Aalto University
School of Science

Digital Audio Effects 2016
Brno, Czech Republic
8th of September

Keynote

Parametric Spatial Room Impulse Response Analysis and Synthesis: A High-Resolution Approach

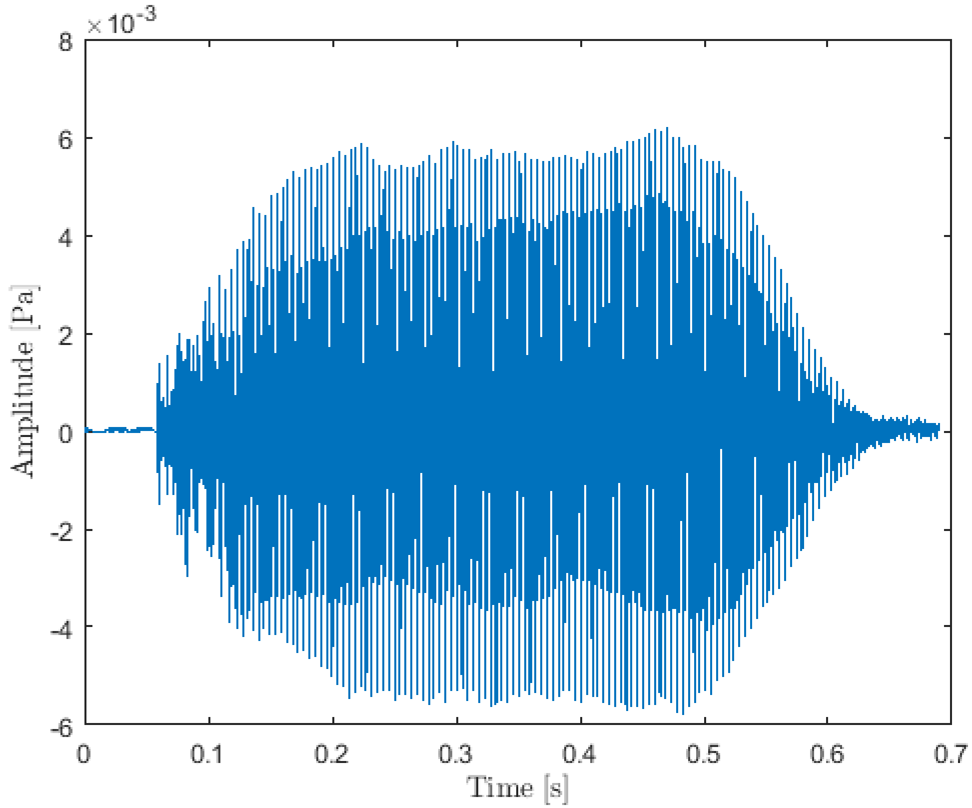
Sakari Tervo, Post-doctoral researcher
Department of Computer Science
Aalto University School of Science, Finland

Introduction to parametric estimation

Aalto University
School of Science

Parametric estimation

Consider a note played with a bowed violin

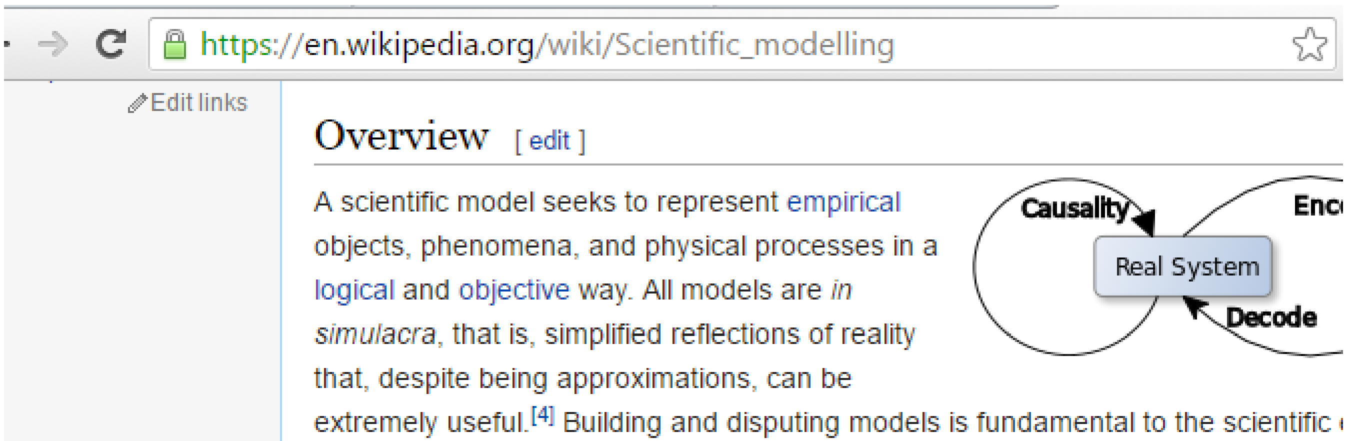


The player played the note A3

Aalto University
School of Science

The model

Parametric estimation requires a model of the physical world.



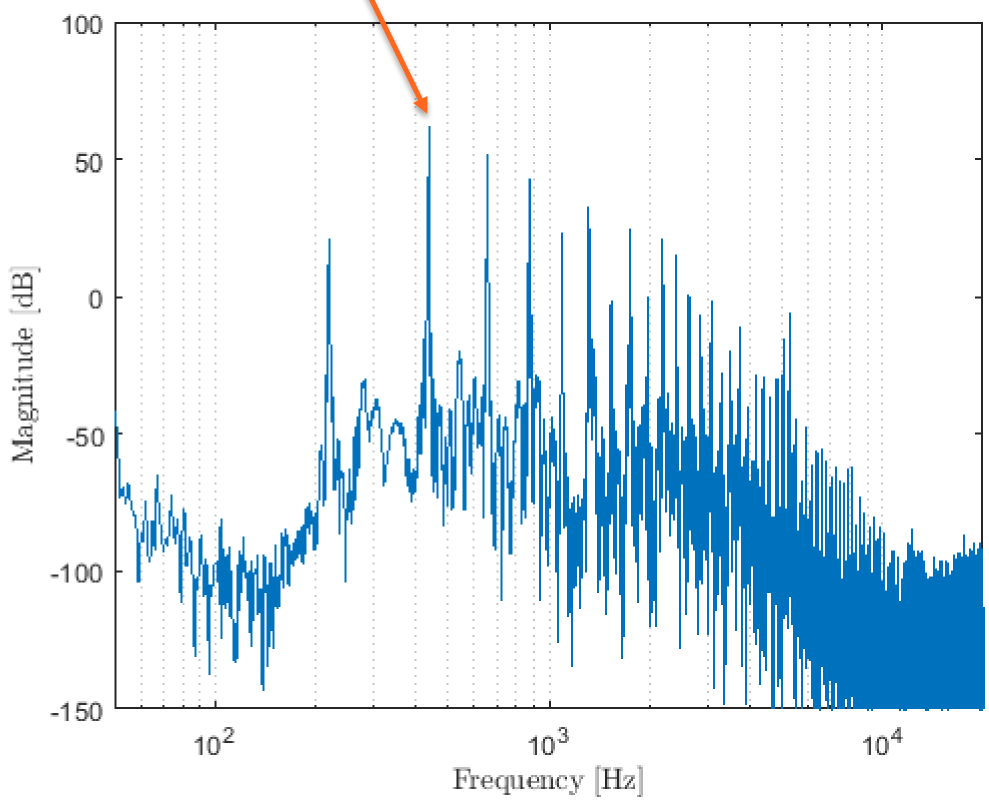
What defines a note?

Is it the fundamental frequency?

Which model describes the fundamental frequency?

Parametric estimation

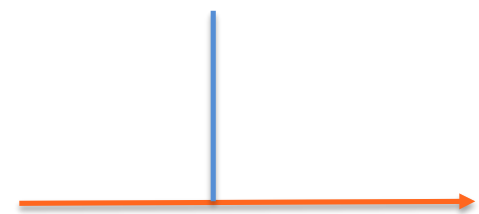
F0 ~ 440 Hz



Model 0

"The strongest mode"

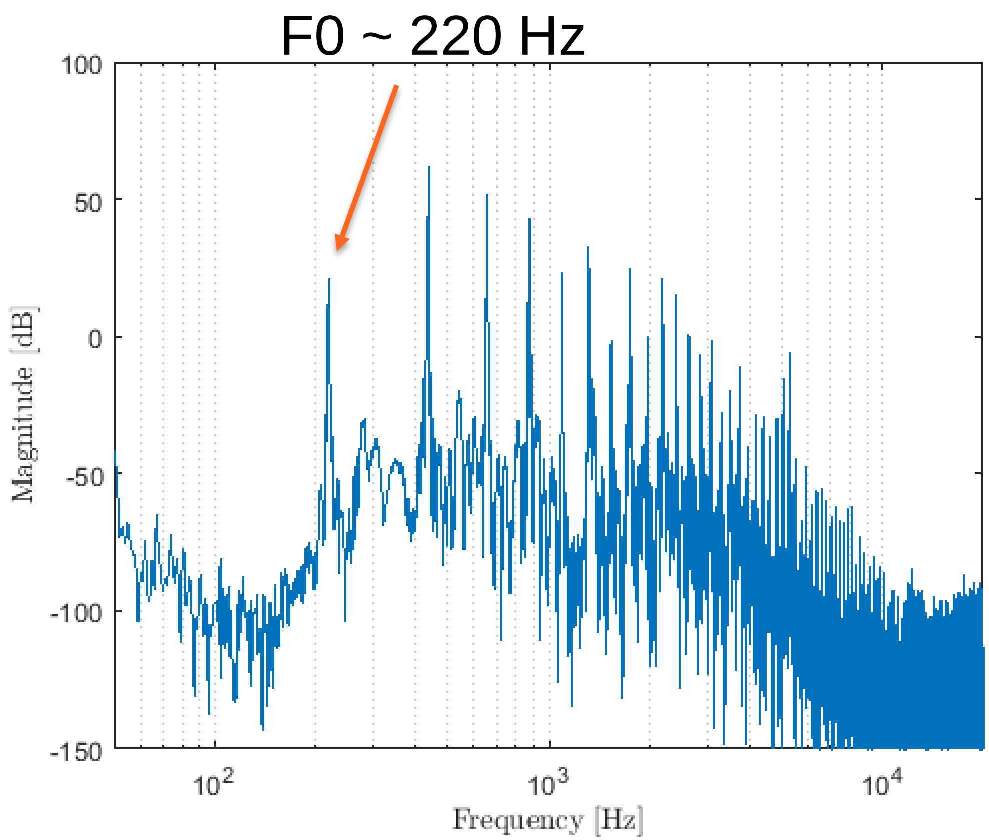
Probability



Frequency

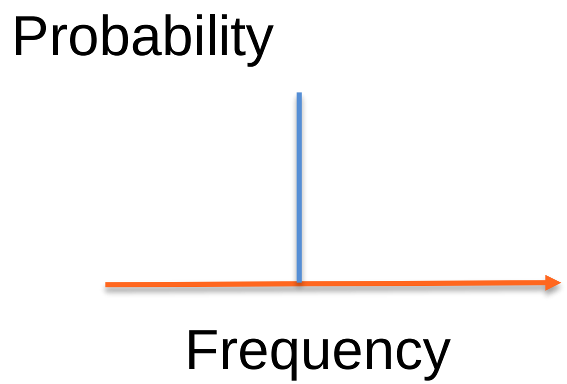
Parametric estimation

Refining the model



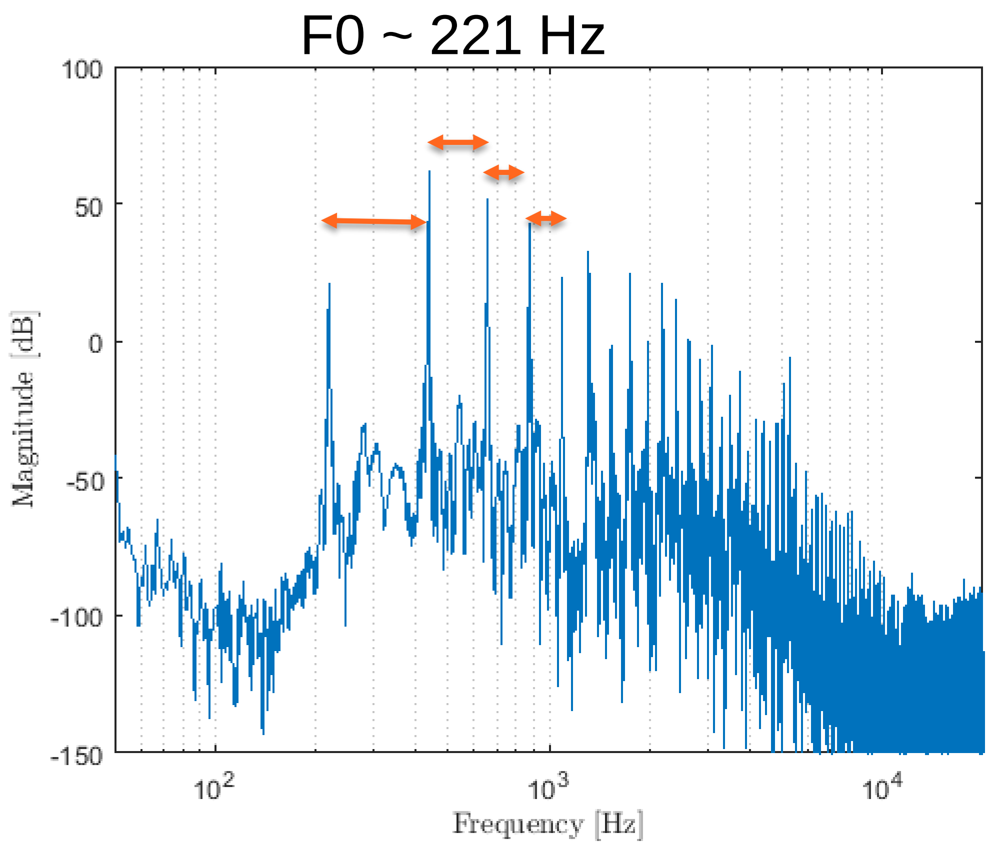
Model 1

”First strong mode”



