Translation of a Higher Order Ambisonics Sound Scene Based on Parametric Decomposition

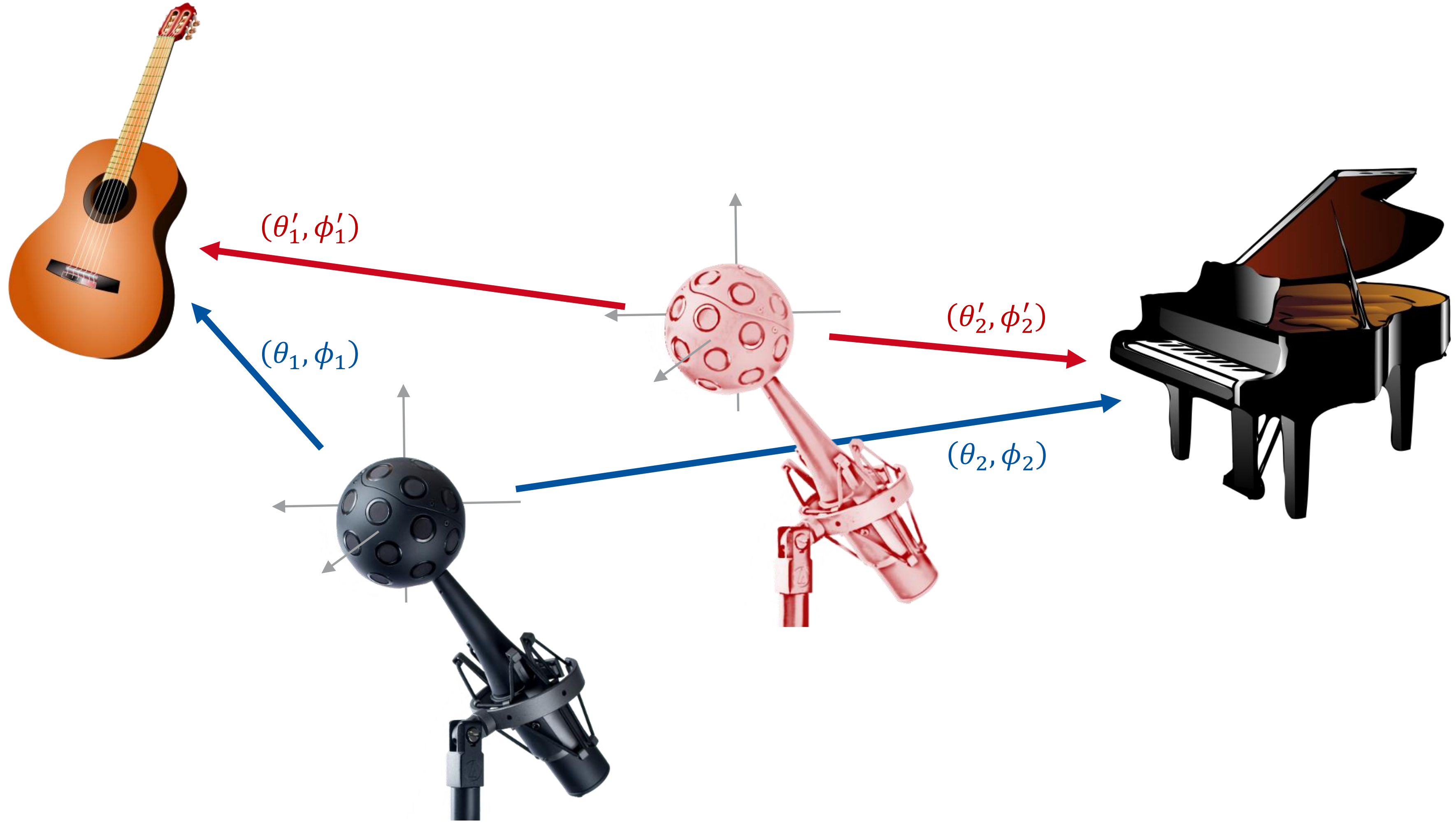
Maximilian Kentgens, Andreas Behler, Peter Jax

IEEE ICASSP 2020

Institute of Communication Systems
Prof. Dr.-Ing. Peter Jax
RWTH AACHEN UNIVERSITY

