Model-Based and Learning-Based Approaches for Speech Enhancement and Source Localisation

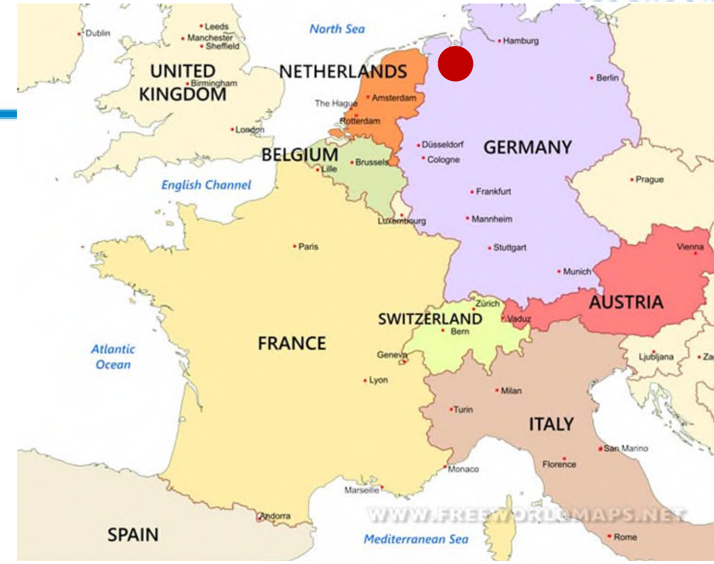
Prof. Dr. Simon Doclo

University of Oldenburg, Dept. of Medical Physics and Acoustics
and Cluster of Excellence Hearing4all

Aalborg Universitet – March 13, 2023

University of Oldenburg

- **Founded in 1973**
- **Named after Carl von Ossietzky**
- **16.000 students**
- **250 professors,**
  1.300 scientific staff
- **Research profile:**
  - Humans & Technology (Hearing Research, Sensory Neuroscience)
  - Environment & Sustainability
  - Society & Education



Hearing research in Oldenburg

since 1993

Dept. for Medical Physics and Acoustics

- Basic research
- Education

10 Professors
20+ Postdocs
50+ PhD students

since 2000

Institute of Hearing technology and Audiology

- Education
- Application-oriented research

since 1996

Hörzentrum gGmbH

- Application-oriented research (hearing devices)
- Audiological consulting
- Evaluation studies

since 2008

Branch Hearing, Speech and Audio Technology

- Application-oriented research (consumer electronics)

Research facilities

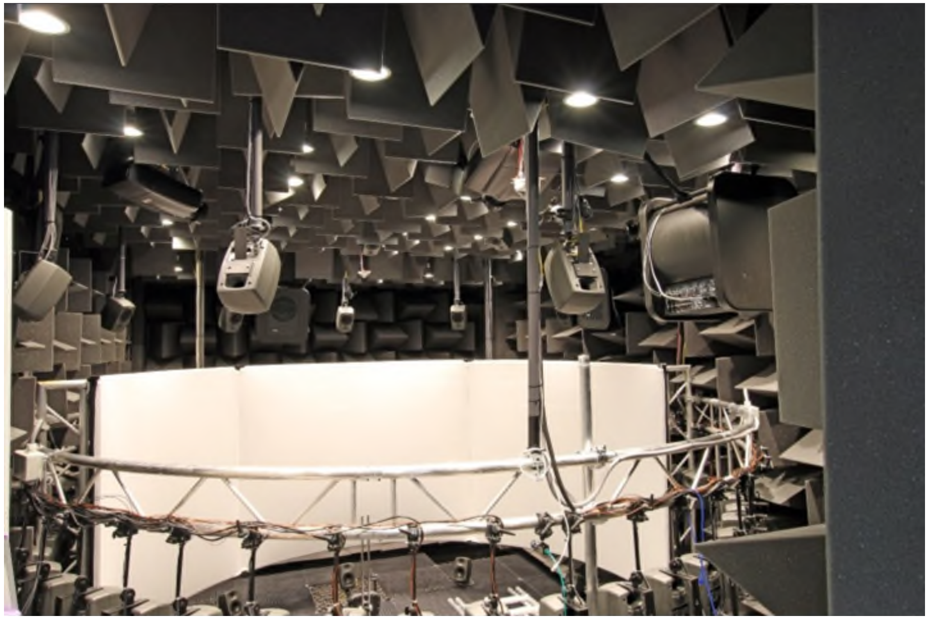
- **Free-field and sound-proof listening booths**
- **Anechoic chamber** (8,5m x 7m x 4m; $f_c \approx 50$ Hz)

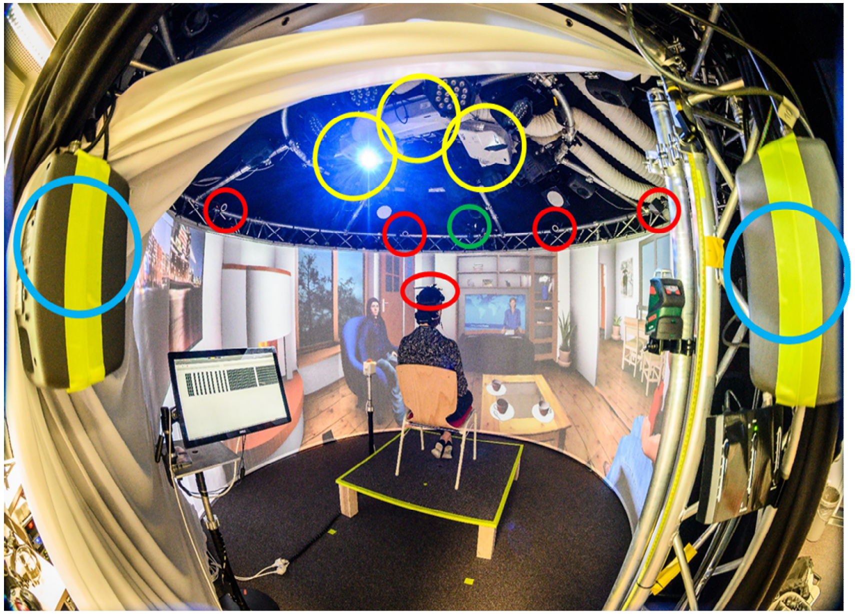
Research facilities

- **Communication acoustics simulator** (active system, 16 microphones + 24 loudspeakers, T_{60}: 0.4 – 4 sec)
- **Variable acoustics lab** (passive, T_{60}: 0.2 – 1 sec)

Research facilities

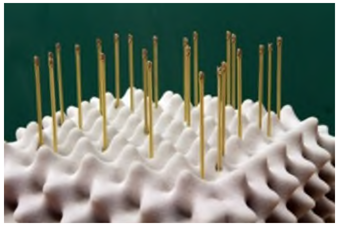
- **Virtual reality lab** (3D Ambisonics, 86 loudspeakers, cylindrical screen video projection)
- **Gesture lab** (interactive audio-visual scenes, motion/head tracking, eye movement/EOG)





Research topics

- **Single- and multi-microphone speech enhancement**
  - **Noise reduction** (DNN-based, exploiting interframe correlation)
  - **Dereverberation** (spectral enhancement, multi-channel equalization, blind probabilistic model-based)
  - **Acoustic sensor networks** (spatially distributed microphones, sampling rate offset estimation, distributed processing)
  - **Computational acoustic scene analysis** (CASA, localization)
  - **Beamformer design** (e.g., virtual artificial head)



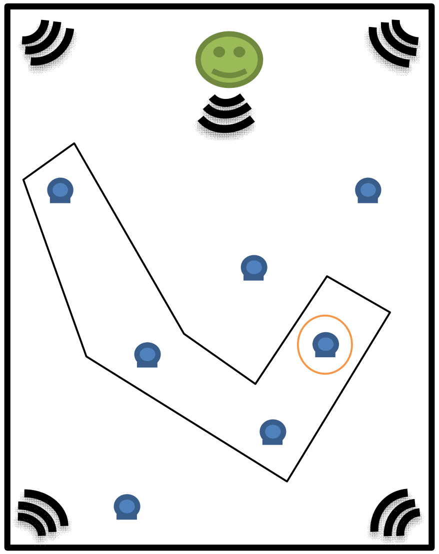
- **Signal processing for ear-mounted communication devices**
  - **Binaural noise reduction**, aiming at preserving spatial impression of acoustic scene (binaural cues)
  - Open-fitting hearing devices: **acoustic transparency**, **feedback cancellation** and **active noise/occlusion control**
  - EEG-based **auditory attention decoding** for steering beamformers

I. Acoustic sensor networks

Acoustic sensor networks

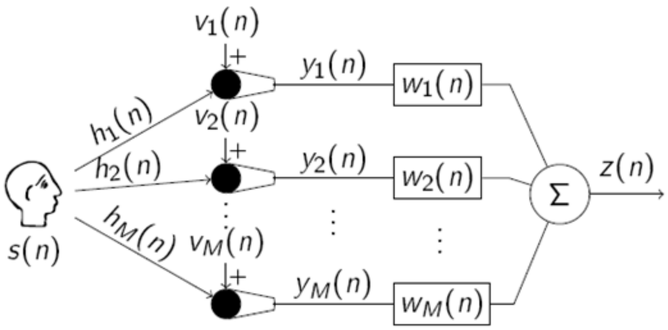
- Exploit spatial diversity of **spatially distributed microphones** for improved speech enhancement and source localisation

- Previous and current research:
  - Low-complexity method to **estimate relative transfer function (RTF)** vector of target speaker for hearing aids + external microphone(s)
  - Improved trade-off between **noise reduction** and **binaural cue preservation**
  - (Binaural) **source localization** exploiting external microphones
  - **Dereverberation** using weighted prediction error method with microphone-dependent prediction delay
  - Microphone utility and **subset selection**
  - **Sampling rate offset** estimation



Blind multi-microphone speech enhancement

- **Filter-and-sum structure :** $z = \mathbf{w}^H \mathbf{y}$



Blind multi-microphone speech enhancement

- **Filter-and-sum structure :** $\boxed{z = \mathbf{w}^H \mathbf{y}}$
- **"Workhorse algorithm": parametric Multi-channel Wiener filter (MWF)**

Goal: estimate desired speech component in reference microphone + trade off interference reduction and speech distortion

$$\min_{\mathbf{w}} \mathcal{E}\{|\mathbf{w}^H\mathbf{x} - x_1|^2\} + \mu\mathcal{E}\{|\mathbf{w}^H\mathbf{i}|^2\} \quad \Rightarrow \quad \mathbf{w}_{MWF} = (\mathbf{\Phi}_x + \mu\mathbf{\Phi}_i)^{-1}\mathbf{\Phi}_x\mathbf{e}$$

