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

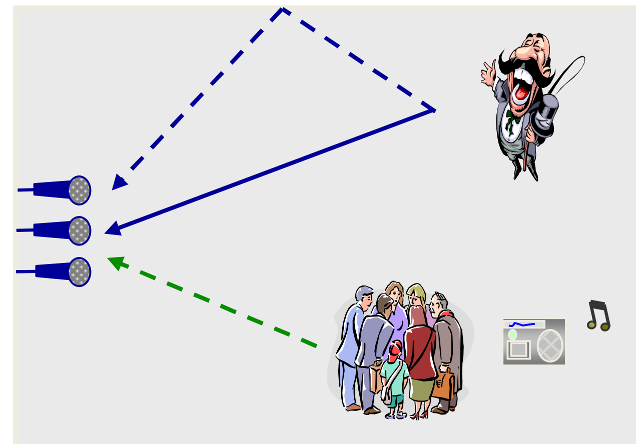
Prof. Dr. Simon Doclo

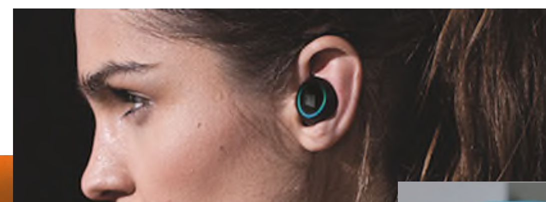
University of Oldenburg, Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all

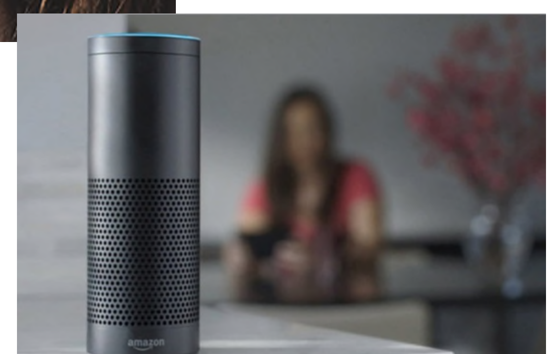
EUSIPCO 2023 – September 8, 2023

Introduction

- **Speech communication applications**
- **Acoustic environment** : target speaker + ambient noise, competing speakers, reverberation
- Degradation of **speech quality/intelligibility** and speech recognition performance

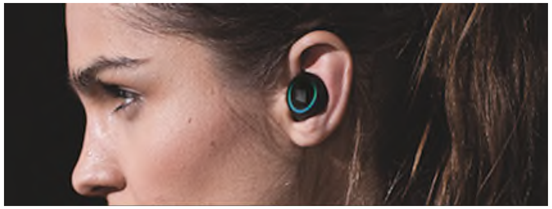






Introduction

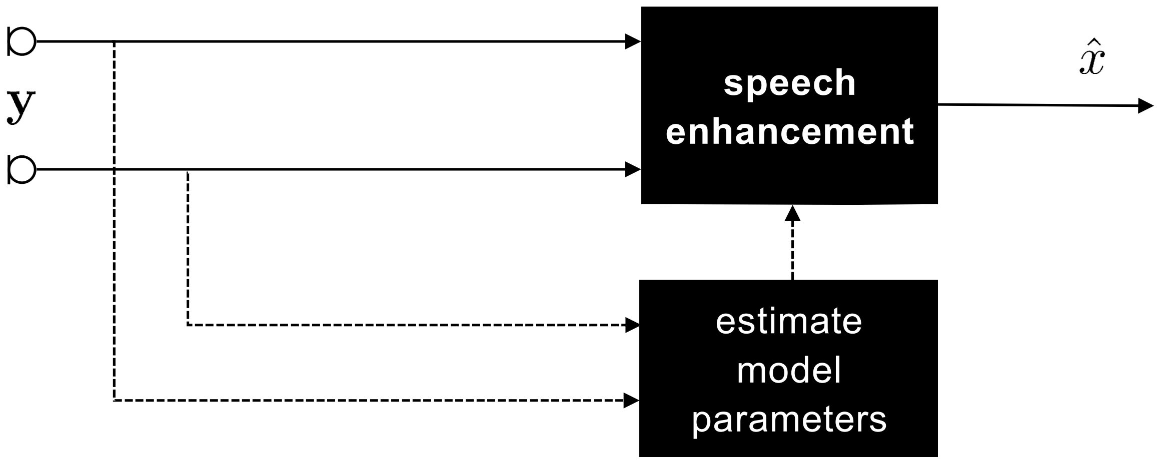
- **Speech enhancement algorithms:** extract target speaker by performing
  - Noise reduction
  - Dereverberation
  - Source separation



- **Requirements** for speech communication applications:
  - Low speech distortion
  - On-line processing (low-latency)
  - Generalization / robustness to varying acoustic conditions (moving sources/microphones, SNRs, …)
  - Computational complexity

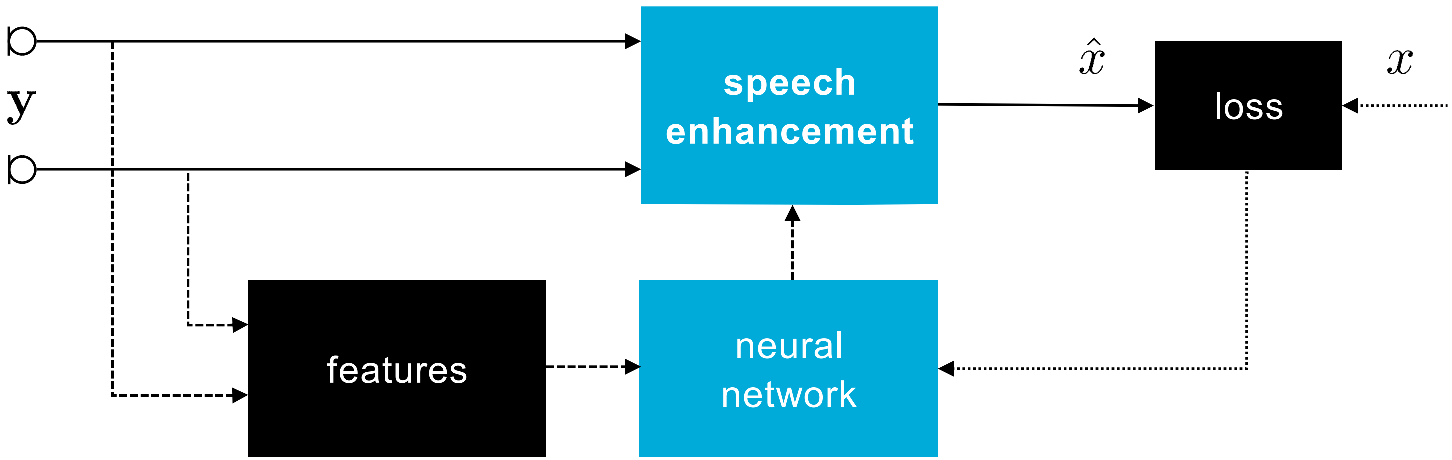
Introduction

- **Speech enhancement algorithms:**
  - single microphone (spectro-temporal) $\rightarrow$ multiple microphones (spatial)
  - model-based approaches (estimation of model parameters)



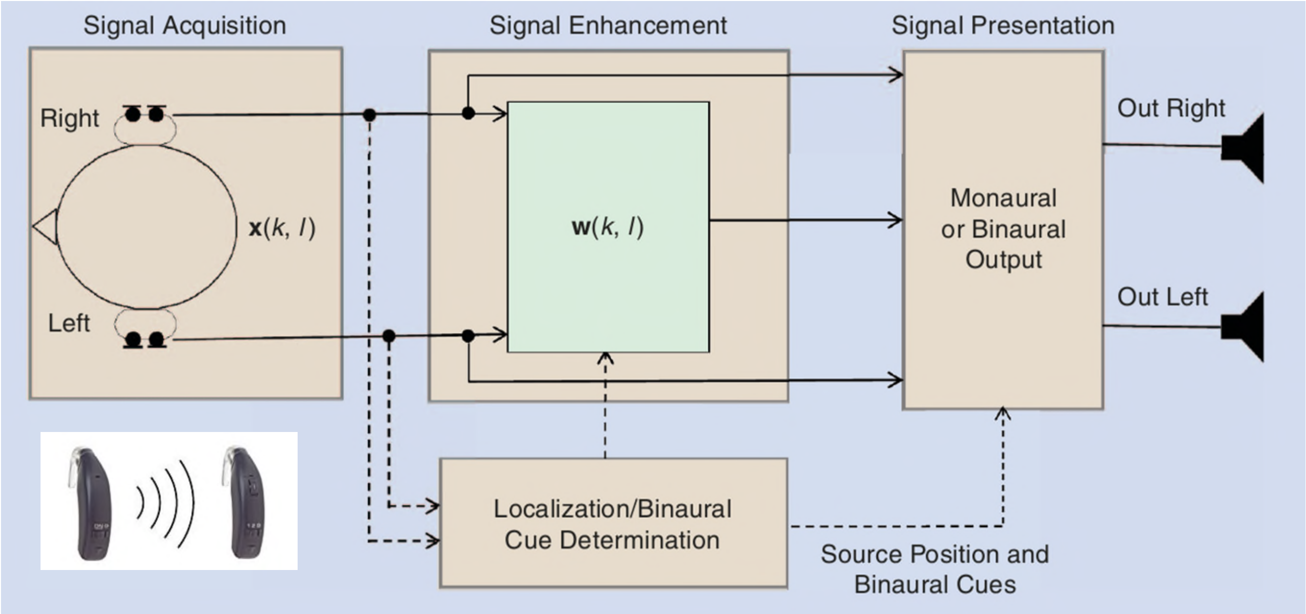
Introduction

- **Speech enhancement algorithms:**
  - single microphone (spectro-temporal) $\rightarrow$ multiple microphones (spatial)
  - model-based approaches (estimation of model parameters)
  - learning-based approaches (supervised learning using deep neural networks)
  - hybrid approaches (combination of model-based and learning-based)