From Single Spot Microphone Array Recording…

…to User-Movement-Enabled Immersive Reproduction



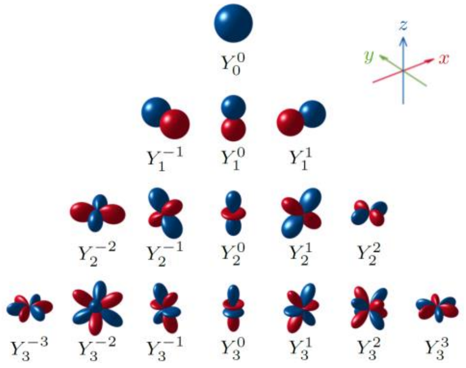
θ inclination angle
ϕ azimuth angle

Signal Model

- Higher Order Ambisonics (HOA) signal $\boldsymbol{x}(\lambda,\mu) \in \mathbb{C}^{(N+1)^2}$ with Spherical Harmonics truncation order N:

$$\boldsymbol{x}(\lambda,\mu) = \boldsymbol{x}_s(\lambda,\mu) + \boldsymbol{x}_a(\lambda,\mu)$$

λ: frame index
μ: frequency bin



  - **Direct part $\boldsymbol{x}_s(\lambda,\mu)$**: variable number of $I(\lambda,\mu) \in \{0,1,2,\dots,(N+1)^2\}$ plane wave sources:

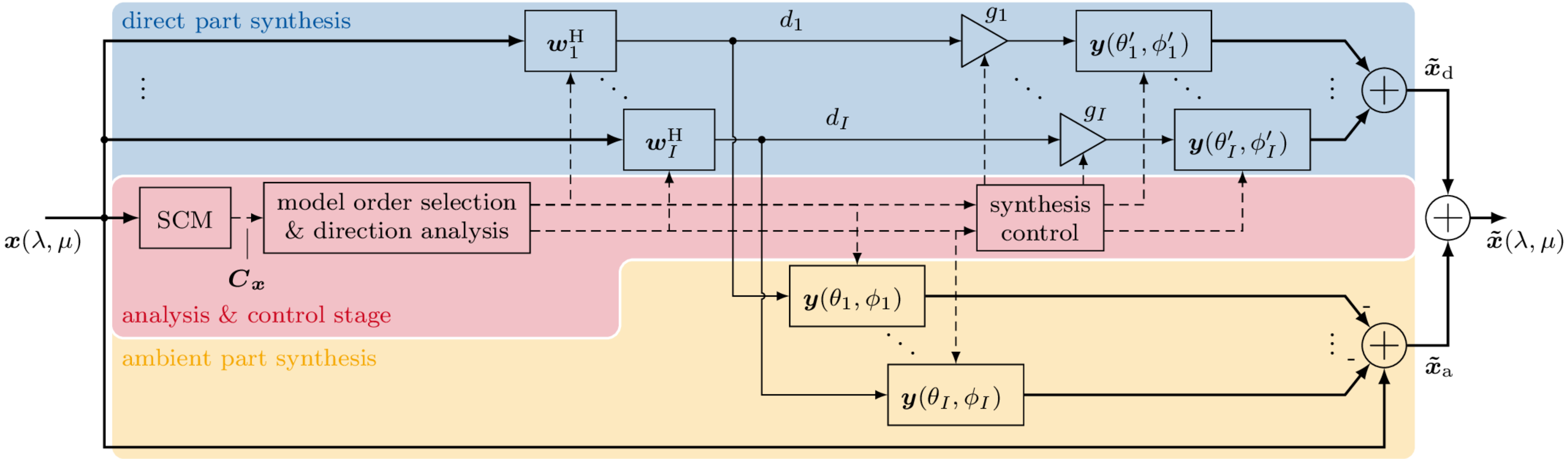
$$\boldsymbol{x}_s(\lambda,\mu) = \underbrace{\left[\,\boldsymbol{y}(\theta_1,\phi_2)\,|\,\dots\,|\,\boldsymbol{y}(\theta_{I(\lambda,\mu)},\phi_{I(\lambda,\mu)})\,\right]}_{=\text{ array manifold matrix } \boldsymbol{Y}_s} \cdot \begin{pmatrix} s_1(\lambda,\mu) \\ \vdots \\ s_{I(\lambda,\mu)}(\lambda,\mu) \end{pmatrix}$$

  - **Ambient part $\boldsymbol{x}_a(\lambda,\mu)$**: spatially diffuse

- Spatial covariance matrix (SCM):

