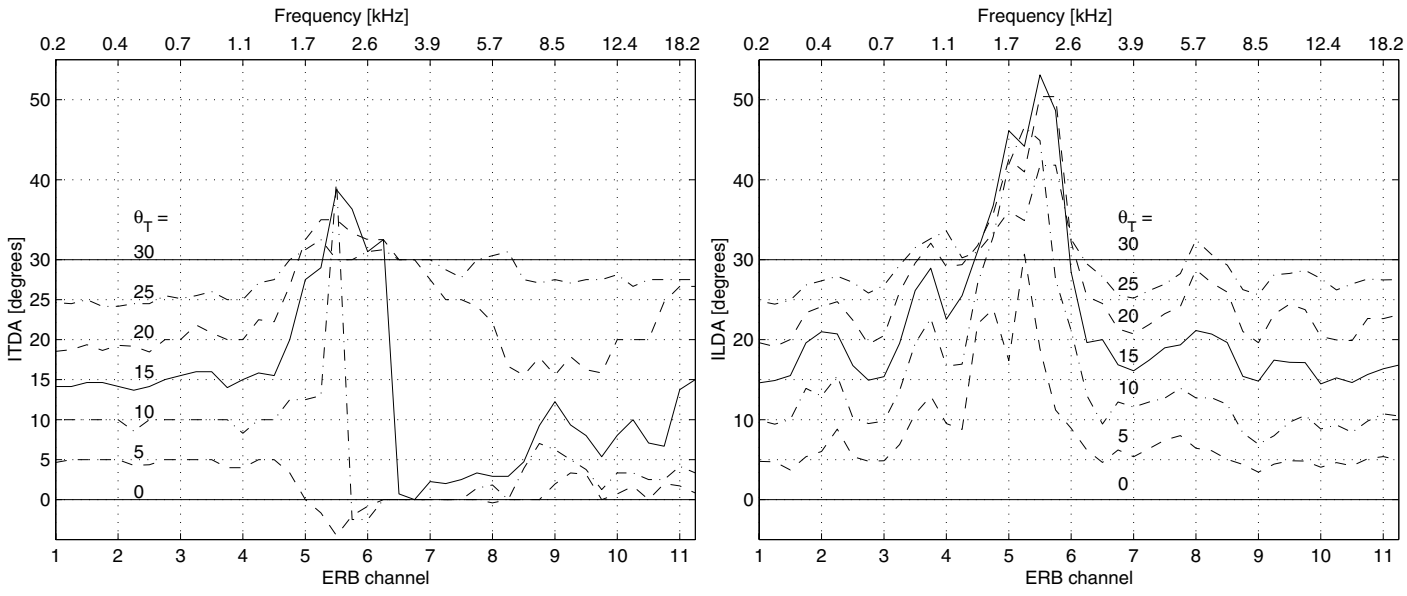
Amplitude panning



ITD = 0 (?)

Amplitude panning, mechanism behind formation of perceived direction

- loudspeaker amplitude difference changes to interaural time difference at low frequencies
- loudspeaker amplitude difference changes to interaural level difference at high frequencies

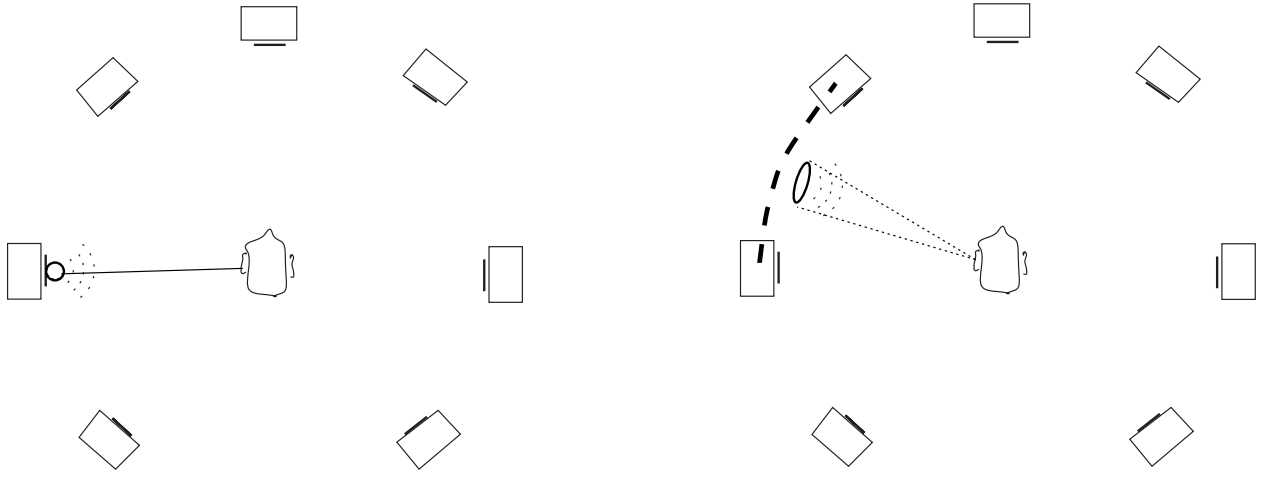


Pulkki, Ville. Spatial sound generation and perception by amplitude panning techniques. PhD thesis. Helsinki University of Technology, 2001.

Spread issue

Perceived spread of amplitude-panned virtual sources depends on virtual source direction **p**

- When **p** is coincident with loudspeaker direction, "point-like"
- When **p** is in-between loudspeakers, "more or less spread"
- Frequency-dependent ITD and ILD cues do not match with real source

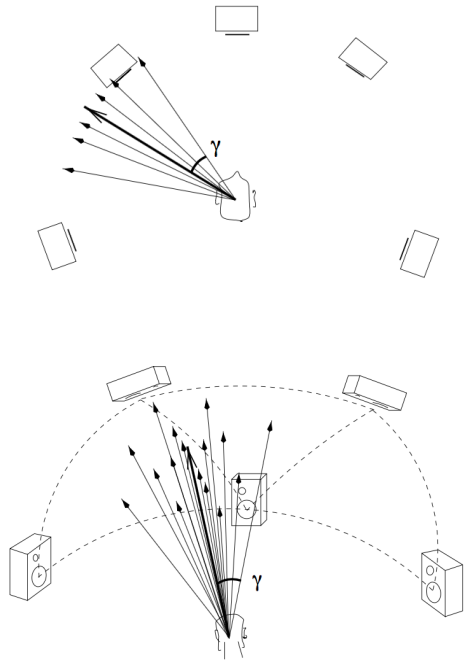


Multiple-direction amplitude panning

Make the spread even!

- Define $\mathbf{p}$
- Define a number of vectors $\mathbf{p}_i$ around $\mathbf{p}$ within angular range of γ around $\mathbf{p}$
- Compute $\mathbf{g}_i$ for for each $\mathbf{p}_i$,
- Sum over i: $\mathbf{g} = \sum_i \mathbf{g}$

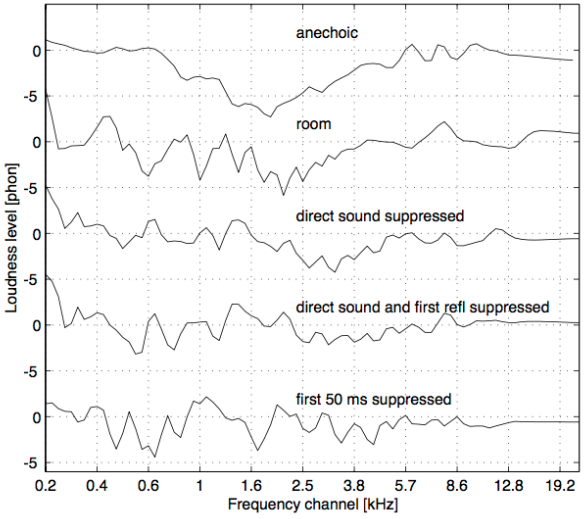
Result: always more than one LS has considerable gain.

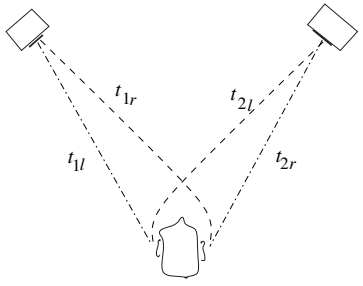


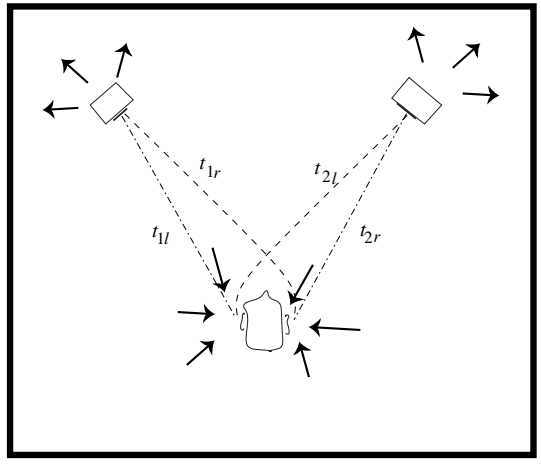
Max demo

Pulkki, V. (1999). Uniform spreading of amplitude panned virtual sources. In Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on (pp. 187-190). IEEE.

Coloration of amplitude-panned sources





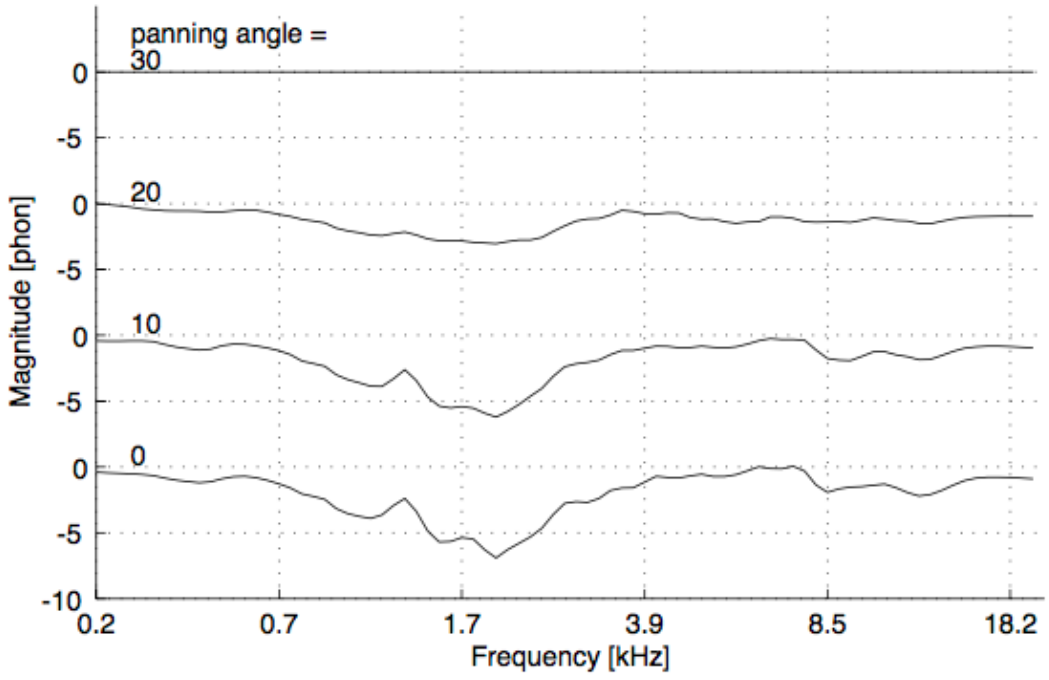


- Direct sounds from loudspeakers interfere $\rightarrow$ comb filter effect $\rightarrow$ audible coloration
- Reflected and reverberated sound paths arrrive at ear canals in incoherent manner, and no comb filter effect occurs
- $\rightarrow$ Amplitude-panned sources are not perceived colored in normal rooms

Coloration of amplitude-panned sources

- Amplitude-panned sources are colored in anechoic listening
- Isn't anechoic listening just a niche that nobody cares?
- Headphone listening with virtual loudspeakers + panning: that is anechoic listening
- (Sony playstation VR + many other VR applications)
- We should do something for this
- An easy solution is to utilize more loudspeakers: when the angle between LS is smaller, traveling time difference is smaller, and comb-filter effect migrates to higher frequencies and becomes less salient

Coloration of amplitude-panned sources in anechoic listening



Characteristics

- Dip around 1-2 kHz
- At high frequencies a bit lower level
- effect depends on
  - panning angle
  - loudspeaker directions
  - room effect

Can we compensate this by equalizing / other means?