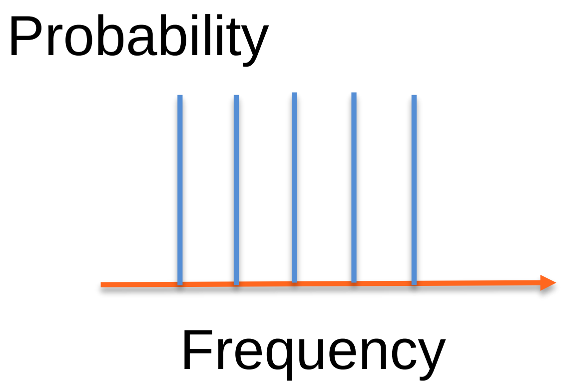
Parametric estimation

Using a more accurate signal model



Model 2

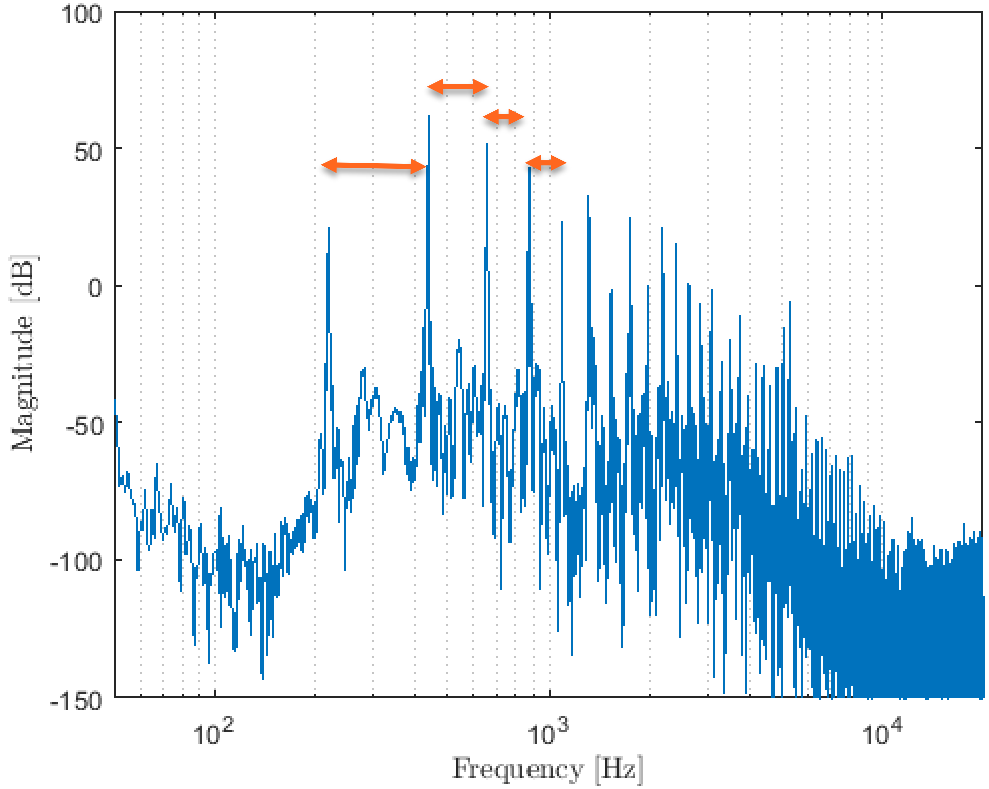
"The average difference between strong modes"



Parametric estimation

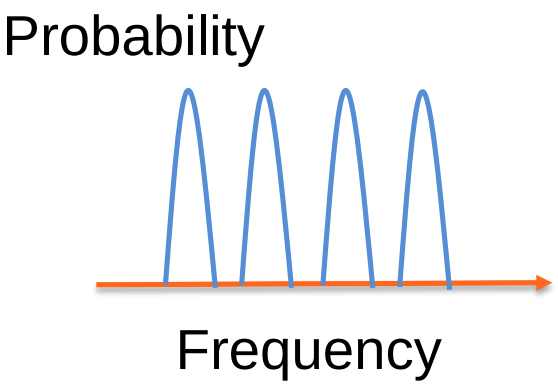
Using a probability distribution

F0 ~ 221.5 Hz



Model 3

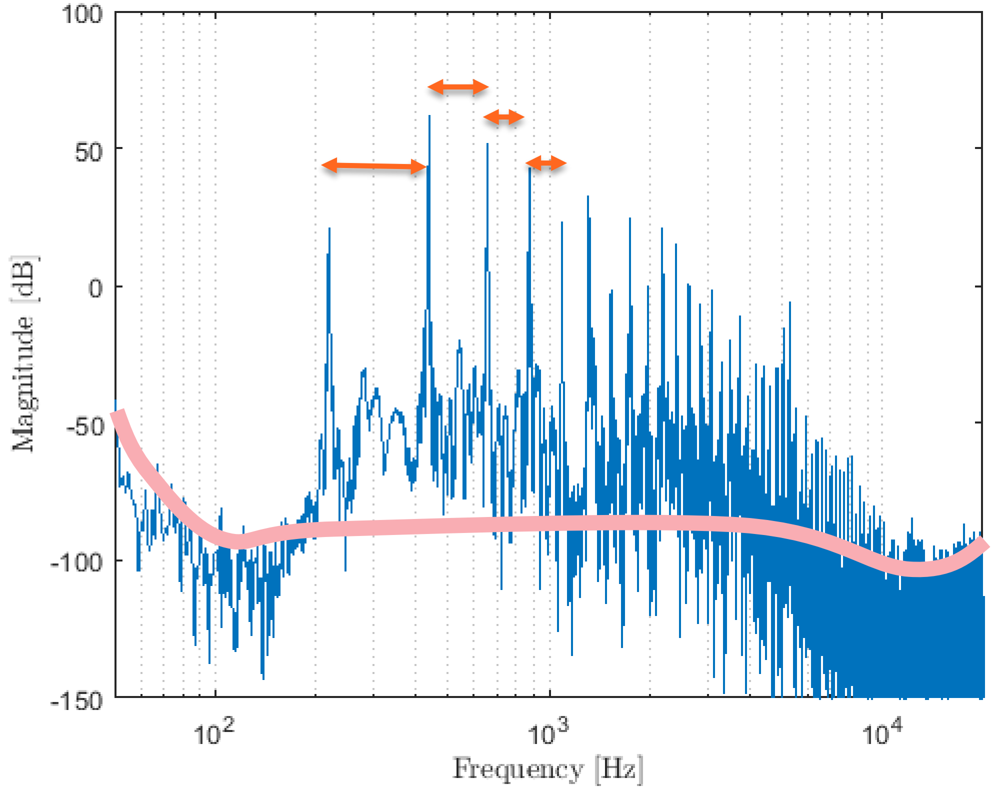
”The average difference between the mean of strong modes”



Parametric estimation

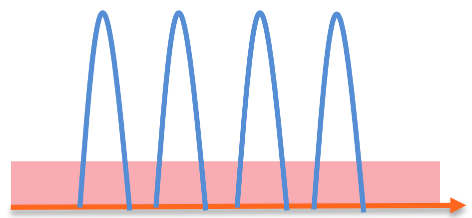
Using a probability distribution with a noise

F0 ~ 221.7 Hz



Model 4

”The average difference between the mean of strong modes, given the background noise”



Parametric estimation

https://www.youtube.com/watch?v=6JeyiM0YNo4

Sakari.Tervo@aalto.fi

Parametric estimation: Conclusions

Requires a model for the *signal* and the *noise*

The model is defined by a set of parameters, which are to be estimated.

The selected model depends on the application.

In reality, after George Box

All models are wrong, some are less wrong

Typically the trade-off is

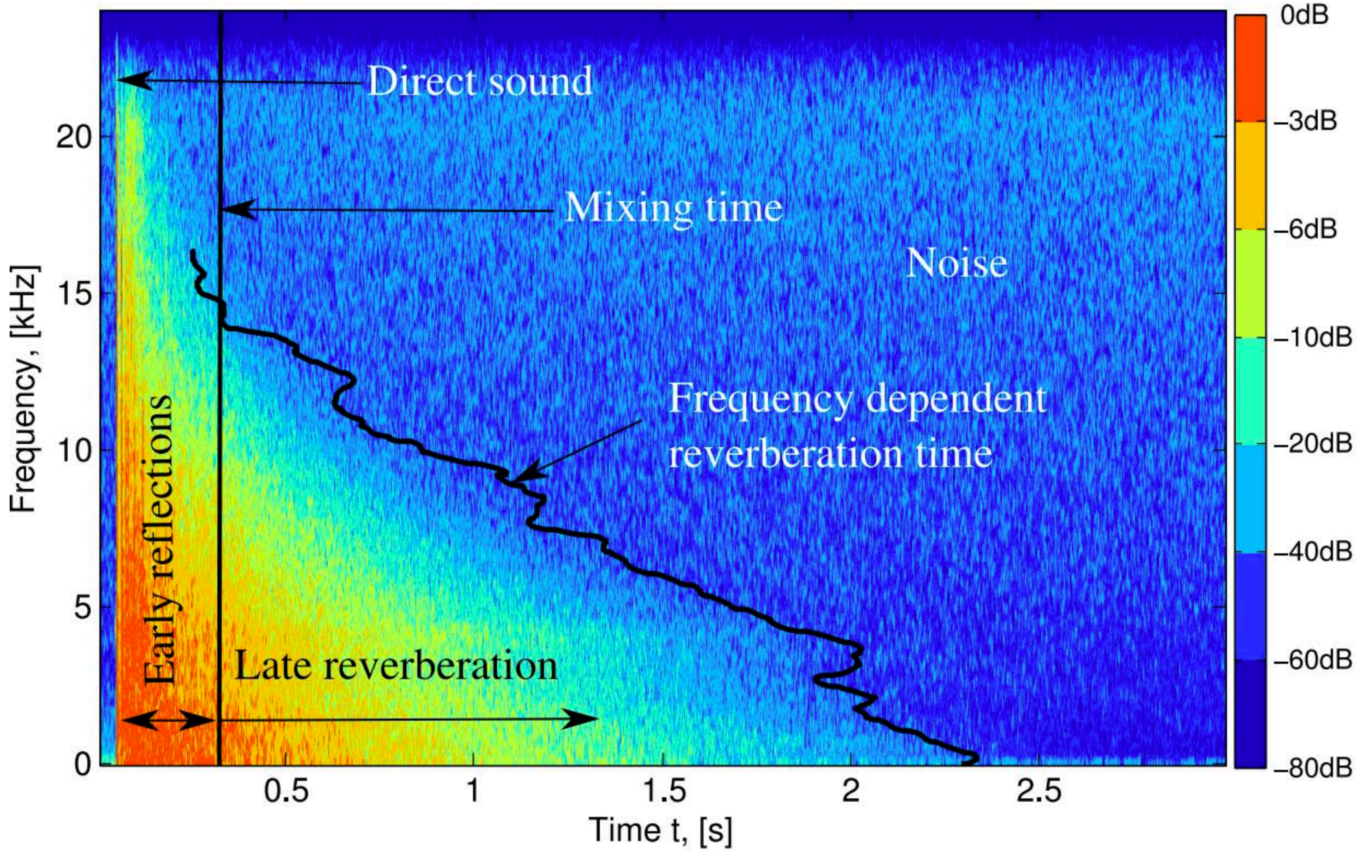
Complexity vs accuracy

Sakari.Tervo@aalto.fi

Parameterization of room impulse responses

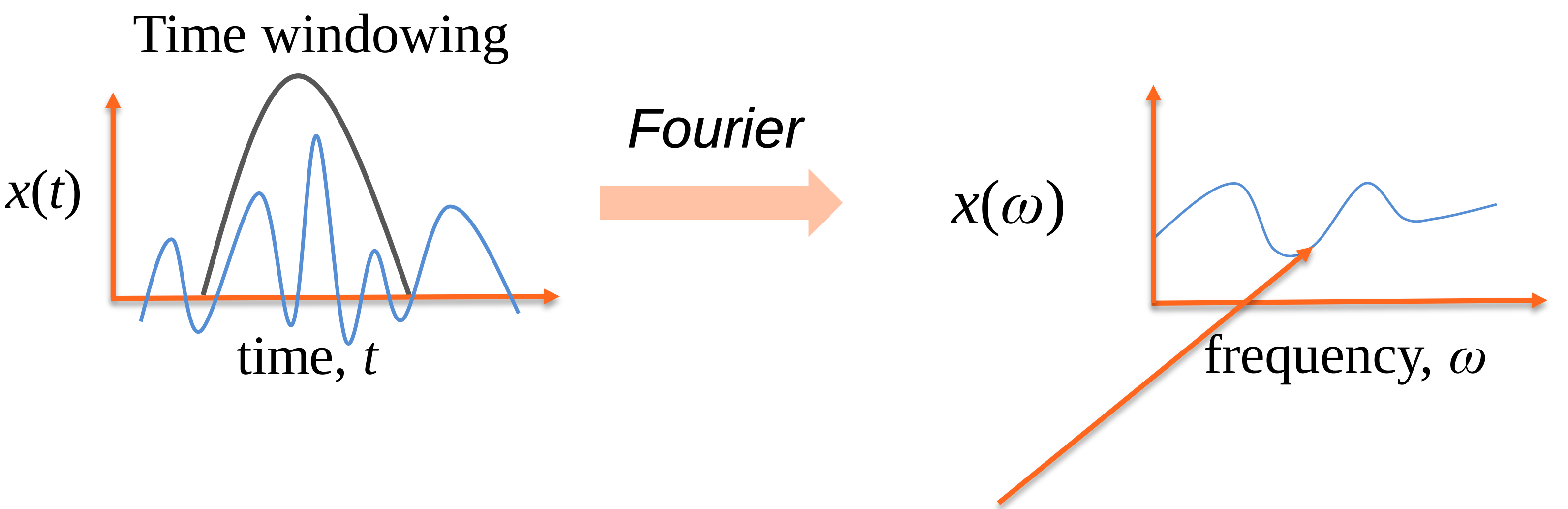
Time-frequency dependency

Room impulse responses are time and frequency dependent.



High-resolution analysis

Given a room impulse response x, what is the highest time-frequency resolution the analysis can have?



Inspect the signal in one time and frequency instant $x(t_0, \omega_0)$

The time window of length L should include at least three wavelengths of the inspected frequency $L > 3 \times 2\,\pi/\omega$

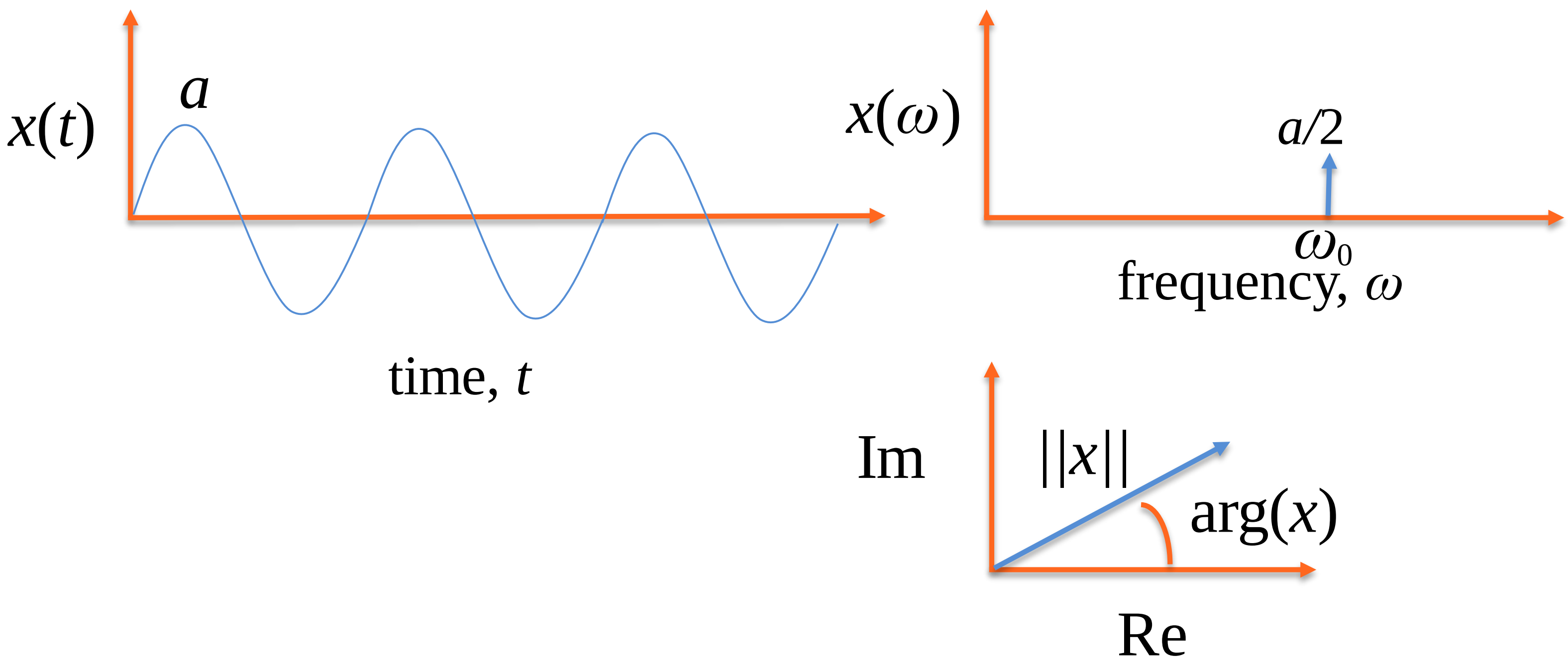
Sakari.Tervo@aalto.fi

High-resolution analysis

The signal at (t_0, ω_0) in the continuous domains is therefore

$x(t) = a \sin(\omega_0 t)$ in the time-domain and

$x(\omega) = a \exp(-i\omega_0 + \varphi_0)\, \delta(\omega - \omega_0)$ in the frequency domain.