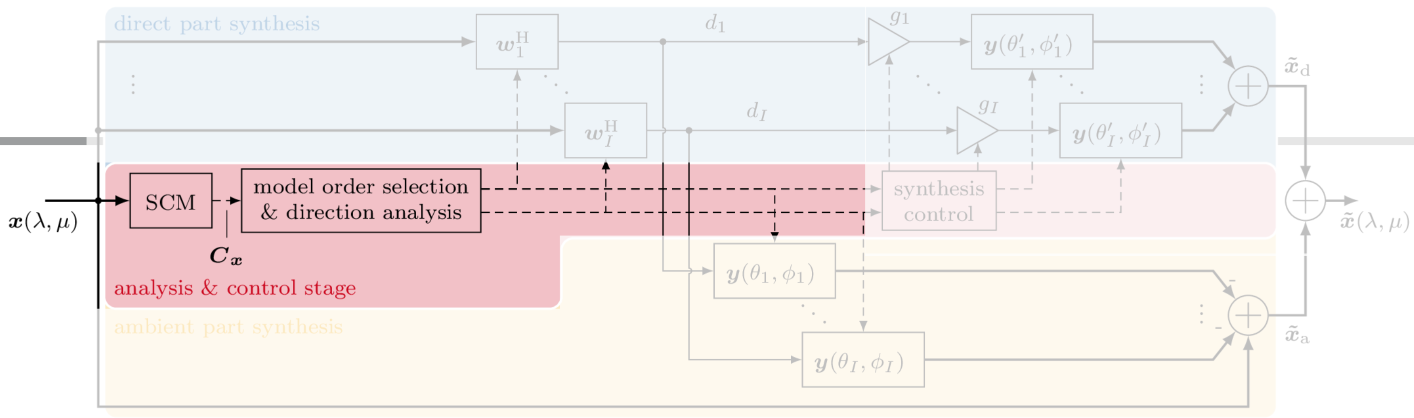
$$\boldsymbol{C}_{\boldsymbol{x}}(\lambda,\mu) = \mathrm{E}\{\boldsymbol{x}(\lambda,\mu)\boldsymbol{x}^H(\lambda,\mu)\}$$
$$= \mathrm{E}\{\boldsymbol{x}_s(\lambda,\mu)\boldsymbol{x}_s^H(\lambda,\mu)\} + \mathrm{diag}(\boldsymbol{\sigma}_a^2)$$

Proposed System Overview



SCM	Estimation of Spatial Covariance Matrix
(θ,ϕ)	Angles: (inclination, azimuth)
λ	Frame index
μ	Frequency index
$\boldsymbol{w}_i$	Beamforming weights
$\boldsymbol{y}(\theta,\phi)$	Vector of Spherical Harmonics evaluated at (θ,ϕ)

Analysis Stage

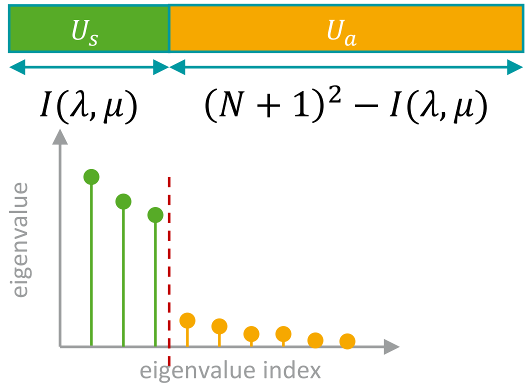


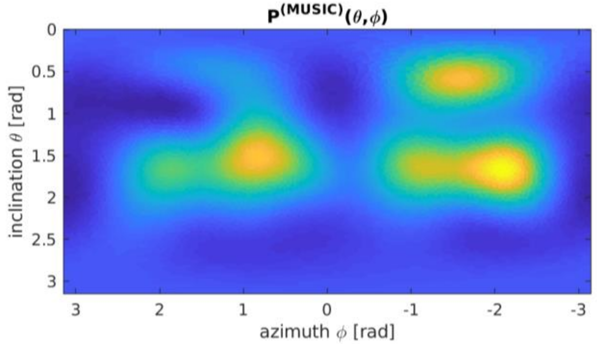
Segregation into multiple direct time-frequency components & ambient residuum

- **Variant 1:** Subspace approach using **SORTE & MUSIC**
  Adopted from [Politis, Tervo & Pulkki, ICASSP'18]