Coloration of amplitude-panned sources in anechoic listening

- Gain factor normalization $\mathbf{g}_{\mathrm{norm}} = \mathbf{g}/(\sum_i g_i^q)$
  - $q = 1$ or $q = 2$
  - In anechoic listening only frequencies below about 800Hz satisfy $q = 1$ condition
  - At frequencies above 2kHz it is not intuitively clear what happens.
- Lets make q to depend on frequency and listening-room-response

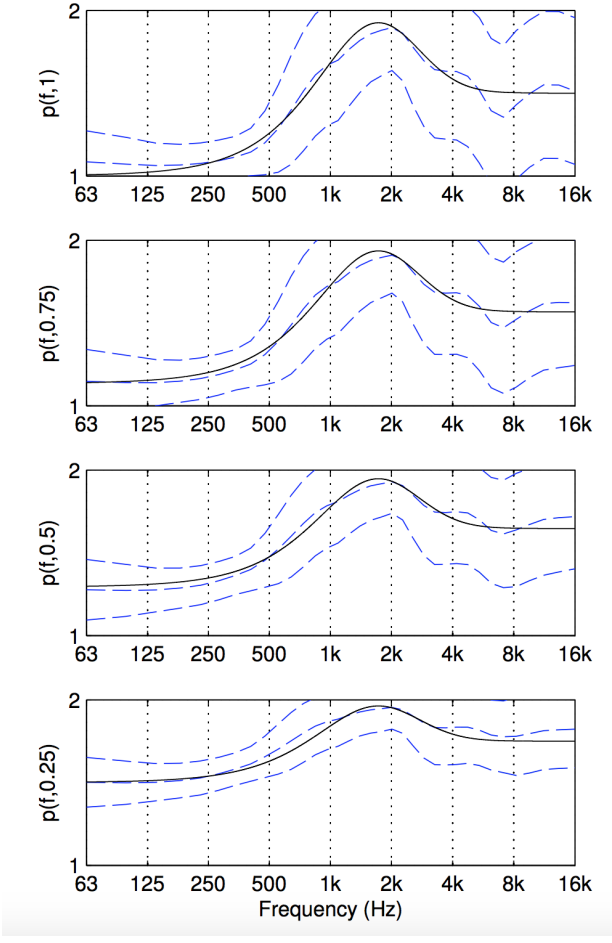
Coloration of amplitude-panned sources in anechoic listening

A solution has been proposed in [1]

- Gain factor normalization $\mathbf{g}_{\mathrm{norm}} = \mathbf{g}/(\sum_i g_i^q)$
- A solution has been numerically obtained using auditory models and room measurements
- $q(f, \mathrm{DTT}) = (p_0(f))\sqrt{\mathrm{DTT}} + 2$
- $q_0(f) = 1.5 - 0.5\cos\left[4.7\tanh\left(a_1 f\right)\max\left(0, 1 - (a_2 f)\right]\right.$
- where $a_1 = 0.00045$ and $a_2 = 0.000085$
- DTT is direct-to-total energy ratio

[1] Laitinen, M. V., Vilkamo, J., Jussila, K., Politis, A., & Pulkki, V. (2014, August). Gain normalization in amplitude panning as a function of frequency and room reverberance. In Audio Engineering Society Conference: 55th International Conference: Spatial Audio. Audio Engineering Society.

Coloration of amplitude-panned sources in anechoic listening



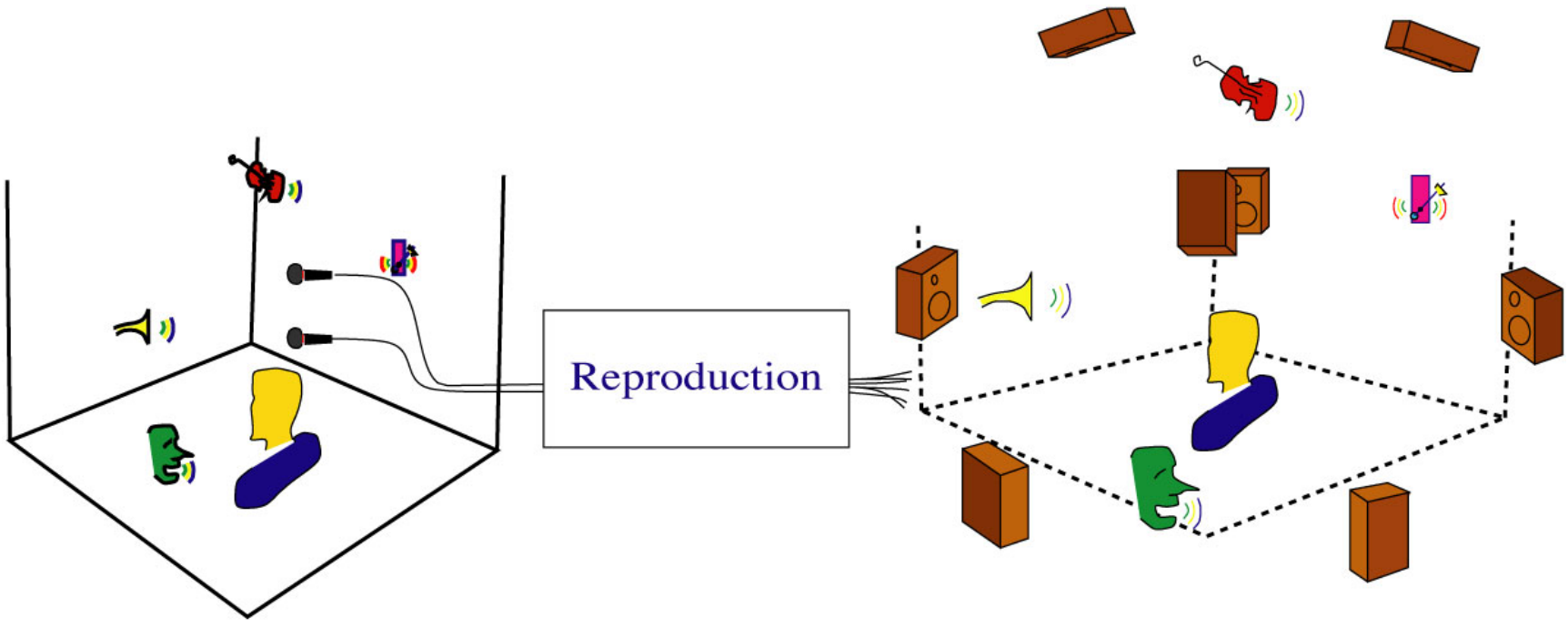
- The figures show $q(f, \mathrm{DTT})$ values
- Results simulated with large number of listening conditions with loudspeaker span from 30° to 80°
- Requires frequency-domain implementation of panning
- Mitigates coloration issues
- Readily implementable in time-frequency-domain processing, such as in DirAC
- Can be implemented with IIR filters (?)

Directional audio coding (DirAC)

Developed in Ville Pulkki's research group 2001 —
Pulkki, Merimaa, Laitinen, Ahonen, Vilkamo, Pihlajamaki, Politis etc

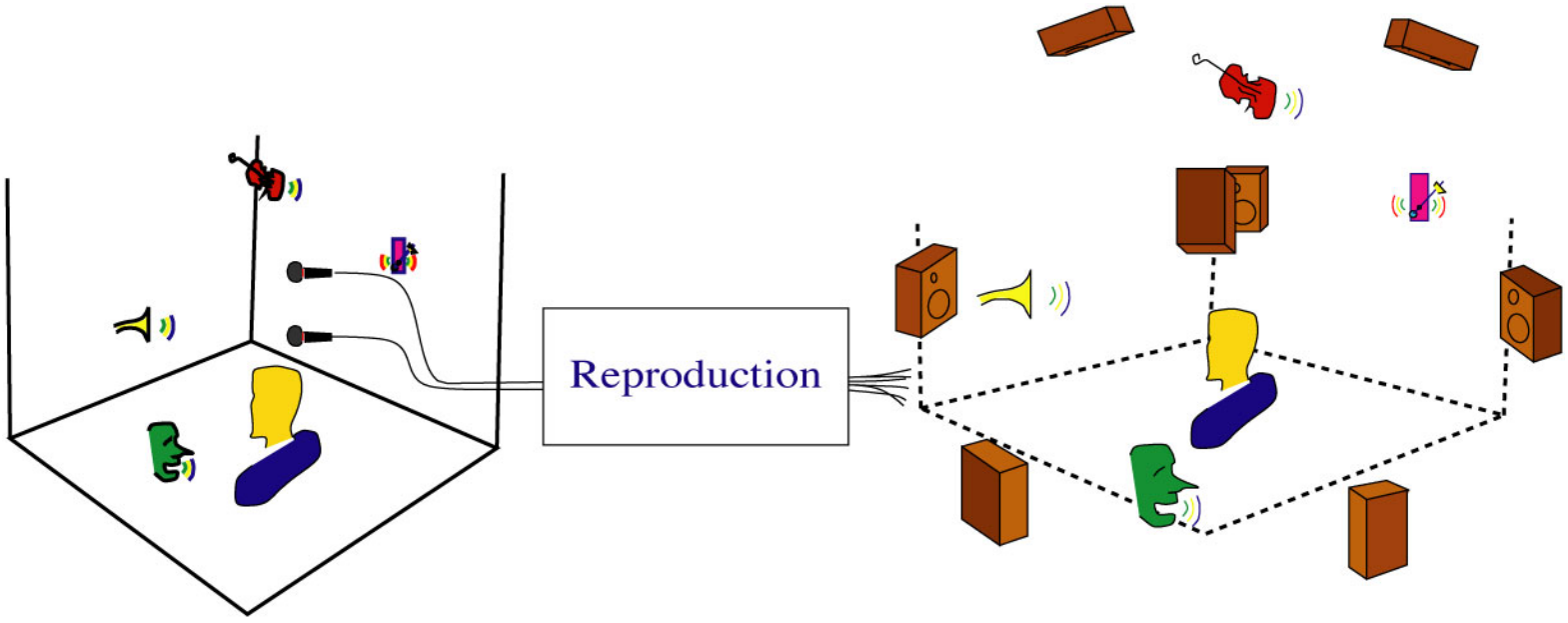
- Method to reproduce/process/compress recorded spatial sound
- Several patents
- Patents sold to Fraunhofer IIS in 2007, largest patent sale in Helsinki University of Technology
- Commercialized as "Fraunhofer UpHear"

How could a sound field be reproduced



Problems with existing techniques

Spatial sound reproduction



Target: relay the perception of sound!