This presentation

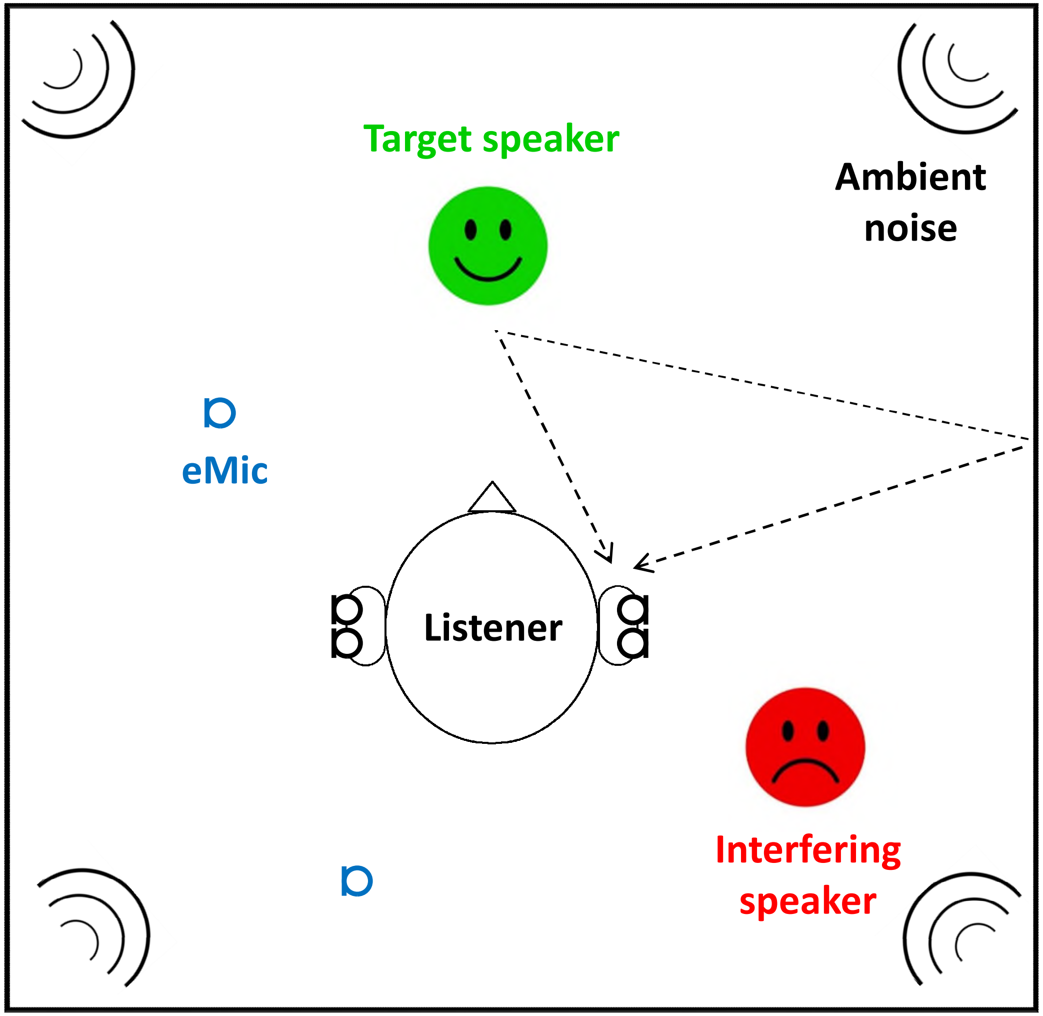
- Focus on binaural **assistive listening devices**
- On-line approaches (model-based, deep learning-based, hybrid) for **multi-microphone noise reduction and source localization**
- Exploit spatially distributed microphones in **acoustic sensor networks**



Multi-microphone speech enhancement

Acoustic scenario

- Assistive listening device with M microphones
- Multiple speakers in noisy and reverberant environment
- External microphones (eMics)

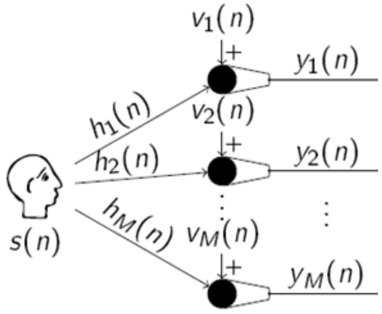


Signal model

- **Transform / encoder domain (e.g. short-time Fourier transform)**

$$\begin{aligned} y_m(k,l) &= h_m(k,l) \star s(k,l) + i_m(k,l) + v_m(k,l) \\ &= a_m(k,l)x_1(k,l) + x_{r,m}(k,l) + i_m(k,l) + v_m(k,l) \end{aligned}$$

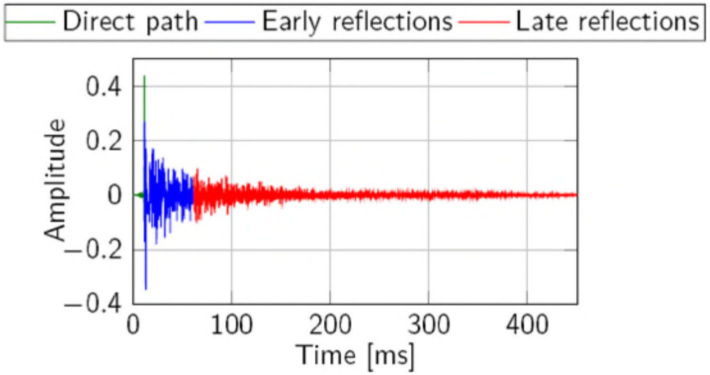
direct and early reverberation | **late reverb** | **interfering source(s)** | **ambient noise**



$$\boxed{\mathbf{y}(k,l) = \mathbf{a}(k,l)x_1(k,l) + \mathbf{u}(k,l)}$$

$\mathbf{a}(k,l)$ = vector of **relative transfer functions (RTFs)** of target speaker

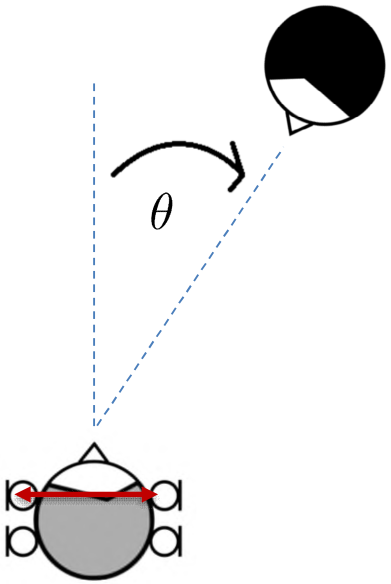
$\mathbf{u}(k,l)$ = vector of **undesired components**

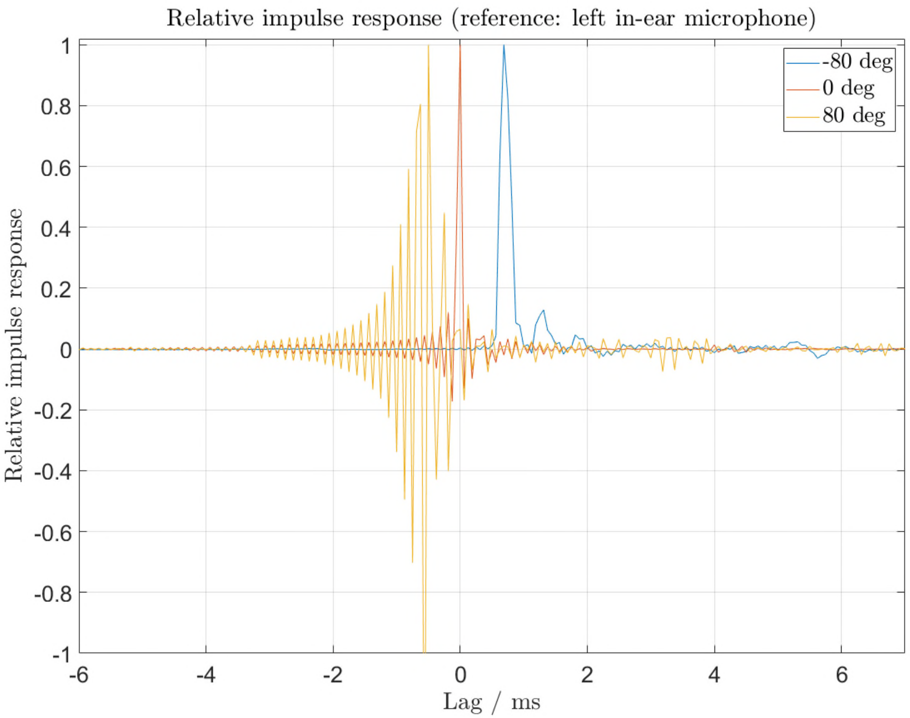


m : microphone index (1…*M*)
k : frequency index
l : time / frame index

Relative transfer functions

$$\mathbf{y}(k,l) = \mathbf{a}(k,l)x_1(k,l) + \mathbf{u}(k,l)$$





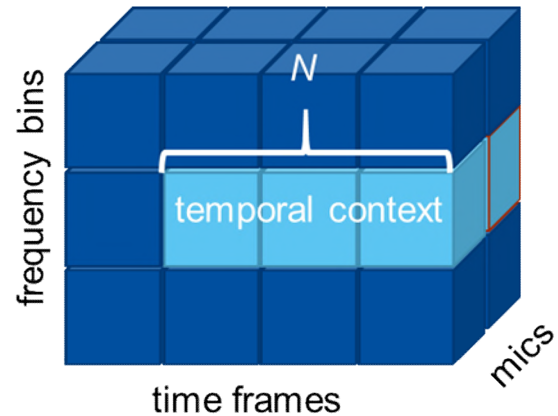
RTF vector $\mathbf{a}(k,l)$ encodes **direction-of-arrival** (DOA) θ of source

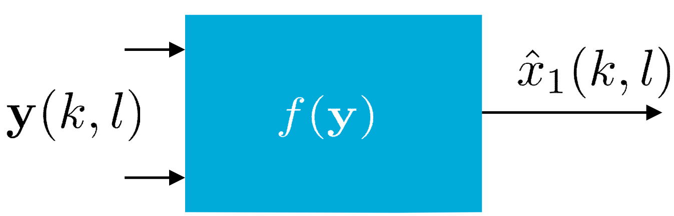
Multi-microphone speech enhancement

$$\mathbf{y}(k,l) = \mathbf{a}(k,l)x_1(k,l) + \mathbf{u}(k,l)$$

- **Objective:** estimate clean speech component in reference microphone $x_1(k,l)$ from noisy and reverberant microphone signals $\mathbf{y}(k,l)$

1. **Non-linear vs. linear filtering (with/without filter structure)**
2. **"Traditional" statistical estimation methods vs. supervised learning methods**
3. **Single-frame vs. multi-frame input vector**

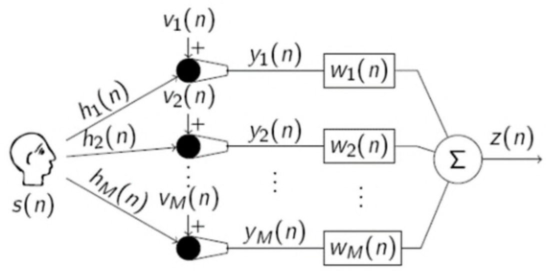




Multi-microphone speech enhancement

- **Parametric multi-channel Wiener filter (MWF):** linear filtering based on filter-and-sum structure

  Objective: estimate speech component + trade off speech distortion vs. reduction of undesired component



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

$\rightarrow$ **requires** estimate of covariance matrices (= *model parameters*)

Use signal model to decompose as **minimum-variance-distortionless-response (MVDR) beamformer and spectral postfilter**

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

$\rightarrow$ **requires** estimate of undesired covariance matrix, relative transfer function (RTF) vector of target speaker, and power spectral densities (PSDs) of speech and undesired components (= *model parameters*)

[Doclo, Kellermann, Makino, Nordholm, *IEEE Signal Processing Magazine*, 2015] [Gannot et al., *IEEE/ACM TASLP*, 2017]

Multi-microphone speech enhancement

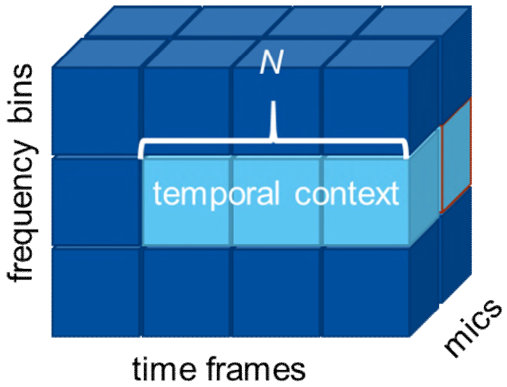
- **Multi-frame extension:**
  - Consider multiple frames: current and past frames (on-line processing)

$$\bar{\mathbf{y}}_m(k,l) = \left[\, y_m(k,l) \;\; y_m(k,l-1) \;\; \ldots \;\; y_m(k,l-N+1) \,\right]$$