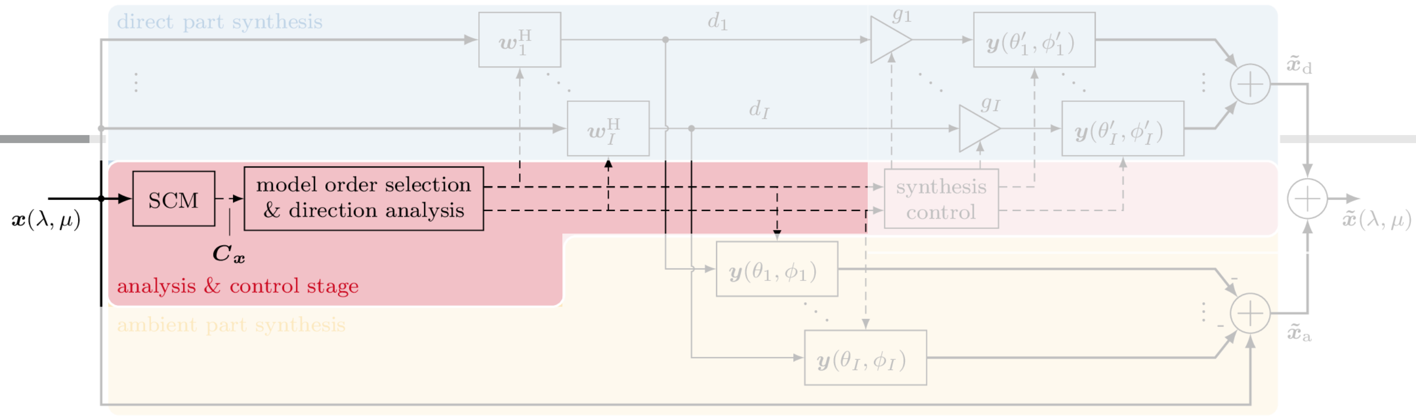
  - Use eigenvalue decomposition to segregate $\boldsymbol{C}_{\boldsymbol{x}}(\lambda,\mu)$ into
    - direct components subspace $U_s(\lambda,\mu)$
    - residual subspace $U_a(\lambda,\mu)$
  - Model order selection: determine number $I(\lambda,\mu)$
    - find gap in eigenvalue sequence using *second-order statistic of the eigenvalues* (SORTE)
  - Use MUSIC to find $I(\lambda,\mu)$ source directions which are local maxima in

$$P^{(\mathrm{MUSIC})}(\theta,\phi) = \frac{1}{\left\|\boldsymbol{y}^H(\theta,\phi)\boldsymbol{U}_a\right\|_2^2}$$





Analysis Stage

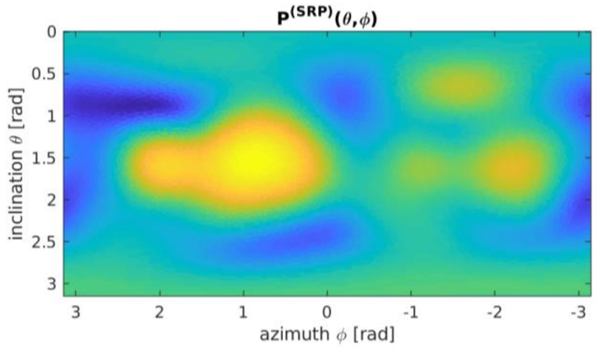


Segregation into multiple direct time-frequency components & ambient residuum

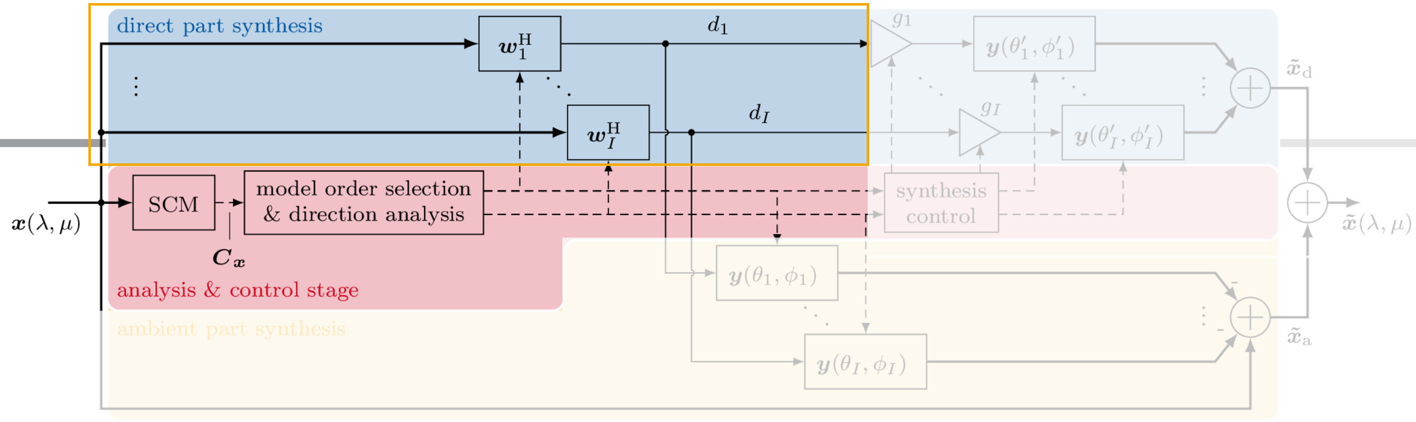
- **Variant 2:** Steered Response Power (**SRP**) Map
  - Source directions: find all local maxima in