Analogy with video

How does a video camera work?

- Lens
- Light from distinct direction is projected to one position at CCD
- CCD encodes the light energy at three frequency channels (RGB)
- Visible light wave lengths 380 nm - 780 nm (less than one octave)
- Very similar with eye

Spatial sound reproduction

Could we do the same with sound than with video camera

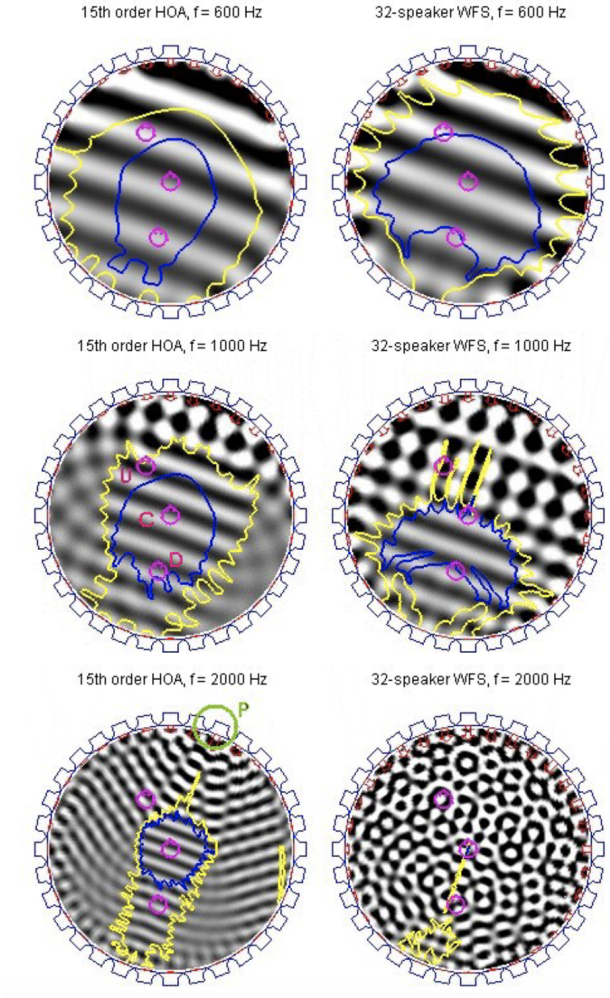
- Create narrow beam for each loudspeaker
- Audible sound includes wave lengths from 2 cm to about 30 m
- Impossible to build a microphone having constant narrow beam width without coloration and noise problems
- Higher-order Ambisonics / beam steering try to do it

Spatial sound reproduction

Holography then, perhaps?

- Lots of spaced microphones
- Lots of loudspeakers
- Wave field synthesis
- Problems
  - High price
  - Directivity of microphones should be matched with directivity of loudspeakers - is it possible?

Sound fields reproduced with WFS and HOA



- monochromatic plane waves reproduced
- valid sound field only in limited listening area "sweet spot"
- at high frequencies huge errors

Daniel, JÈrÙme, Sebastien Moreau, and Rozenn Nicol. "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging." Audio Engineering Society Convention 114. Audio Engineering Society, 2003. This

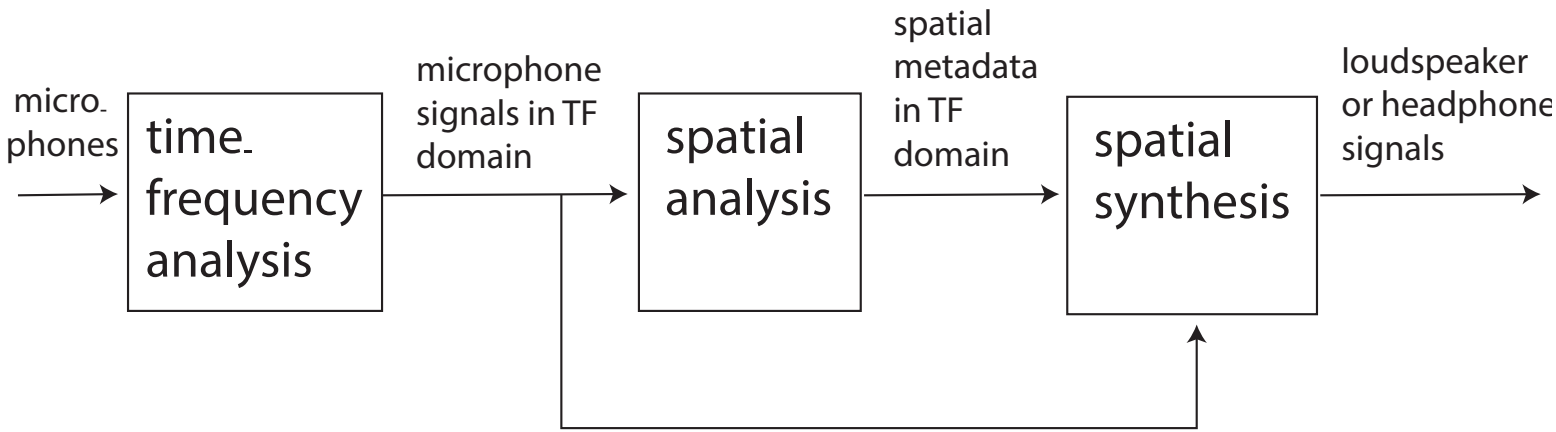
Parametric spatial sound reproduction

Are there any workarounds?

- Human spatial hearing can be fooled easily
- E.g. two coherent sources produce one virtual source in the middle
- Compare with vision: coherent sources do not produce virtual sources
- Assumption: at one frequency band humans perceive only one direction and one coherence cue

Parametric spatial sound reproduction

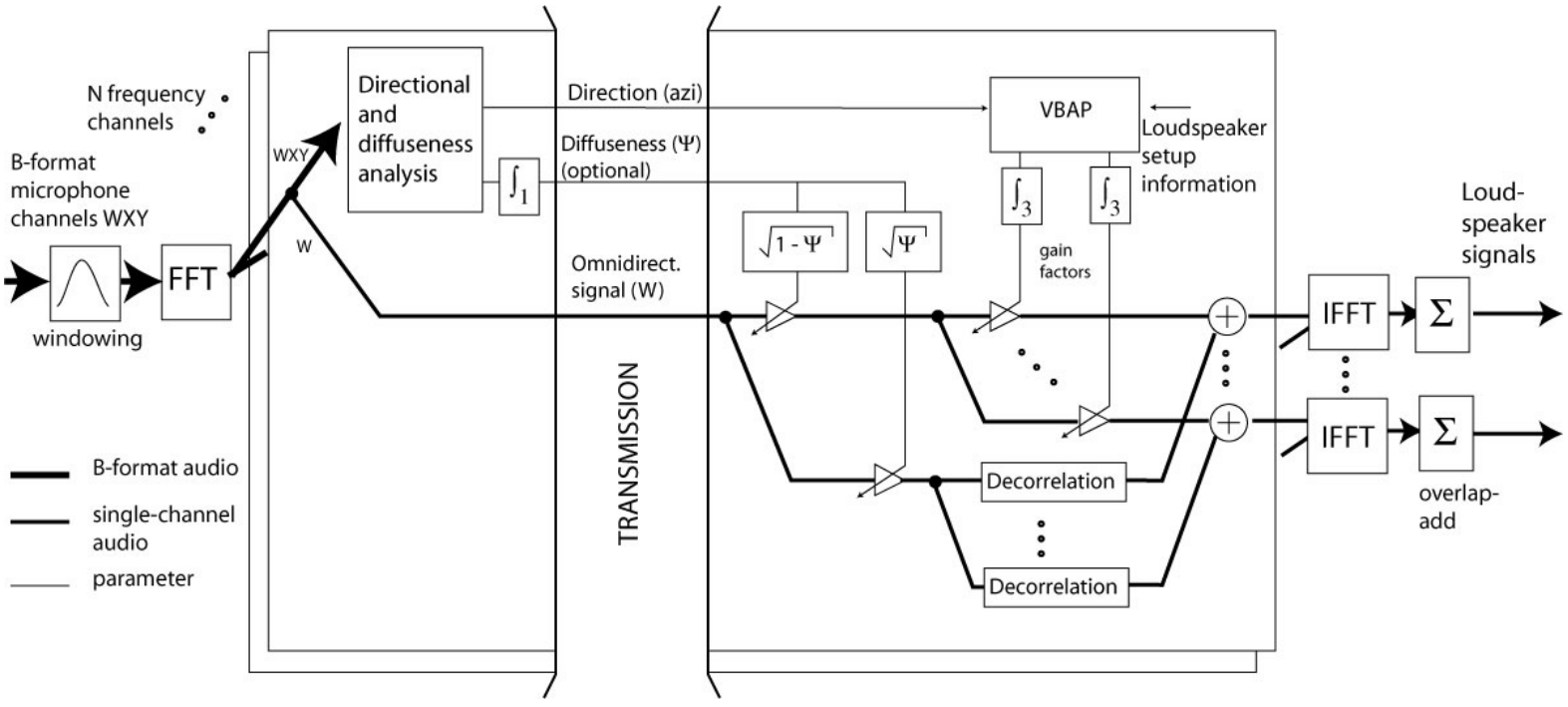
- Capture the sound
- Analyze spatial parameters
- Reproduce the sound in a way which recreates the spatial parameters



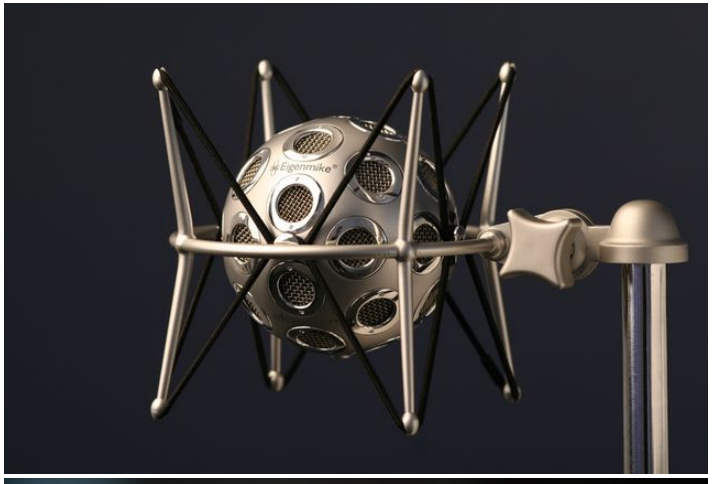
Assumptions in DirAC

- Assumption 1: listener is able to localize only one sound object at one time-frequency position
- Assumption 2: good reproduction quality is obtained, if we reproduce correctly the
  - direction,
  - diffuseness, and
  - spectrum of sound