  - Multi-frame speech vector $\bar{\mathbf{x}}_m(k,l)$ can be decomposed into **temporally correlated and uncorrelated components**:

$$\bar{\mathbf{x}}_m(k,l) = \gamma_x(k,l) x_m(k,l) + \bar{\mathbf{x}}'_m(k,l) \qquad \gamma_x(k,l) = \frac{\mathcal{E}\{\bar{\mathbf{x}}_m(k,l) x_m^*(k,l)\}}{\mathcal{E}\{|x_m(k,l)|^2\}}$$

  - **Speech interframe correlation vector** $\gamma_x(k,l)$ depends on sound (e.g., voiced vs. unvoiced) $\rightarrow$ **highly time-varying**



Multi-microphone speech enhancement

- **Multi-frame extension:**
  - Consider multiple frames: current and past frames (on-line processing)

$$\bar{\mathbf{y}}_m(k,l) = \left[\, y_m(k,l) \;\; y_m(k,l-1) \;\; \ldots \;\; y_m(k,l-N+1) \,\right]$$

  - Multi-frame speech vector $\bar{\mathbf{x}}_m(k,l)$ can be decomposed into **temporally correlated and uncorrelated components**:

$$\bar{\mathbf{x}}_m(k,l) = \boldsymbol{\gamma}_x(k,l) x_m(k,l) + \bar{\mathbf{x}}'_m(k,l) \qquad \boldsymbol{\gamma}_x(k,l) = \frac{\mathcal{E}\{\bar{\mathbf{x}}_m(k,l) x_m^*(k,l)\}}{\mathcal{E}\{|x_m(k,l)|^2\}}$$

  - **Speech interframe correlation vector** $\boldsymbol{\gamma}_x(k,l)$ depends on sound (e.g., voiced vs. unvoiced) $\rightarrow$ **highly time-varying**
  - **Signal model:**

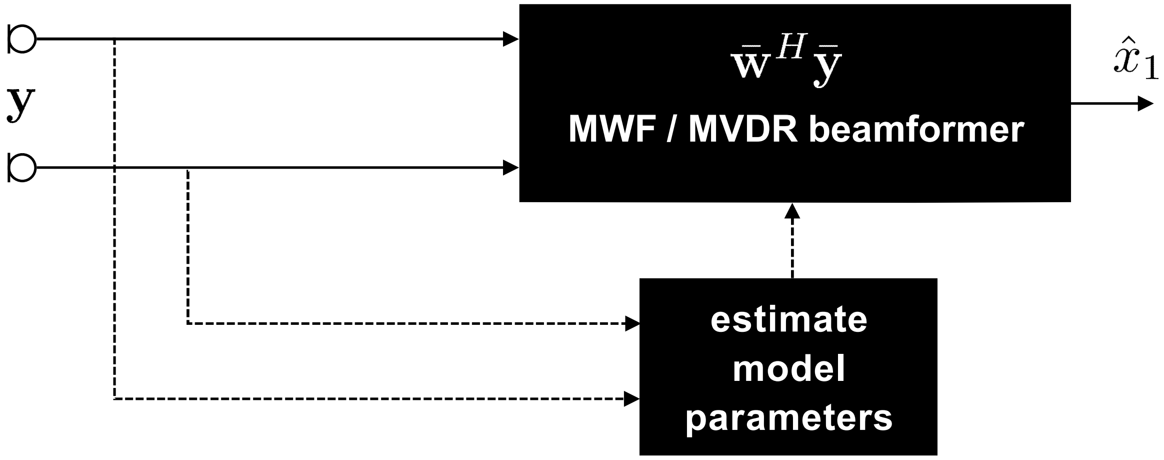
$$\boxed{\bar{\mathbf{a}}(k,l) = \underbrace{\mathbf{a}(k,l)}_{\text{RTF vector of target speaker}} \otimes \underbrace{\boldsymbol{\gamma}_x(k,l)}_{\text{Interframe correlation vector}}}$$

RTF vector of target speaker
(time-varying, acoustic environment)

Interframe correlation vector
(highly time-varying, speech)

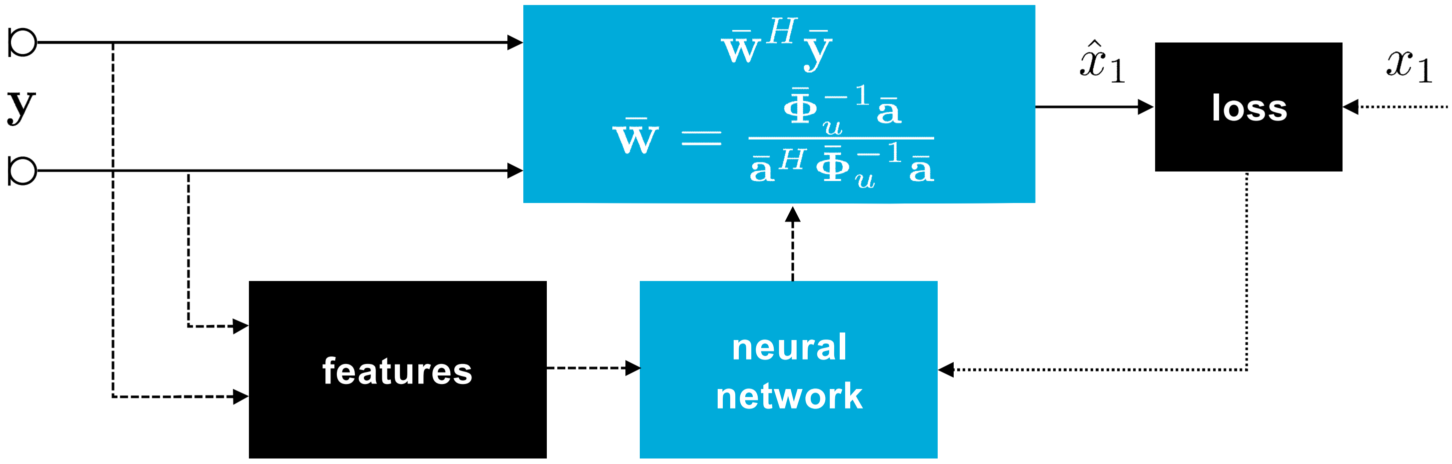
Multi-microphone speech enhancement

- **"Traditional" statistical estimation of parameters** (requiring assumptions)
  - *Covariance matrices, power spectral densities*, e.g., assuming that undesired component is more stationary than speech component, reverberation is diffuse
  - *Relative transfer function (RTF) vector of target speaker*, e.g., assuming anechoic propagation, known source activity, spatially distributed microphones and uncorrelated undesired component
  - *Speech interframe correlation vector (IFC)*, e.g., using subspace-based estimators but difficult to accurately estimate since highly time-varying



Multi-microphone speech enhancement

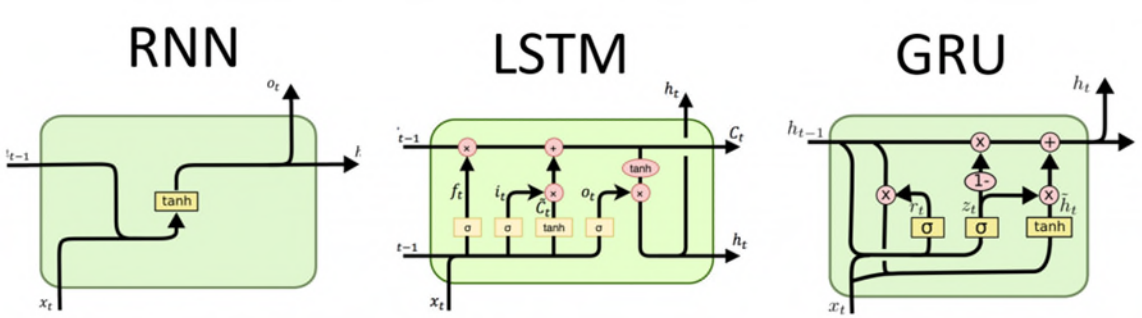
- **Supervised learning** by minimizing loss function (assumptions in training data)
  - Directly estimate filter coefficients: single-frame/masking or multi-frame/deep filtering, e.g. [Mack 2019]
  - ***Hybrid approach* : impose filter structure and estimate parameters in end-to-end fashion**, e.g. ADL-MVDR [Zhang 2021], mask-based neural beamforming [Ochiai 2023], deep MFMVDR [Tammen 2023], DeepFilterNet [Schröter 2023]



Multi-microphone speech enhancement

- **Supervised learning** by minimizing loss function (assumptions in training data)
  - Directly estimate filter coefficients: single-frame/masking or multi-frame/deep filtering, e.g. [Mack 2019]
  - ***Hybrid approach* : impose filter structure and estimate parameters in end-to-end fashion**, e.g. ADL-MVDR [Zhang 2021], mask-based neural beamforming [Ochiai 2023], deep MFMVDR [Tammen 2023], DeepFilterNet [Schröter 2023]

- **Neural network architectures:**
  - Long short-term memory (LSTM), transformer, temporal convolutional networks (TCN)
  - For computational complexity reasons often gated recurrent units (GRU)



Multi-microphone speech enhancement

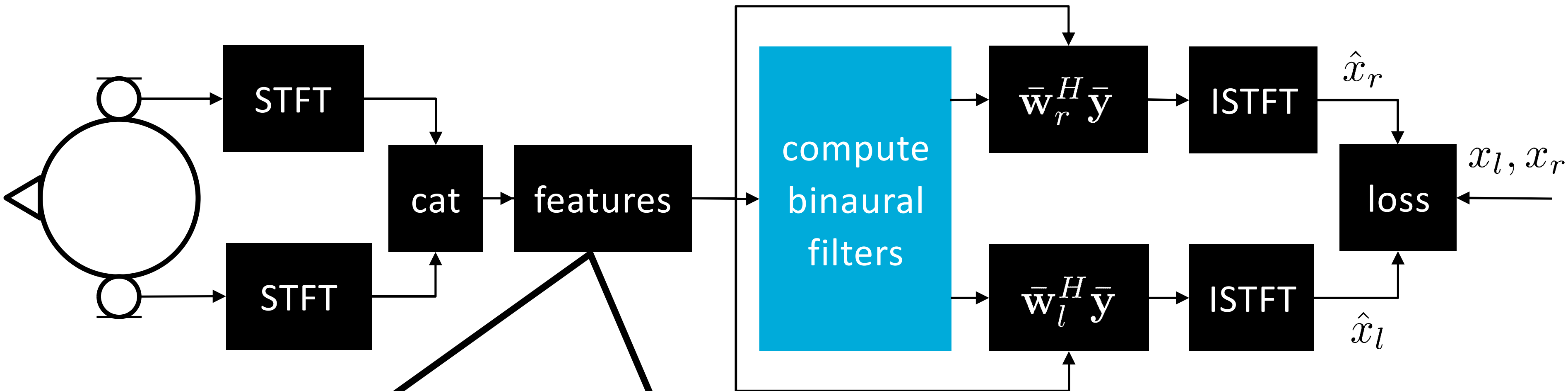
- **Supervised learning** by minimizing loss function (assumptions in training data)
  - Directly estimate filter coefficients: single-frame/masking or multi-frame/deep filtering, e.g. [Mack 2019]
  - ***Hybrid approach* : impose filter structure and estimate parameters in end-to-end fashion**, e.g. ADL-MVDR [Zhang 2021], mask-based neural beamforming [Ochiai 2023], deep MFMVDR [Tammen 2023], DeepFilterNet [Schröter 2023]