$$P^{(\mathrm{SRP})}(\theta,\phi) = \boldsymbol{y}^{\mathrm{H}}(\theta,\phi)\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{y}(\theta,\phi)$$

  - Model order selection: I = # minima found



Analysis Stage



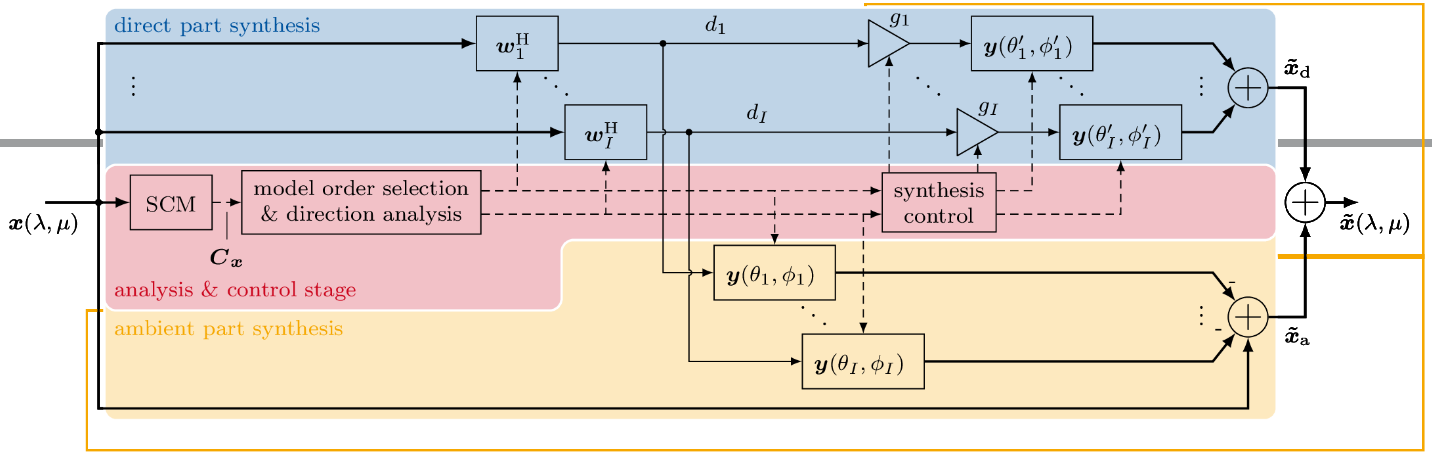
- **Beamforming** to extract direct components $i = 1,2, \dots, I$

$$d_i(\lambda, \mu) = \boldsymbol{w}_i^H(\lambda, \mu)\boldsymbol{x}(\lambda, \mu)$$

weights $\boldsymbol{w}_i$: distortionless-response constraint in direction (θ_i, ϕ_i) and null constraints for $(\theta_j, \phi_j),\ j \neq i$

- No semantic relation between direct components in time and frequency!