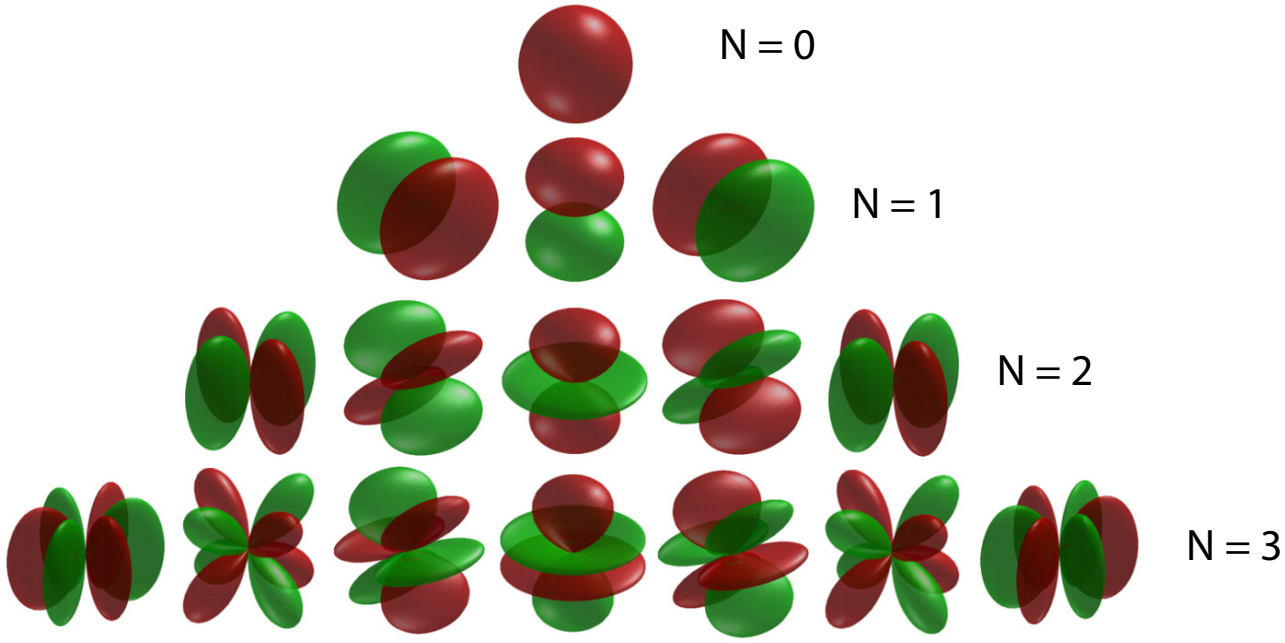
Example implementation



B-format microphones



B-format directional patterns



N denotes the order of patterns (and microphone)
0th-order (omni) microphones capture pressure signal [W]
1st-order dipole microphones capture volume velocity signals [X,Y,Z]

DirAC details, some of them

- Time-frequency transform
  - Filter banks
  - STFT
  - The system can be seen a filter changing weights fast in time, aliasing issues have to be taken into account
- $p \propto$ W is pressure signal, $\mathbf{u} \propto$ [X Y Z] is 3D velocity vector
- $\mathbf{I}_a = \Re[p^*(k,n)\,\mathbf{u}(k,n)]$(active intensity vector)
- $e = \frac{\rho_0}{2}||\mathbf{u}||^2 + \frac{|p|^2}{2\rho_0 c^2}$ (energy density)
- Direction of arrival DOA $= -\boldsymbol{I}_a$
- Diffuseness $\psi = 1 - \frac{||\mathrm{E}[\mathbf{I}_a]||}{c\mathrm{E}[e]}$
- Temporal integration of parameters
  - Short constants for DOA and Diffuseness
  - Longer for loudspeaker gains

Matlab code for directional analysis

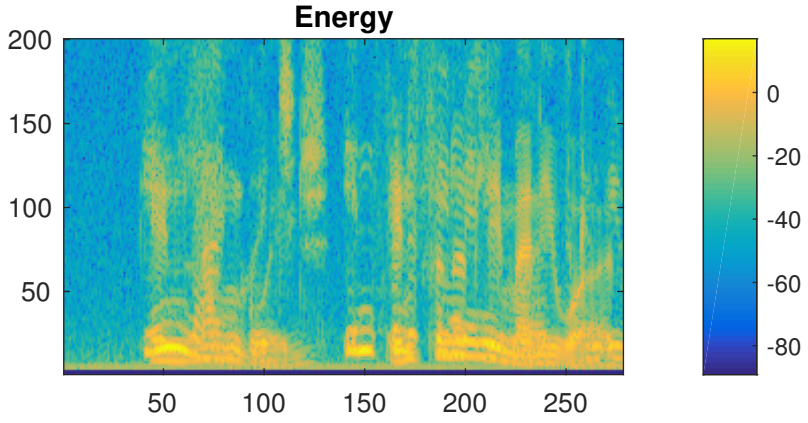
```
% diranalysis.m
% Author: V. Pulkki
% Example of directional analysis of simulated B-format recording
Fs=44100; % Generate signals
sig1=2*(mod([1:Fs]',40)/80-0.5) .* min(1,max(0,(mod([1:Fs]',Fs/5)-Fs/10)));
sig2=2*(mod([1:Fs]',32)/72-0.5) .* min(1,max(0,(mod([[1:Fs]+Fs/6]', Fs/3)-Fs/6)));
% Simulate two sources in directions of 50 and 170 degrees
w=(sig1+sig2)/sqrt(2);
x=sig1*cos(50/180*pi)+sig2*cos(-170/180*pi);
y=sig1*sin(50/180*pi)+sig2*sin(-170/180*pi);

% Add fading in diffuse noise with 36 sources evenly in the horizontal plane 43 for dir=0:10:350
    noise=(rand(Fs,1)-0.5).*(10.\^(([1:Fs]'/Fs)-1)*2));
    w=w+noise/sqrt(2);
    x=x+noise*cos(dir/180*pi);
    y=y+noise*sin(dir/180*pi);
end
hopsize=256; % Do directional analysis with STFT
winsize=512; i=2; alpha=1./(0.02*Fs/winsize);
Intens=zeros(hopsize,2)+eps; Energy=zeros(hopsize,2)+eps;
```

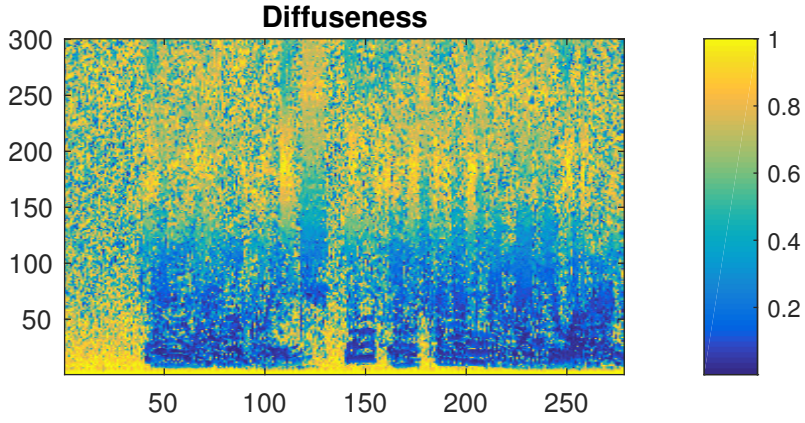
Pulkki, Ville, Tapio Lokki, and Davide Rocchesso. "Spatial effects." DAFX: Digital Audio Effects, Second Edition (2011): 139-183.

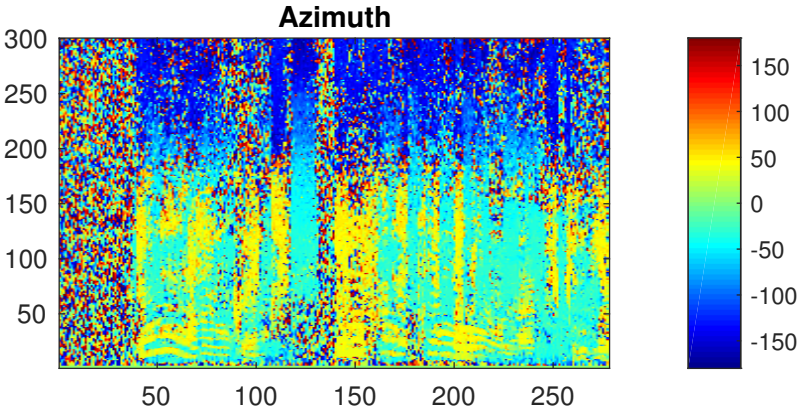
Matlab code for directional analysis

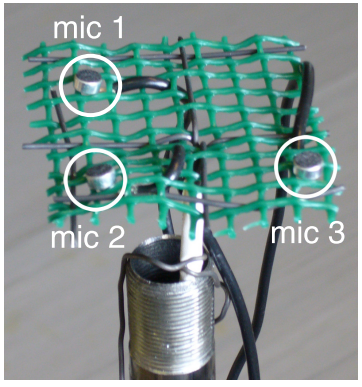
```
for time=1:hopsize:(length(x)-winsize)
        % moving to frequency domain
        W=fft(w(time:(time+winsize-1)).*hanning(winsize));
        X=fft(x(time:(time+winsize-1)).*hanning(winsize));
        Y=fft(y(time:(time+winsize-1)).*hanning(winsize));
        W=W(1:hopsize);X=X(1:hopsize);Y=Y(1:hopsize);

        %Intensity computation
        tempInt = real(conj(W) * [1 1 ] .* [X Y])/sqrt(2);%Instantaneous
        Intens = tempInt * alpha + Intens * (1 - alpha); %Smoothed

        % Compute direction from intensity vector
        Azimuth(:,i) = round(atan2(Intens(:,2), Intens(:,1))*(180/pi)); %Energy computation
        tempEn=0.5 * (sum(abs([X Y]).^2, 2) * 0.5 + abs(W).^2 + eps);%Inst
        Energy(:,i) = tempEn*alpha + Energy(:,(i-1)) * (1-alpha); %Smoothed

        %Diffuseness computation
        Diffuseness(:,i) = 1 - sqrt(sum(Intens.^2,2)) ./ (Energy(:,i)); i=i+1;
end

% Plot variables
figure(1); imagesc(log(Energy)); title('Energy');
set(gca,'YDir','normal'); xlabel('Time frame'); ylabel('Freq bin');
figure(2); imagesc(Azimuth);colorbar;
set(gca,'YDir','normal') title('Azimuth'); xlabel('Time frame'); ylabel('Freq bin');
figure(3); imagesc(Diffuseness);colorbar;
set(gca,'YDir','normal'); title('Diffuseness'); xlabel('Time frame'); ylabel('Freq bin');
```