$\rightarrow$ **requires** estimate of covariance matrices

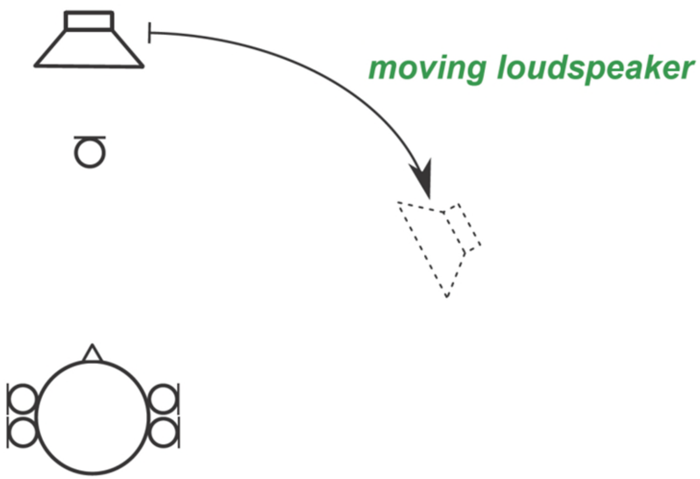
Can be decomposed as **MVDR beamformer and spectral postfilter**

$$\boxed{\mathbf{w}_{MWF} = \frac{\mathbf{\Phi}_i^{-1}\mathbf{a}}{\mathbf{a}^H\mathbf{\Phi}_i^{-1}\mathbf{a}} \frac{\phi_{x_1}}{\phi_{x_1} + \mu(\mathbf{a}^H\mathbf{\Phi}_i^{-1}\mathbf{a})^{-1}}}$$

$\rightarrow$ **requires** estimate/model of interference covariance matrix, estimate/model of relative transfer function (RTF) vector of desired speaker, and PSDs of speech and interference components at MVDR output

RTF vector estimation exploiting external microphone

- **Estimate RTF vector of target speaker** to steer binaural MVDR beamformer
- **Spatial coherence (SC) method:** assume that noise components in hearing aid microphones and external microphone are uncorrelated, e.g., when external microphone is spatially separated from HA microphones + diffuse noise field



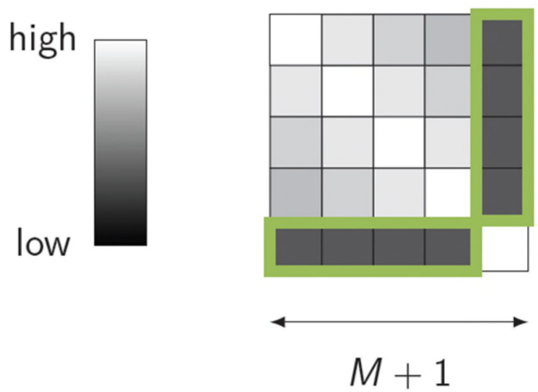
$$\mathbf{w}_L = \frac{\mathbf{R}_v^{-1}\mathbf{a}_L}{\mathbf{a}_L^H \mathbf{R}_v^{-1}\mathbf{a}_L}$$

RTF vector estimation exploiting external microphone

- **Estimate RTF vector of target speaker** to steer binaural MVDR beamformer
- **Spatial coherence (SC) method:** assume that noise components in hearing aid microphones and external microphone are uncorrelated, e.g., when external microphone is spatially separated from HA microphones + diffuse noise field

  → correlate noisy HA microphone signals with noisy external microphone signal and normalize by reference element
- **Low computational complexity** with similar (even better in practice) performance than state-of-the-art covariance whitening (CW) approach

$$\mathbf{w}_L = \frac{\mathbf{R}_v^{-1}\mathbf{a}_L}{\mathbf{a}_L^H \mathbf{R}_v^{-1}\mathbf{a}_L}$$



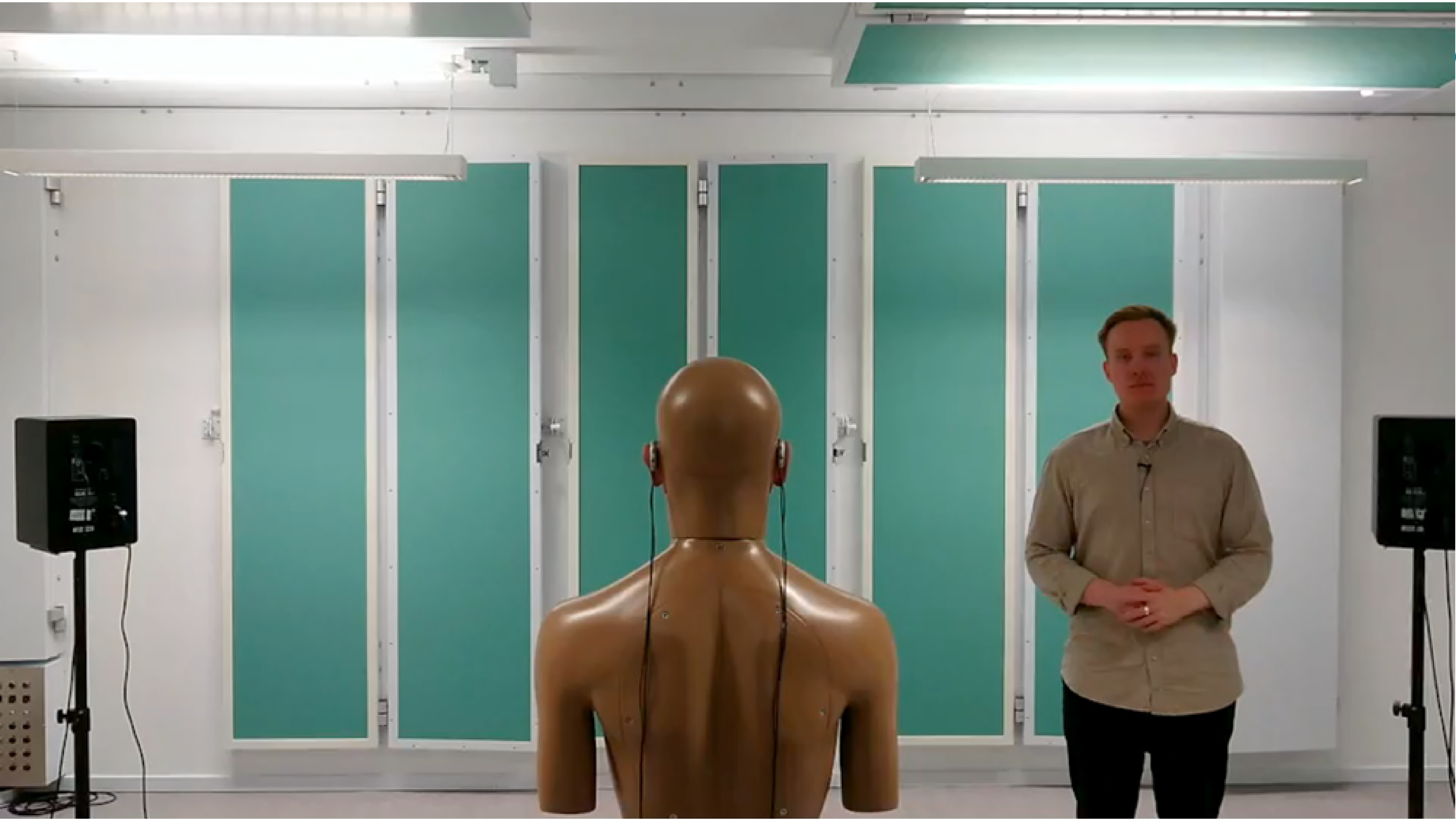
Low spatial coherence between noise components in local array signals and external microphone signal

$$\bar{\mathbf{a}}_{\mathrm{L}}^{\mathrm{SCE}} = \frac{\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}{\mathbf{e}_{\mathrm{L}}^T\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}, \quad \bar{\mathbf{a}}_{\mathrm{R}}^{\mathrm{SCE}} = \frac{\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}{\mathbf{e}_{\mathrm{R}}^T\bar{\mathbf{R}}_{\mathrm{y}}\mathbf{e}_{\mathrm{E}}}$$

$$\bar{\mathbf{w}}_{\mathrm{L}}^{\mathrm{SCE}} = \begin{bmatrix} \alpha \cdot [\mathbf{I}_{2M},\ \mathbf{0}_{2M\times 1}]\,\bar{\mathbf{w}}_{\mathrm{L}} \\ \alpha(1+\beta)\cdot \mathbf{e}_{\mathrm{E}}^T\bar{\mathbf{w}}_{\mathrm{L}} \end{bmatrix}$$

[Gößling, Doclo, *Proc. IWAENC* 2018] [Gößling, Doclo, *Proc. ICASSP 2019*] [Gößling, Marquardt, Doclo, *IEEE/ACM TASLP*, 2021]

Audio Demo



Extension: Multiple external microphones

- Each external microphone yields (different) RTF estimate
- **Linear combination/selection** of RTF estimates (per frequency)

$$\boxed{\boldsymbol{a}_{\mathrm{L}}^{\mathrm{SC-C}} = \frac{\boldsymbol{A}_{\mathrm{L}}^{\mathrm{SC}}\boldsymbol{c}}{\boldsymbol{e}_{\mathrm{L}}^{T}\boldsymbol{A}_{\mathrm{L}}^{\mathrm{SC}}\boldsymbol{c}}}$$

1. Input SNR-based selection

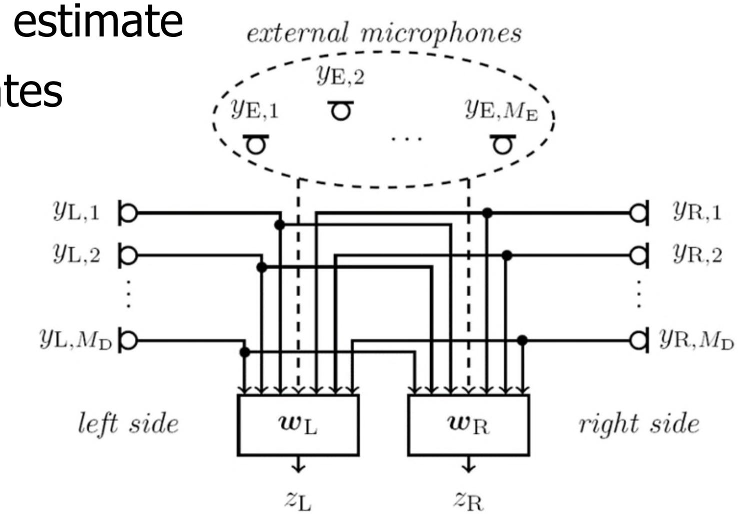
$$\boldsymbol{c}^{\mathrm{iSNR}} = \boldsymbol{e}_{\mathrm{E},\hat{i}}\,, \quad \hat{i} = \arg\max_{i} \frac{\boldsymbol{e}_{\mathrm{E},i}^{T}\boldsymbol{R}_{\mathrm{y}}\boldsymbol{e}_{\mathrm{E},i}}{\boldsymbol{e}_{\mathrm{E},i}^{T}\boldsymbol{R}_{\mathrm{n}}\boldsymbol{e}_{\mathrm{E},i}}$$

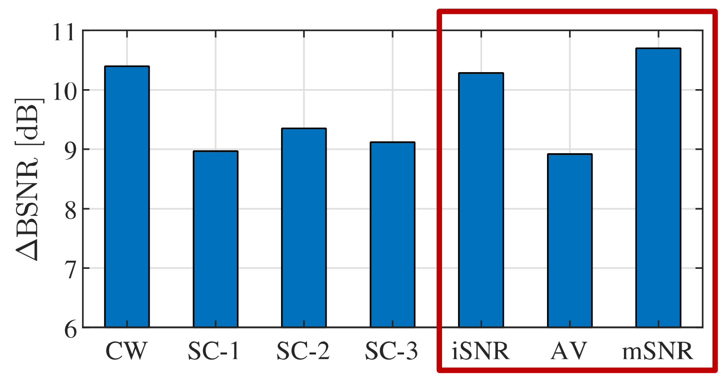
2. Simple averaging

$$\boldsymbol{c}^{\mathrm{AV}} = \left[\frac{1}{M_{\mathrm{E}}}, \ldots, \frac{1}{M_{\mathrm{E}}}\right]^{T}$$