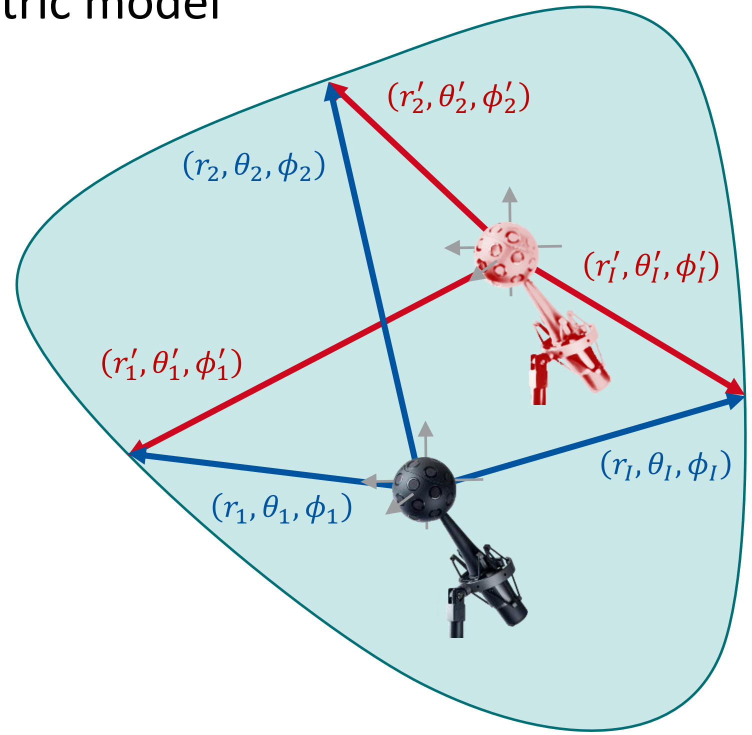
Synthesis Stage



- **Direct components:** repanning & gain adjustment according to geometric model

$$\tilde{\boldsymbol{x}}_d = [\, \boldsymbol{y}(\theta_1', \phi_1') \mid \dots \mid \boldsymbol{y}(\theta_I', \phi_I') \,] \cdot \begin{pmatrix} \frac{r_1}{r_1'} & & 0 \\ & \ddots & \\ 0 & & \frac{r_I}{r_I'} \end{pmatrix} \cdot \begin{pmatrix} d_i(\lambda, \mu) \\ \vdots \\ d_I(\lambda, \mu) \end{pmatrix}$$

a priori information on source distances $r_1, r_2, \dots, r_I$ needed

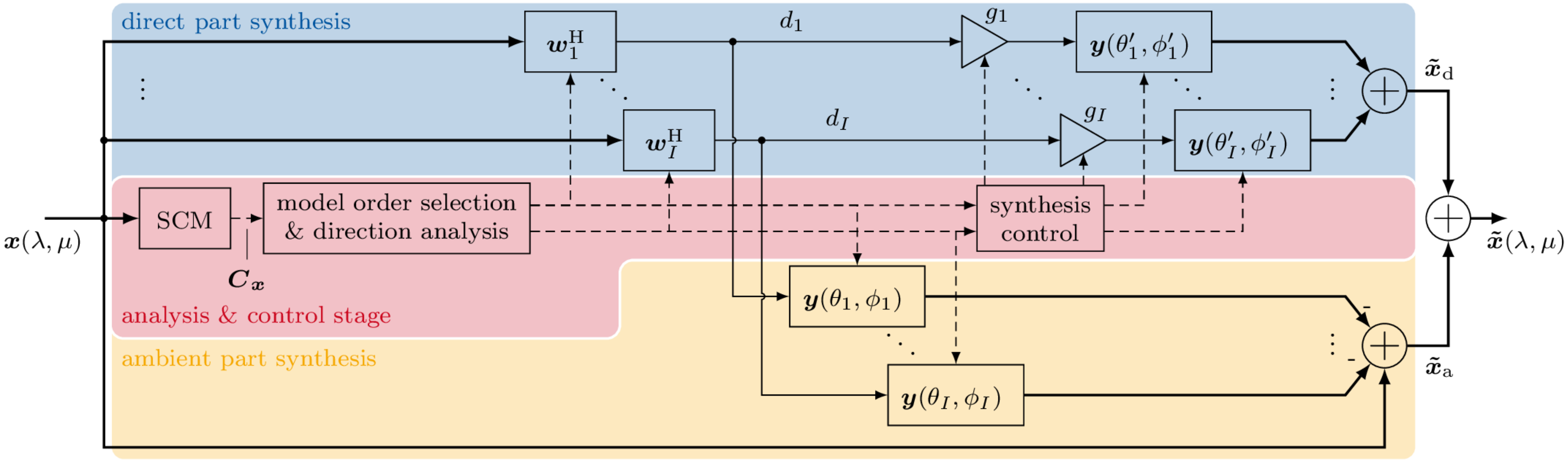


- **Ambient part:** cancel out direct components from input signal $\boldsymbol{x}(\lambda, \mu)$

- In total:

$$\tilde{\boldsymbol{x}}(\lambda, \mu) = \tilde{\boldsymbol{x}}_d(\lambda, \mu) + \tilde{\boldsymbol{x}}_a(\lambda, \mu)$$