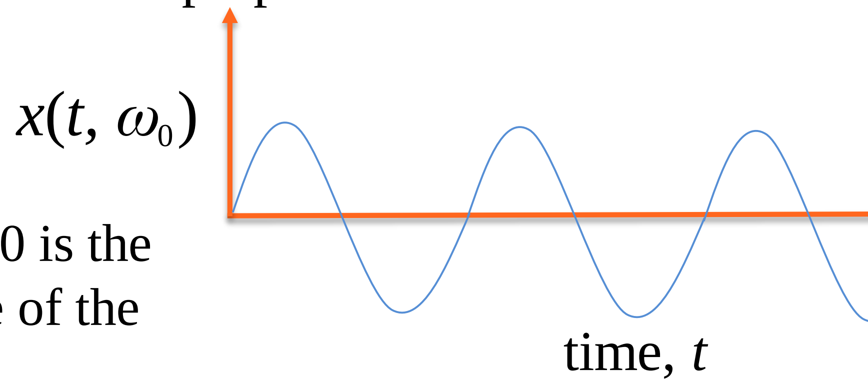


Sakari.Tervo@aalto.fi

Parameterization of a room impulse response

What kind of model can we assume for a time instant at one frequency? Total response is expressed as the superposition of several sinusoids

$$x(\omega_0) = \sum_i c_i(\omega_0), \text{ where}$$



i is the index of the acoustic event, e.g. $i = 0$ is the direct sound, and c is the complex response of the event.

Since the room impulse response is linear we can write:

$$x(\omega) = \sum_i c_i(\omega),$$

which describes the whole impulse response as a sum of sinusoidal signals.

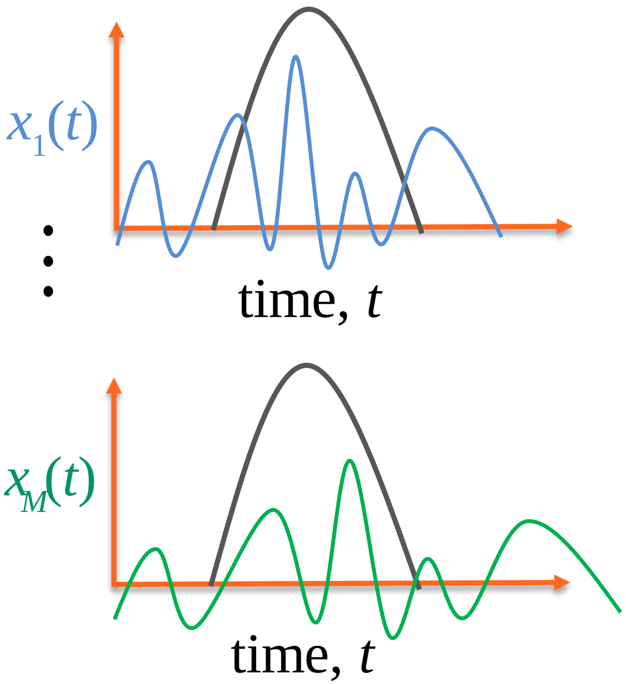
Parameterization of spatial room impulse responses

Aalto University
School of Science

Spatial room impulse response

A room impulse measured with a microphone array.

At t_0 and in ω_0 we have



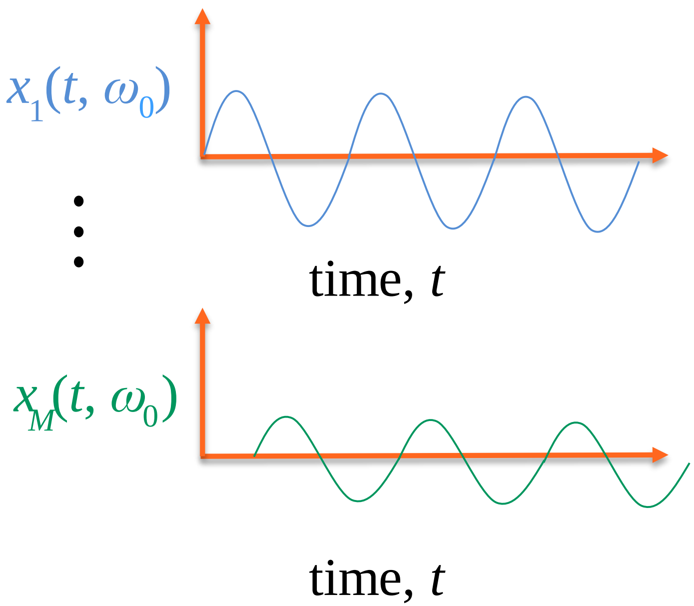
$$\begin{bmatrix} x_1(t_0, \omega_0) \\ \vdots \\ x_M(t_0, \omega_0) \end{bmatrix} = \mathbf{x}(t_0, \omega_0)$$

Where M is the number of microphones.

Image from mhacoustics.com

Spatial room impulse response

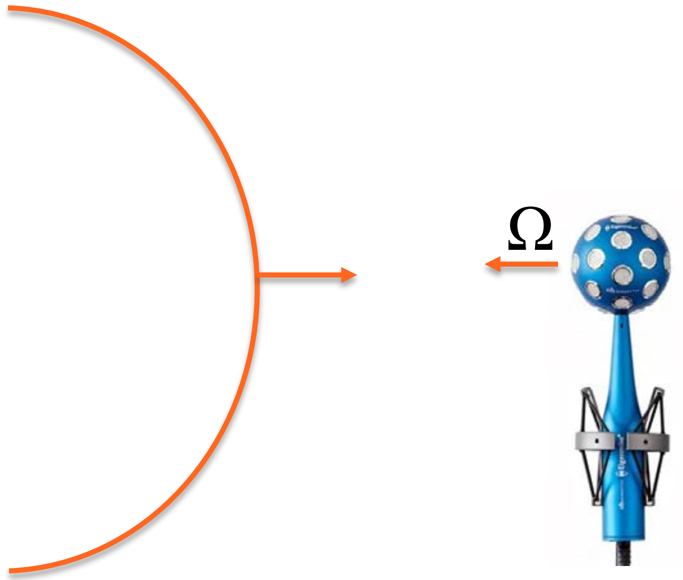
Again, in the time-domain, the signals are sinusoids, but with different scaling and delay.



Sakari.Tervo@aalto.fi

Direction of arrival (DOA) Ω

Since the microphones are in different positions we can obtain the direction of arrival of a sound wave.

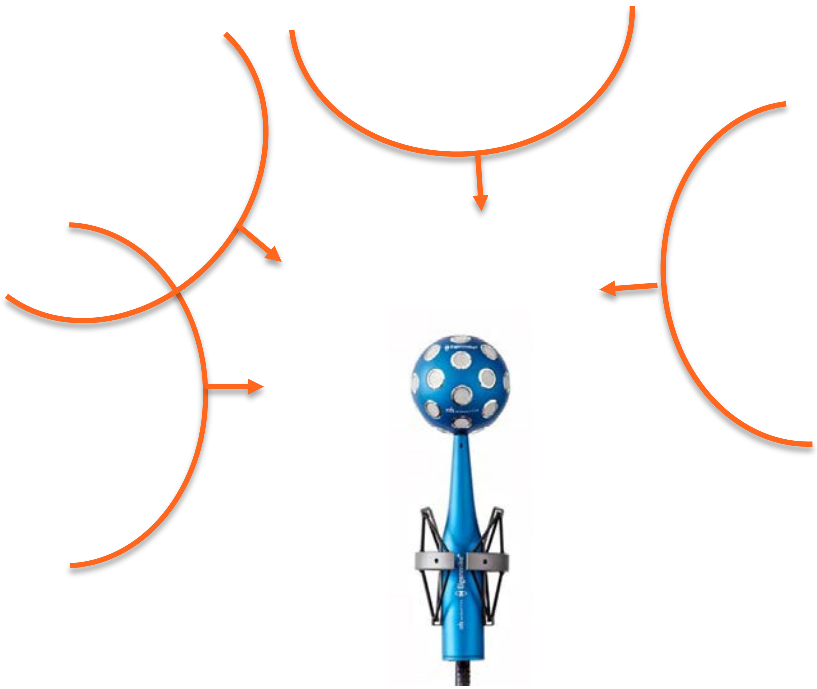


In general we require that $M > 3$, in order to obtain the Direction of Arrival Ω

Sakari.Tervo@aalto.fi

Direction of Arrival

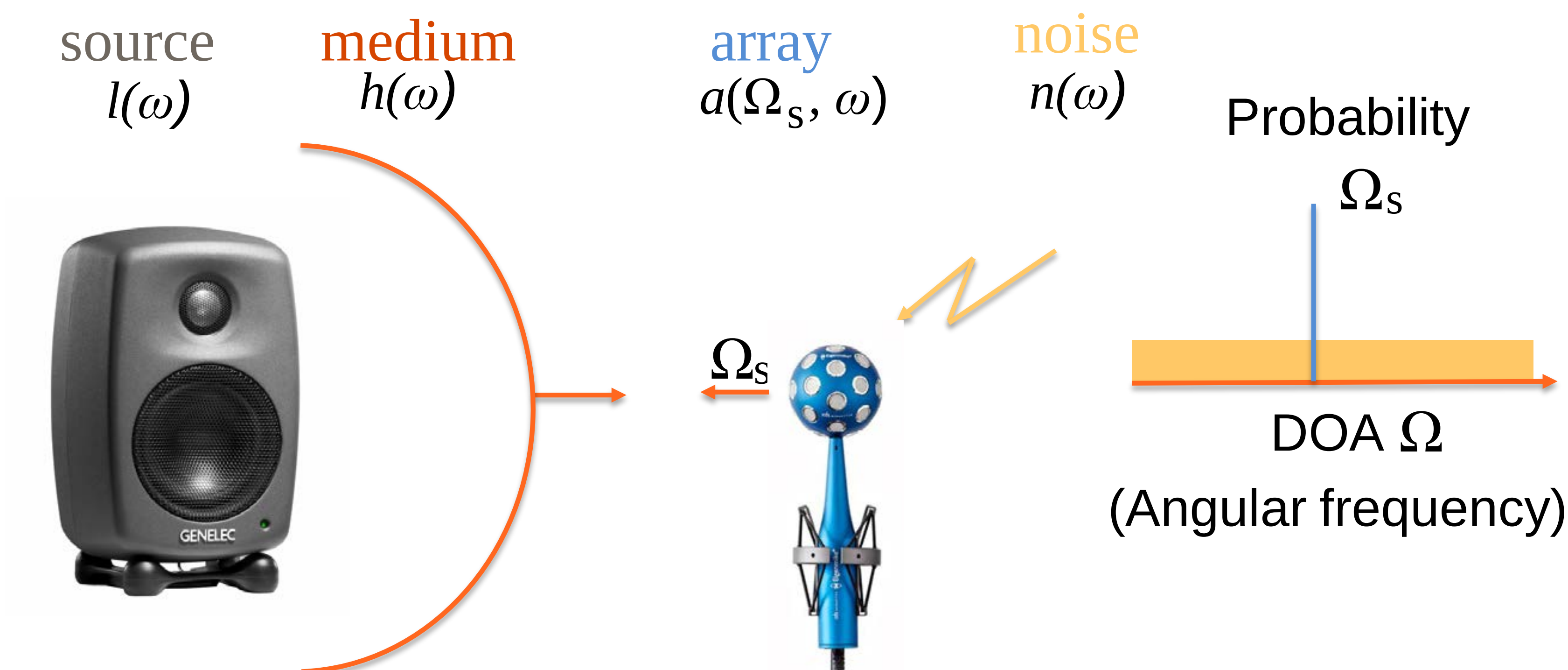
How many DOAs D can we find from a single measurement?



In general, $D < M$, but depends on the array and frequency.

Spatial room impulse response

The measurement is affected by



Images from genelec.com and mhacoustics.com

Source response

Source response is different in each direction and at each distance $l(\Omega, r, \omega)$

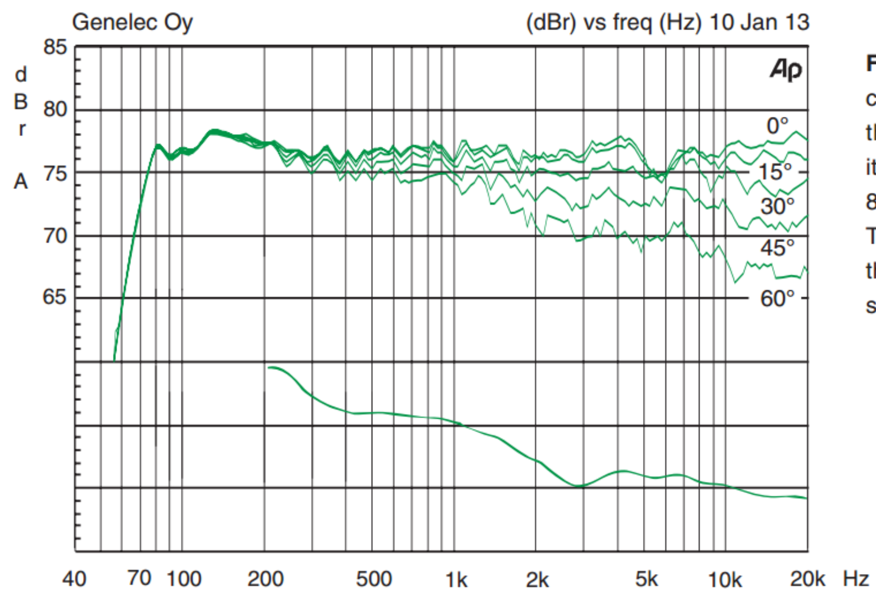


Figure 5. The upper curve group shows the horizontal directivity characteristics of the 8010A measured at 1 m. The lower curve shows the system's power response.

Images from genelec.com and mhacoustics.com

Sakari.Tervo@aalto.fi

Air absorption

Depends on the frequency and travelled distance

Affected by the composition of air, temperature, humidity, etc.