3. Output SNR-maximizing combination

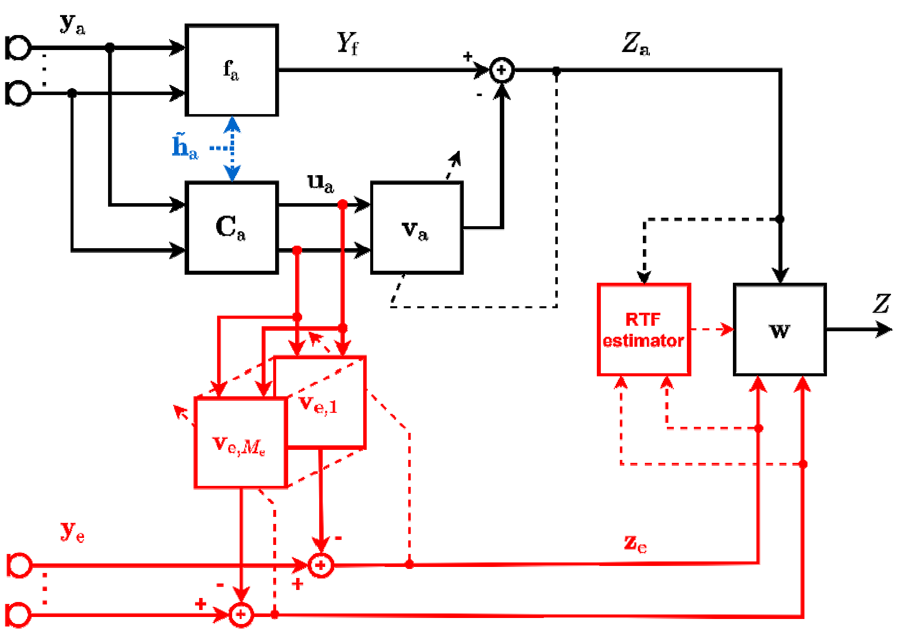
$$\boxed{\boldsymbol{c}^{\mathrm{mSNR}} = \arg\max_{\boldsymbol{c}}\ \mathrm{SNR}_{\mathrm{BMVDR,L}}^{\mathrm{out}} = \mathcal{P}\{\boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Lambda}_1\}}$$





Extension: Partial RTF vector estimation

- **Partial RTF vector estimation** in general acoustic scenario (e.g. interfering speaker and noise)



- **Assumption:** part of RTF vector is known (e.g. anechoic steering vector for hearing aids)

$$\mathbf{a} = \begin{bmatrix} \tilde{\mathbf{a}}_{\text{known}} \\ \mathbf{a}_{\text{unknown}} \end{bmatrix}$$

- **GSC-ESR structure:** create external speech references by removing undesired components (interference, noise) in external microphone signals using noise+interference references of Generalized Sidelobe Canceller structure

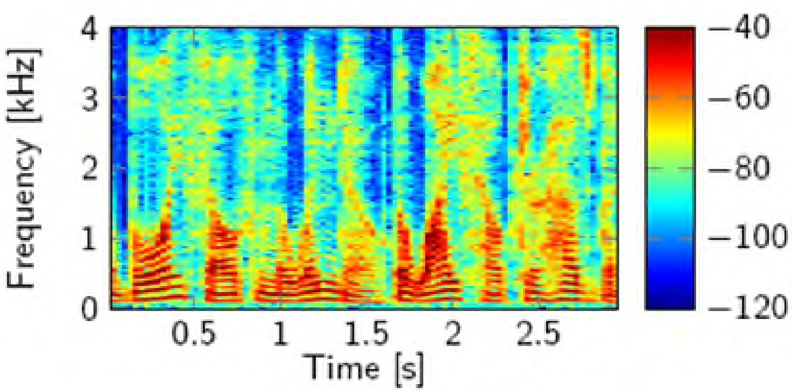
$$\mathbf{v}_{e,m_e} = \left(\mathbf{C}_a^H \mathbf{R}_{n,a} \mathbf{C}_a\right)^{-1} \mathbf{C}_a^H \mathbf{E}_a \mathbf{R}_n \mathbf{e}_{e,m_e} \qquad \mathbf{v}_{e,m_e} = \left(\mathbf{C}_a^H \mathbf{R}_{y,a} \mathbf{C}_a\right)^{-1} \mathbf{C}_a^H \mathbf{E}_a \mathbf{R}_y \mathbf{e}_{e,m_e}$$

Dereverberation: Weighted prediction error

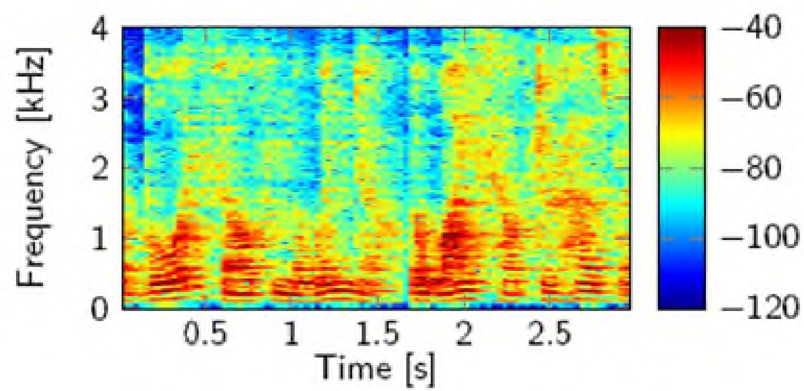
- **Goal**: estimate clean speech STFT coefficients $s(k,l)$ from reverberant (and noisy) STFT coefficients $y_m(k,l)$ by subtracting late reverberant component

$$y_m(k,l) = \underbrace{h_m(k,l) \star s(k,l)}_{x_m(k,l)} + v_m(k,l)$$

  - Probabilistic estimation using (statistical) **models of desired speech signal and reverberation**
  - Exploit **sparsity properties** of speech in STFT-domain



Clean



Reverberant

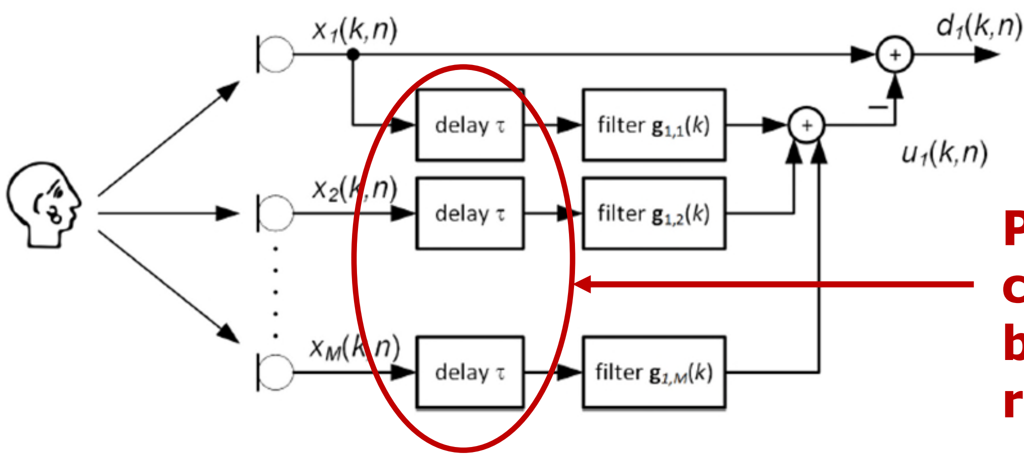
Dereverberation: Weighted prediction error

- **Weighted prediction error (WPE) method for dereverberation**

$$\mathbf{x}_1(k) = \mathbf{d}(k) + \mathbf{X}_\tau(k)\mathbf{g}(k)$$

$$\hat{\mathbf{d}}(k) = \mathbf{x}_1(k) - \mathbf{X}_\tau(k)\hat{\mathbf{g}}(k)$$

predicted reverberation



Prediction delay plays crucial role / trade-off between residual reverberation and distortion

- **Prediction delay** is usually chosen based on correlation properties of speech, i.e. microphone-independent

Dereverberation: Weighted prediction error

- **Generalization of original WPE approach** [Nakatani et al., 2010]
  - STFT coefficients of desired signal are assumed to be modelled using **circular sparse/super-Gaussian prior with time-varying variance** $\lambda(n)$

$$\rho(d(n)) = \max_{\lambda(n)>0} \mathcal{N}_{\mathbb{C}}(d(n);0,\lambda(n))\boxed{\psi(\lambda(n))}$$

    Scaling function $\psi(.)$ can be interpreted as **hyper-prior on variance**

  - **Maximum-Likelihood Estimation**

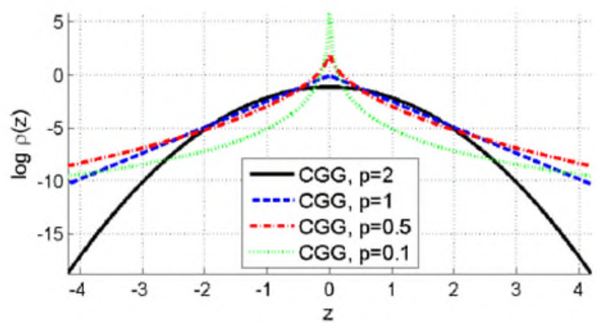
$$\mathcal{L}(\mathbf{g}) = \prod_{n=1}^{N} \rho(d(n)) \Rightarrow \min_{\boldsymbol{\lambda}>0,\mathbf{g}} \sum_{n=1}^{N} \left(\frac{|d(n)|^2}{\lambda(n)} + \log \pi\lambda(n) - \boxed{\log \psi(\lambda(n))} \right)$$

  - **Alternating optimization procedure**
    1. Estimate **prediction vector**

$$\hat{\mathbf{g}}^{(i+1)} = \left(\mathbf{X}_\tau^H \mathcal{D}^{-1}_{\hat{\boldsymbol{\lambda}}^{(i)}} \mathbf{X}_\tau\right)^{-1} \mathbf{X}_\tau^H \mathcal{D}^{-1}_{\hat{\boldsymbol{\lambda}}^{(i)}} \mathbf{x}_1$$

    2. Estimate **variances** (assuming complex generalized Gaussian prior with shape parameter p)

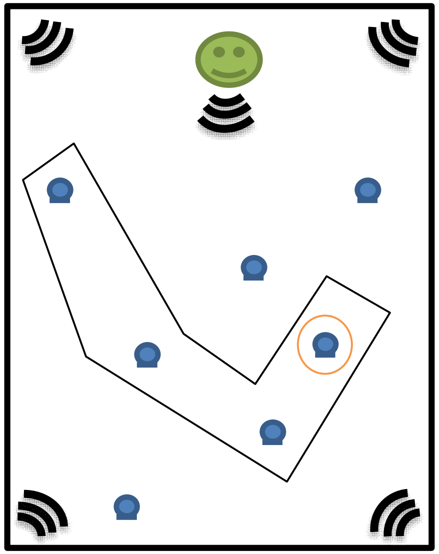
$$\boxed{\hat{\lambda}^{(i+1)}(n) = |\hat{d}^{(i+1)}(n)|^{2-p},}$$