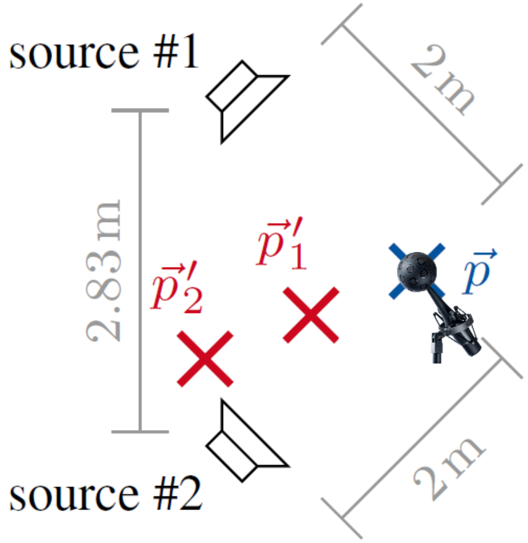
Discussion



- **Early reflections** not considered in signal model
  - Likely to be modified as if they were direct sound
  - **Precedence effect:** early reflections do not contribute to perceived source direction

Subjective Evaluation

MUSHRA Listening Test for Localization & Overall Audio Quality | 18 Participants



T_{60}	0.4 s
source #1	pop music
source #2	male speaker
noise	diffuse (−30 dB)
sample duration	8 s

- HOA scene (order N=4) with two sound sources
- Translation to $\vec{p}_1'$ and $\vec{p}_2'$

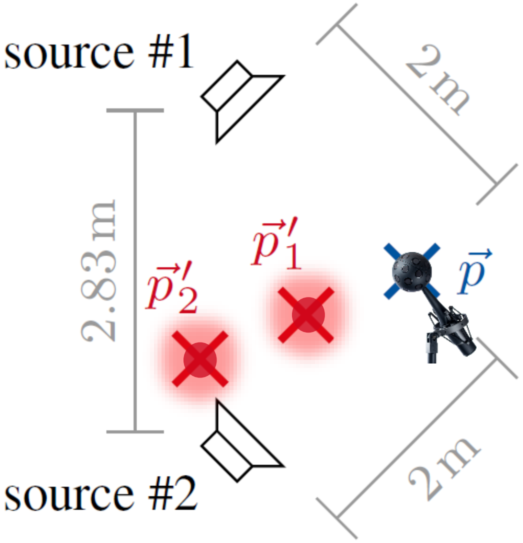
- Two variants of the proposed algorithm:
  - **MU**SIC/SORTE
  - **SR**P-based

- For reference:
  - **OR**ACLE: direction estimates replaced by actual source directions
  - Low anchor (**AL**): omni-directional sound, 4 kHz low-pass filtered
  - Mid anchor (**AM**): non-translated input signal $\boldsymbol{x}(\lambda, \mu)$
  - Hidden reference (**HR**): ground truth at $\vec{p}_1'$ and $\vec{p}_2'$, respectively