- **Neural network architectures:**
  - Long short-term memory (LSTM), transformer, temporal convolutional networks (TCN)
  - For computational complexity reasons often gated recurrent units (GRU)

- **Loss functions:**
  - Mostly defined in time-domain after reconstruction (iSTFT) / decoding
  - Mean-square error (MSE), scale-invariant signal-to-distortion ratio (SI-SDR), mean absolute spectral error, psycho-acoustically motivated loss function

$\|\hat{\mathbf{x}} - \mathbf{x}\|^2$ $\quad 10\log_{10}\frac{\|\alpha\mathbf{x}\|^2}{\|\hat{\mathbf{x}}-\alpha\mathbf{x}\|^2}$ $\quad \beta\,|\hat{x}(k,l) - x(k,l)| + (1-\beta)\,\big||\hat{x}(k,l)| - |x(k,l)|\big|$

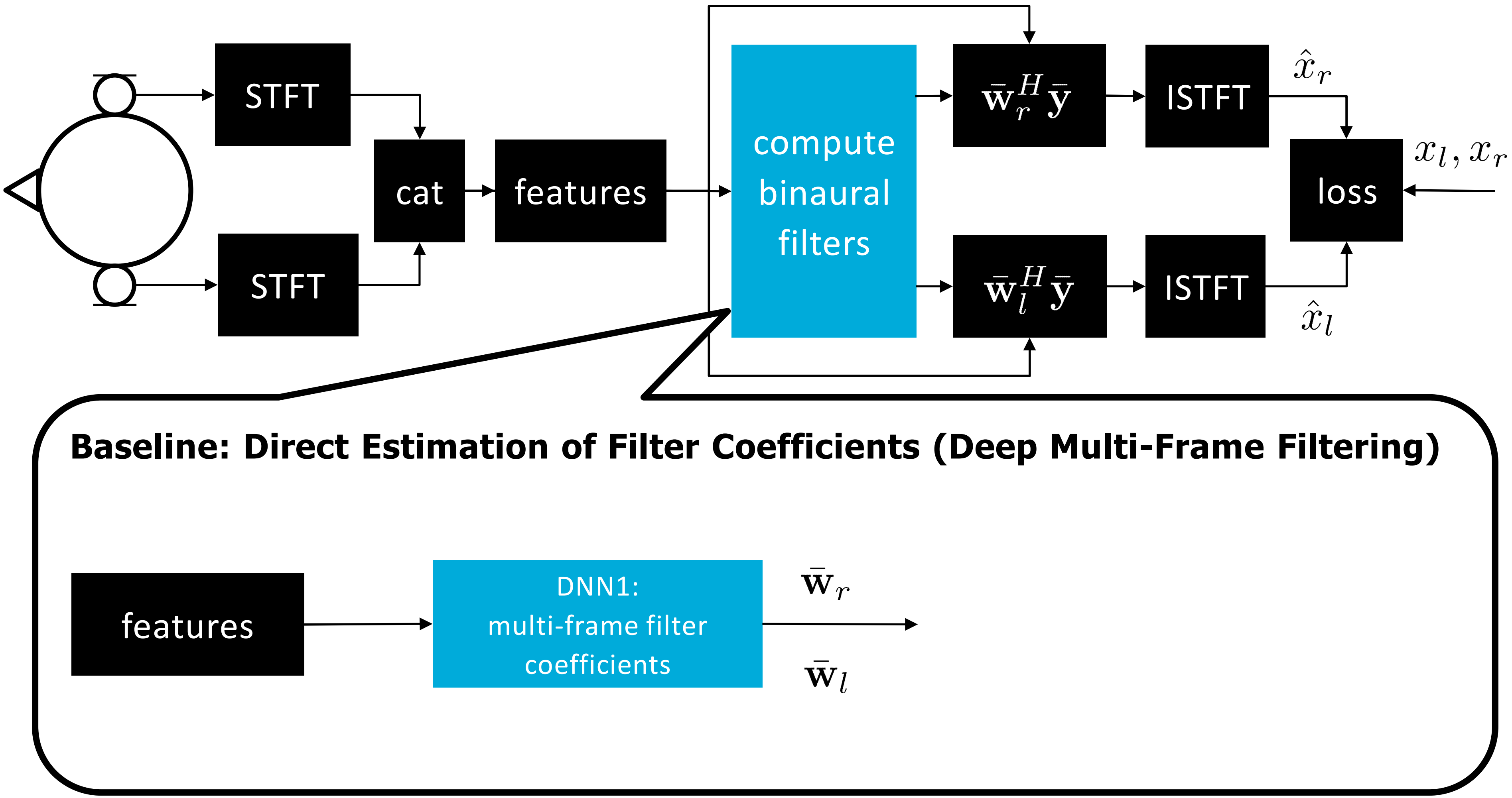
Application to binaural hearing devices



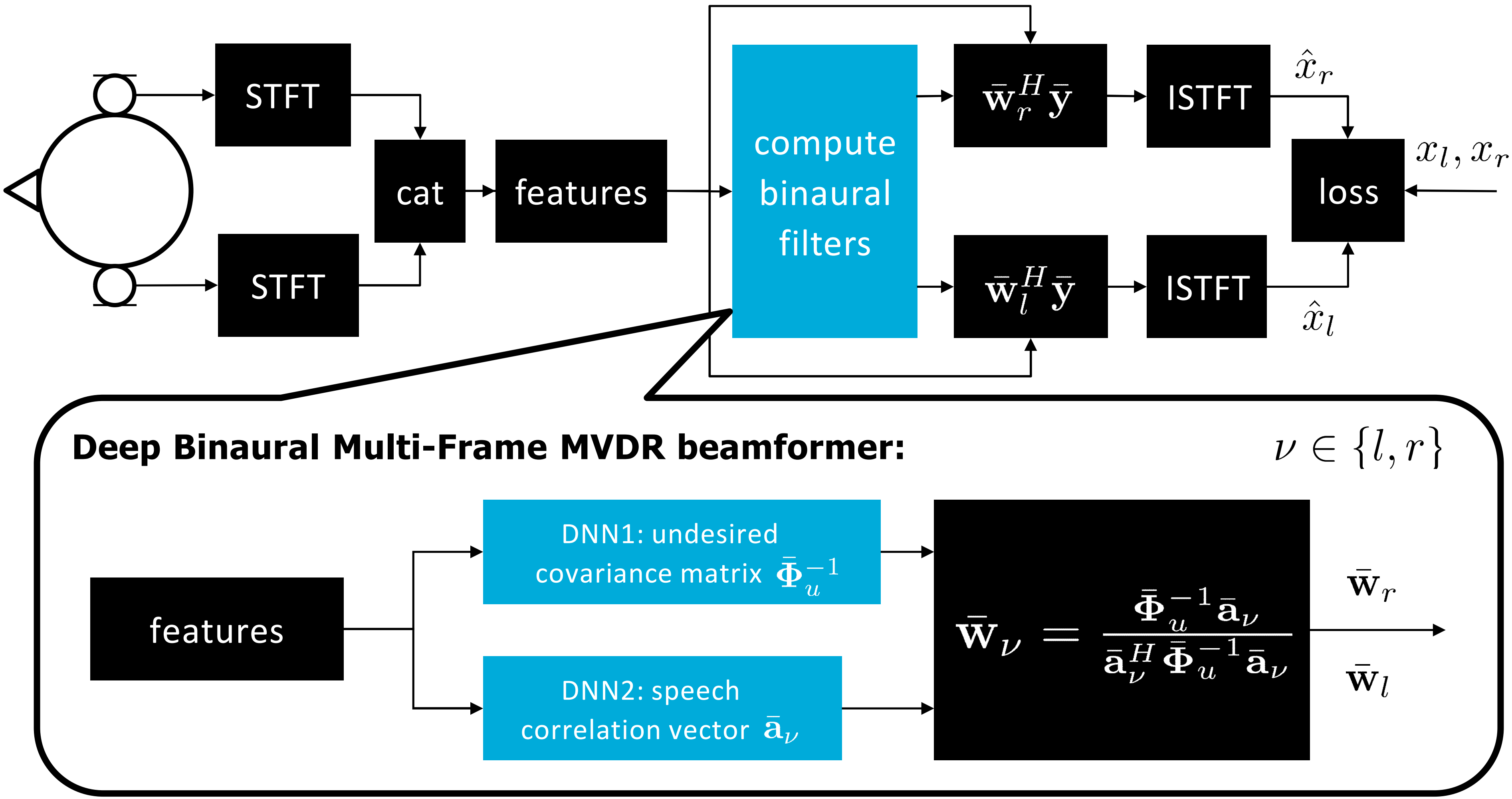
Features: multi-channel concatenation of

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

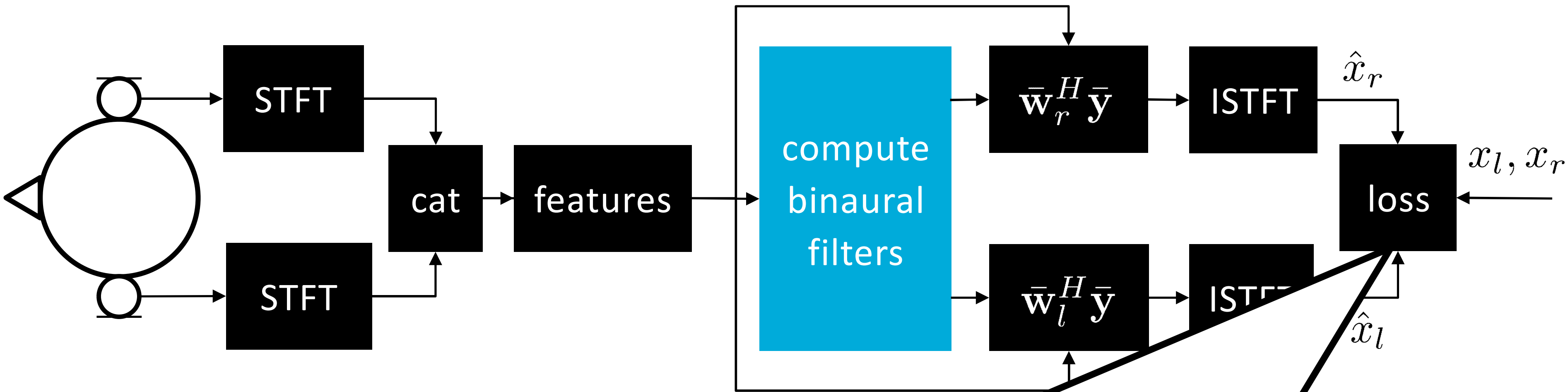
Application to binaural hearing devices



Application to binaural hearing devices



Application to binaural hearing devices



Loss: Combined Mean Absolute Spectral Error [Wang, 2020]