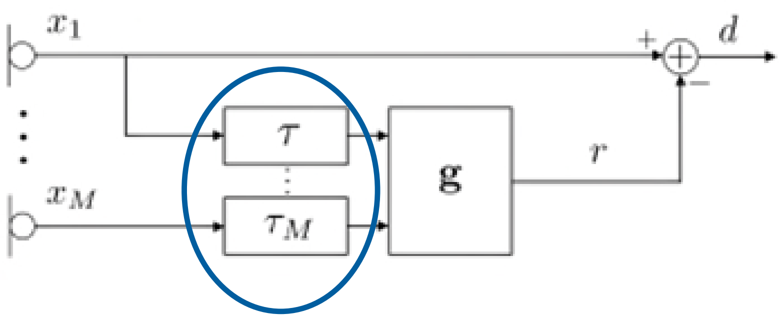
WPE for acoustic sensor networks

- When microphones are spatially distributed, time differences of arrival (TDOAs) between microphones may be large and diverse
- When using WPE with a **fixed prediction delay**, this may lead to distortion or excessive reverberation

  ⇒ apply TDOA compensation to WPE input, leading to **microphone-dependent prediction delays**

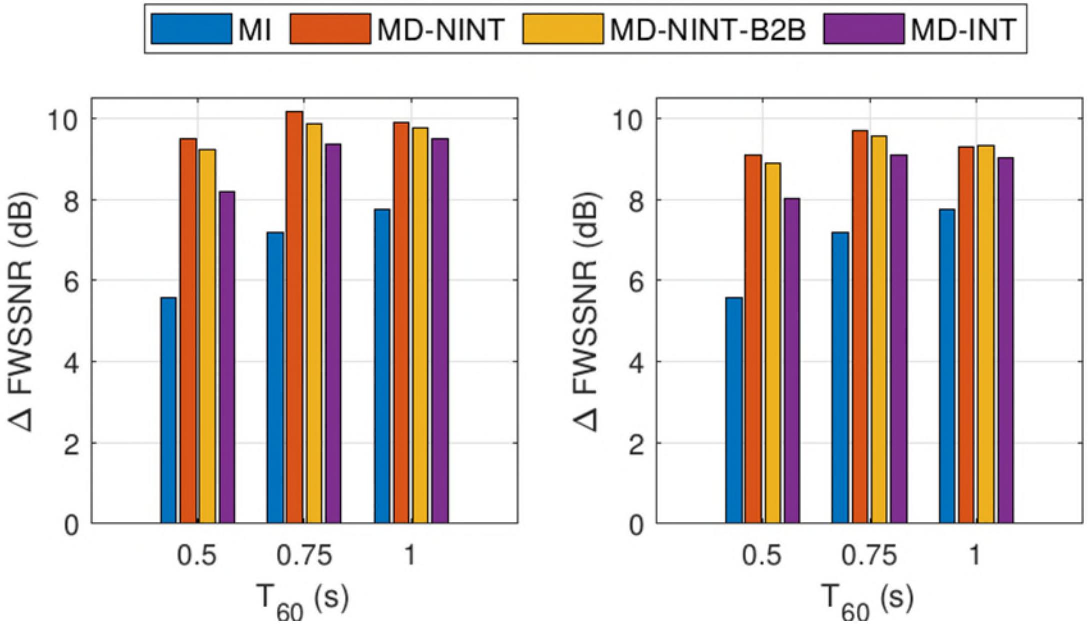


- Different schemes to implement prediction delays
  - non-integer prediction delays with crossband filters (NINT)
  - non-integer prediction delays with band-to-band approximation (NINT-B2B)
  - (coarse) integer prediction delays (INT)



WPE for acoustic sensor networks

- **Simulation results:**
  - Fixed prediction delay (MI) may result in low speech quality, depending on position of speech source
  - Microphone-dependent prediction delays: NINT performs best, closely followed by NINT-B2B; INT performs worse than NINT, however at significantly lower computational complexity

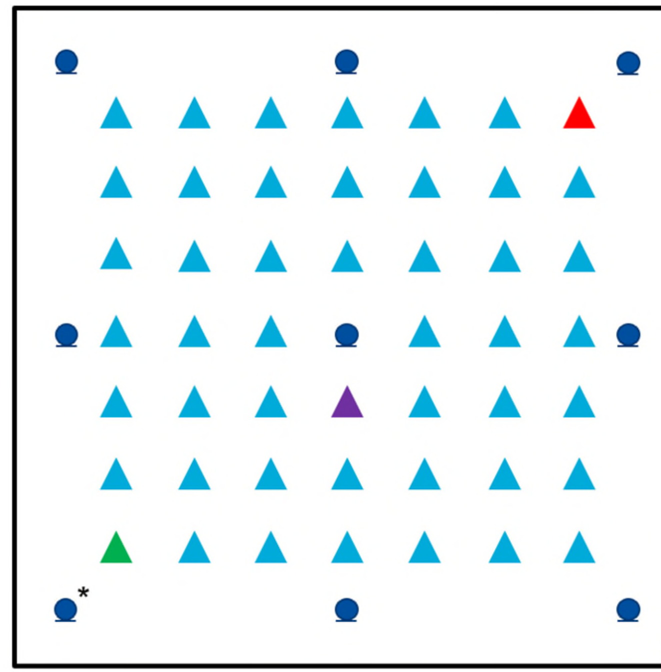


M=9, fs=16 kHz; STFT: 64ms (overlap 16ms); WPE: L_g=12, τ=2, p=0.5

WPE for acoustic sensor networks

- **Simulation results:**

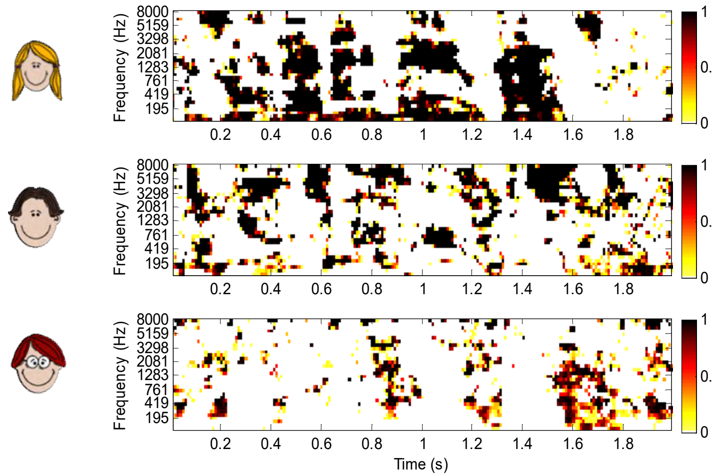
	Position 1	Position 2
Reverberant microphone signal		
Fixed prediction delay		
Microphone-dependent prediction delay (NINT)		



$T_{60} \approx 750$ms, M=9, fs=16 kHz; STFT: 64ms (overlap 16ms); WPE: L_g=12, τ=2, p=0.5; estimated TDOAs (GCC-PHAT)

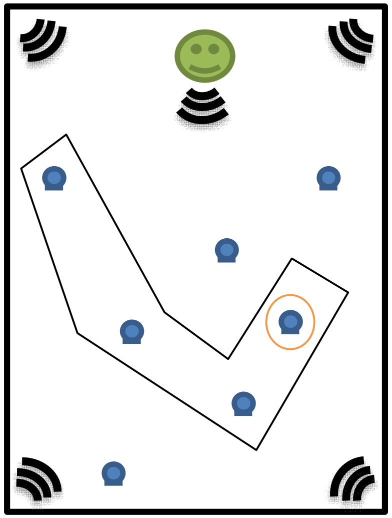
Current/future work

- **Complex and time-varying scenarios**: incorporate CASA into control path of algorithms, switch between keeping all speakers or removing undesired speakers



- **Smart speaker scenario:** multiple nodes with multiple microphones

- **WPE-based dereverberation in acoustic sensor networks:** microphone utility, microphone subset selection, reference microphone selection



- **(Binaural) source localisation** exploiting external microphones

- **Sampling rate offset estimation and compensation** for distributed noise reduction (DANSE)

II. Deep multi-frame noise reduction

Carl von Ossietzky
Universität
Oldenburg
Signal Processing Group

Deep Multi-Frame Noise Reduction

March 13, 2023

Marvin Tammen, Simon Doclo

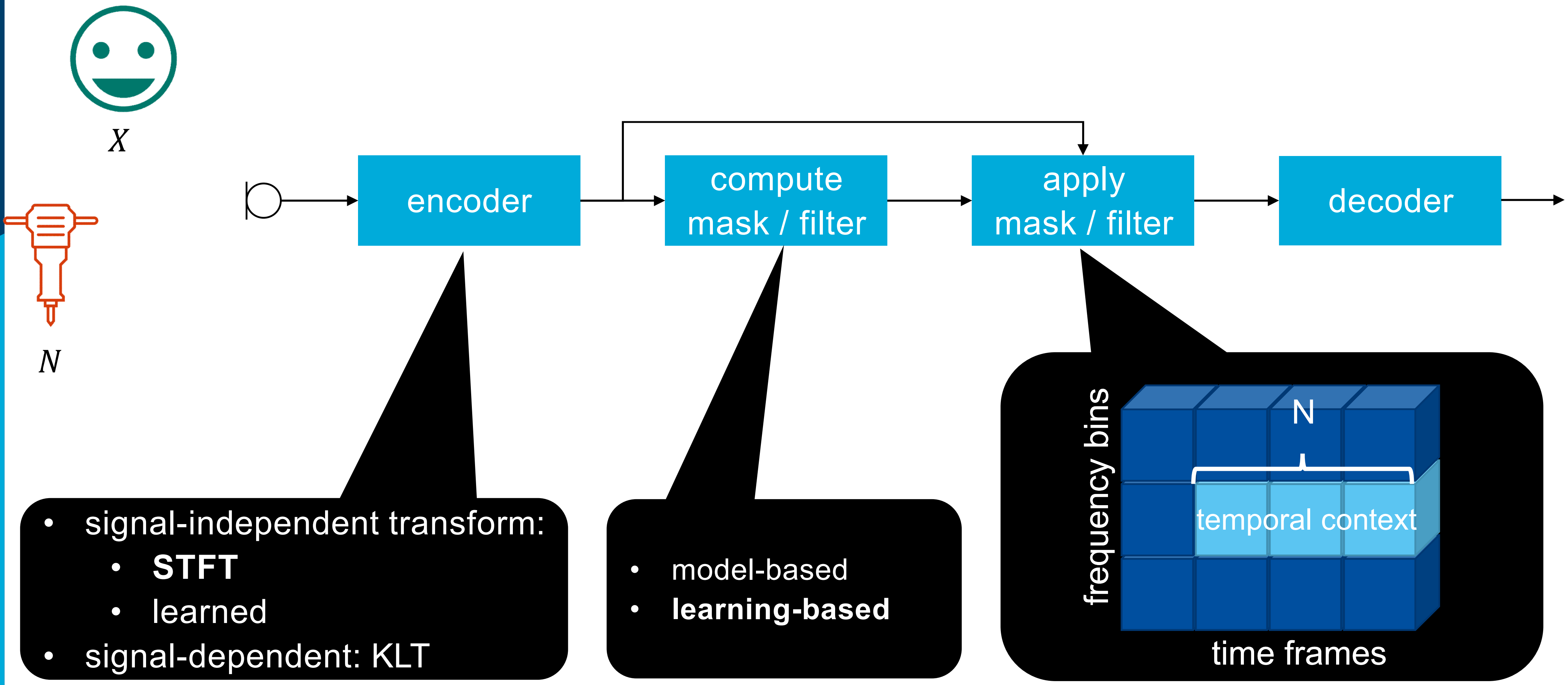
Outline

Deep Multi-Frame Noise Reduction for Single-Microphone Speech Enhancement

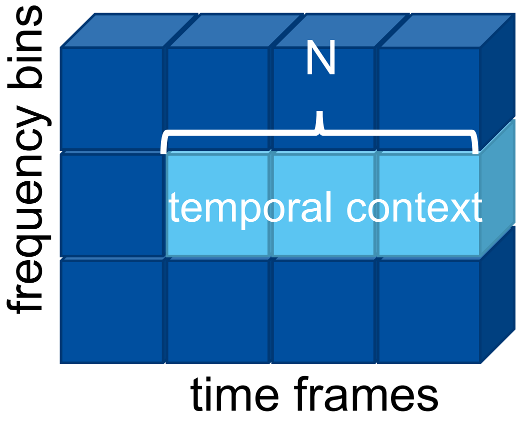
- Problem Statement
- Multi-Frame MVDR Filter