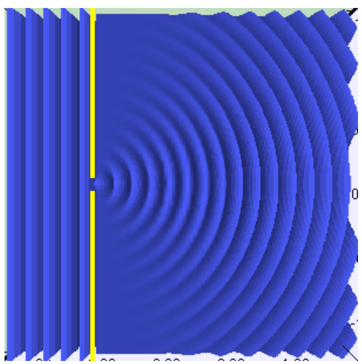
Air absorption

Images from Wikipedia

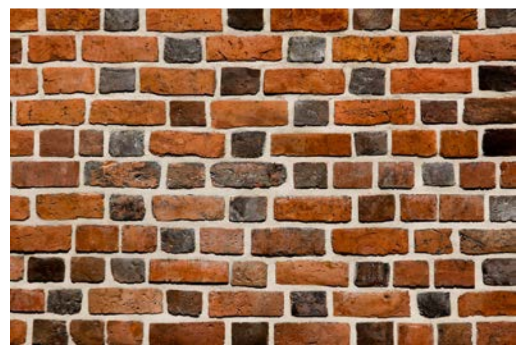
The acoustic path

The sound in the acoustic path is altered by other phenomena besides air absorption, for example.

Diffraction



Reflections on the boundaries



Images from Wikipedia

Array response, i.e., steering vector or array manifold

The transfer function $a(\Omega, \omega)$, i.e., response of each microphone w.r.t. frequency and DOA

Can be measured or modelled. Measurement is recommended in the general case.

Image from researchgate

Noise

Spatially white, independently and identically distributed.

Cause by the thermal noise in the electronic devices, A/D conversion etc.

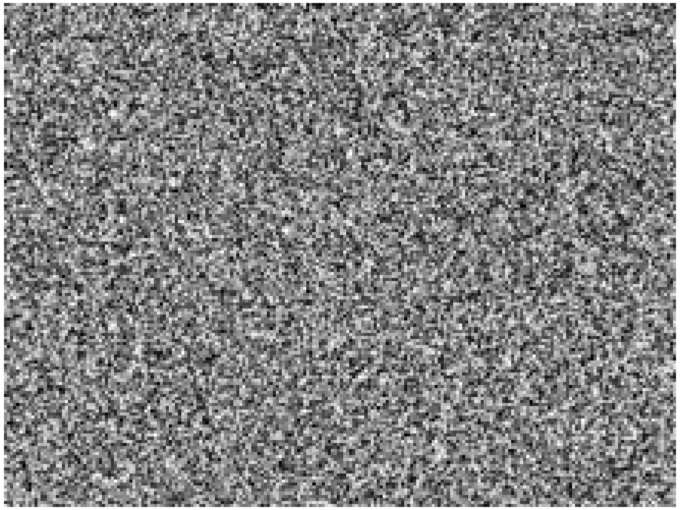
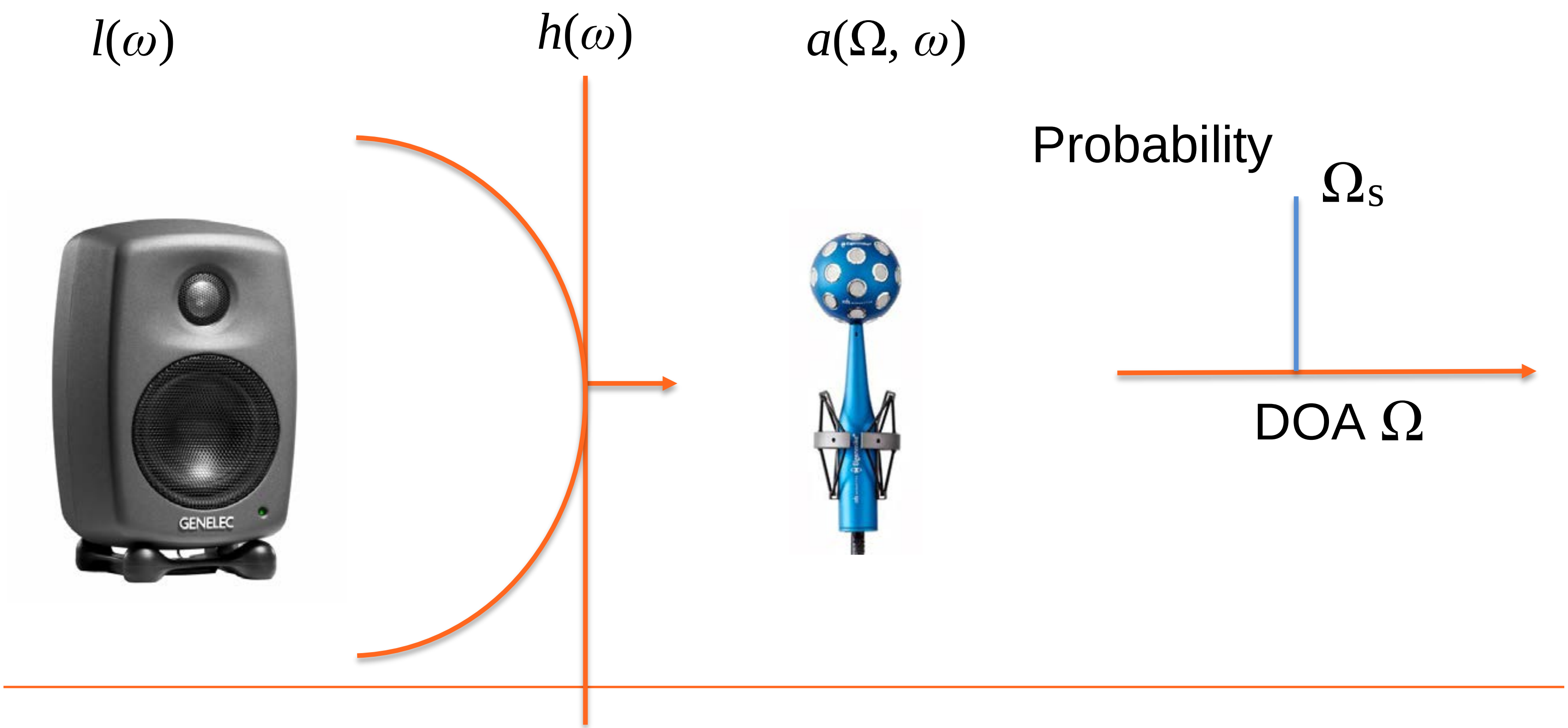


Image from Wikipedia

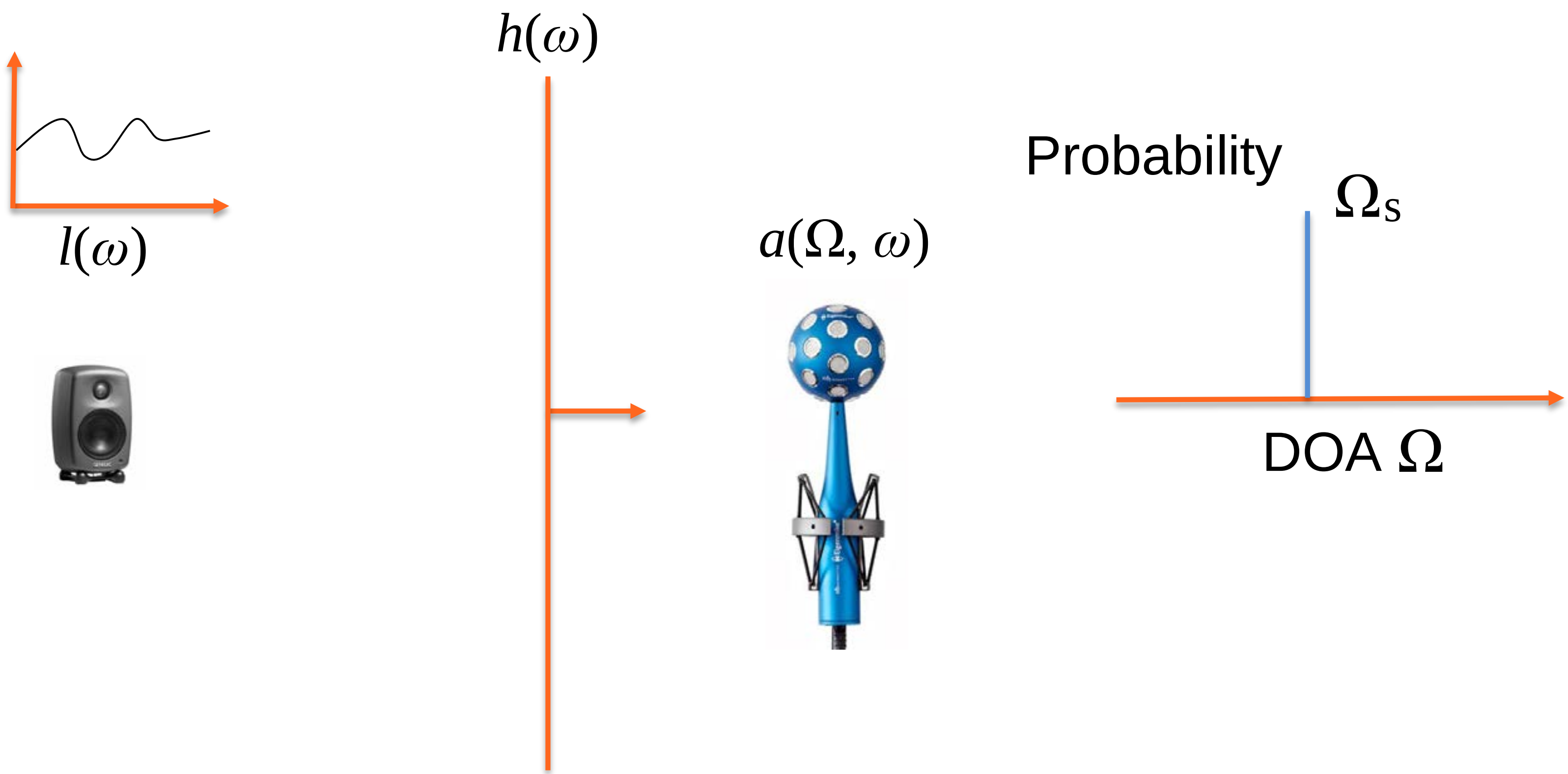
Plane-wave assumption

When the source is in the far-field, the sound field is a plane wave from the array's perspective



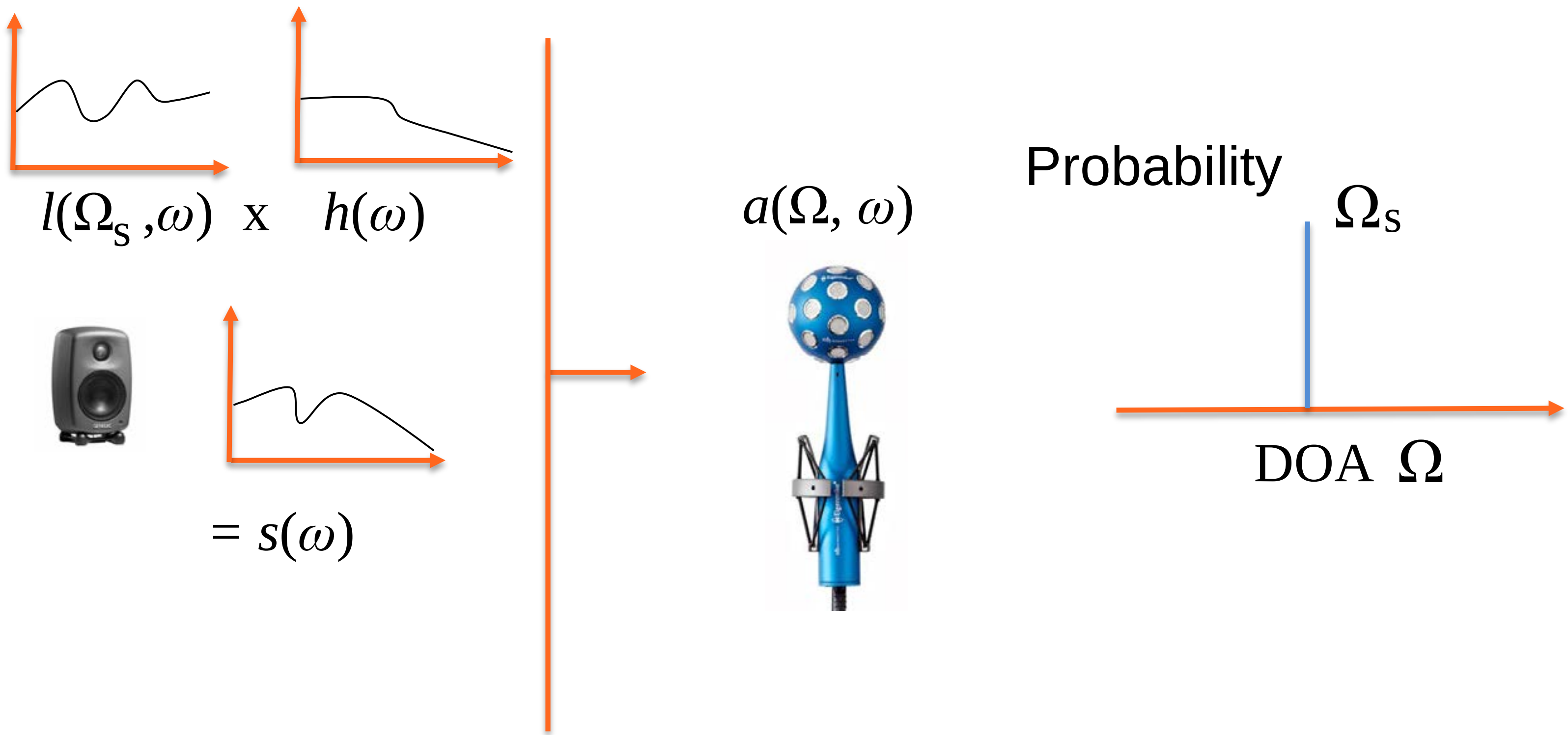
Spatial room impulse response

What is simplest way of parameterizing this problem?