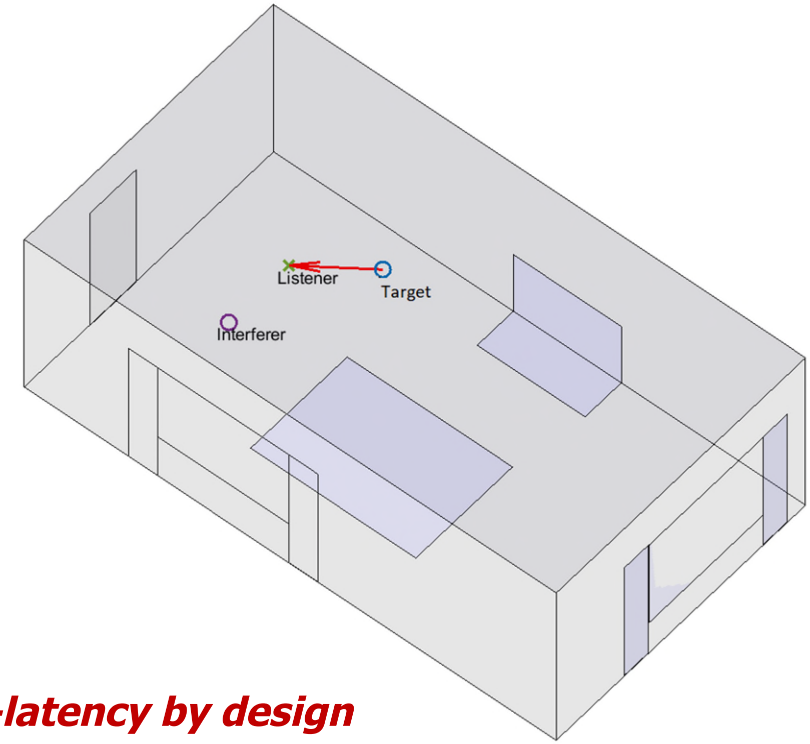
$$\frac{1}{2} \sum_{\mu \in \{l,r\}} \beta \left|\hat{x}_\nu(k,l) - x_\nu(k,l)\right| + (1-\beta) \Big| |\hat{x}_\nu(k,l)| - |x_\nu(k,l)| \Big|$$

- STFT re-analysis after overlap-add
- emphasizes spectral magnitude (β=0.4)

Simulations

- **Datasets and settings**

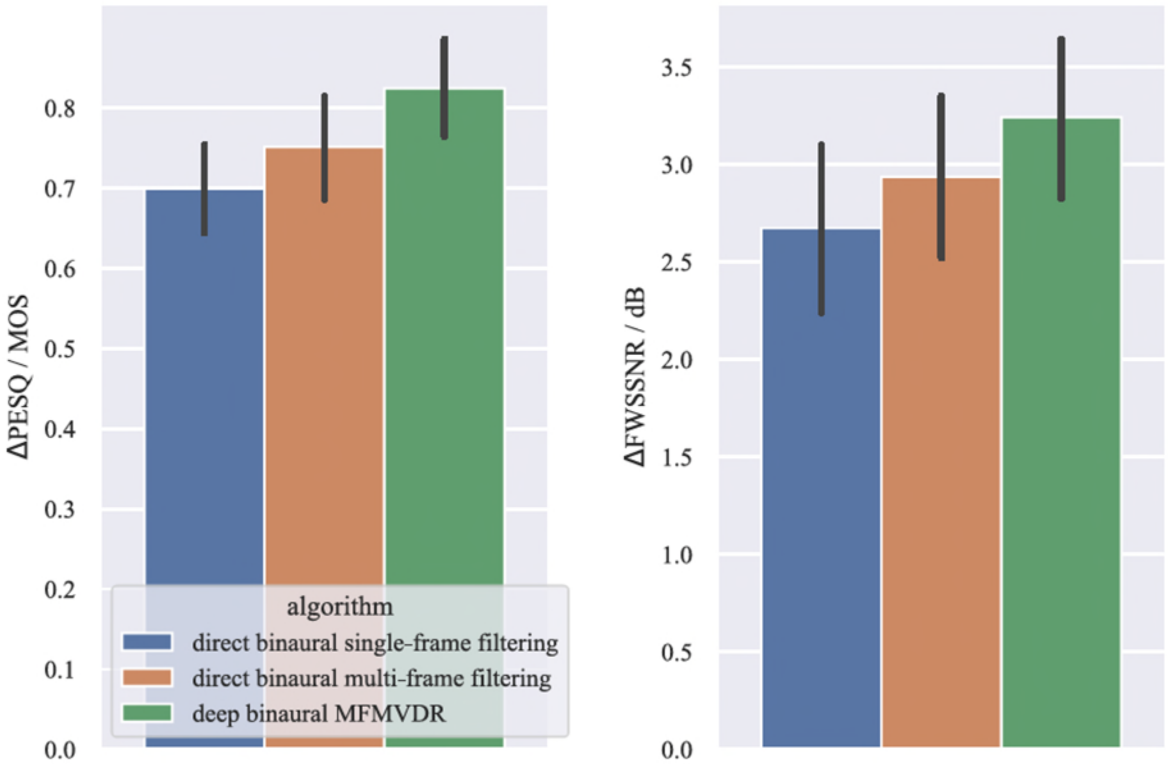
	training	testing
speech	DNS3 train set	DNS1 test set
noise		
Room impulse responses	Clarity challenge (simulated)	Kayser database (measured)
Reverberation time (T_{60})	200 – 400 ms	
SNRs	0 - 15 dB	−5 – 20 dB
length	100 h	17 min



- f_s = 16 kHz, STFT: 8 ms frames, 2 ms shift → ***low-latency by design***
- N = 5 frames for mult-frame filter → 16ms context
- DNNs: causal temporal convolutional networks
- Adam optimizer, learning rate: 0.0003, training for 150 epochs (early stopping)

Simulations

- **Objective performance metrics**



- Benefit of multi-frame filtering vs. single-frame filtering
- Benefit of imposing multi-frame MVDR filter structure

Audio demo

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

[Tammen & Doclo, *Proc. IWAENC 2022*]

Audio Demo

Simulations

- **Computational complexity**

algorithm	trainable weights / M	bottleneck dimension	memory / MB	RTF	RTF contribution, MFMVDR / %
deep MFMVDR (SPP)	5.3	231	195.6	0.176	54.9
deep MFMVDR (RS)	4.9	128	137.3	0.167	47.9
deep MFMVDR (CD)	5.3	128	110.9	0.139	39.0
deep MFMVDR (PDT)	5.1	128	93.3	0.170	43.4
deep MFMVDR (R1)	5.1	128	85.2	0.100	7.5
masking (real)	5.0	226	30.2	0.075	0.0
masking (complex)	5.0	226	28.5	0.077	0.0
DMFF	5.2	226	29.5	0.079	0.0

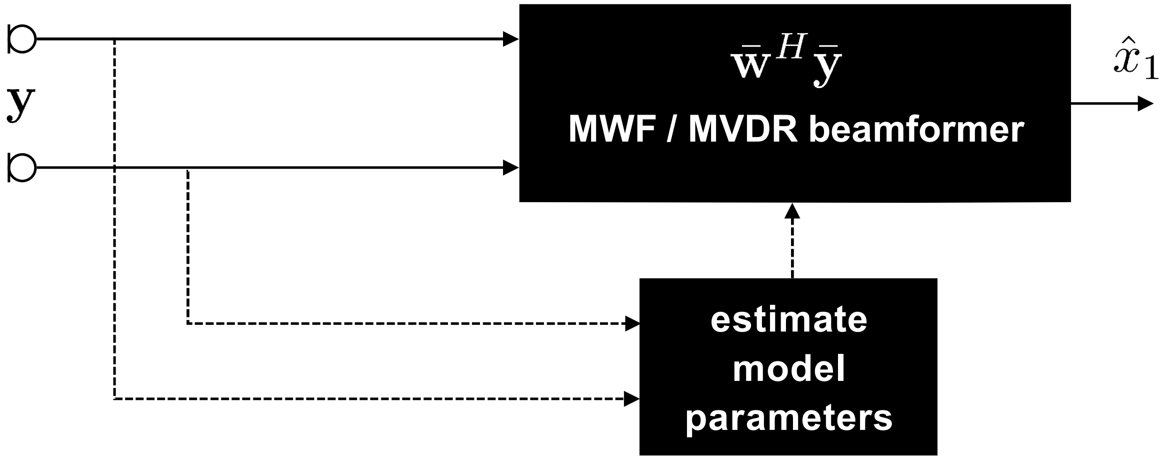
- Real-time capability of all algorithms
- Deep MFMVDR filter computationally more complex than direct multi-frame filter (DMFF), mainly due to additional linear algebra operations

  → can be alleviated by assuming rank-1 (R1) structure

single core of AMD EPYC 7443P CPU clocked at 3.8 GHz; 10 s-long signals

Multi-microphone speech enhancement

- **"Traditional" statistical estimation of parameters** (requiring assumptions)
  - *Covariance matrices, power spectral densities*, e.g., assuming that undesired component is more stationary than speech component, reverberation is diffuse
  - *Relative transfer function (RTF) vector of target speaker*, e.g., assuming anechoic propagation, known source activity, spatially distributed microphones and uncorrelated undesired component
  - *Speech interframe correlation vector (IFC)*, e.g., using subspace-based estimators but difficult to accurately estimate since highly time-varying



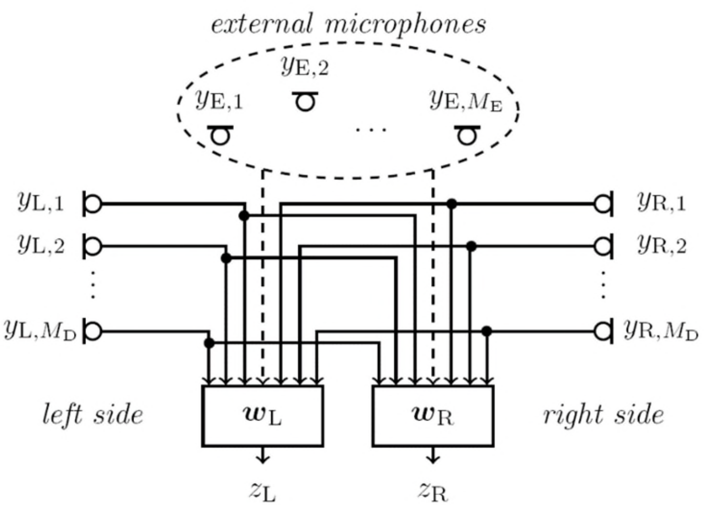
External microphones

- Exploit the availability of one or more external microphones **(acoustic sensor network)** with hearing aids

  [Bertrand 2009, Szurley 2016, Yee 2018, Farmani 2018, Kates 2018, Ali 2019, Corey 2021, Gößling 2021]