Extension Towards Binaural Noise Reduction

Deep Multi-Frame Noise Reduction



Signal Model



- noisy multi-frame vector: $\boldsymbol{y}_t = [Y_t \quad \dots \quad Y_{t-N+1}]^T = \boldsymbol{x}_t + \boldsymbol{n}_t$
- multi-frame speech vector $\boldsymbol{x}_t = [X_t \quad \dots \quad X_{t-N+1}]^T$
- $\boldsymbol{x}_t$ can be decomposed into **temporally correlated and uncorrelated components** w.r.t. X_t:

$$\boldsymbol{x}_t = \boldsymbol{\gamma}_{x,t}\, X_t + \boldsymbol{x}'_t, \qquad \boldsymbol{\gamma}_{x,t} = \frac{\mathcal{E}\{\boldsymbol{x}_t X_t^*\}}{\mathcal{E}\{|X_t|^2\}} \in \mathbb{C}^N$$

→ highly time-varying **speech interframe correlation (IFC) vector $\boldsymbol{\gamma}_{x,t}$**

→ depends on sound (e.g. voiced vs. unvoiced)

Multi-Frame MVDR Filter

- minimize output noise PSD while preserving temporally correlated speech component:

$$\mathbf{w}_t^{MFMVDR} = \min_{\mathbf{w}} \mathbf{w}^H \boldsymbol{\Phi}_{n,t}\, \mathbf{w}, \;\; \text{s.t. } \mathbf{w}^H \boldsymbol{\gamma}_{x,t} = 1$$

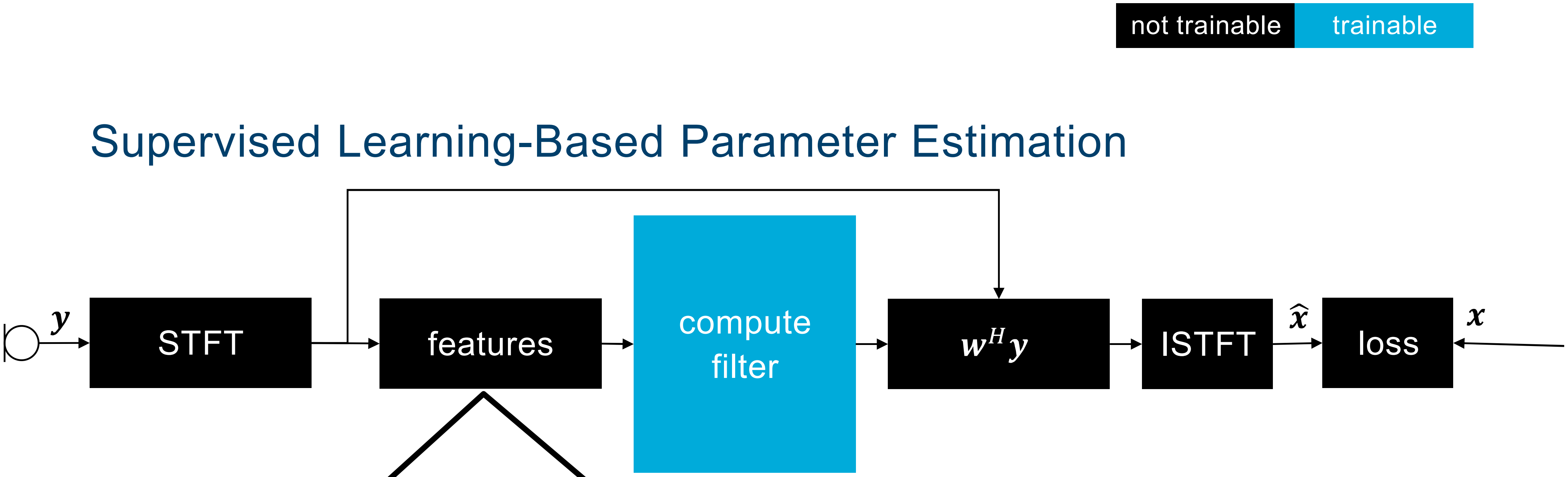
- solved by multi-frame MVDR (MFMVDR) filter:

$$\boxed{\mathbf{w}_t^{MFMVDR} = \frac{\boldsymbol{\Phi}_{n,t}^{-1}\boldsymbol{\gamma}_{x,t}}{\boldsymbol{\gamma}_{x,t}^{H}\boldsymbol{\Phi}_{n,t}^{-1}\boldsymbol{\gamma}_{x,t}}}$$

  - requires estimate of inverse noise covariance matrix $\boldsymbol{\Phi}_{n,t}^{-1}$ and speech IFC vector $\boldsymbol{\gamma}_{x,t}$
  - **Deep MFMVDR filter**: estimate quantities by integrating fully differentiable MFMVDR filter into supervised learning framework, minimizing time-domain loss function at output of MFMVDR filter