A simple parameterization

A source signal $s(\omega)$ in direction Ω_S



Aalto University
School of Science

Sakari.Tervo@aalto.fi

Parametric model

Noise is additive, the responses are convolved with each other

$$\begin{bmatrix} x_1(\omega) \\ \vdots \\ x_M(\omega) \end{bmatrix} \begin{matrix} = \\ \\ = \end{matrix} \begin{bmatrix} a_1(\Omega_S, \omega)s(\omega) + n_1(\omega) \\ \vdots \\ a_M(\Omega_S, \omega)s(\omega) + n_M(\omega) \end{bmatrix}$$

In vector format

$$\mathbf{x}(\omega) = s(\omega)\mathbf{a}(\Omega_S,\omega) + \mathbf{n}(\omega)$$

More than one “source”

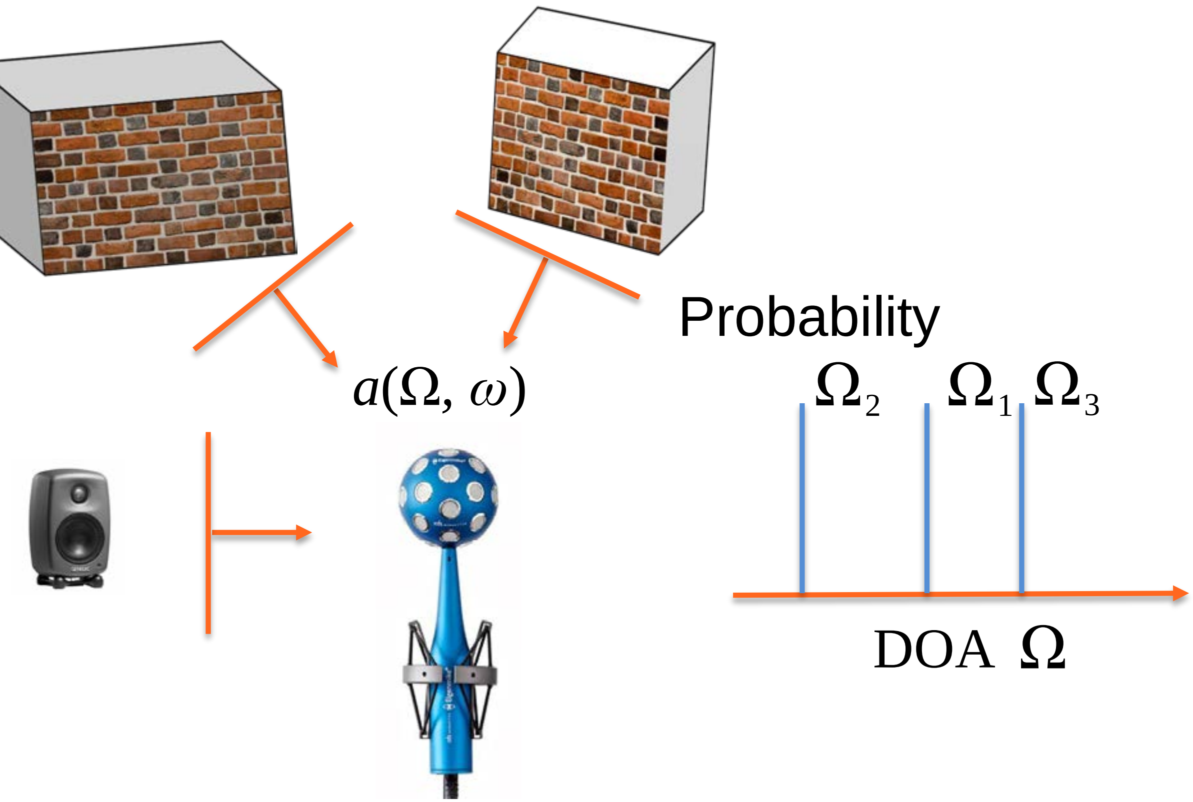
$$\mathbf{x} = \mathbf{s}\mathbf{A}(\boldsymbol{\Omega}) + \mathbf{n},$$

where

$$\boldsymbol{\Omega} = [\Omega_1, \Omega_2, \Omega_3]$$

$$\mathbf{A}(\boldsymbol{\Omega}) = [\mathbf{a}(\Omega_1), \mathbf{a}(\Omega_2), \mathbf{a}(\Omega_3)]$$

$$\mathbf{s} = [s_1, s_2, s_3]$$



For simplicity ω is omitted from the above notation.

How general is the model?

The impulse response will always follow the deterministic model at a time instant and in a single frequency given that the array is in the far-field and that the array response is accurate.

Very General!

Parametric estimation

Unknown parameters

Unknown parameters in the model are DOAs $\boldsymbol{\Omega}$, source signals $\mathbf{s}$, and noise variance σ^2.

$$\mathbf{x}(\omega) = s(\omega)\mathbf{A}(\boldsymbol{\Omega},\omega) + \mathbf{n}(\omega)$$

Assumptions on the covariances

The noise is uncorrelated and identically distributed with a variance σ^2

$$\mathbf{E}[\mathbf{n}(\omega)\mathbf{n}^{\mathrm{H}}(\omega)] = \sigma^2\mathbf{I}.$$

The source signal is deterministic

$$\mathbf{E}[\boldsymbol{x}(\omega)] = \mathbf{A}(\boldsymbol{\Omega}, \omega)s(\omega),$$

$$\text{since} \quad \mathbf{x}(\omega) = \mathbf{A}(\boldsymbol{\Omega}, \omega)s(\omega) + \mathbf{n}(\omega).$$

Thus

$$\mathbf{E}[(\mathbf{x}(\omega)\text{-}\mathbf{E}[\mathbf{x}(\omega)])\,(\mathbf{x}(\omega)\text{-}\mathbf{E}[\mathbf{x}(\omega)])^{\mathrm{H}}] = \sigma^2\mathbf{I},$$

and

$$\mathbf{E}[\mathbf{x}(\omega)\mathbf{x}^{\mathrm{H}}(\omega)] = \mathbf{A}(\boldsymbol{\Omega}, \omega)s(\omega)s^{\mathrm{H}}(\omega)\mathbf{A}^{\mathrm{H}}(\boldsymbol{\Omega}, \omega) + \sigma^2\mathbf{I}$$

Maximum Likelihood (ML) Estimation

If we assume that the noise is of Gaussian shape we can write a likelihood function, for simplicity ω is omitted

$$p(\boldsymbol{\Omega}, \Sigma, \mathbf{s}) = |\pi\boldsymbol{\Sigma}|^{-1}\exp([-(\mathbf{x}-\mathbf{A}(\Omega)s)\,\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{A}(\Omega)s)^{H}]).$$

Since we have $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ the likelihood is presented as

$$p(\boldsymbol{\Omega}, \sigma^2, \mathbf{s}) = |\pi\sigma^2\mathbf{I}|^{-1}\exp(-[(\mathbf{x}-\boldsymbol{A}(\Omega)s)\,\sigma^{-2}(\mathbf{x}-\mathbf{A}(\Omega)s)^{H}])$$

The minimum argument of the negative log-likelihood gives the ML estimates

$$\hat{\boldsymbol{\Omega}}, \hat{\sigma}^2, \hat{\mathbf{s}} = \underset{\boldsymbol{\Omega},\, \sigma^2, s}{\arg\min}\left(\log(|\pi\sigma^2\mathbf{I}|) + \log([(\mathbf{x}-\mathbf{A}(\Omega)s)\,\sigma^{-2}\,(\mathbf{x}-\mathbf{A}(\Omega)s)^{H}])\right)$$

This function is in general non-linear multidimensional and has multiple minima.

Concentrated ML estimation

One can only find the estimates via non-convex optimization.

The source signal *s* is difficult to obtain via optimization because of the circularity of the phase.

When we assume that *Ω* and *s* are fixed, the noise is given as

$$\hat{\sigma}^2(\Omega) = 1/M\ \mathrm{Tr}(\mathbf{P}_{\mathbf{A}}^{\perp}(\Omega)\hat{\mathbf{R}}), \quad (1) \quad \text{where}$$

$\mathbf{P}_{\mathbf{A}}^{\perp}(\Omega) = \mathbf{I} - \mathbf{A}(\Omega)\mathbf{A}^{\dagger}(\Omega)$, is the orthogonal projector to the null-space and

$\hat{\mathbf{R}} = \mathbf{x}\,\mathbf{x}^{H}$, is the estimated covariance matrix of the microphones signals.

Inserting the (1) into the likelihood gives a quadratic function

$$\hat{\Omega}, \hat{\mathbf{s}} = \arg\min_{\Omega, s}\left((\mathbf{x}-\mathbf{A}(\Omega)s)(\mathbf{x}-\mathbf{A}(\Omega)s)^{H}\right)$$

Concentrated ML estimation

When we minimize the quadratic function w.r.t. to *s* we obtain

$$\hat{\mathbf{s}}(\boldsymbol{\Omega}) = \mathbf{A}^{\dagger}(\boldsymbol{\Omega})\mathbf{x}. \quad (2) \qquad \hat{\sigma}^2(\boldsymbol{\Omega}) = 1/M \operatorname{Tr}(\mathbf{P}_{\mathbf{A}}^{\perp}(\boldsymbol{\Omega})\hat{\mathbf{R}}), \quad (1)$$

Inserting (1) and (2) to the negative log-likelihood

$$\hat{\boldsymbol{\Omega}}, \hat{\sigma}^2, \hat{\mathbf{s}} = \arg\min_{\boldsymbol{\Omega}, \sigma^2, \mathbf{s}} \left(\log(|\pi\sigma^2\mathbf{I}|) + \log([(\mathbf{x}-\mathbf{A}(\boldsymbol{\Omega})\mathbf{s})\, \sigma^{-2}\, (\mathbf{x}-\mathbf{A}(\boldsymbol{\Omega})\mathbf{s})^{H}]) \right)$$

returns

$$\hat{\boldsymbol{\Omega}} = \arg\min_{\boldsymbol{\Omega}} \left(\operatorname{Tr}(\mathbf{P}_{\mathbf{A}}^{\perp}(\boldsymbol{\Omega})\hat{\mathbf{R}}) \right). \quad (3)$$

Concentrated ML in Matlab

```
function [nLogL, sigma2, s] = CML(Omega, M, x)
R = x*x'; % Covariance matrix of the measurements
A = getSteeringVector(Omega); % Array Response
invA = pinv(A); % Pseudo-inverse of A
PIA = eye(M)-A*invA; % Orthogonal projector to the null-space
nLogL = trace(PIA*R); % Equation (3), negative log-likelihood
sigma2 = 1/M*nLogL % Equation (1), variance
s = invA*y; % Equation (2), source signal
```

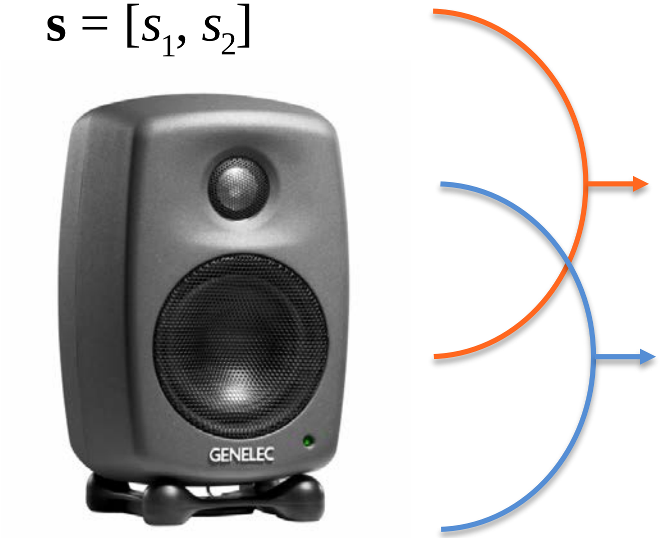
Example, two-way loudspeaker

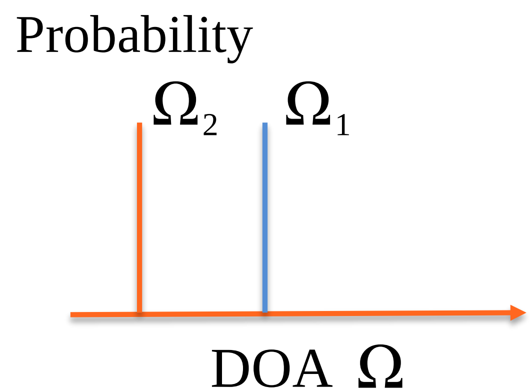
A two-way loudspeaker has two sources, different DOAs and different source signals

$$\boldsymbol{\Omega} = [\Omega_1, \Omega_2]$$

$$\mathbf{A}(\boldsymbol{\Omega}) = [\mathbf{a}(\Omega_1), \mathbf{a}(\Omega_2)]$$

$$\mathbf{s} = [s_1, s_2]$$





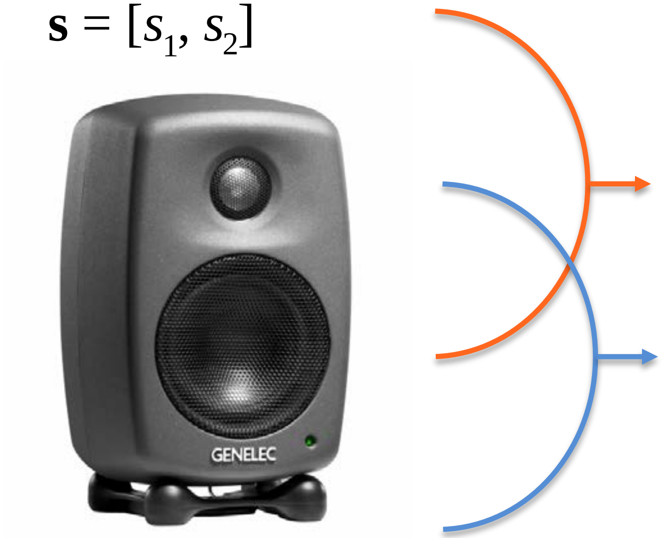
Example, two-way loudspeaker

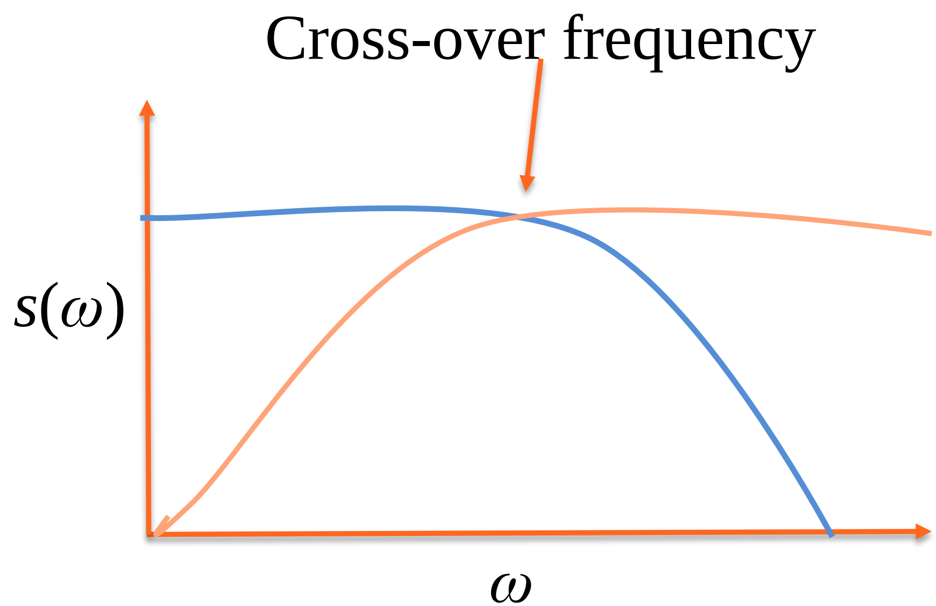
A two-way loudspeaker has two sources, different DOAs and different source signals

$$\mathbf{\Omega} = [\Omega_1, \Omega_2]$$

$$\mathbf{A}(\mathbf{\Omega}) = [\mathbf{a}(\Omega_1), \mathbf{a}(\Omega_2)]$$

$$\mathbf{s} = [s_1, s_2]$$

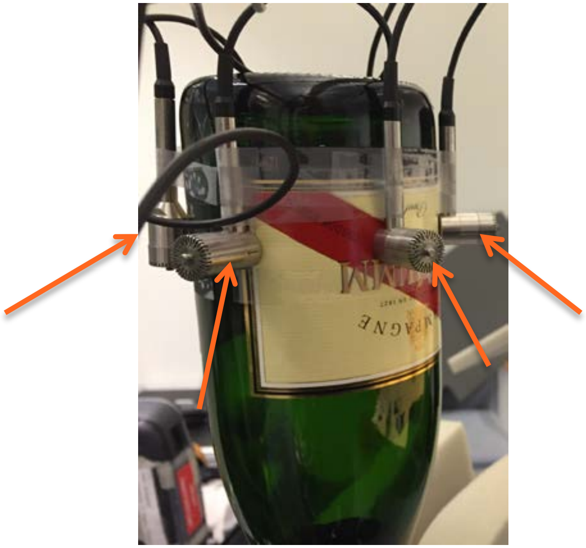




Microphone array

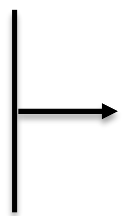
Instead of commercial one we use a DIY array.