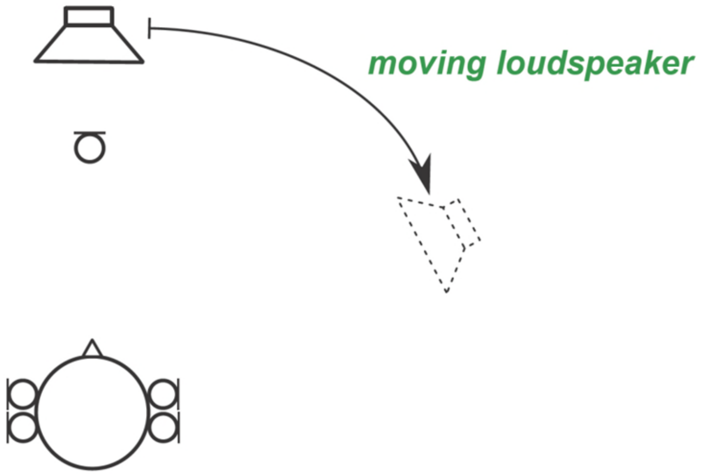
- Integrate external microphone(s) with hearing aid microphones for:
  - Low-complexity method to **estimate relative transfer function (RTF)** vector of target speaker
  - Improved **noise reduction** and **binaural cue preservation** performance

$$\mathbf{w}_L = \frac{\mathbf{\Phi}_v^{-1}\mathbf{a}_L}{\mathbf{a}_L^H\mathbf{\Phi}_v^{-1}\mathbf{a}_L}, \qquad \mathbf{w}_R = \frac{\mathbf{\Phi}_v^{-1}\mathbf{a}_R}{\mathbf{a}_R^H\mathbf{\Phi}_v^{-1}\mathbf{a}_R}$$



RTF vector estimation exploiting external microphone

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

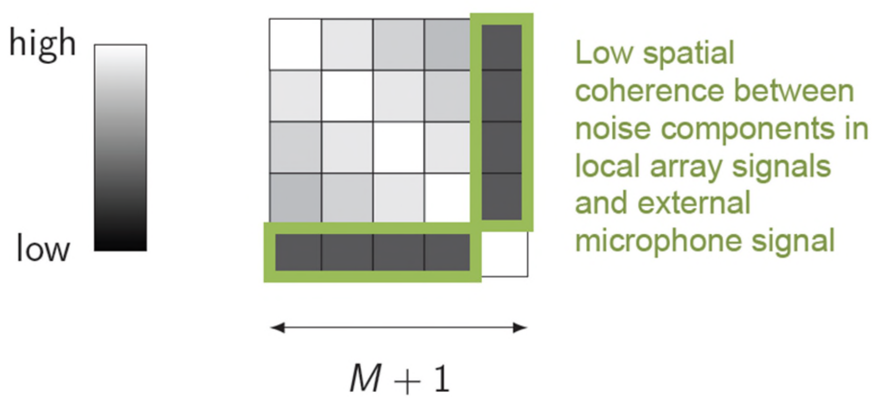


$$\hat{\mathbf{w}}_L = \frac{\hat{\mathbf{\Phi}}_v^{-1}\hat{\mathbf{a}}_L}{\hat{\mathbf{a}}_L^H\hat{\mathbf{\Phi}}_v^{-1}\hat{\mathbf{a}}_L}$$

RTF vector estimation exploiting external microphone

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

  → correlate HA microphone signals with external microphone signals and normalize by reference element

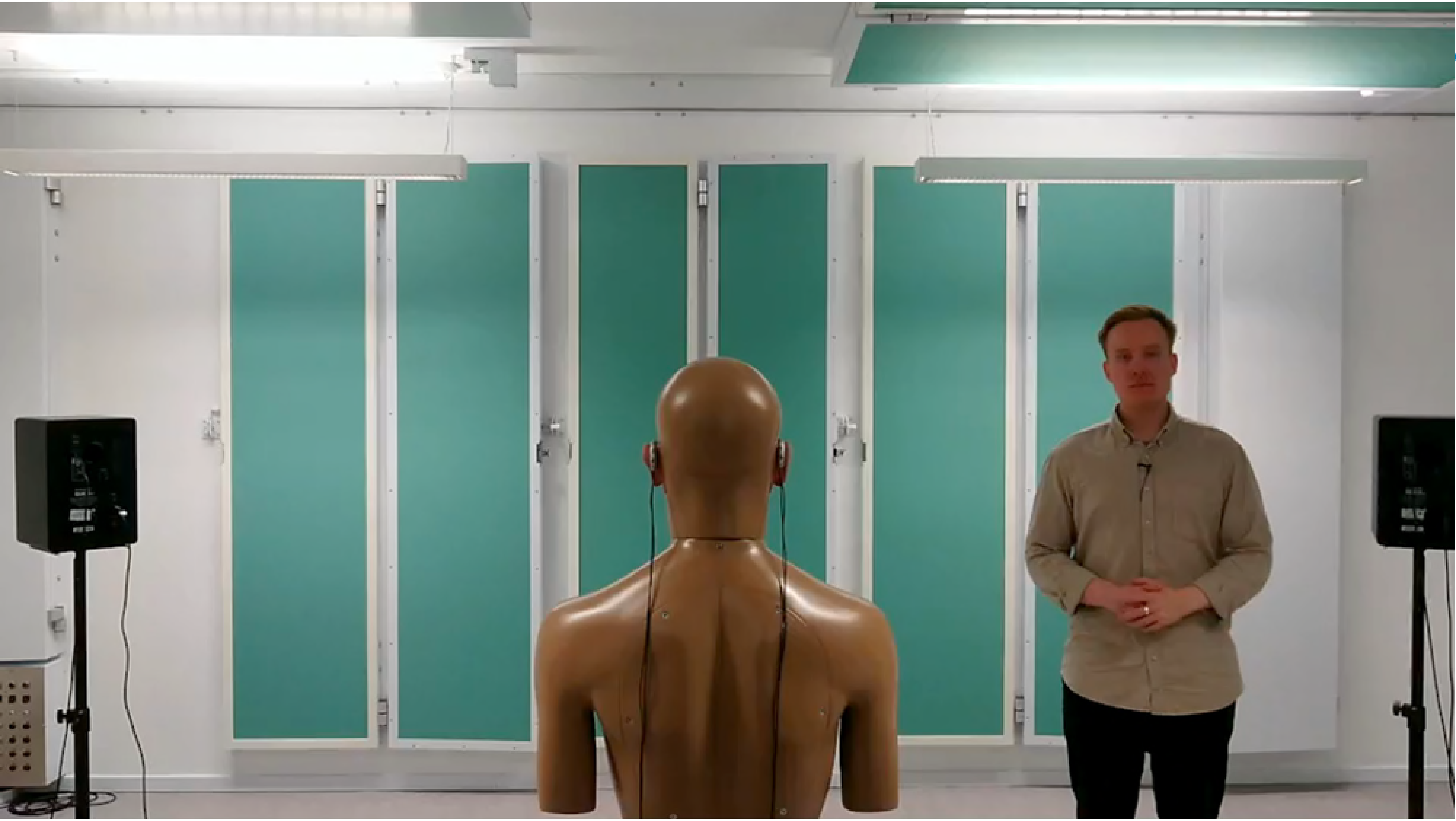


$$\hat{\mathbf{a}}_L = \frac{\hat{\mathbf{\Phi}}_y \mathbf{e}_E}{\mathbf{e}_L^T \hat{\mathbf{\Phi}}_y \mathbf{e}_E}, \qquad \hat{\mathbf{a}}_R = \frac{\hat{\mathbf{\Phi}}_y \mathbf{e}_E}{\mathbf{e}_R^T \hat{\mathbf{\Phi}}_y \mathbf{e}_E}$$

Unbiased estimate of elements corresponding to HA microphones

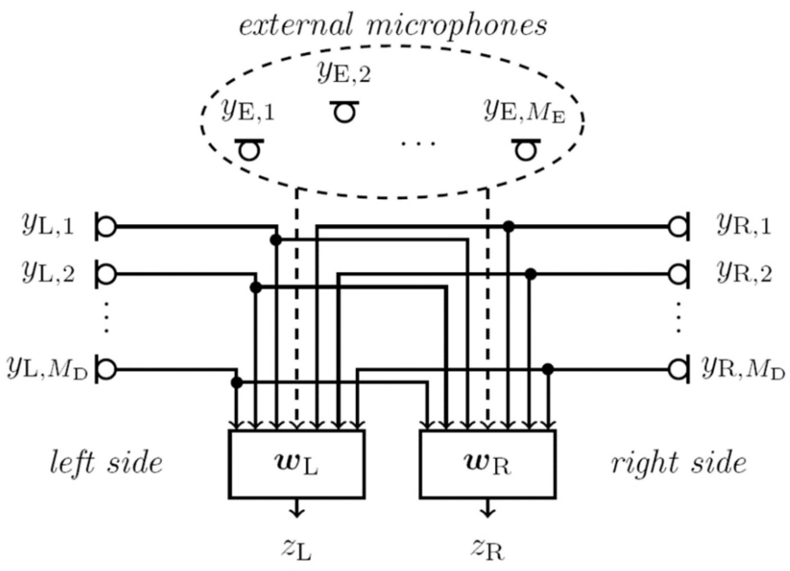
- **Low computational complexity** with similar (even better in practice) performance than state-of-the-art covariance whitening approach

Audio Demo



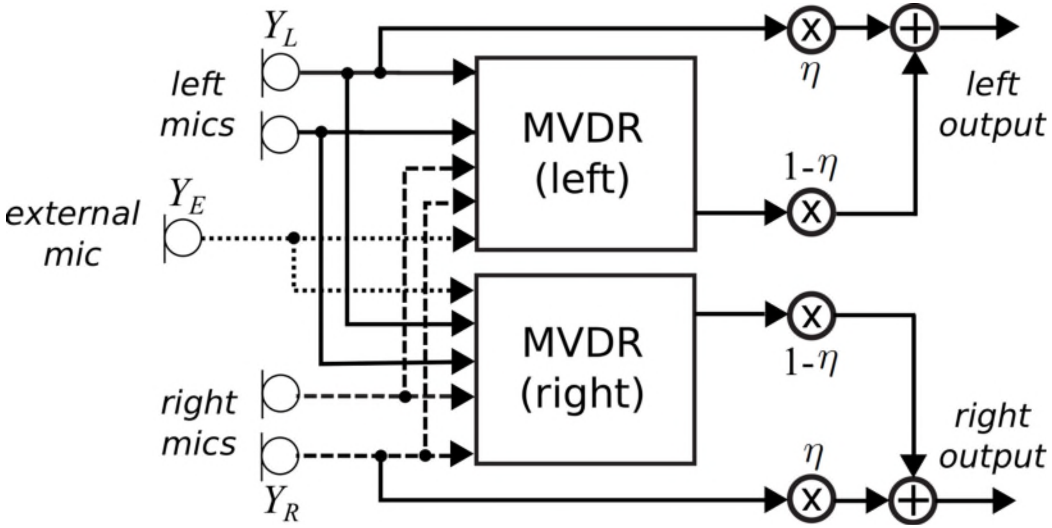
MVDR beamformer exploiting external microphones