Carl von Ossietzky
Universität
Oldenburg
Signal Processing Group

Supervised Learning-Based Parameter Estimation

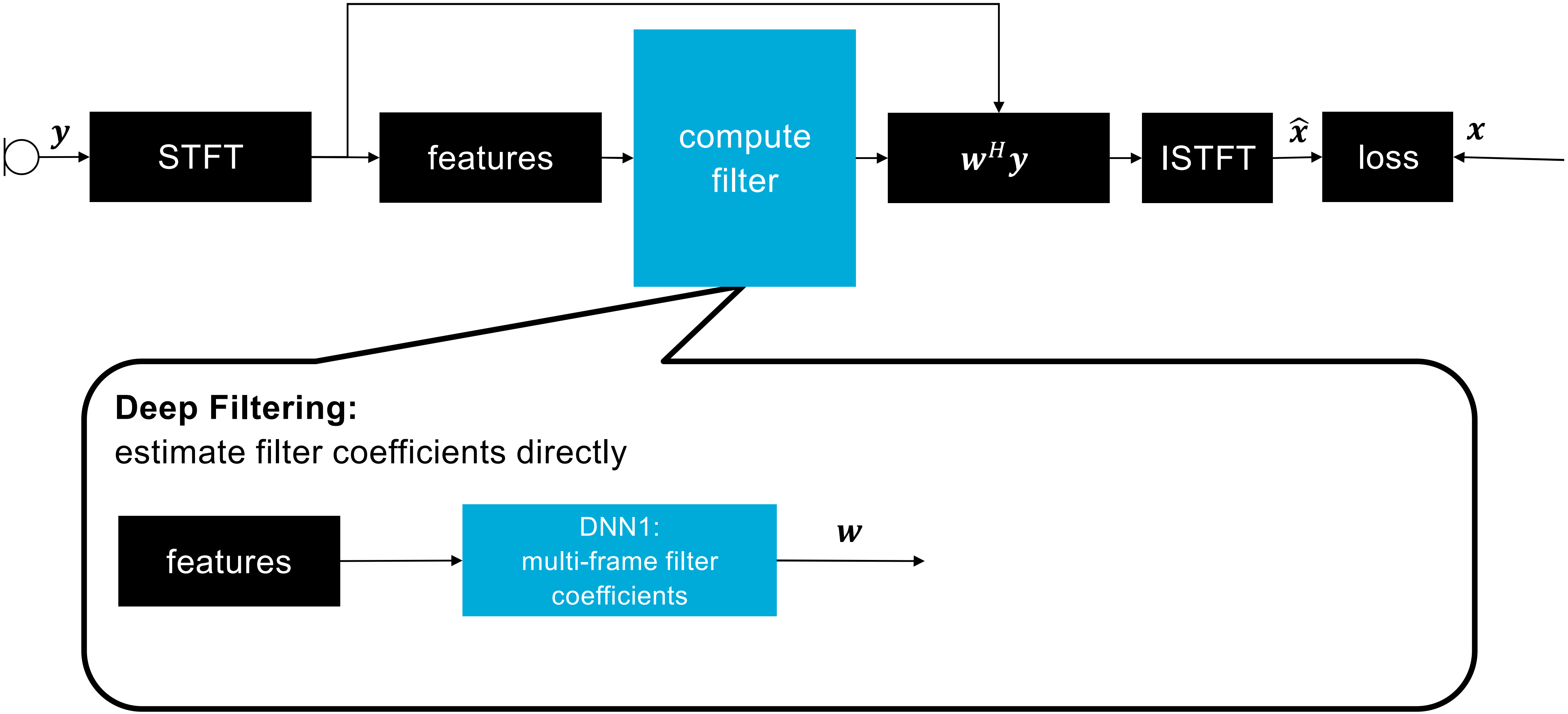


Features: concatenation of

1. logarithm of noisy magnitude
2. cosine of noisy phase
3. sine of noisy phase

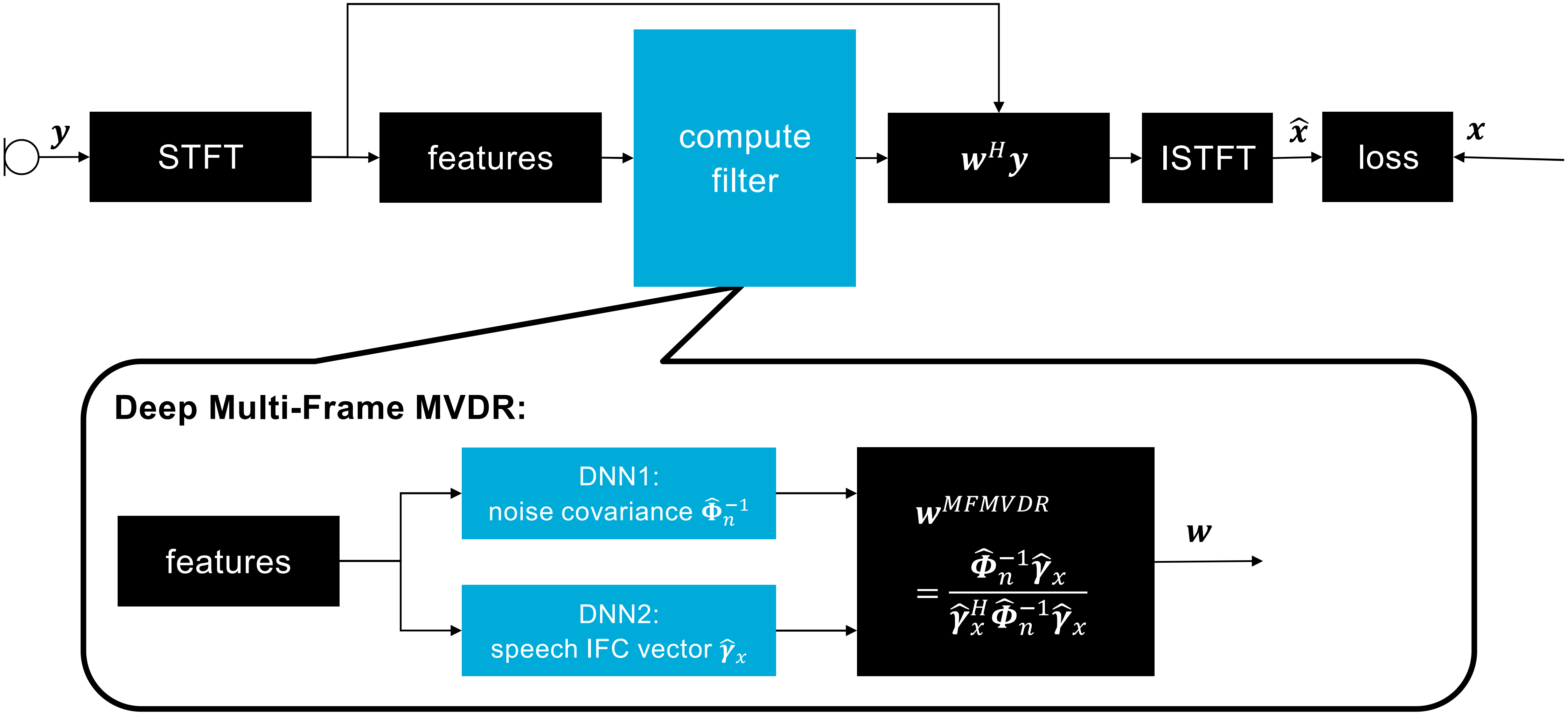
not trainable | trainable

Supervised Learning-Based Parameter Estimation



not trainable | trainable

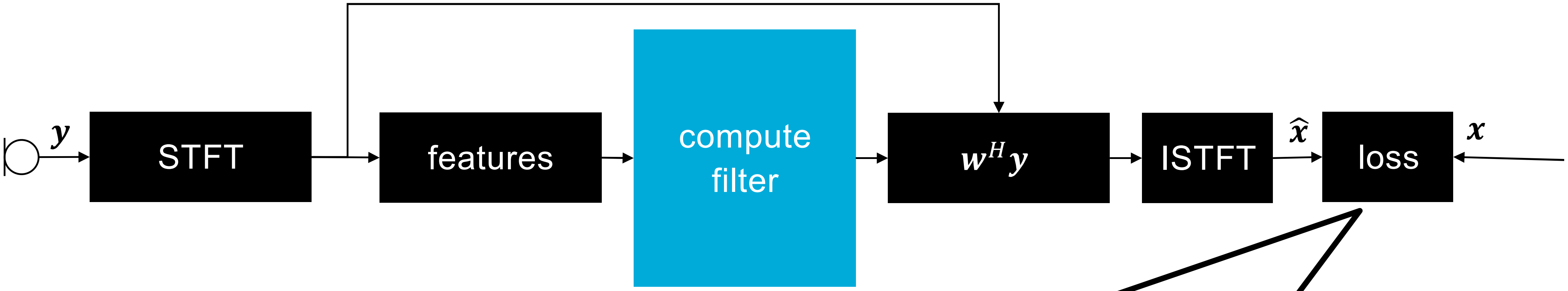
Supervised Learning-Based Parameter Estimation



Carl von Ossietzky Universität Oldenburg
Signal Processing Group

not trainable | trainable

Supervised Learning-Based Parameter Estimation



Loss: Scale-Invariant Signal-to-Distortion Ratio (SI-SDR)

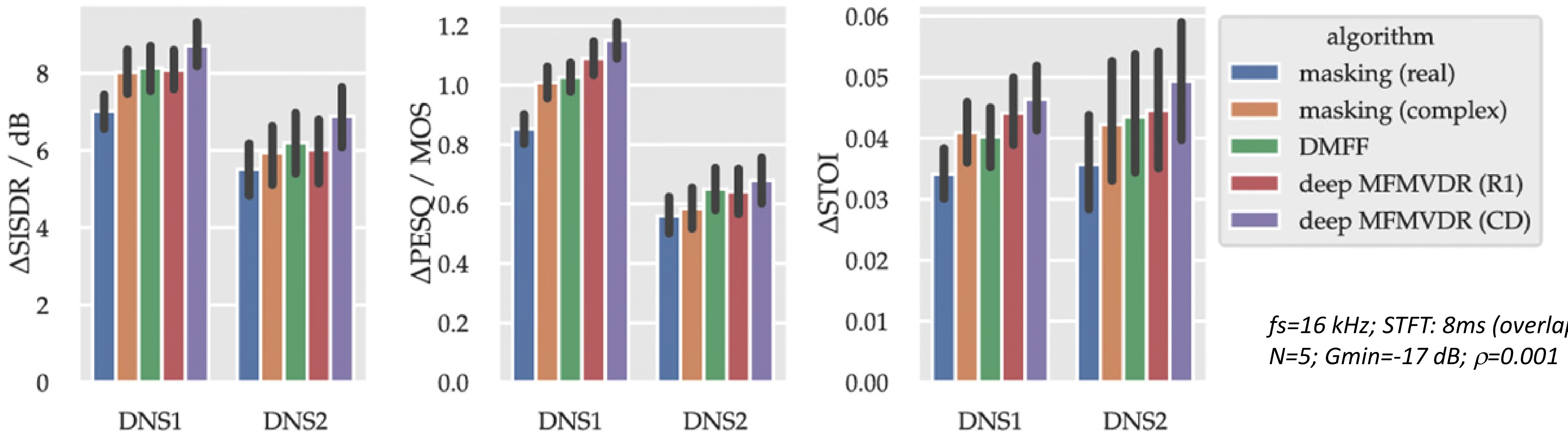
$$L = 10 \log_{10}\left(\frac{\|\boldsymbol{x}\|^2}{\|\boldsymbol{x}-\alpha\hat{\boldsymbol{x}}\|^2}\right), \alpha = \frac{\hat{\boldsymbol{x}}^T \boldsymbol{x}}{\|\boldsymbol{x}\|^2}$$

- [J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, in *Proc. 2019 ICASSP]*
- popular time-domain loss for speech enhancement and separation algorithms

Carl von Ossietzky
Universität
Oldenburg
Signal Processing Group

Simulation Results

- **Deep Noise Suppression (DNS) challenge datasets**: diverse speech and noise sources
- DNN architecture: causal **temporal convolutional network (TCN)**: 2 stacks of 4 layers each, kernel size 3 → temporal receptive field of 61 frames (128 ms)



fs=16 kHz; STFT: 8ms (overlap 6ms); N=5; Gmin=-17 dB; ρ=0.001

- Performance benefit of
  - **complex-valued masking** vs. **real-valued masking**
  - **MFMVDR structure** vs. **direct filtering**

Simulation Results

- **Real-time factor**

algorithm	RTF
masking (real)	0.151 ± 0.001
masking (complex)	0.152 ± 0.002
DMFF	0.157 ± 0.004
deep MFMVDR (R1)	0.230 ± 0.003
deep MFMVDR (CD)	0.324 ± 0.013

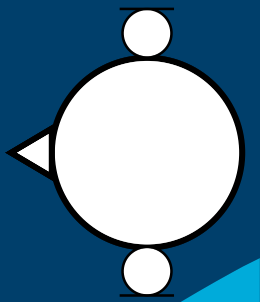
- **Network size**

algorithm	bottleneck dimension	weights / (1×10^6)
deep MFMVDR (SPP)	231	5.3
masking (real)	226	5.0
masking (complex)	226	5.0
DMFF	226	5.2
deep MFMVDR (RS)	128	4.9
deep MFMVDR (CD)	128	5.3
deep MFMVDR (PDT)	128	5.1
deep MFMVDR (R1)	128	5.1

Simulation Results - Audio examples

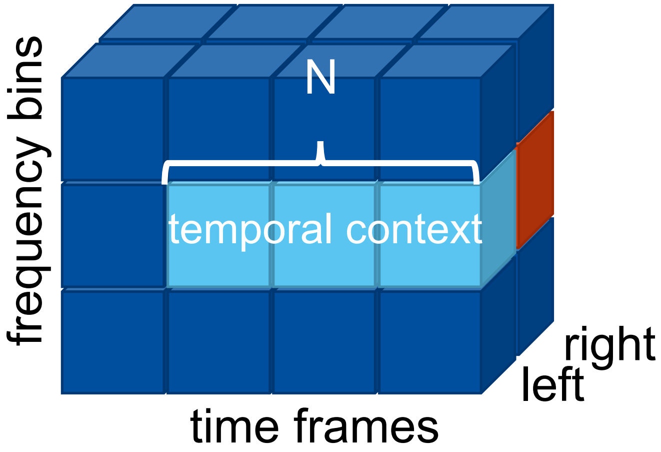
noisy		
single-frame mask, complex		
multi-frame filter, direct estimation		
multi-frame filter, MFMVDR structure		

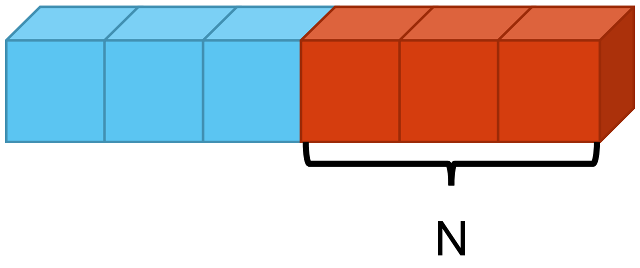
Carl von Ossietzky
Universität
Oldenburg
Signal Processing Group



Extension Towards Binaural (Multi-Microphone) Noise Reduction

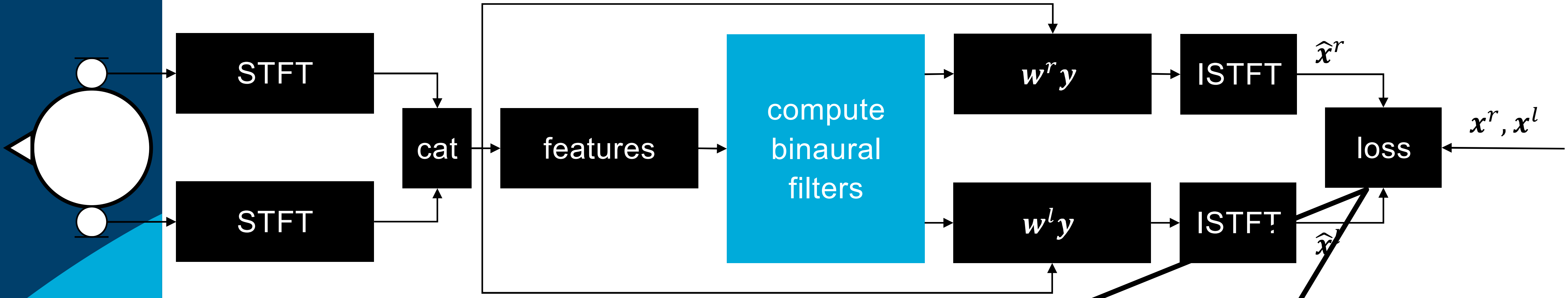
	monaural	binaural
signal vector	$\boldsymbol{y}_t = [Y_t \quad \dots \quad Y_{t-N+1}]^T$	$\boldsymbol{y}_t = [Y_t^l \quad \dots \quad Y_{t-N+1}^l \quad Y_t^r \quad \dots \quad Y_{t-N+1}^r]^T$
target signal	X_t	X_t^l, X_t^r
used correlations	temporal	spatio-temporal