1. Buy a bottle of champagne
2. Share it with our colleagues
3. Attach microphones on the surface



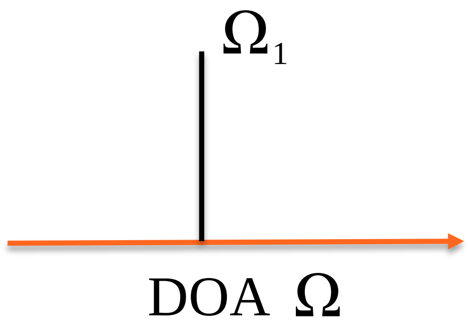
Example, measurement of array response

The steering vectors are measured in the far-field. Measurement is implemented in an office space, and windowing is applied to avoid reflections.



$a(\Omega, \omega)$

Probability



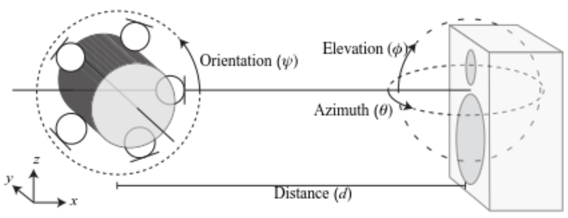
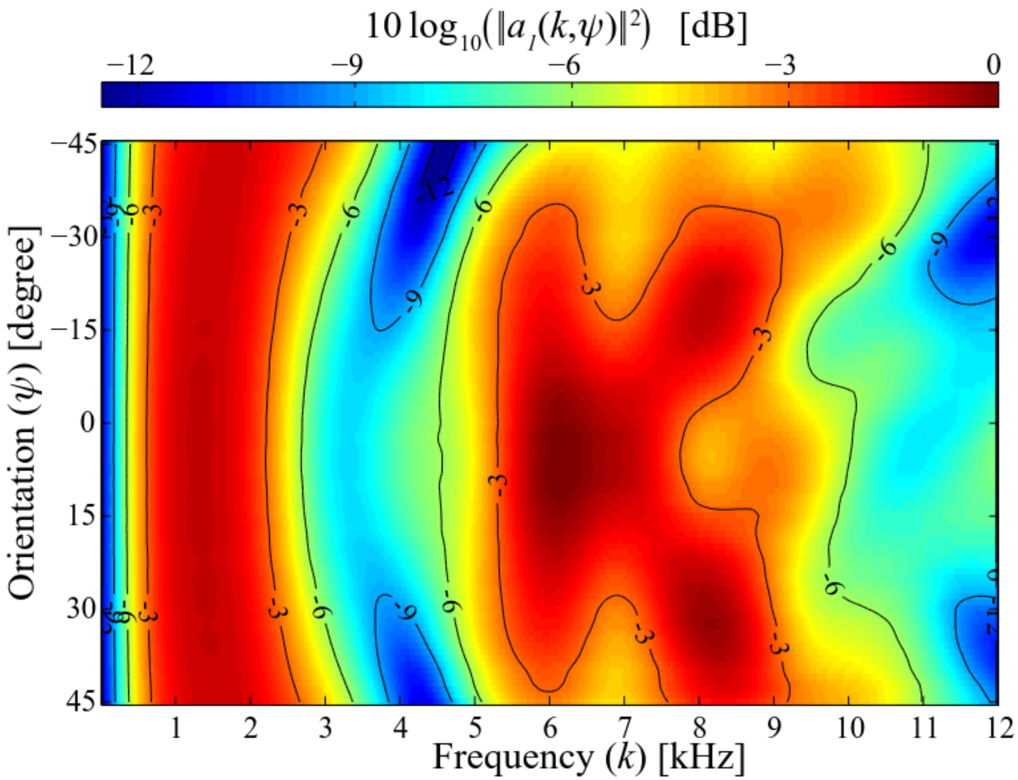


Figure 1: Diagram of the measurement setup.

Example, measurement of steering vectors

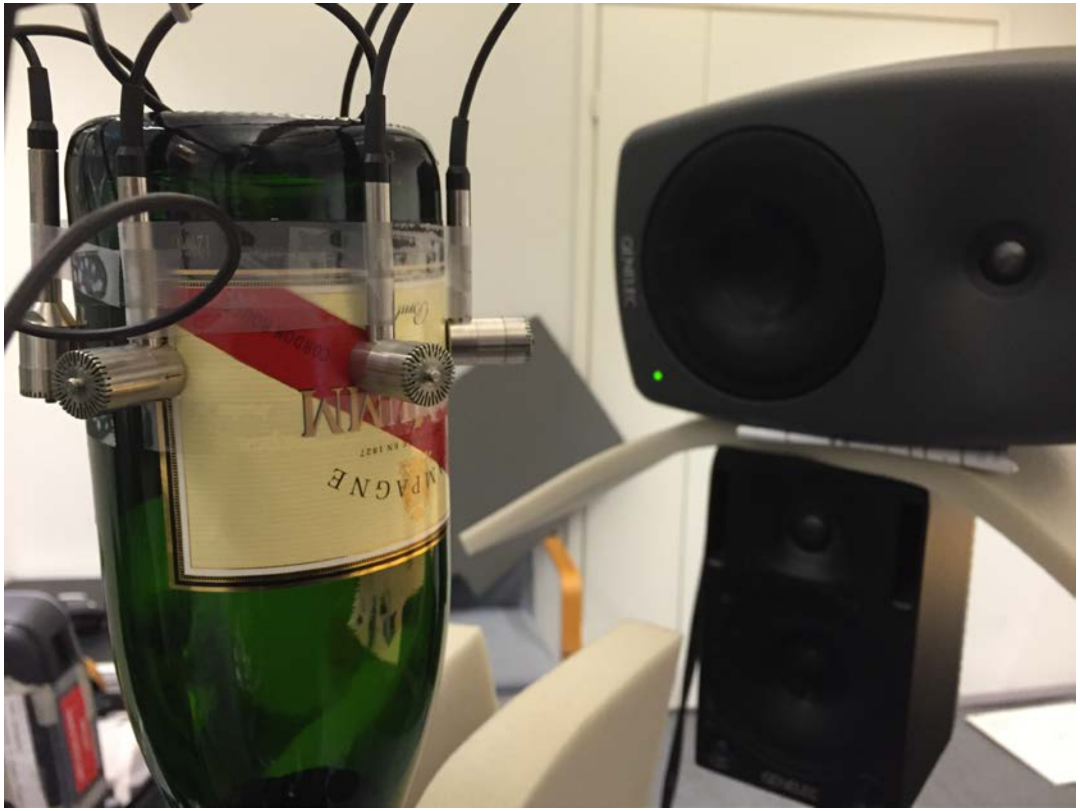
Steering vector $a(\Omega, \omega)$ magnitude response for a microphone