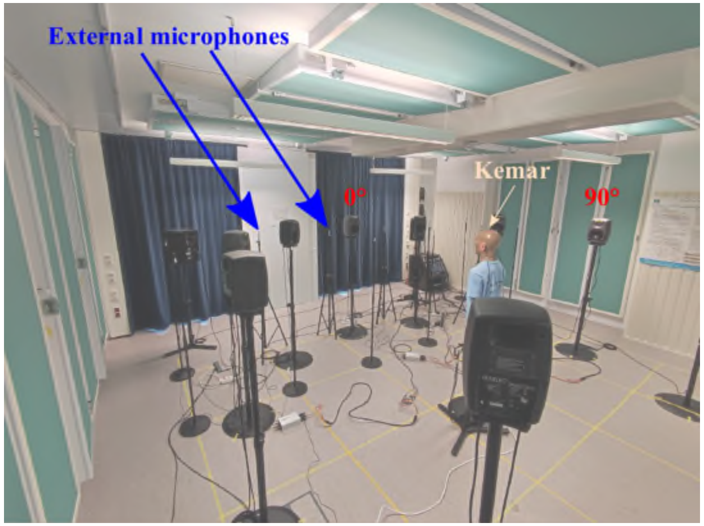
- **Extensions** for multiple external microphones, acoustic scenarios with multiple competing speakers and smart speaker scenario



MVDR beamformer exploiting external microphones

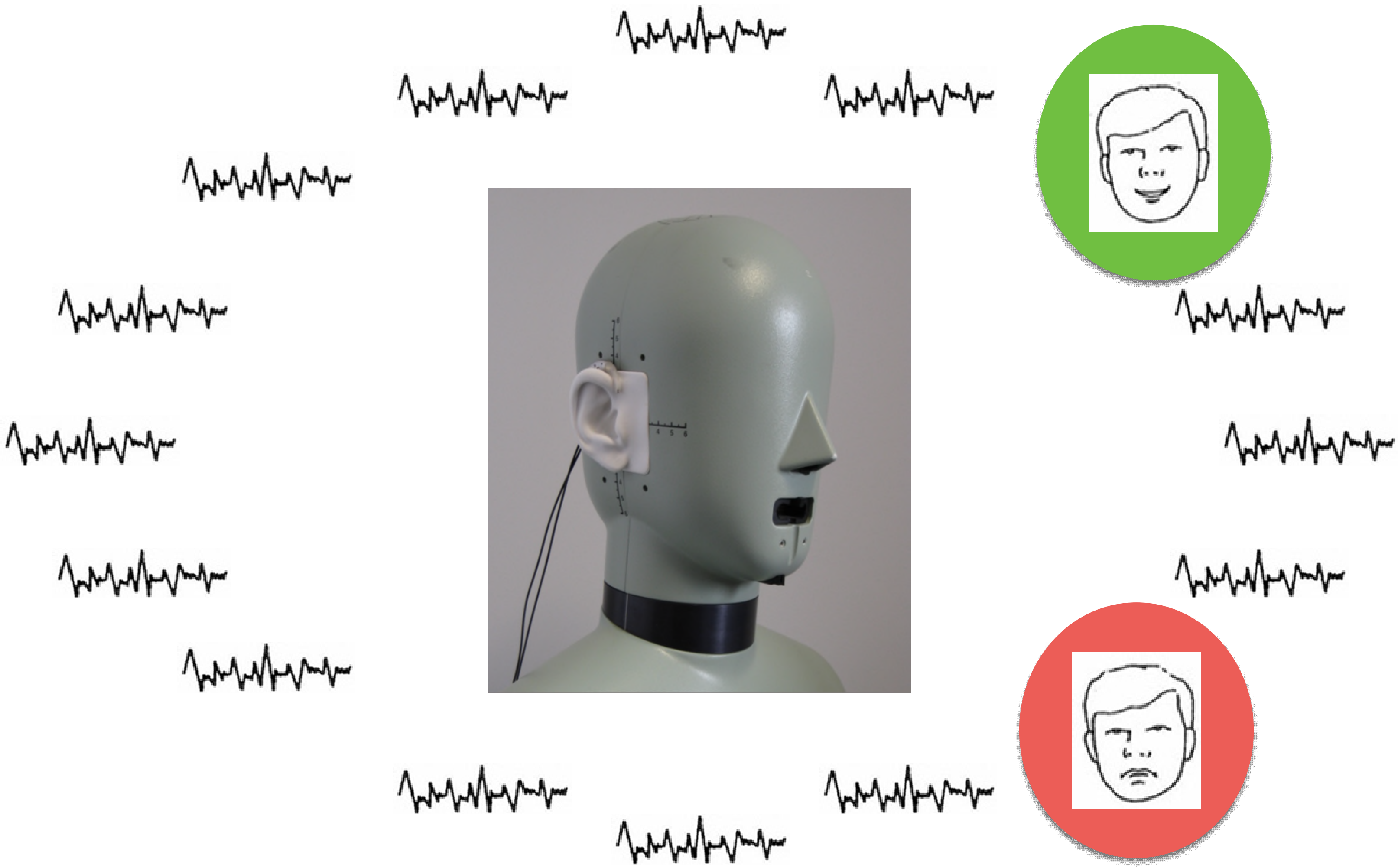
- **Extensions** for multiple external microphones, acoustic scenarios with multiple competing speakers and smart speaker scenario
- **Binaural cue preservation** of complete acoustic scene by using partial noise estimation
- **Publicly available database** with hearing aids and spatially distributed microphones (https://zenodo.org/record/7986447)





Sound source localization

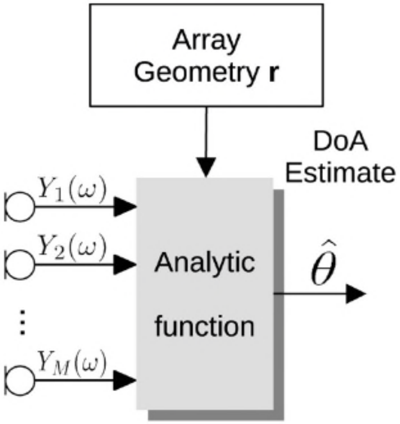
Sound source localization



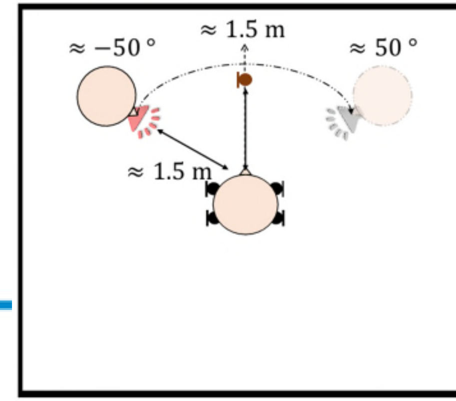
Sound source localization

1. Model-based approaches

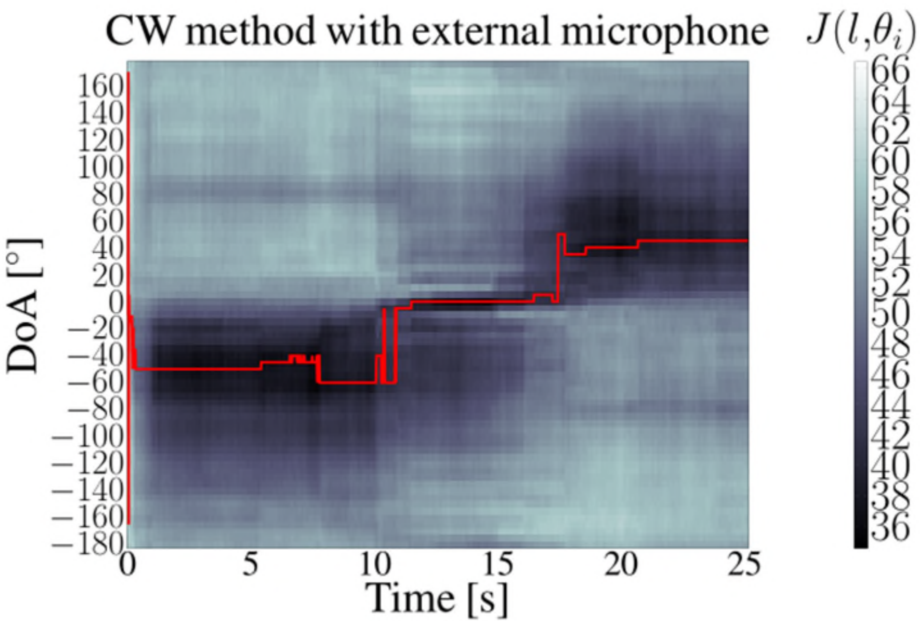
- Computation of **analytical function** (spatial pseudo-spectrum), typically based on prototype anechoic (relative) transfer functions $\tilde{\mathbf{a}}(k,\theta_i)$
  - *Beamforming*, e.g. steered response power [DiBiase 2000, Zouhourian 2018]
    $$p(k,l,\theta_i) = \frac{\tilde{\mathbf{a}}^H(k,\theta_i)\hat{\mathbf{\Phi}}_y(k,l)\tilde{\mathbf{a}}(k,\theta_i)}{\|\tilde{\mathbf{a}}(k,\theta_i)\|_2^2\,\|\mathbf{y}(k,l)\|_2^2}$$
  - *Subspace-based*, e.g. MUSIC [Schmidt 1986], [Dmochowski 2007]
    $$p(k,l,\theta_i) = \frac{1}{\|\hat{\mathbf{Q}}_u^H(k,l)\tilde{\mathbf{a}}(k,\theta_i)\|_2}$$
  - *Relative transfer function matching* [Braun 2015, Fejgin 2022]
    $$p(k,l,\theta_i) = \arccos\frac{|\tilde{\mathbf{a}}^H(k,\theta_i)\hat{\mathbf{a}}(k,l)|}{\|\tilde{\mathbf{a}}(k,\theta_i)\|_2\,\|\hat{\mathbf{a}}(k,l)\|_2}$$
- Requires **frequency integration/fusion** mechanism
- Prototype (relative) transfer functions can be computed from microphone array geometry/characteristics
  $\rightarrow$ **flexibility towards different array geometries**