Example with low-end 3-microphone 2D array





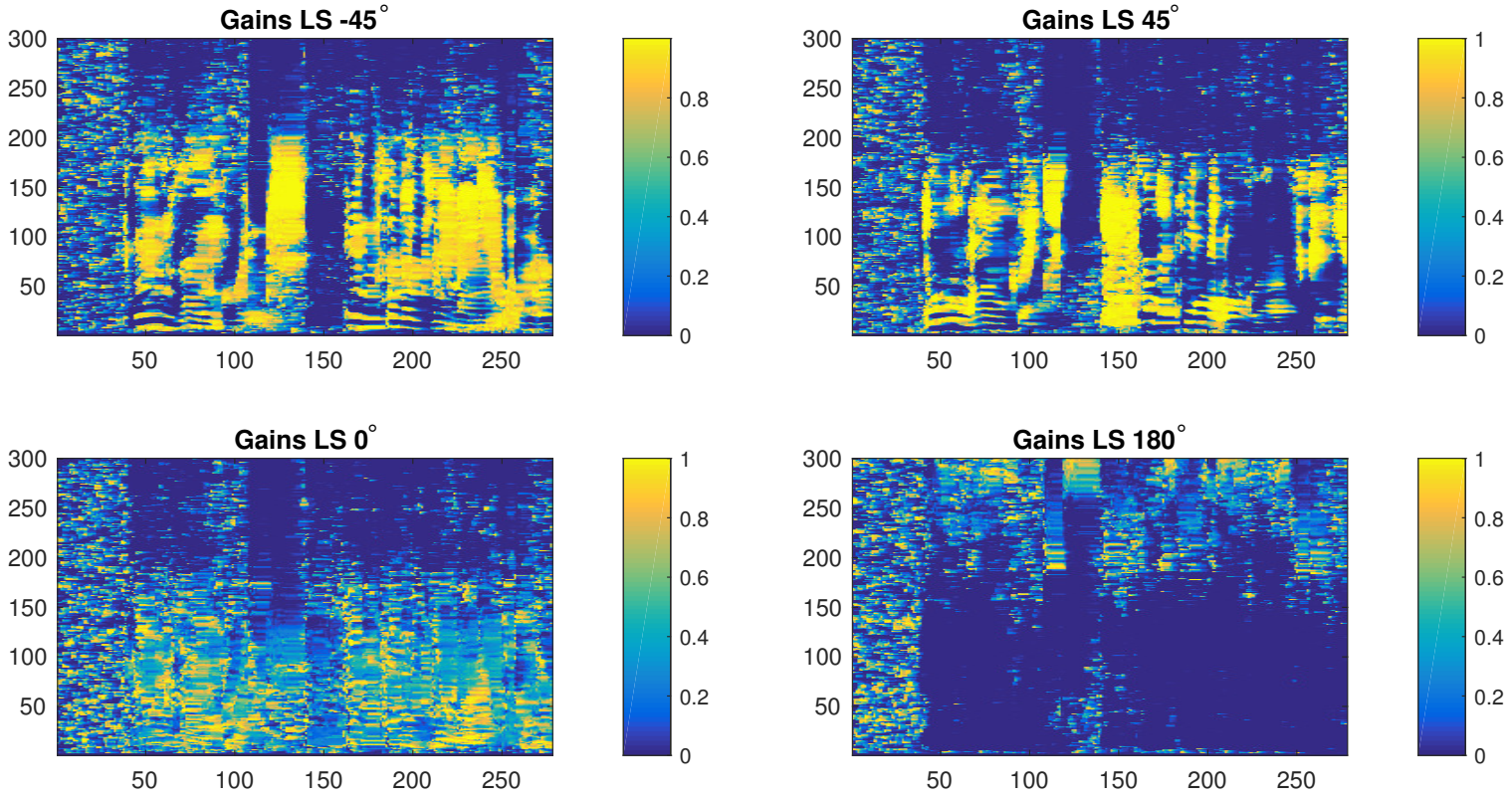




Two speech sources in $\pm 45°$, anechoic chamber

▷ Sound captured with one of the microphones

Examples of soft masks for non-diffuse stream



Rendering to 8-channel octagonal loudspeaker setup

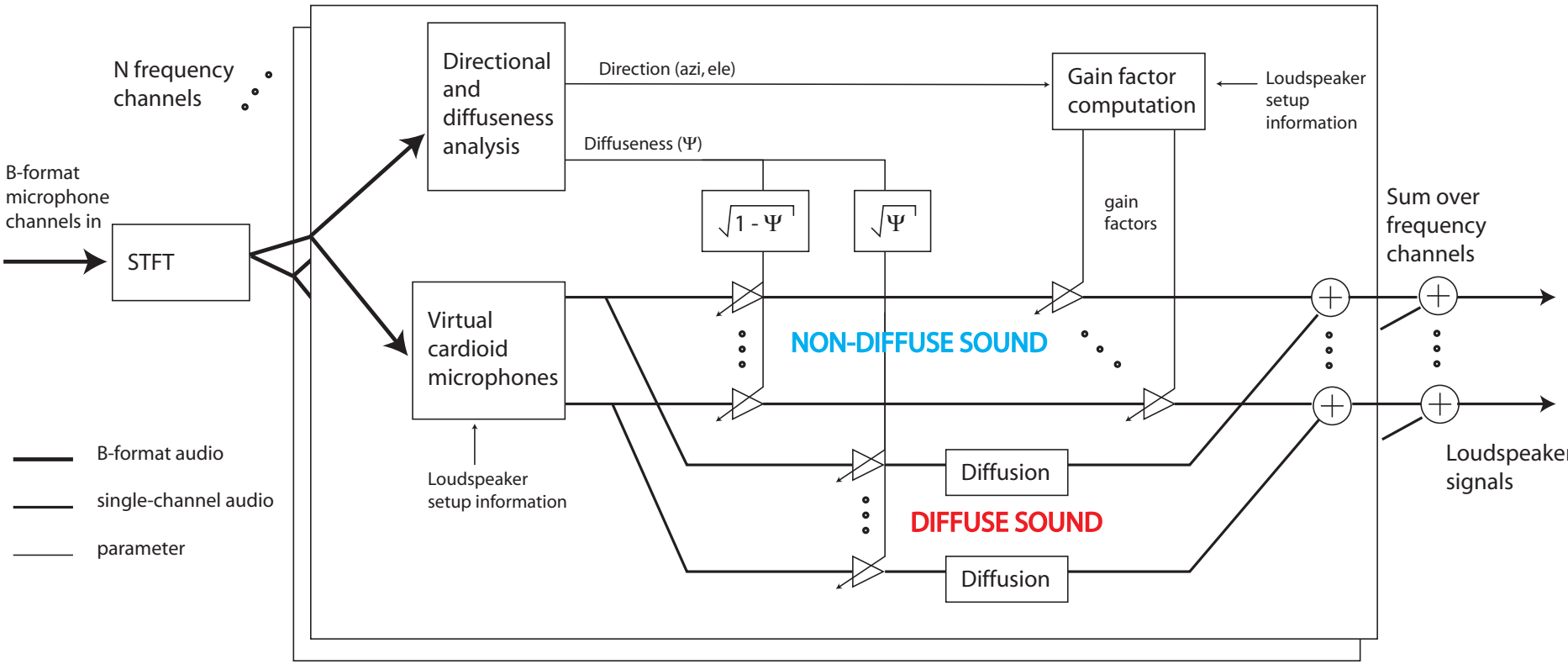
▷ Mono ▷ ND2ch ▷ D2ch ▷ ND+D2ch ▷ ALL CHAN

Properties of teleconference-DirAC

- Very good quality with spectrally non-overlapping sources in free field
- Diffuse reverberation subject to spatial and timbral artifacts
- Sources on opposite sides of the microphone with spectral overlap
  - Severe model mismatch
  - Timbral artifacts: "added room effect", "smearing of transients"
  - Spatial artifacts: "sources pull each other"
  - Sources near noise masking threshold: impossible to localize

Why these artifacts?

"HQ" implementation



This works better, but dont exactly know why (2007).

DirAC research 2004-2019

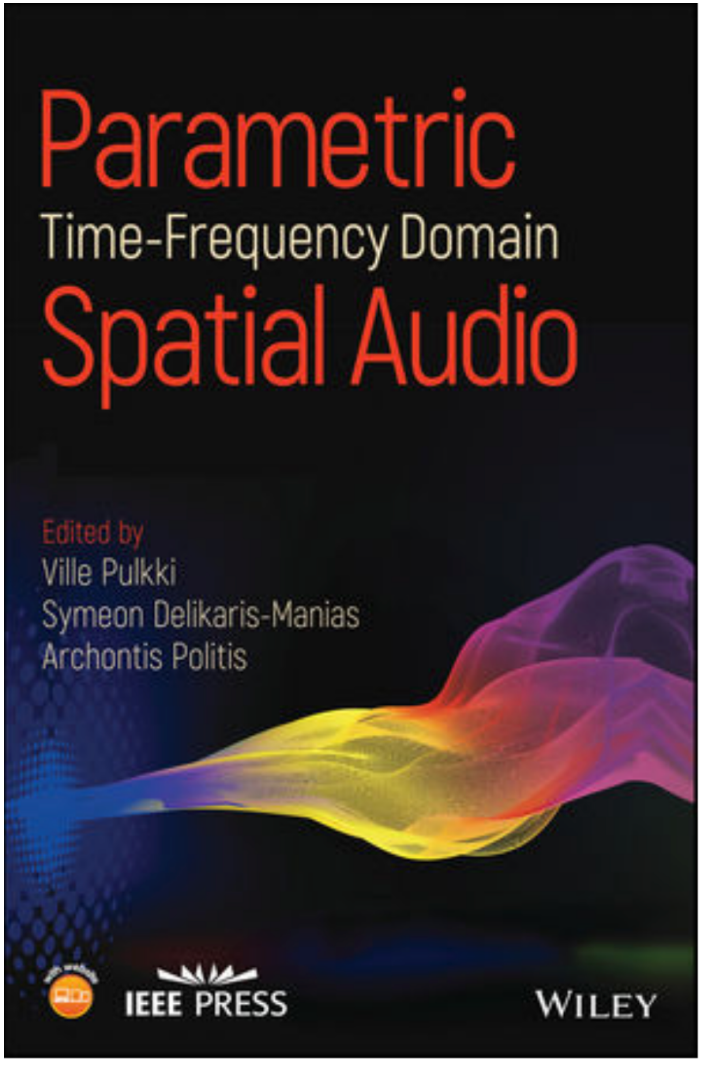
- Work on microphone arrays
- Headphone reproduction / Loudspeaker reproduction
- Virtual reality audio engine
- Methods to optimize audio quality

Related spatial audio techniques

- Coding of 5.1 audio into stereo+metadata and back to 5.1 (MPEG Surround)
- Upmixing of stereo recordings into surround
- New types of directive microphones
- Spatial audio effects
- Rendering of room impulse responses into loudspeaker sets