Sakari.Tervo@aalto.fi

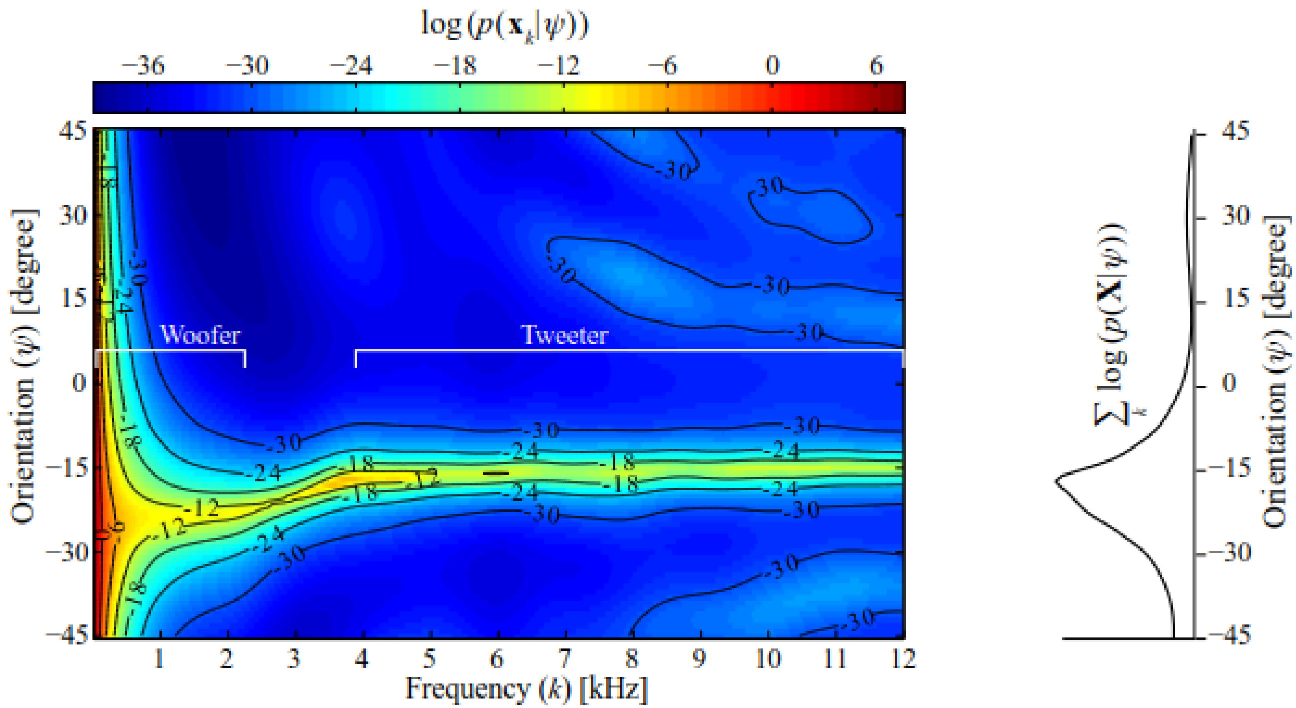
Example, measurement setup

Near-field measurements of the loudspeaker



Example, log-likelihood

Log-likelihood (DML) with D = 1 source

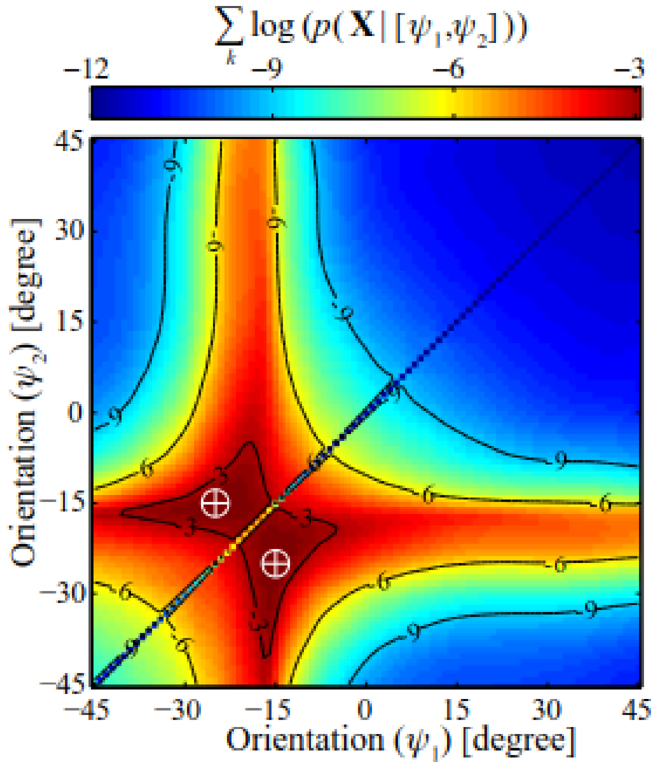


a) DOA estimation *w.r.t.* frequency for *D=1*

b) DOA estimation over $k \in [1,6]$ kHz, *D=1*

Example, log-likelihood

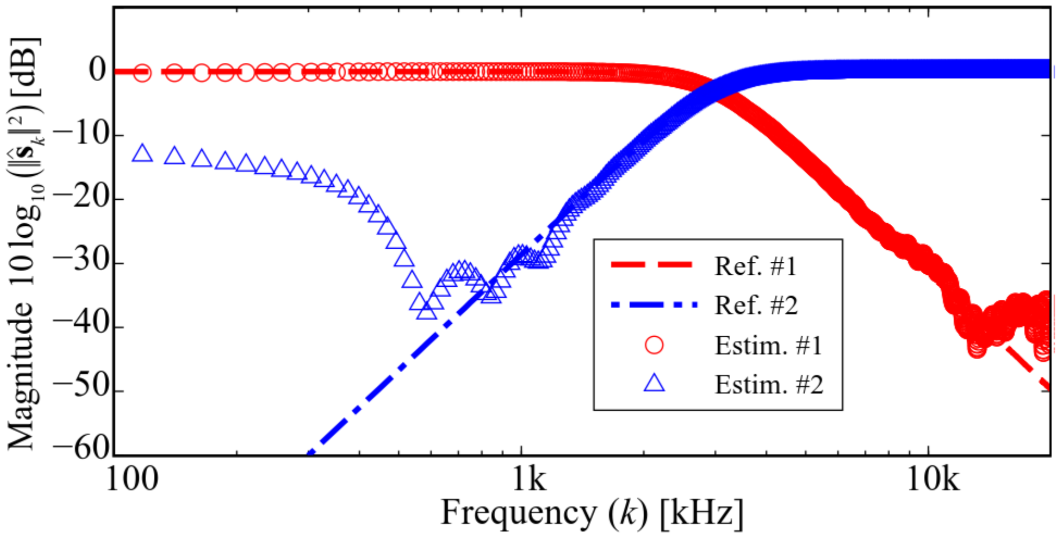
Log-likelihood (DML) with D = 2 sources



c) DOA estimation over
$k \in [1,6]$ kHz, D=2

Example, two-way loudspeaker

Simulated two-way loudspeaker

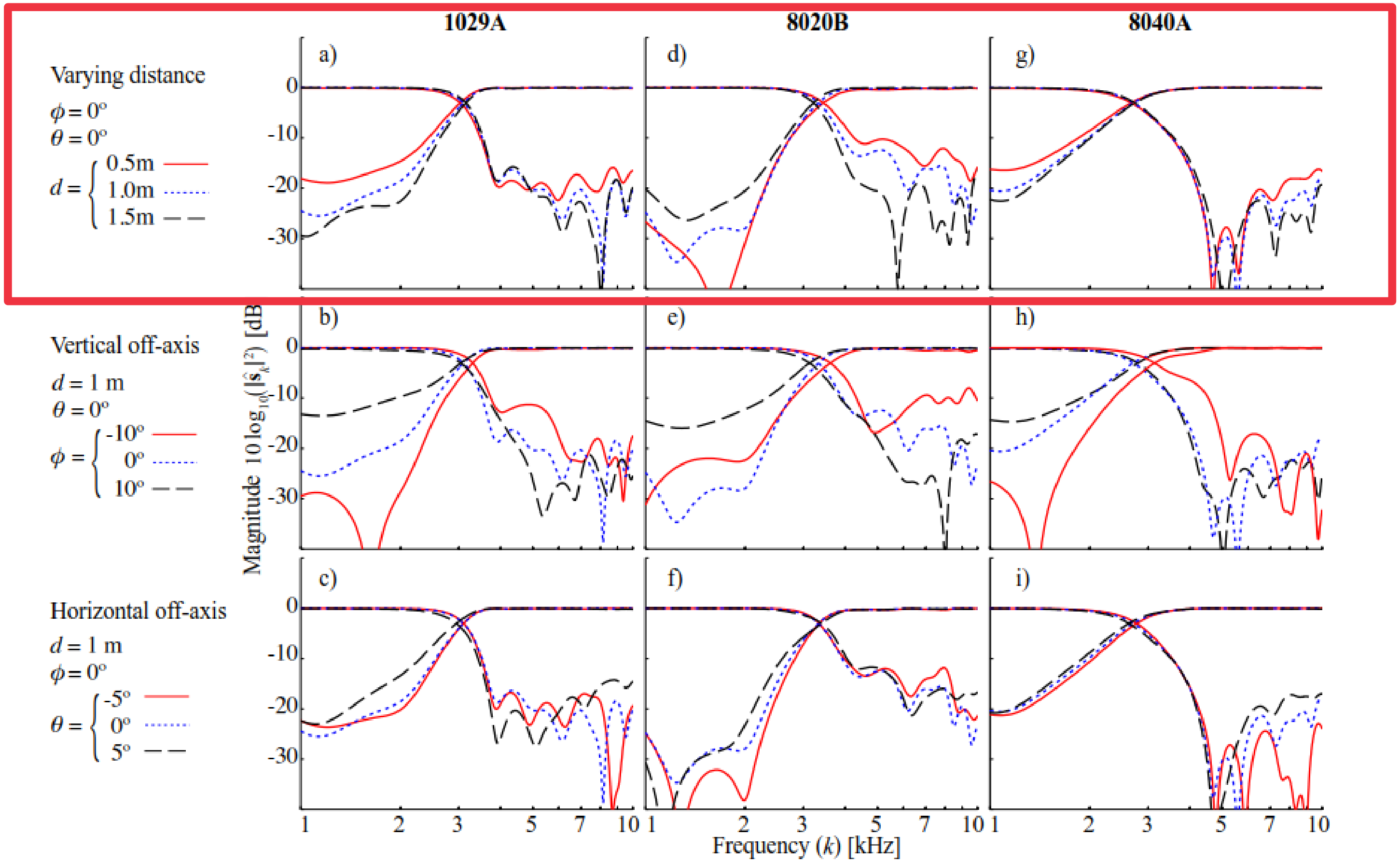


Aalto University
School of Science

Sakari.Tervo@aalto.fi

Example, different two-way loudspeakers

Transfer functions at different distance and angles with 3 loudspeakers types

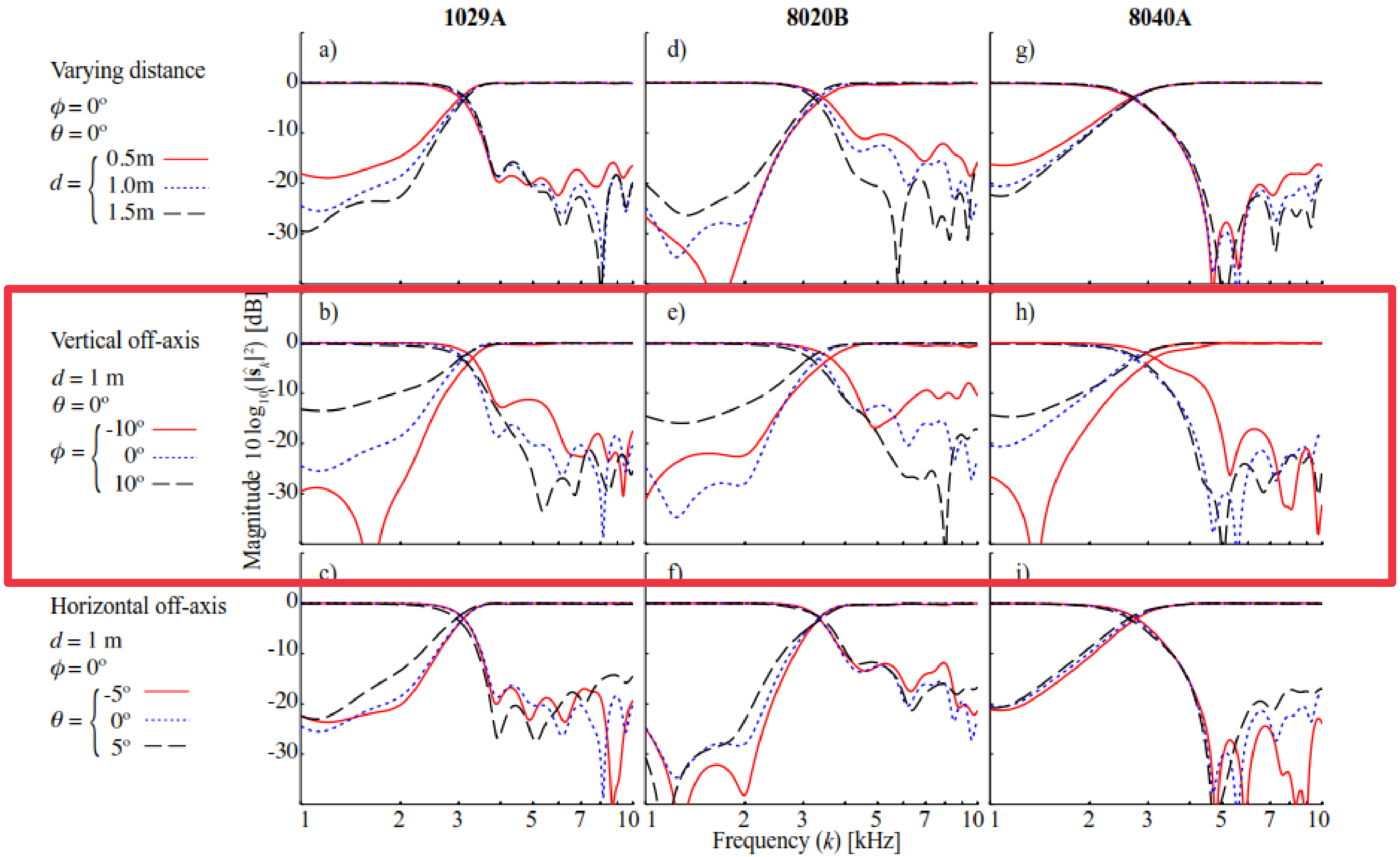


Aalto University
School of Science

Sakari.Tervo@aalto.fi

Example, different two-way loudspeakers

Transfer functions at different distance and angles with 3 loudspeakers types

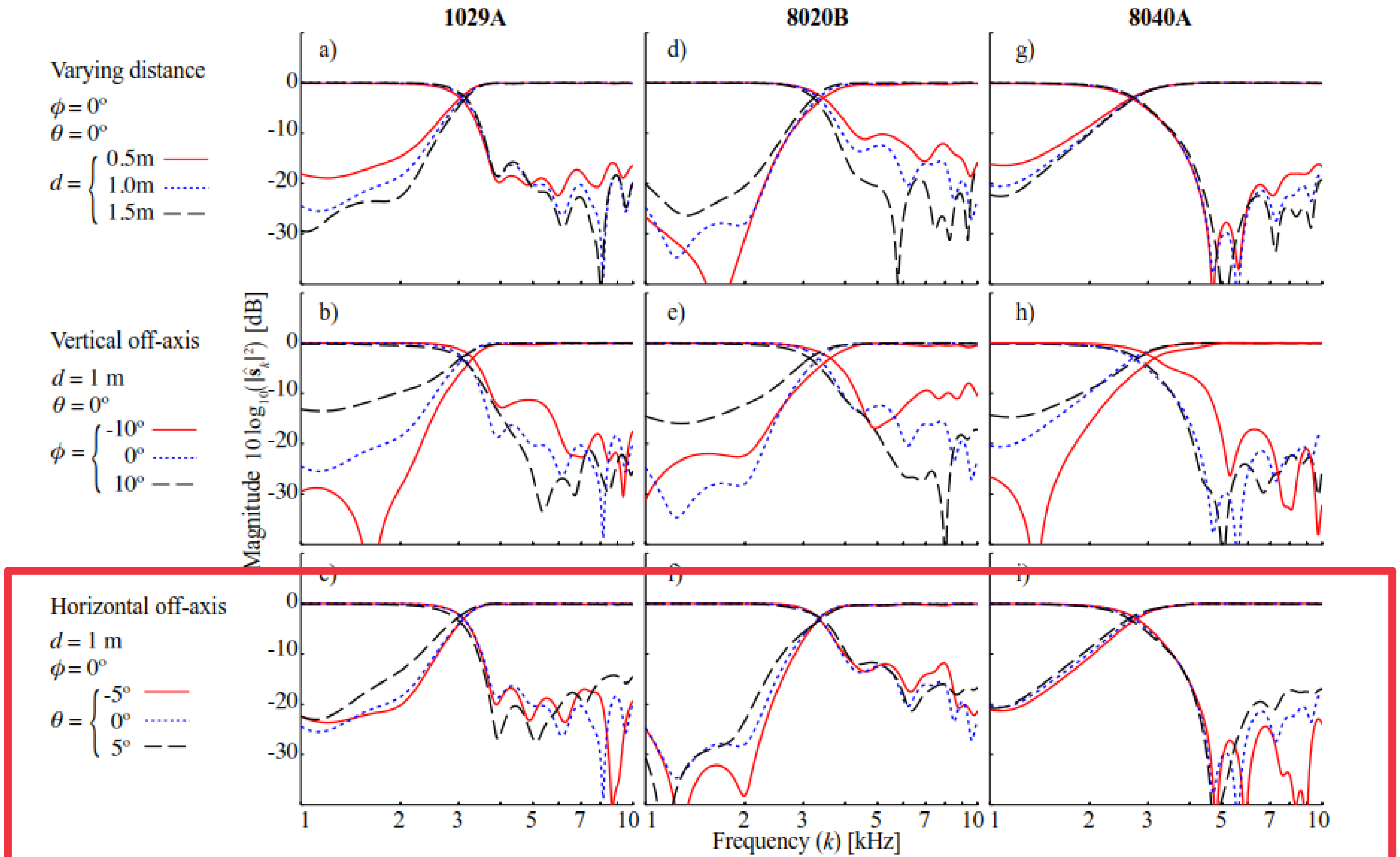


Aalto University
School of Science

Sakari.Tervo@aalto.fi

Example, different two-way loudspeakers

Transfer functions at different distance and angles with 3 loudspeakers types



Aalto University
School of Science
Sakari.Tervo@aalto.fi

Detection of reflections

Aalto University
School of Science

Detection of reflections

Parametric estimation requires the knowledge of the number of reflections/sources.

The number cannot be assumed to be known a priori in the general case, although, so far, it is always assumed in the literature.

Detection of reflections

Simultaneous detection and estimation detection methods are the only approaches that can be applied to the determenistic model in the coherent case.

Simultaneous estimation and detection algorithm:

Initialize D = 0

1. D = D + 1
2. Estimate the parameters
3. Calculate the test statistic for the parameters
4. Decide based on a priori distribution if the model fits the data
   1. If the zero hypothesis cannot be accepted Goto 1
   2. Else stop iteration

Likelihood based detection

Cochran's Theorem states that a normalized ML estimate of variance follows Chi-squared distribution

$$R = 2M\,\hat{\sigma}^2 / \dot{\sigma}^2 = 2\mathrm{Tr}(\,\mathbf{P}_{\mathrm{A}}^{\perp}(\hat{\Omega})\hat{\mathbf{R}})/\dot{\sigma}^2 \sim \chi^2(v),$$

where $\dot{\sigma}^2$ is a consistent noise estimate and v is the degrees of freedom. The decision H_0 is accepted if

$$D_e(H_0) = 1 \text{ if } R < \text{Chi-Inv-}\chi^2(v, \gamma),$$

where γ is a pre-defined significance level, e.g, 99 %.

Likelihood-based detection in Matlab

```
function D = LBD(x, sigma2c, M)
D = 0;
while 1
    D = D + 1;
    v = 2*(M-2*D); % Degrees of freedom
    fun = @(Omega) CML(Omega, M, x); % Find DOAs
    Omega = findMinimum(fun, [zeros(2*d,1)]);
    nLogL = CML(Omega, M, x); % Likelihood
    if nLogL/sigma2c < chi2inv(.99, v)
        break; % Stop iteration if criteria met
    end
end
```

Example: likelihood based estimation and detection

Measurement of a corner in a semi-anechoic room.

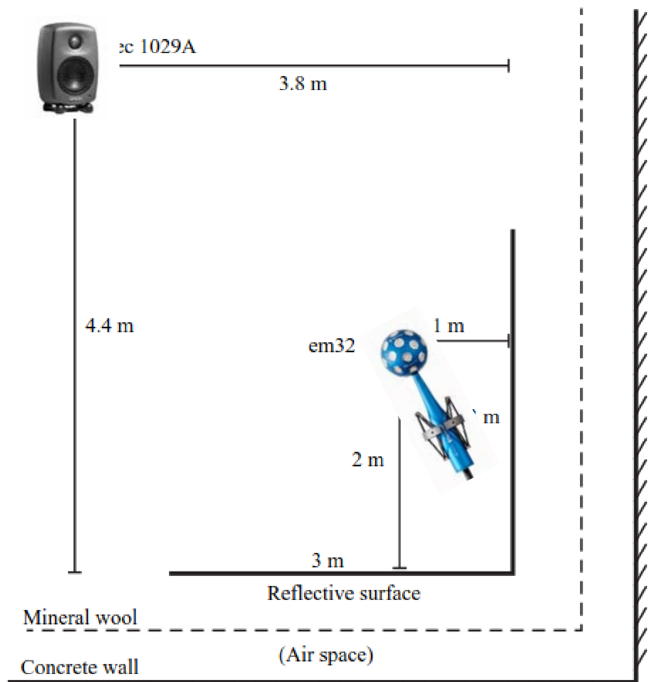


Fig. 5: The measurement setup in the semi-anechoic room. After [12].

Example: Obtaining a reference

The reference is obtained by analyzing the impulse response with short time windows 64 samples at 48 kHz and assuming D = 1.

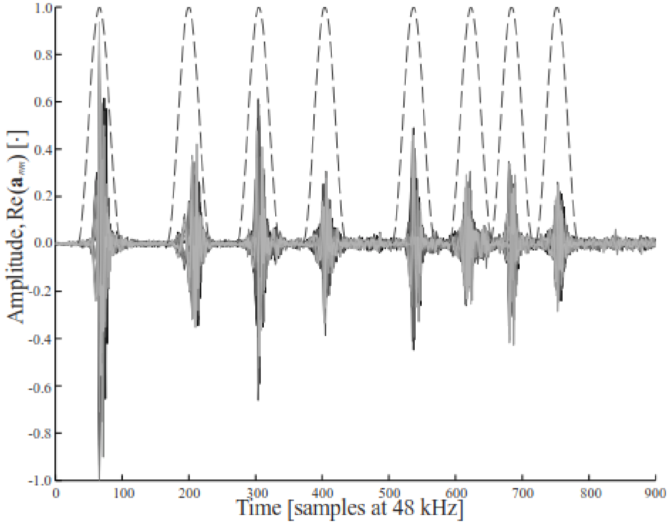


Fig. 6: Real part of SH coefficients $\mathbf{a}_{nm}$ in the time domain, and the windowing (---) applied to derive the reference values. Note that the $(m+1)^2 = 16$ SH coefficients overlap in the visualization heavily. After [12].

Example: likelihood based estimation and detection

In the evaluation we use a rectangular window length of 256 samples with an overlap of 255 samples at 48 kHz, and analyze the response at 4 kHz.