not trainable | trainable

Supervised Learning-Based Parameter Estimation



Loss: Combined Mean Absolute Spectral Error

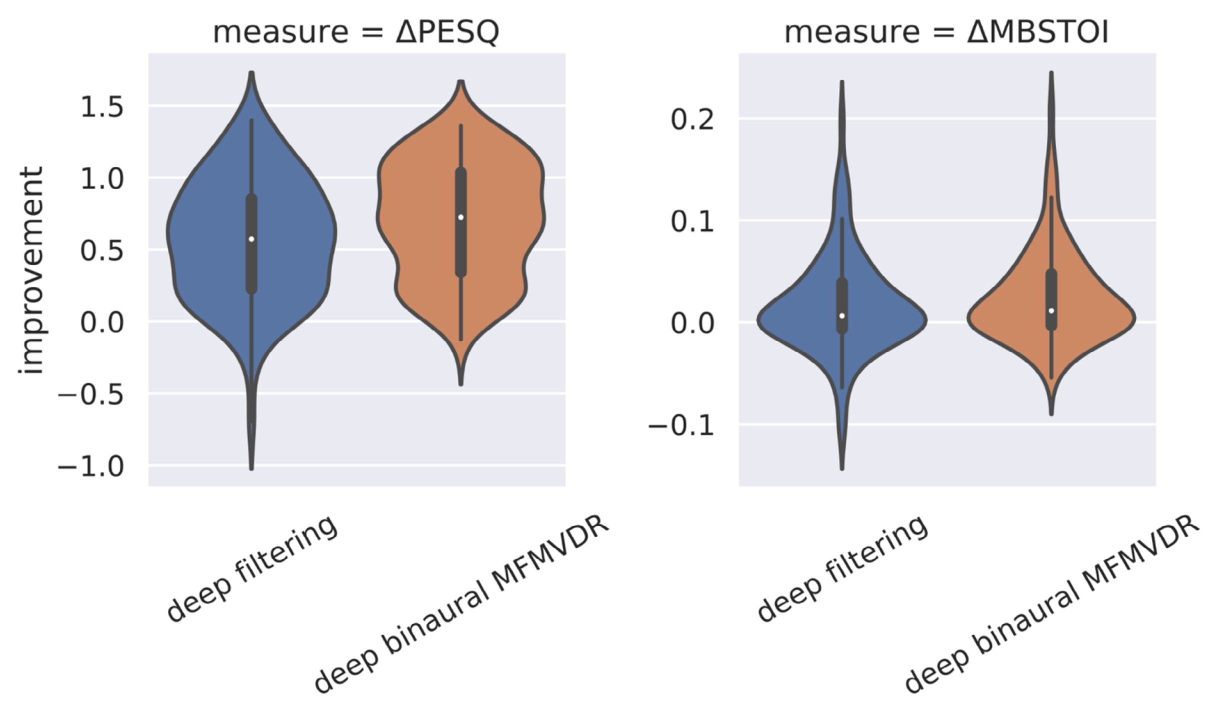
$$\frac{1}{2} \sum_{\nu \in \{l,r\}} \beta \left| X^\nu - \hat{X}^\nu \right| + (1-\beta) \left| |X^\nu| - |\hat{X}^\nu| \right|$$

- more robust against reverberation than SI-SDR
- [Z.-Q. Wang, P. Wang, and D. Wang, *IEEE/ACM TASLP*, 2020]

Carl von Ossietzky Universität Oldenburg
Signal Processing Group

Simulation Results

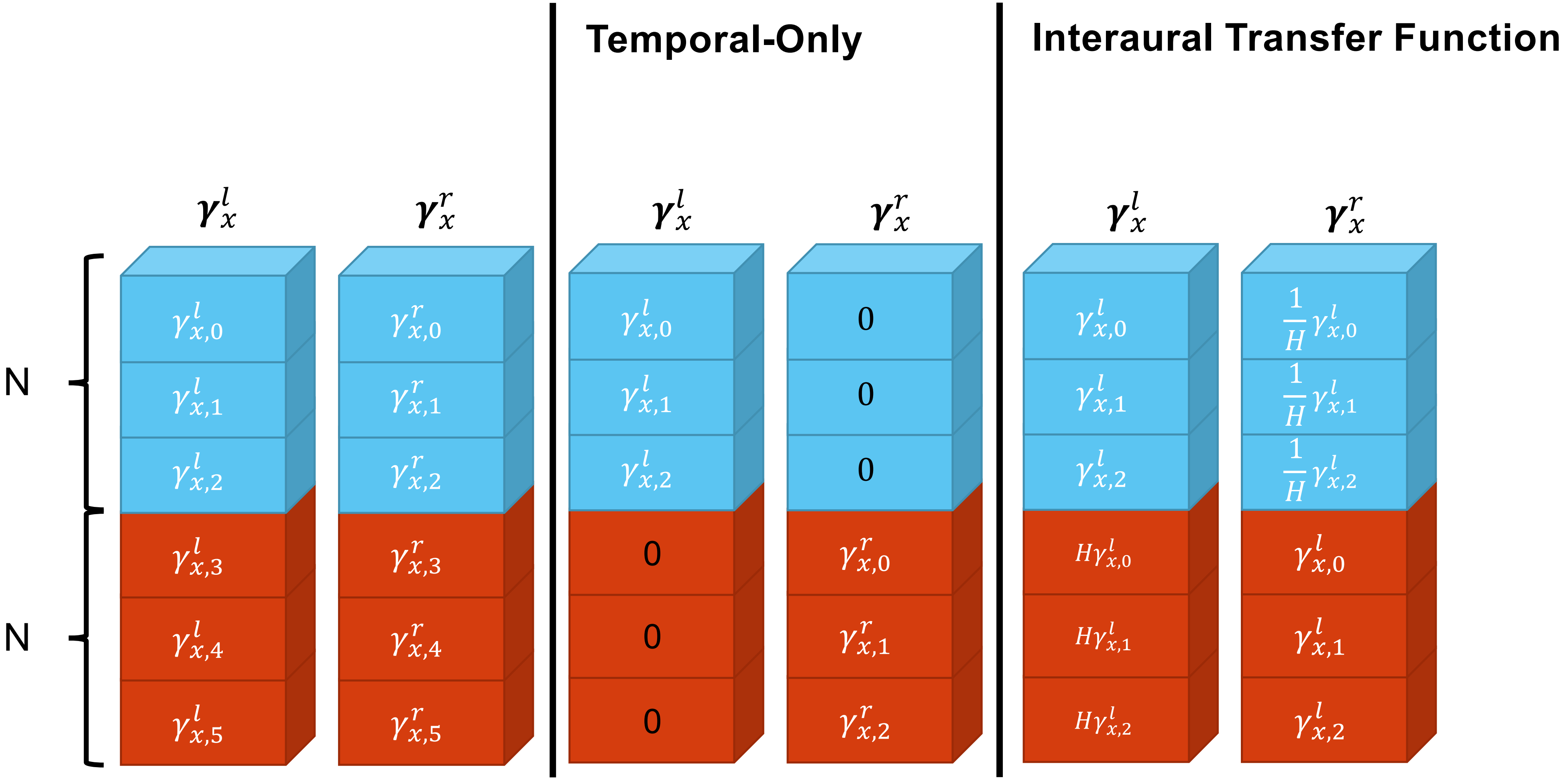
- dataset based on **Clarity Enhancement Challenge**
  - diverse localized speech and noise sources
  - simulated binaural RIRs, mild reverberation
- DNN architecture: causal **temporal convolutional network (TCN)**
- performance benefit of using MFMVDR structure vs. direct filtering



Simulation Results – Audio Examples

clean		
noisy		
binaural multi-frame filter, **direct estimation**		
binaural multi-frame filter, **MFMVDR structure**		

Possible Simplifications (speech IFC vector)



Conclusions

- **Considerable monaural and binaural noise reduction** performance using supervised learning-based approaches
- Consistent **benefit by imposing multi-frame MVDR structure**
- Complexity of deep binaural MFMVDR filter can be reduced by
  - assuming a quasi-stationary interaural transfer function
  - preserving only temporal target correlations
- **Current/future research:**
  - Investigation of deep (multi-microphone) binaural MFMVDR filter for dynamic acoustic scenarios
  - Joint noise reduction and binaural cue preservation of complete acoustic scene using deep learning-based approaches

References

- M. Tammen, D. Fischer, B. T. Meyer, S. Doclo, *DNN-based speech presence probability estimation for multi-frame single-microphone speech enhancement*, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 191-195.
- M. Tammen, S. Doclo, *Deep multi-frame MVDR filtering for single-microphone speech enhancement*, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 8443-8447.
- M. Tammen, S. Doclo, *Deep Multi-Frame MVDR Filtering For Binaural Noise Reduction*, in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022.
- M. Tammen, S. Doclo, *Parameter Estimation Procedures for Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement*, *IEEE/ACM Trans. Audio, Speech and Language Processing*, 2022, submitted.

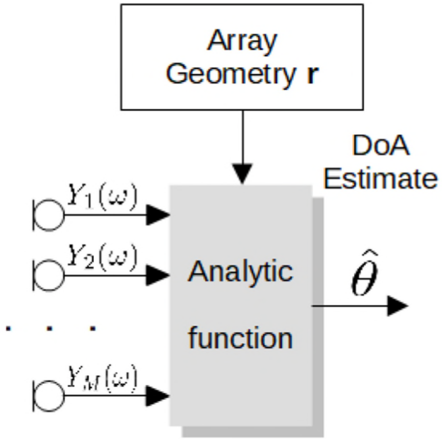
III. Geometry-aware sound source localisation

Sound source localisation

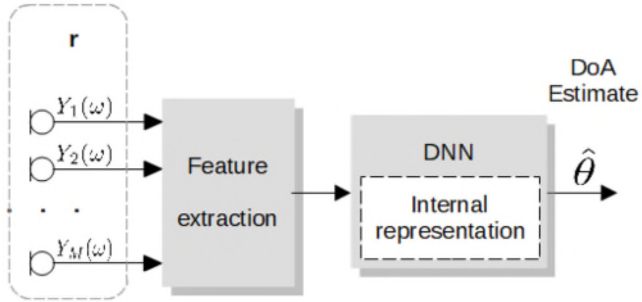
- **Model-based approaches** (e.g. SRP-PHAT, MUSIC)
  - Computation of analytical function, which explicitly depends on microphone array geometry
    $\rightarrow$ **flexibility towards different array geometries**

$$P(\theta, \mathbf{r}) = 2\pi \sum_{k=1}^{M} \sum_{l=1}^{M} \int_{-\infty}^{\infty} \Gamma_{k,l}(\omega) e^{j\omega \tau_{k,l}(\theta)} d\omega$$

$$P(\theta, \mathbf{r}) = \frac{1}{||\mathbf{a}^H(\theta, \mathbf{r}) \mathbf{E}_N \mathbf{E}_N^H \mathbf{a}(\theta, \mathbf{r})||}$$