Book

- 15 chapters, 416 pages
- Matlab code
- Published Dec 2017



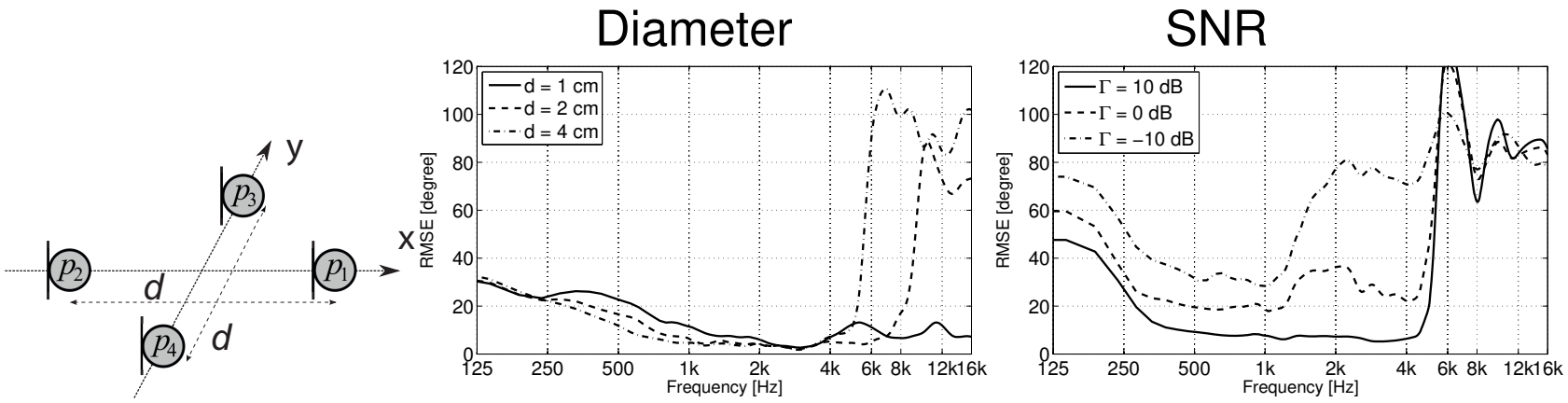
Summary

- Amplitude panning
  - apply sound with different gains to loudspeakers
  - Principles, explanations, quality considerations
- Parametric time-frequency-domain spatial audio
  - Analyze the properties of sound field in TF-domain, and utilize the analyzed parameters in rendering of spatial sound
  - Directional audio coding

Limitations of differential arrays

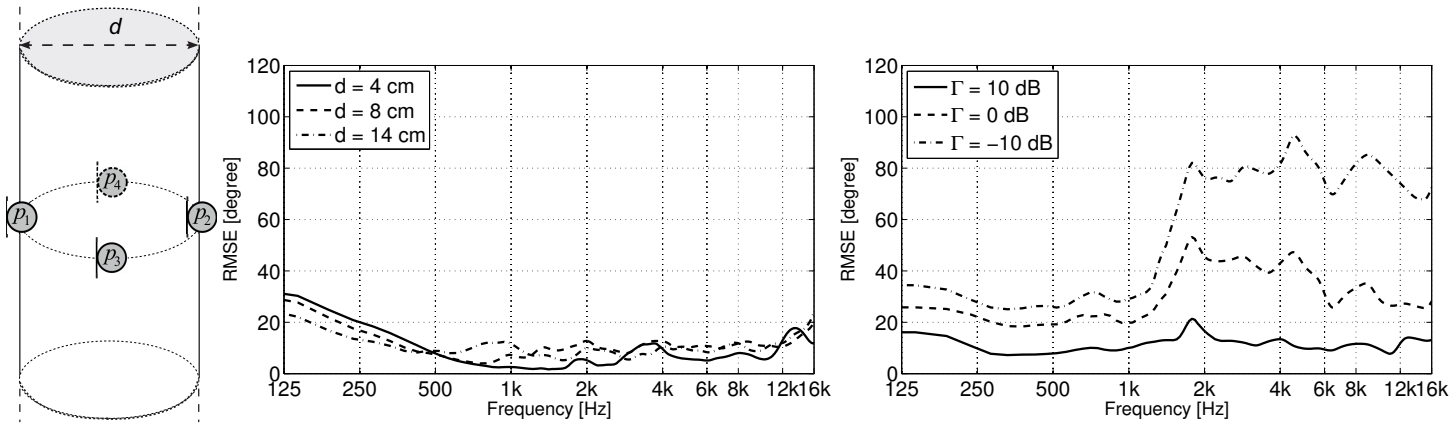
- Low-frequency noise
  - Instable direction and diffuseness estimation
  - Mitigated with temporal integration
- Spatial aliasing
  - Highly biased directional values
  - In some cases can be mitigated

Square array



- Pressure gradient
- Square arrays of omni microphones, B-format microphones
- LF noise, HF aliasing
- Ahonen, del Galdo et al [JAES 2012]

Arrays with shadowing



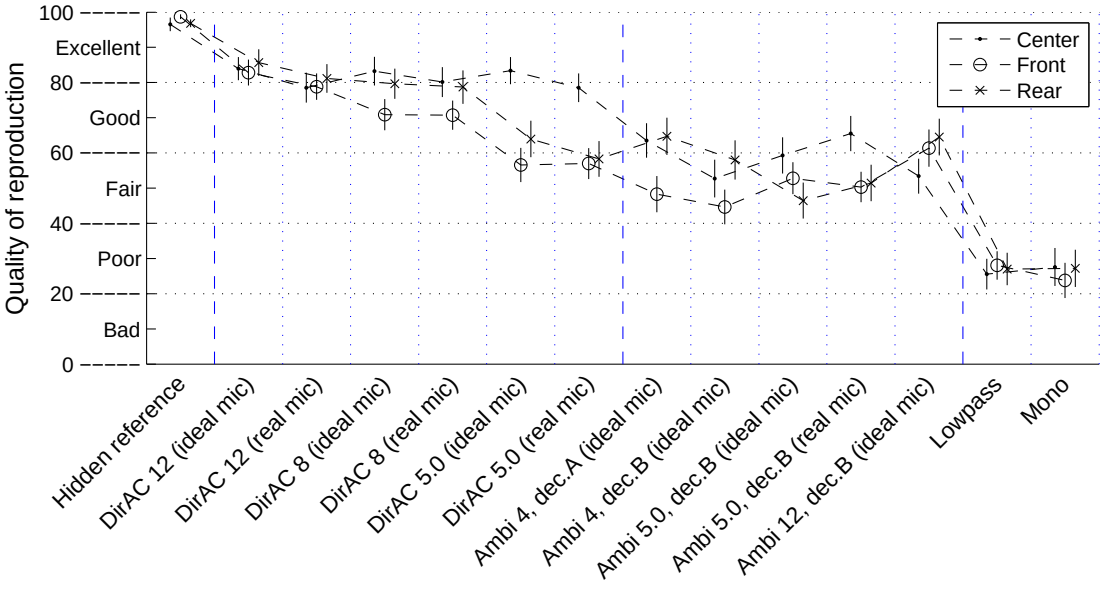
$$\widetilde{i}_x(n,k) = |p_1(n,k)|^2 - |p_2(n,k)|^2,$$
$$\widetilde{i}_y(n,k) = |p_3(n,k)|^2 - |p_4(n,k)|^2,$$

- When capsule signals are available, and some shadowing takes place
- LF+MF: pressure gradient
- HF: energy gradient
- A-format microphones, cylinder arrays and spherical arrays

Other microphone arrays with DirAC

- Multiple microphones in array: ESPRIT by Thiergart, Kratschmer et al, [AES Conv 2011]
- Two microphones, cross-correlation: Kratschmer, Thiergart et al [AES Conv 2012]
- Basically any DOA analysis method can be applied

HQ-DirAC subjective tests



- HQ-DirAC, comparison to reference scenario with 24 loudspeakers
- Largest issues with spatially complex scenarios audible as small timbral artifacts
- Vilkamo [JAES 2009]

Sources of timbral artifacts

- Diffuse sound leaks into non-diffuse stream
  - This has not found to be a problem
- Non-diffuse sound leaks into diffuse stream
  - Major problem
  - Transients are decorrelated, causing annoying smearing
  - Direct sound is decorrelated, "added room effect", or "sources are perceived too far" issues
- Target for development: minimize decorrelated energy!

How to avoid decorrelation

- Processing of transients separately
  - Recognize transients
  - Use better time resolution / bypass decorrelation [Laitinen, Kuech et al. 2011]
- Covariance-domain processing
  - minimize decorrelated energy
- Divide sound field into sectors from higher-order recording
  - perform separate analysis for each sector
- Perform more elaborate analysis to sound field (e.g., multiple DOA values), Thiergart [IEEE TASLP, 2014]

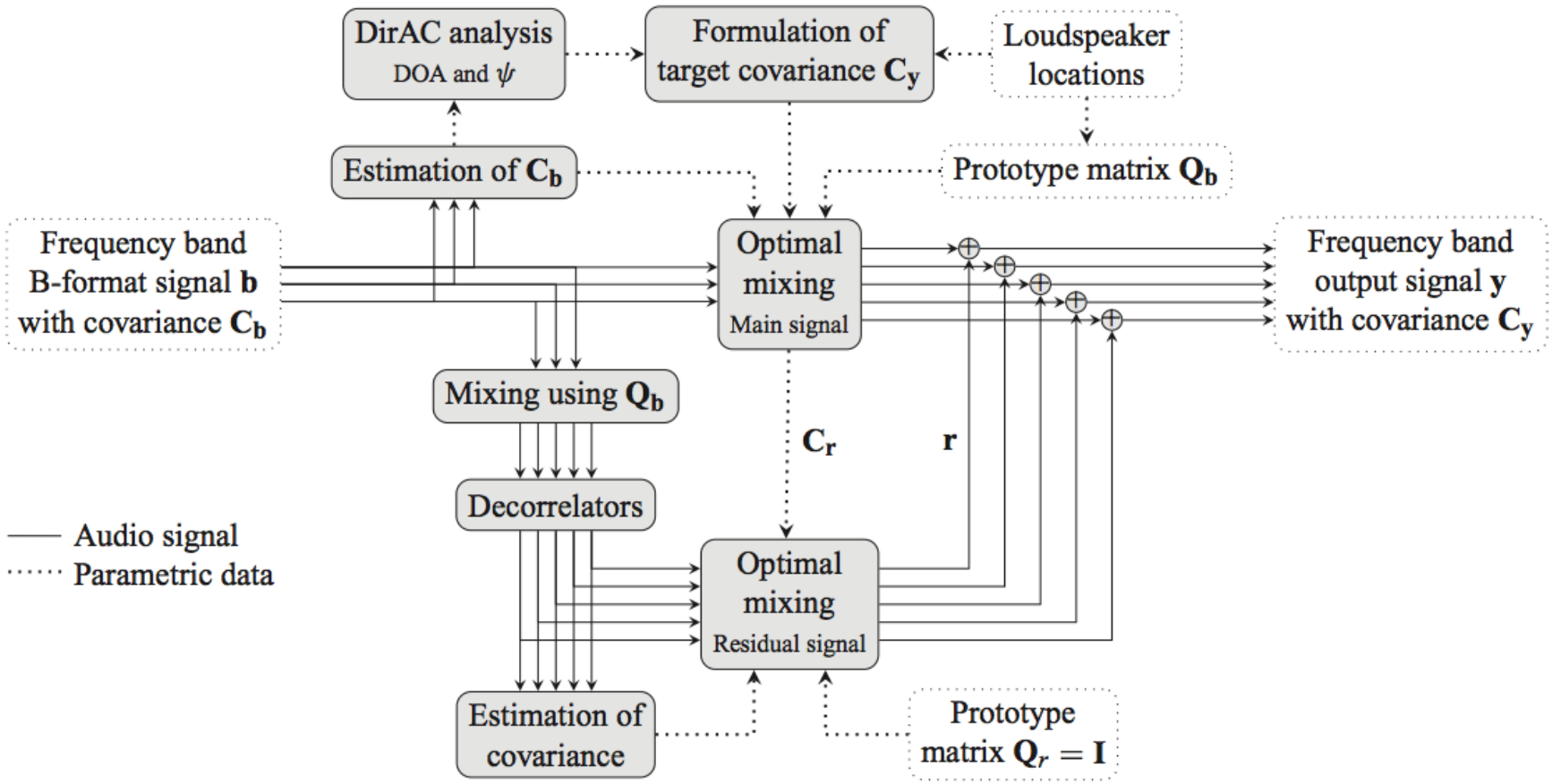
Covariance-domain processing

Least-squares optimized solution for synthesis

- the covariance matrix of output is dictated by directional parameters
- optimized mixing solution leads to minimization of decorrelated energy