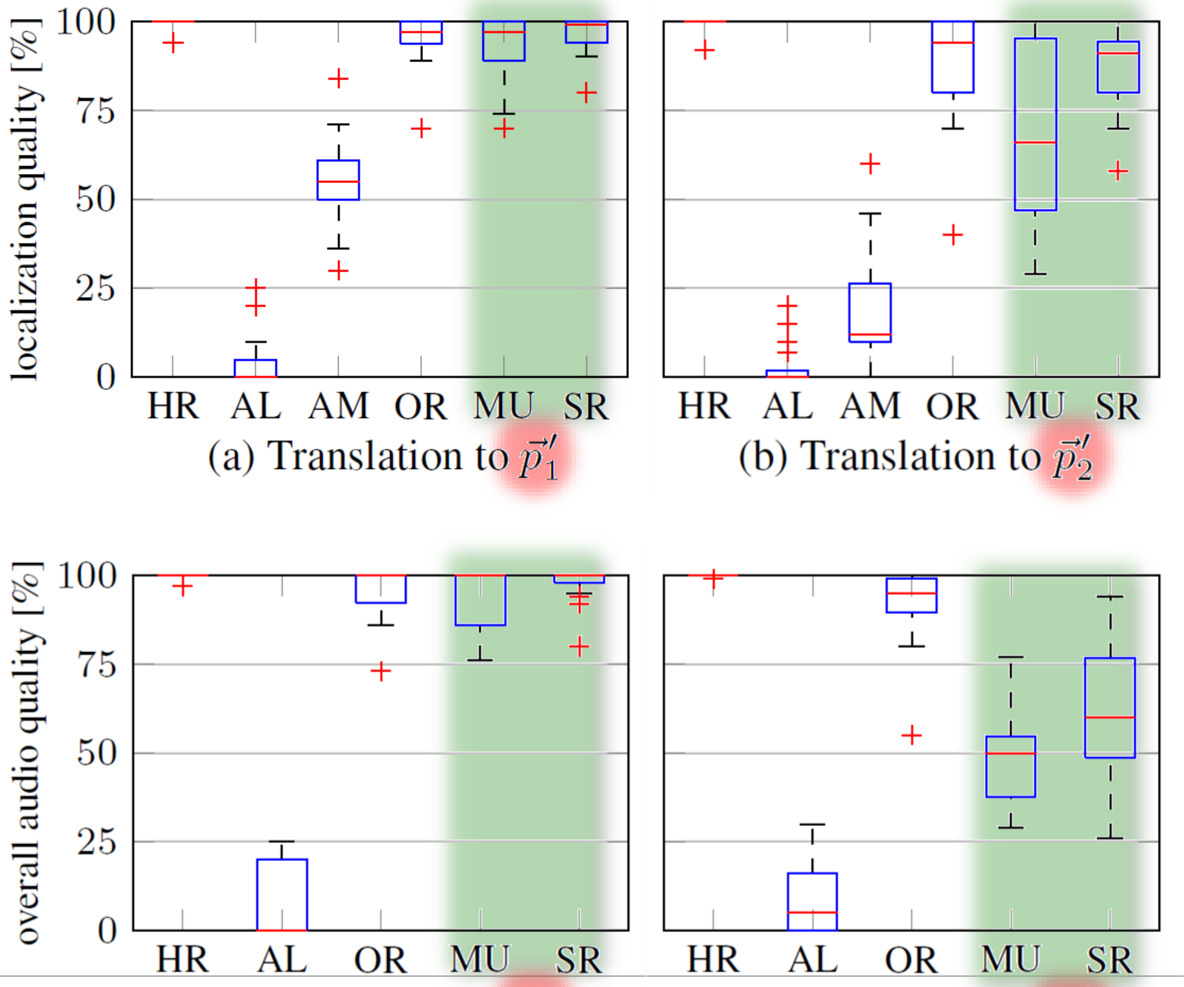
sampling rate	16 kHz
SH order N	4
frame length	512 samples
frame overlap	50 %
sub-band grouping	critical bands
rec. smoothing decay time	0.1 s
# spatial sampling points Q	900
oracle distance r_i	2 m

Subjective Evaluation

MUSHRA Listening Test for Localization & Overall Audio Quality | 18 Participants



T_{60}	0.4 s
source #1	pop music
source #2	male speaker
noise	diffuse (−30 dB)
sample duration	8 s



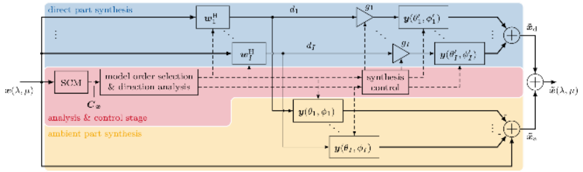
(a) Translation to $\vec{p}_1'$ (b) Translation to $\vec{p}_2'$

(c) Translation to $\vec{p}_1'$ (d) Translation to $\vec{p}_2'$

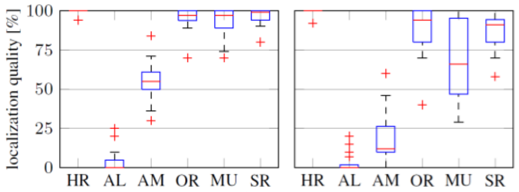
HR	hidden reference
AL	low anchor
AM	mid anchor (input signal)
OR	proposed method with oracle information on source directions
MU	**proposed method with MUSIC**
SR	**proposed method with SRP**

Conclusions

- Novel 3DoF+ system for translation in scene-based HOA content
  - Only single capture device needed



- Excellent subjective performance…
  - … despite model error due to early reflections not treated differently than direct sound (→ psychoacoustic precedence effect)
  - Median subject ratings in localization between 91 % and 99 % (SRP method)



(a) Translation to $\vec{p}_1'$ (b) Translation to $\vec{p}_2'$

- Applications: VR/AR systems, immersive teleconferencing & telepresence

Institute of Communication Systems

RWTH AACHEN UNIVERSITY