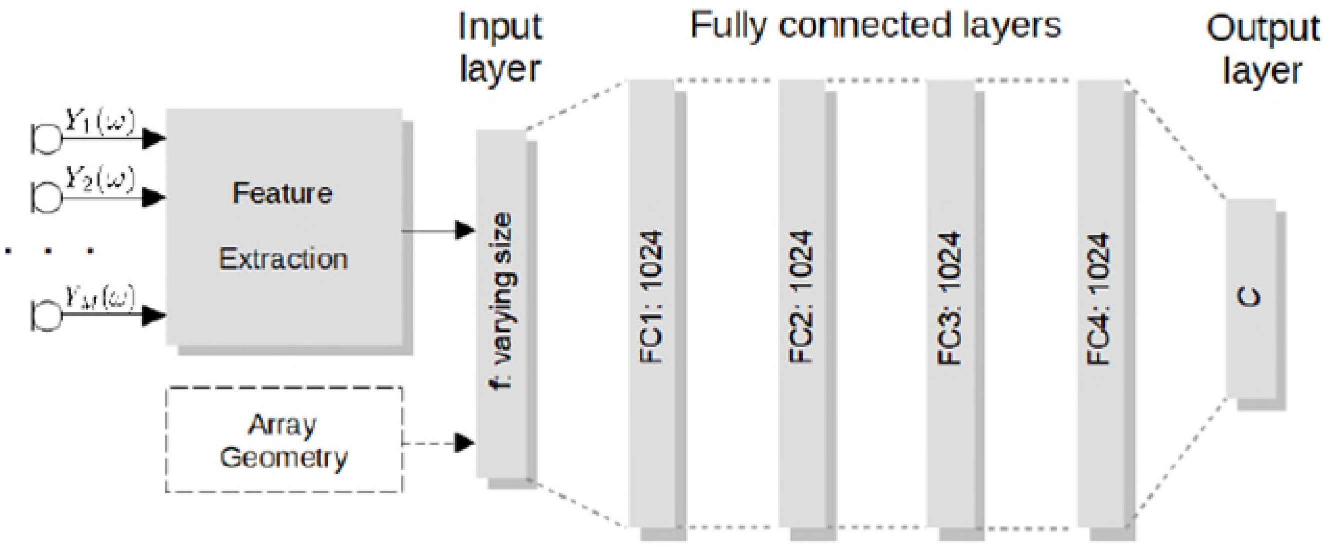


- **Supervised learning-based approaches**
  - Learn relationship between input features and DOA (classification problem)
  - **Training data implicitly based on underlying array geometry** $\rightarrow$ internal representation
  - Substantial performance degradation when applying DNN trained for certain array geometry to other array geometry



Geometry-aware DOA estimation

- **Aim:** supervised learning-based approach that generalizes well to different microphone array geometries
- DNN taking mixed data features as input:
    1. features extracted from microphone signals
    2. microphone array geometry (assumed to be known!)



Geometry-aware DOA estimation

- **Supervised learning systems:**
  1. **CNN:** using signal phases as input features [Chakrabarty & Habets, 2019]
  2. **FC-full:** using time-domain GCC-PHAT between all microphone pairs as input features

$$\gamma_{k,l} = \mathcal{F}^{-1}\left\{\frac{Y_k(\omega)\cdot Y_l^*(\omega)}{|Y_k(\omega)\cdot Y_l^*(\omega)|}\right\} \qquad \mathbf{f}_{full} = [\boldsymbol{\gamma}^{\delta}_{1,2}, \boldsymbol{\gamma}^{\delta}_{1,3}, \ldots, \boldsymbol{\gamma}^{\delta}_{M-1,M}]$$

  3. **FC-max:** reduced feature set only using location of (interpolated) maxima of GCC-PHAT

$$d_{k,l} = \arg\max_{\tau^{\delta}} \boldsymbol{\gamma}^{\delta}_{k,l} \qquad \mathbf{f}_{max} = [\tilde{d}_{1,2}, \tilde{d}_{1,3}, \ldots, \tilde{d}_{M-1,M}]$$

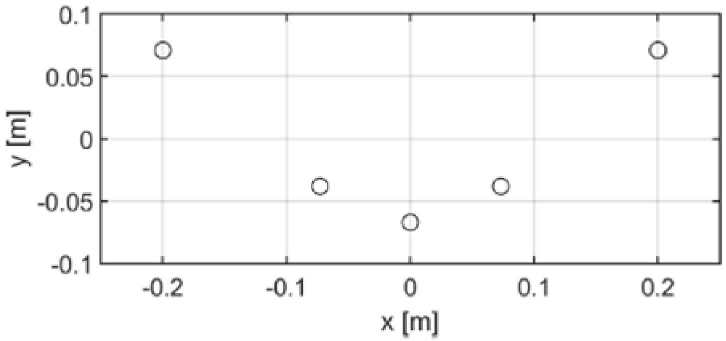
  4. **FC-GA:** using maxima of GCC-PHAT + microphone array geometry as input features

$$\mathbf{f}_r = [x_1, \ldots, x_M, y_1, \ldots, y_M] \qquad \mathbf{f} = [\mathbf{f}_{max}, \mathbf{f}_r]$$

Geometry-aware DOA estimation

- **Simulation results:**
  - Single static sound source in noisy and reverberant environment
  - 72 DOA classes (5° resolution), f_s = 8 kHz, framelength = 32 ms
  - Multi-condition training using simulated microphone signals (speech + white noise as sound source, diffuse babble noise), cross-entropy loss function
    - **CNN, FC-full, FC-max: trained for specific microphone array geometry** (M=5, arc-shaped)
    - **FC-GA: every training sample uses different microphone array geometry** (M=5, planar array, random positions with width and depth of 0.4 m)

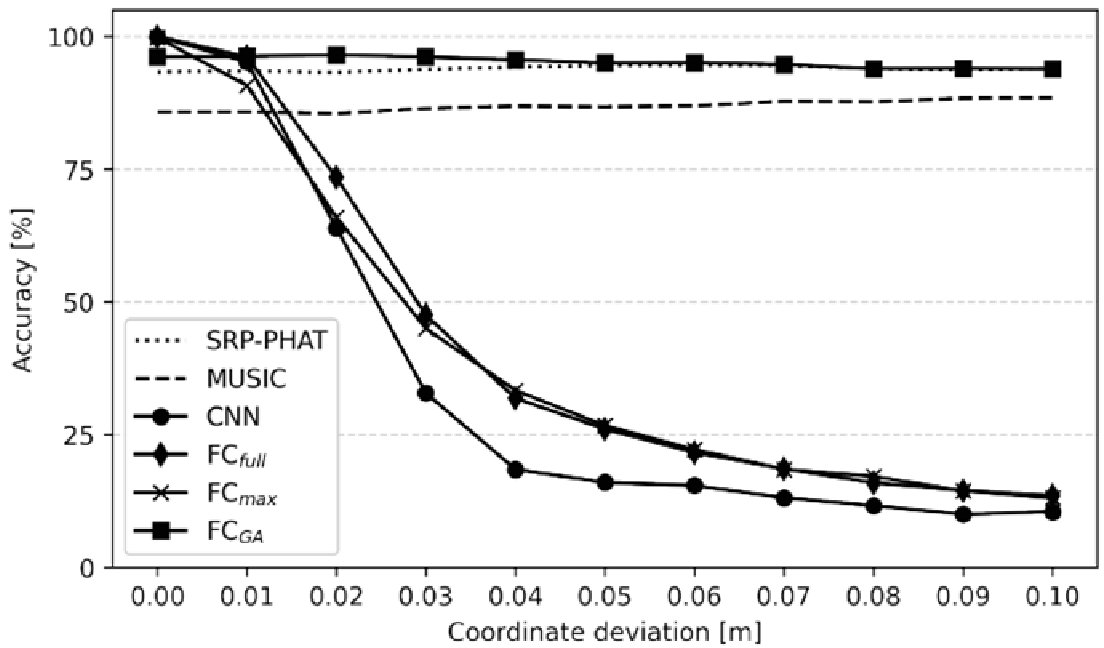
Room dimensions:	$[9.0, 5.0, 3.0]\,\mathrm{m} \pm [1.0, 1.0, 0.5]\,\mathrm{m}$
Array position:	$[4.5, 2.5, 1.5]\,\mathrm{m} \pm [0.5, 0.5, 0.5]\,\mathrm{m}$
Source distance:	1.0 - 3.0 m [within boundaries]
Source direction:	$0° : 5° : 355°$
T_{60}:	0.13 s - 1.0 s
SNR:	0 - 30 dB



Geometry-aware DOA estimation

- **Sensitivity to random coordinate deviations**
  - No deviations: DNN-based systems outperform model-based algorithms
  - **Small deviations: substantial performance degradation for baseline DNN-based systems**
  - **Proposed geometry-aware system robust to deviations**

T_{60} = 500ms, SNR=20 dB

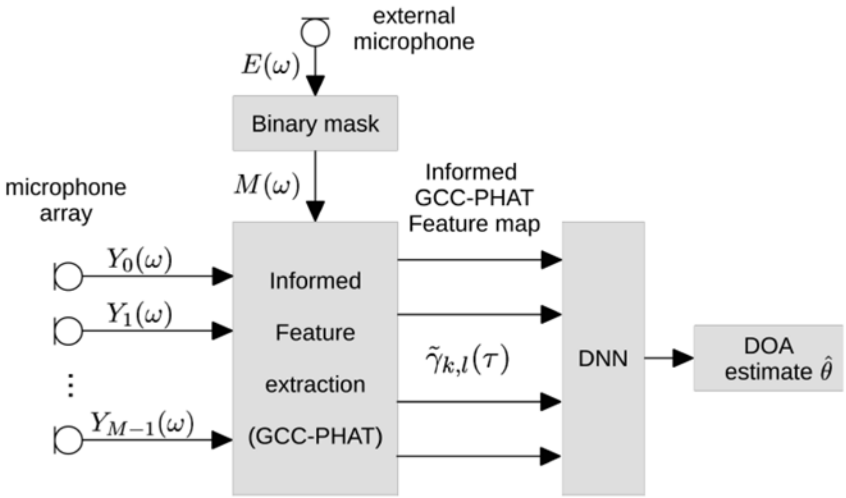


- **Performance for random (perfectly known) planar array geometry**

Algorithm	MAE [°]	Accuracy [%]
SRP-PHAT	2.44	93.5
MUSIC	2.69	86.0
FC_{GA}	**1.47**	**96.1**

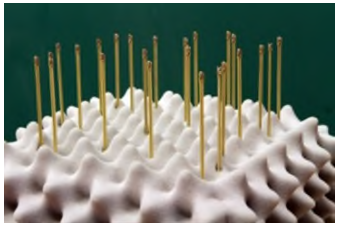
Current/future work

- **Investigate robustness** to inaccuracies in assumed microphone array geometry
- **Improved conditioning** on microphone array geometry (e.g. using feature-wise linear modulation / FiLM)
- **Signal-informed DOA estimation** exploiting external microphone *(Kowalk et al., IWAENC 2022)*



Research topics

- **Single- and multi-microphone speech enhancement**
  - **Noise reduction** (DNN-based, exploiting interframe correlation)
  - **Dereverberation** (spectral enhancement, multi-channel equalization, blind probabilistic model-based)
  - **Acoustic sensor networks** (spatially distributed microphones, sampling rate offset estimation, distributed processing)
  - **Computational acoustic scene analysis** (CASA, localization)
  - **Beamformer design** (e.g., virtual artificial head)



- **Signal processing for ear-mounted communication devices**
  - **Binaural noise reduction**, aiming at preserving spatial impression of acoustic scene (binaural cues)
  - Open-fitting hearing devices: **acoustic transparency**, **feedback cancellation** and **active noise/occlusion control**
  - EEG-based **auditory attention decoding** for steering beamformers

Questions ?

http://www.sigproc.uni-oldenburg.de

YouTube Signal Processing Uni Oldenburg