[Vilkamo, Bäckström, Kuntz: JAES 2013]

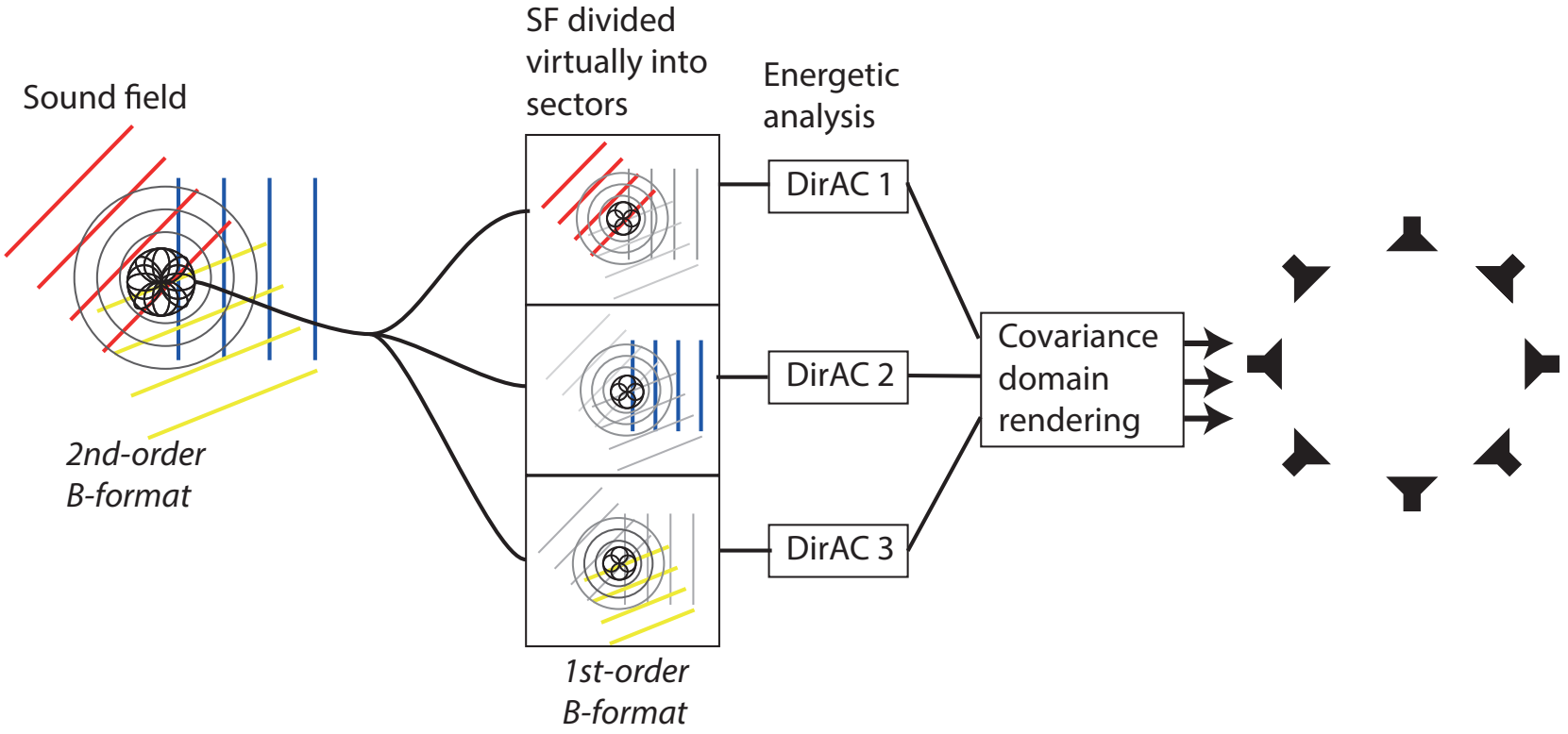
Covariance-domain processing



Solutions with different model of the sound field

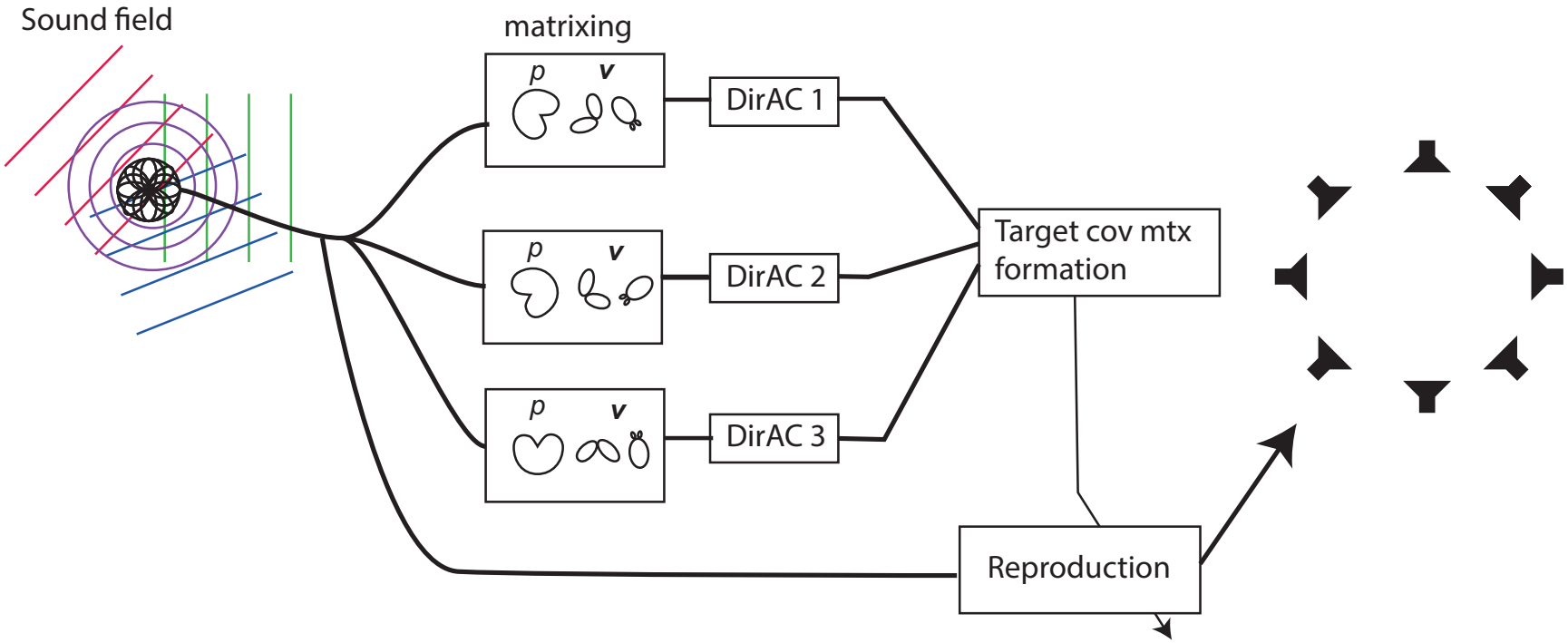
- Higher number of microphones gives more information about sound field
- How to use that information in sound reproduction?
- Divide sound field into sectors (Pulkki, Politis), perform lower-order reproduction for each
- Analyze multiple DOAs, and then reproduce (Thiergart & Habets, Mouchtaris group, Berge)

Sector-based parametric spatial sound reproduction



[Politis et al: IEEE J. Selected Topics Sig Proc 9.5 (2015)]

Sector-based parametric spatial sound reproduction

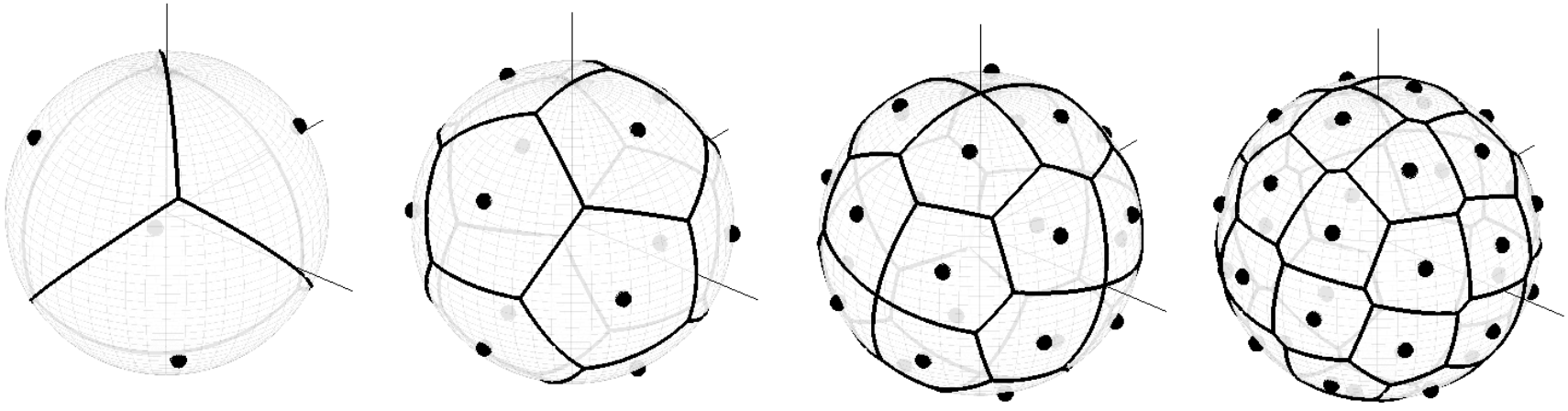


Sector-based parametric spatial sound reproduction

"Higher-order DirAC"

- Challenging acoustical conditions occur rarely within sectors
- Parameters computed with N:th -order input
- Audio signals used in synthesis obtained with (N-1):th -order input
- Self-noise issue of higher-order microphones are also avoided
- System does not lose acoustic energy in any case