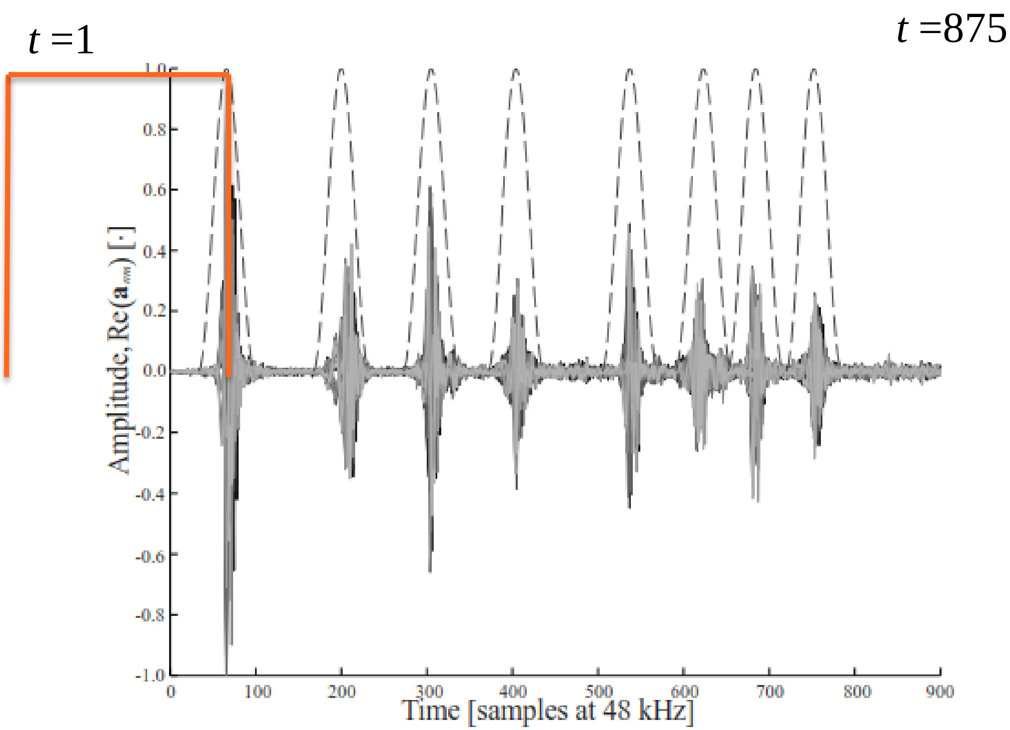
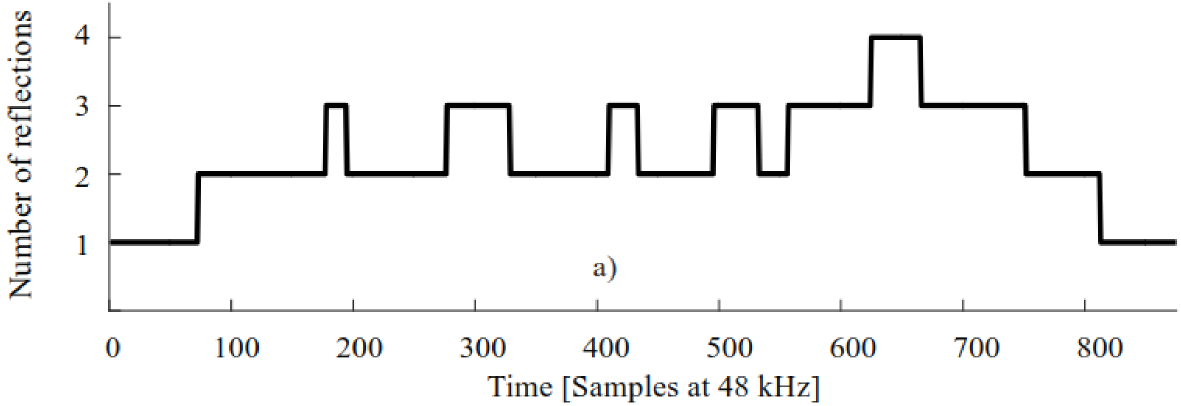


Fig. 6: Real part of SH coefficients $\mathbf{a}_{nm}$ in the time domain, and the windowing (---) applied to derive the reference values. Note that the $(m+1)^2 = 16$ SH coefficients overlap in the visualization heavily. After [12].

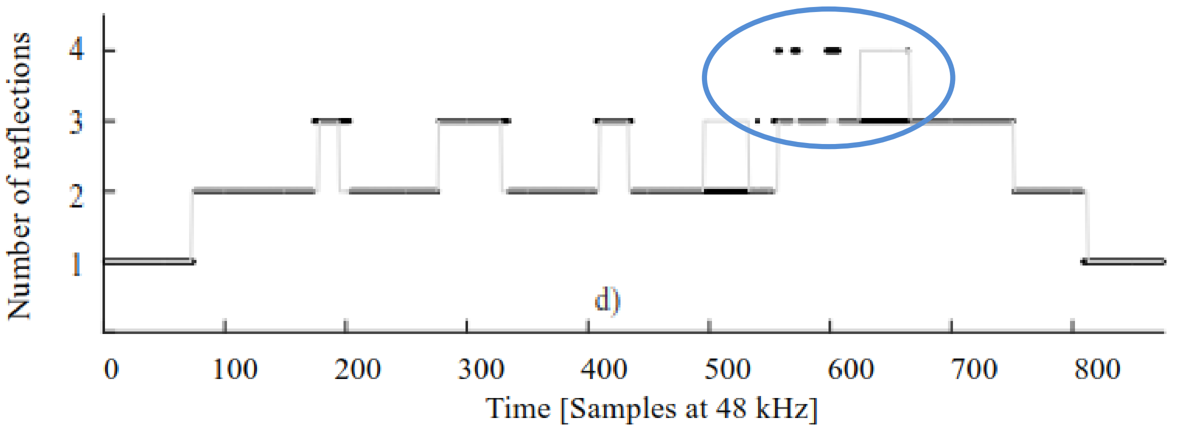
Sakari.Tervo@aalto.fi

Example: likelihood based estimation and detection

The reference for the detection in a window of length $L = 256$



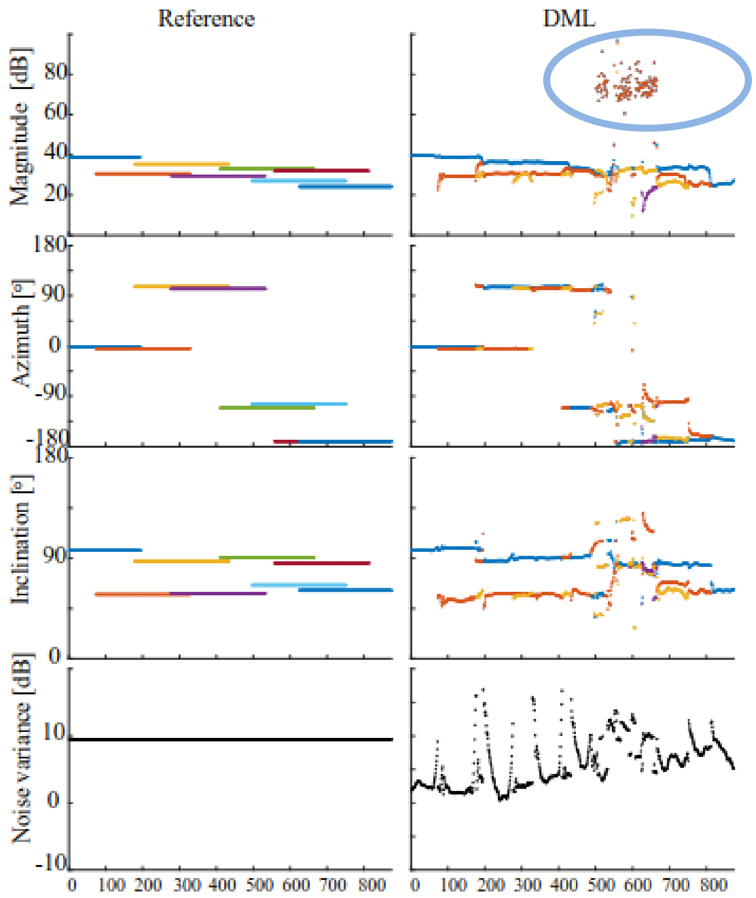
Example: Likelihood based estimation and detection



D	$\hat{D}$, NL [count] 1	2	3	4
1	**133**	2	0	0
2	3	**363**	48	1
3	0	31	**224**	29
4	0	0	19	**22**

Detection rate 92.0 %

Estimation results



Large errors in the reflection signal estimates

Reproduction of spatial sound via room impulse responses

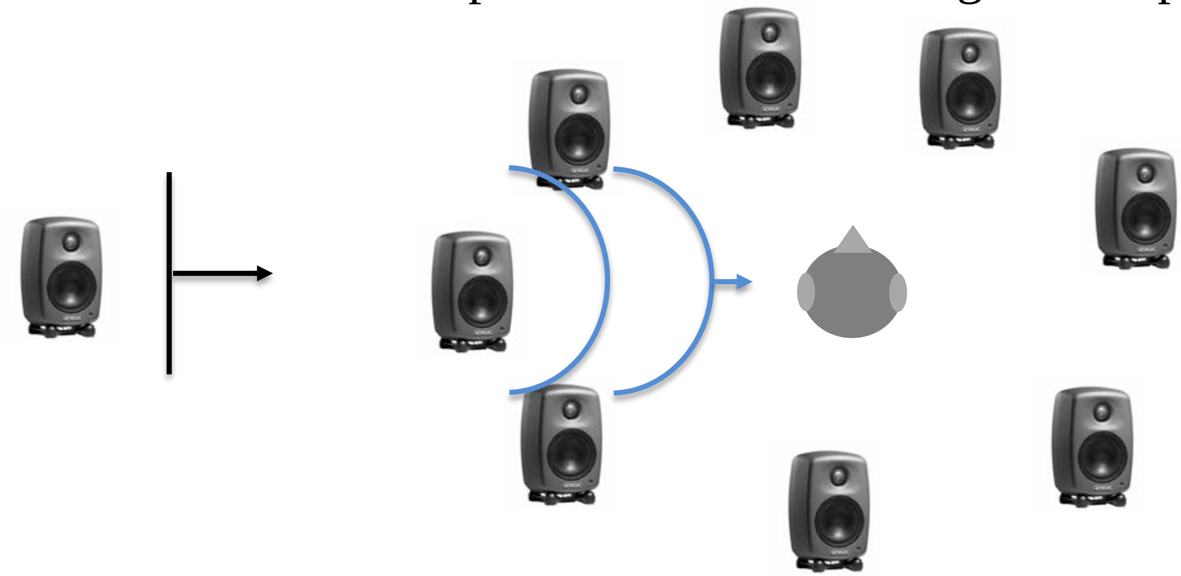
Aalto University
School of Science

Spatial sound reproduction

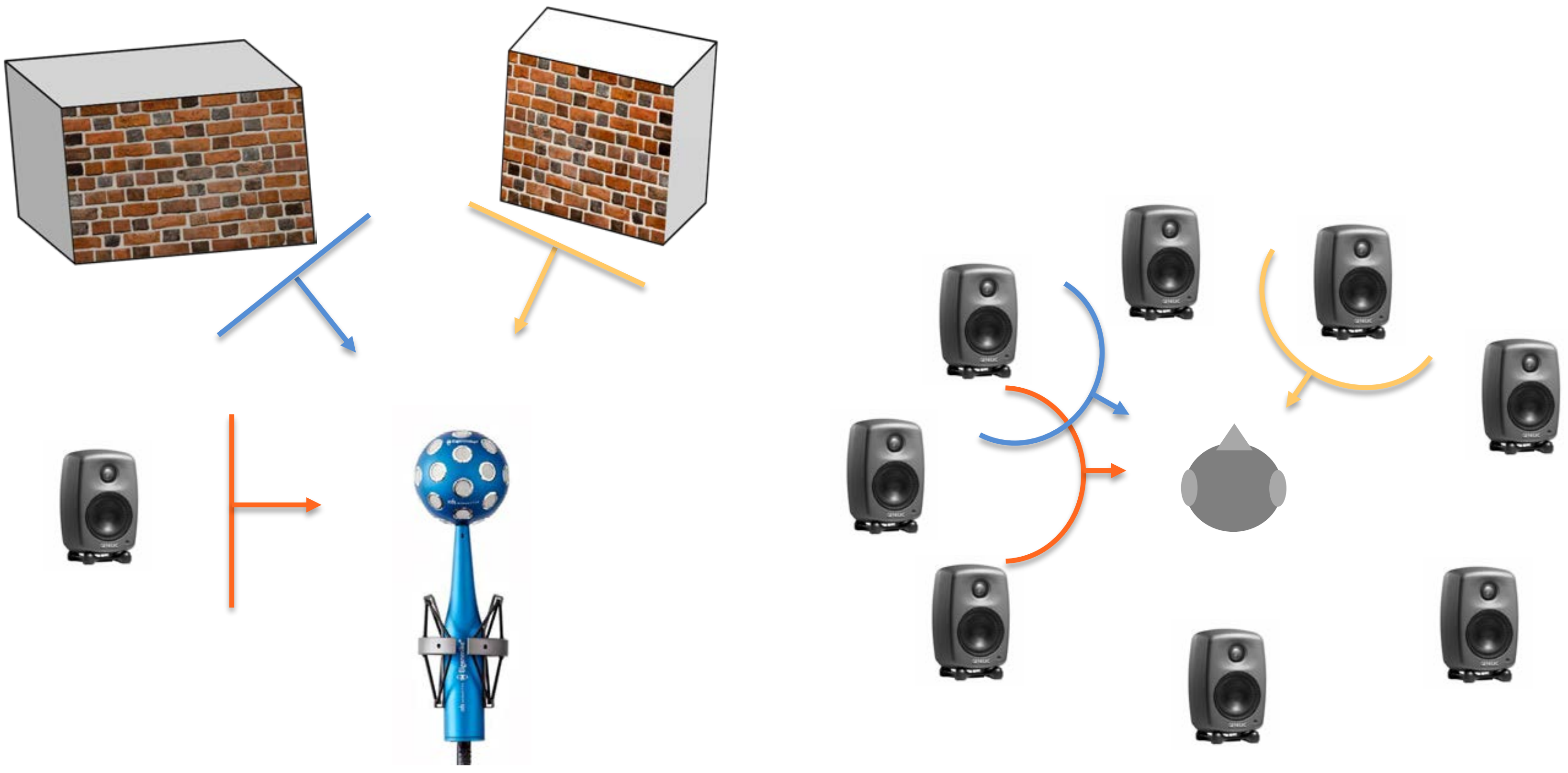
Assign the estimated source signals $\hat{s}$ to the estimated direction $\hat{\Omega}$.

Playback the source signal with vector based amplitude panning, wavefield synthesis, nearest neighbor, ambisonics (whatever order).

Convolve each loudspeaker with a source signal for spatial sound.



Example: Nearest neighbour playback



Thank you for your attention!

Sakari.Tervo@aalto.fi

https://mediatech.aalto.fi/en/research/virtual-acoustics

The research leading to these results has received funding from

- the Academy of Finland, project nos. [257099]

ACADEMY OF FINLAND

References

B. Ottersten *et al.*: “Exact and Large Sample ML Techniques for Parameter Estimation and Detection in Array Processing”, in *Radar Array Processing*, Simon Haykin (ed.), Springer-Verlag, Germany, 1993

S. Tervo and A. Politis: “Direction of Arrival Estimation of Reflections from Room Impulse Responses using a Spherical Microphone Array”, *IEEE/ACM TASLP*, 2015, Volume 23, Issue, 10, Pages 1539-1551

S. Tervo: “Single Snapshot Detection and Estimation of Reflections from Room Impulse Responses in the Spherical Harmonic Domain”, *IEEE/ACM TASLP*, 2016, To appear, 15 pages