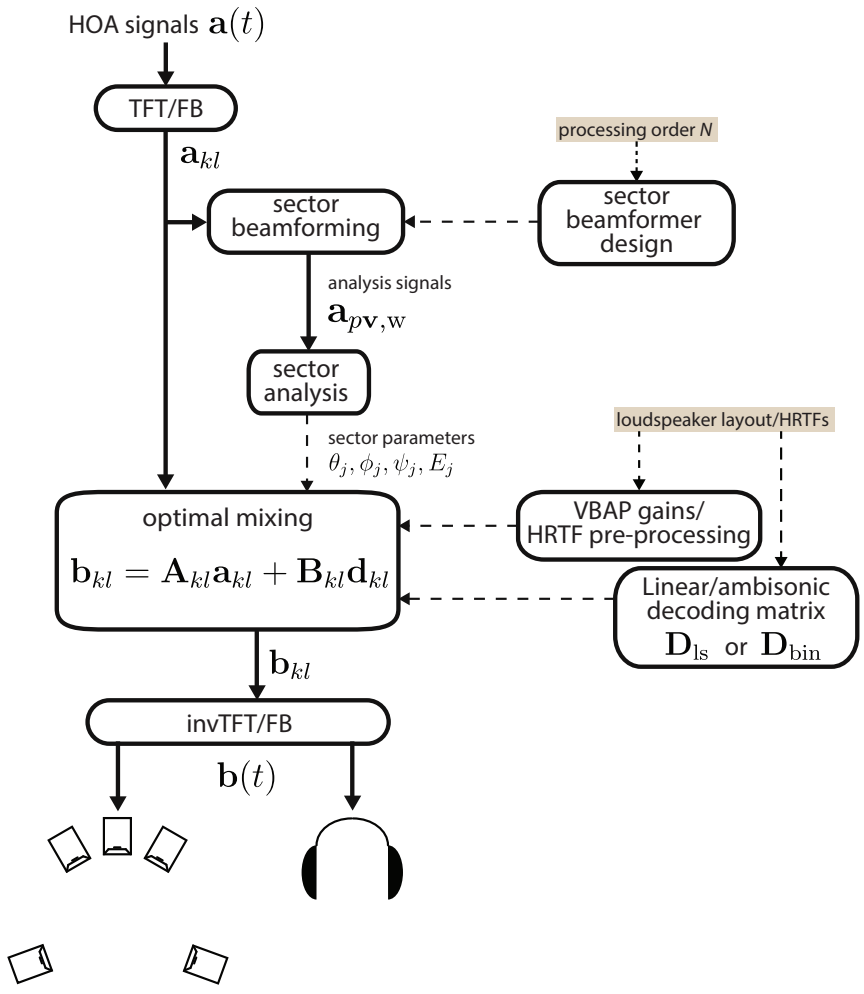
Sectors for HO-microphones



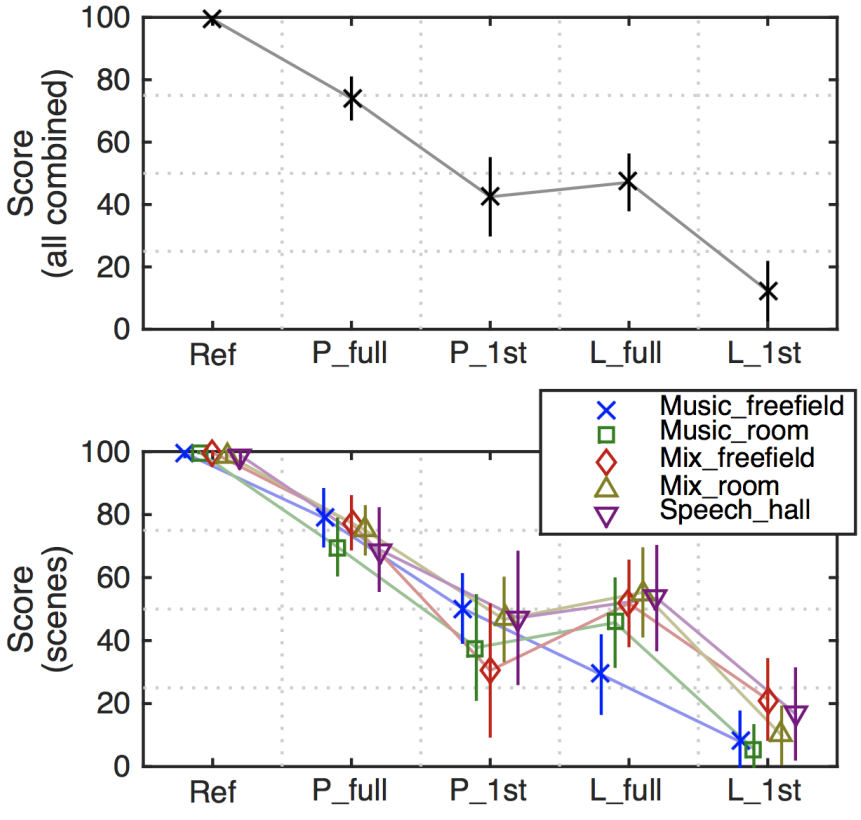
2nd 3rd 4th 5th

Different frequency bands utilize different number of sectors

Processing


HOA signals $\mathbf{a}(t)$
TFT/FB
$\mathbf{a}_{kl}$
processing order *N*
sector beamforming
sector beamformer design
analysis signals
$\mathbf{a}_{p\mathbf{v},\mathrm{w}}$
sector analysis
loudspeaker layout/HRTFs
sector parameters
$\theta_j, \phi_j, \psi_j, E_j$
optimal mixing
$\mathbf{b}_{kl} = \mathbf{A}_{kl}\mathbf{a}_{kl} + \mathbf{B}_{kl}\mathbf{d}_{kl}$
VBAP gains/ HRTF pre-processing
Linear/ambisonic decoding matrix
$\mathbf{D}_{\mathrm{ls}}$ or $\mathbf{D}_{\mathrm{bin}}$
$\mathbf{b}_{kl}$
invTFT/FB
$\mathbf{b}(t)$

Subjective evaluation

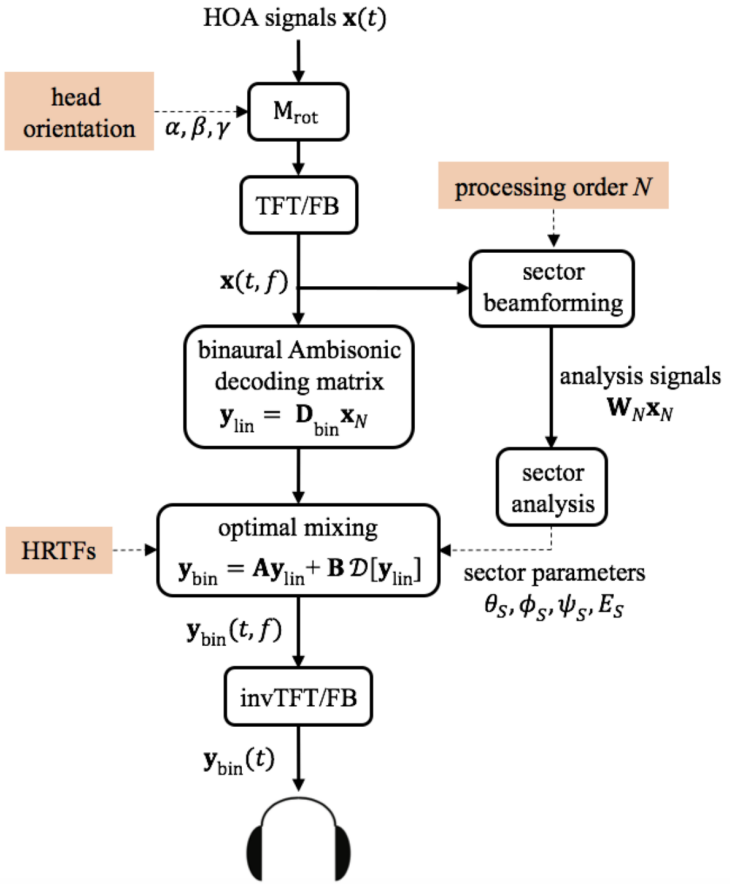


- [Politis & Vilkamo & Pulkki IEEE J. Selected Topics Sig Proc 9.5 (2015)]
- Reference: 28 loudspeakers in anechoic chamber, very challenging 3D sound environments
- Test: Eigenmic recording, playback over HO-DirAC, 1st-order DirAC, 4th-order Ambisonics, 1st-order Ambisonics

Why does HO-DirAC provide better results?

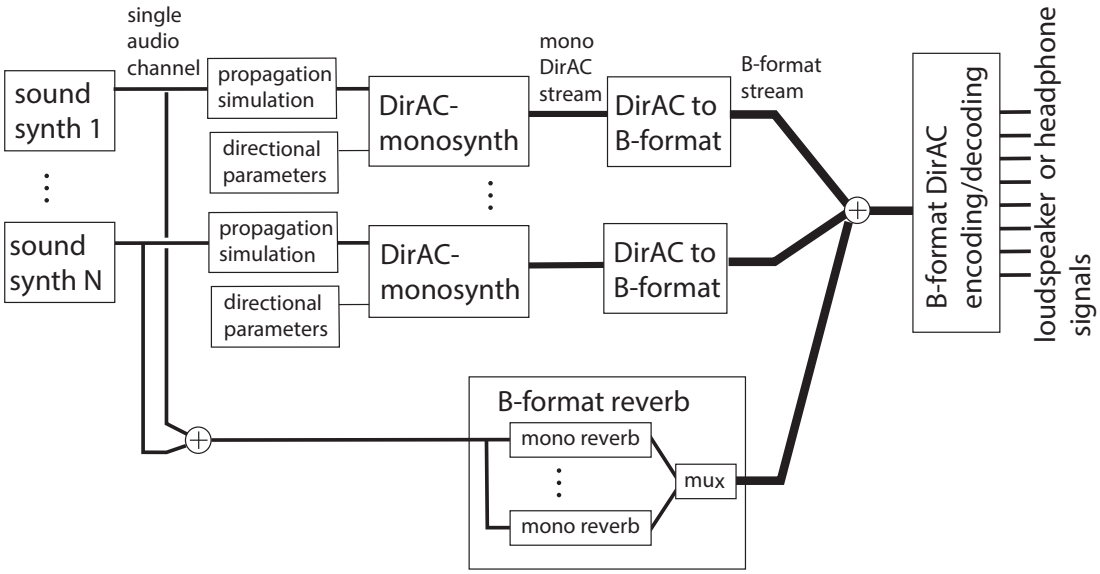
- Spatially separated plane waves sharing the same frequency are processed in different sectors
- Global diffuse field is not diffuse in individual sectors, nevertheless, the combined output is again diffuse
- Avoidance of decorrelation!

HO-DirAC for head-tracked headphones



- Processing optimized for dynamic rendering
- Sector-based computation is used to derive covariance matrices
- WASPAA 2017: Wed 10:30–12:30 *Enhancement of ambisonic binaural reproduction...* Politis, McCormack, Pulkki
- **DEMOS AVAILABLE**

DirAC as virtual reality audio engine



- Control spatial extent of virtual sources
- With headphones: Creation of external - internal sources
- Loudspeaker-setup-independent reverberator
- Efficient transmission of spatial sound