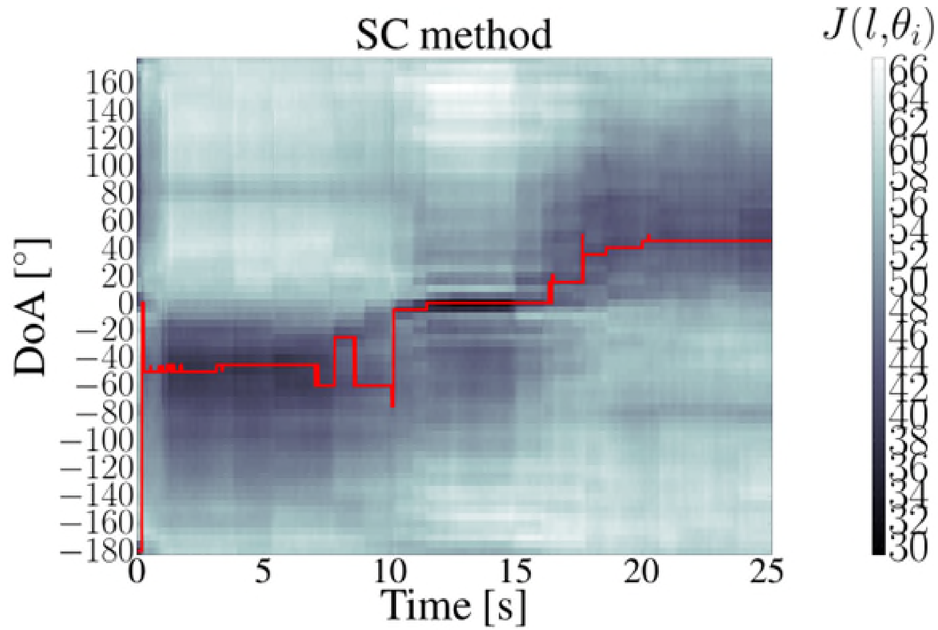


Model-based: RTF-based DOA estimation



- Simulation results with external mic for **moving speaker**



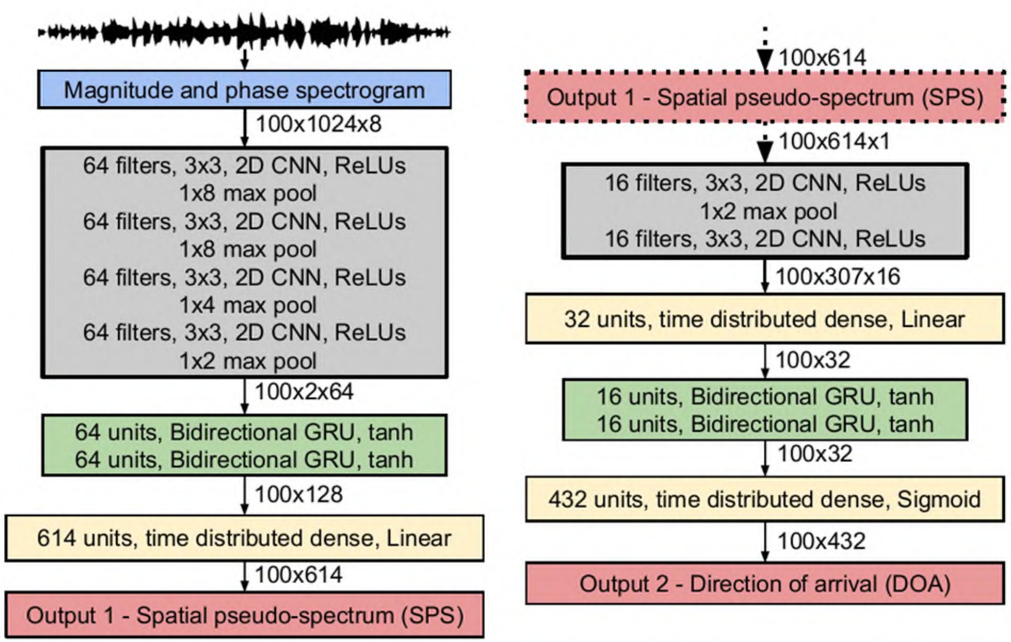


External microphone allows to estimate DOA accurately at low computational complexity without need to estimate noise covariance matrix

T_{60} ≈ 400ms, M=4 (BRIR), recorded diffuse babble noise, SNR = 0 dB; f_s = 16 kHz; STFT: 32ms (overlap 16ms); CW: τ_y=150 ms, τ_v=500 ms; SPP in head-mounted mics

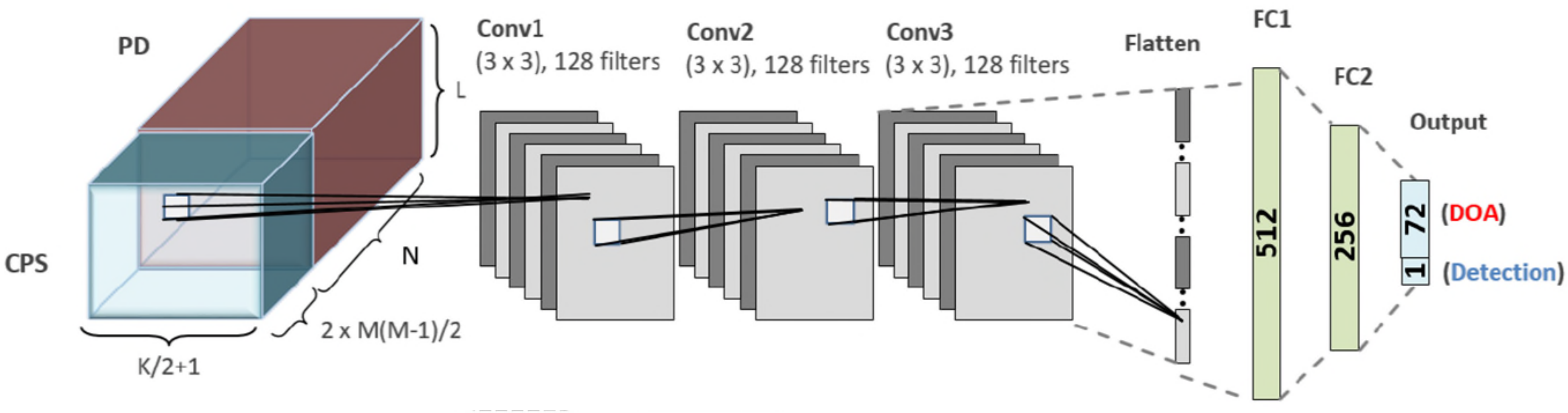
Sound source localization

2. **Learning-based approaches** [Grumiaux et al., JASA 2022]
   - **Learn relationship between input features and DOAs** (classification / regression)
   - Input features: spectrogram, inter-channel features (e.g. relative transfer functions)
   - Neural network architectures: convolutional (recurrent) neural networks, attention-based networks, …



Sound source localization

2. **Learning-based approaches** [Grumiaux et al., JASA 2022]
   - **Learn relationship between input features and DOAs** (classification / regression)
   - Input features: spectrogram, inter-channel features (e.g. relative transfer functions)
   - Neural network architectures: convolutional (recurrent) neural networks, attention-based networks, …

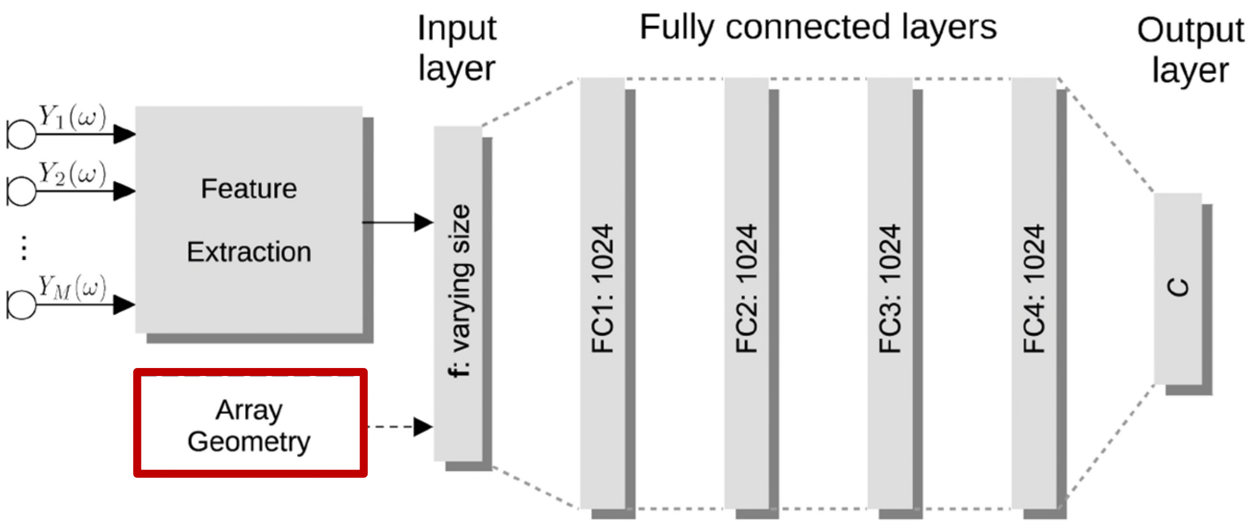


[Varzandeh, Hohmann, Doclo, *EUSIPCO 2023*]

Sound source localization

2. Learning-based approaches [Grumiaux et al., JASA 2022]

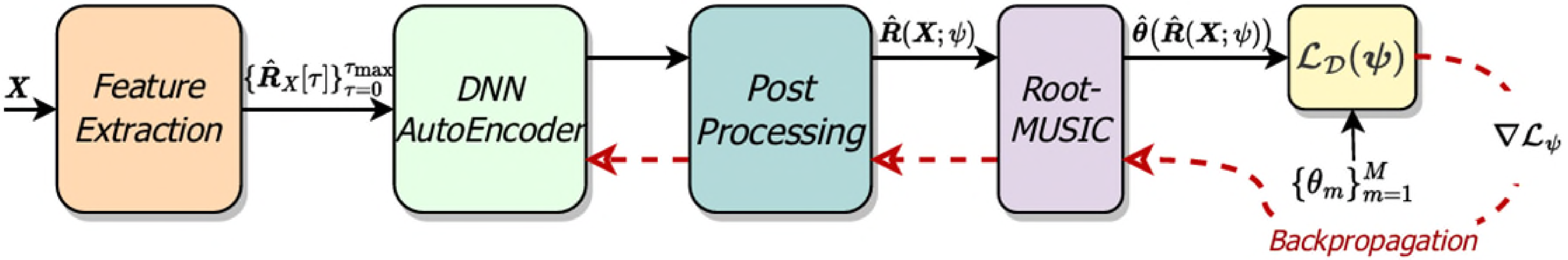
- **Learn relationship between input features and DOAs** (classification / regression)
- Input features: spectrogram, inter-channel features (e.g. relative transfer functions)
- Neural network architectures: convolutional (recurrent) neural networks, attention-based networks, …
- **Training data implicitly based on underlying array geometry**
  → geometry-aware DOA estimation



Sound source localization

3. Hybrid approaches

- **Combination of model-based and learning-based approaches**
  → merge interpretability of model-based approaches with ability to learn from real data
  → more flexible at lower computational complexity
- Examples:
  - End-to-end learning of masks for signal-aware DOA estimation using weighted steered response power method [Wechsler et al., 2022]
  - Deep learning-aided subspace methods [Shmuel et al., 2023]

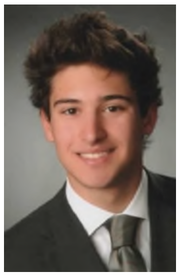


Summary

- **Model-based and learning-based approaches** for multi-microphone speech enhancement and source localization
- **Hybrid approaches** combining models with deep learning:
  - **Interpretability** of model-based approaches without perfectly satisfying model assumptions
  - **Performance** of learning-based approaches
  - **Generalizability** to unseen situations (dynamic acoustic scenes)
  - Especially useful for **low-complexity** applications

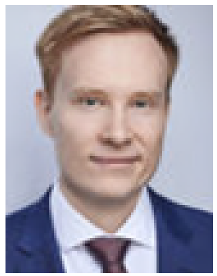
- **Challenges and opportunities:**
  - **Optimal trade-off** between latency, complexity and performance
  - **Best hybrid compromise** between model-based and learning-based approaches
  - **Microphone geometry**-independent/aware learning-based algorithms
  - Explore advantages of **unsupervised/semi-supervised algorithms**

Acknowledgments

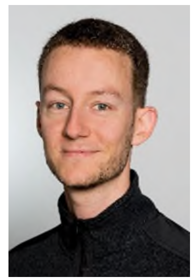


Daniel Fejgin

Henri Gode



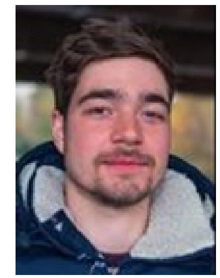
Dr. Nico Gößling



Ulrik Kowalk

Dr. Daniel Marquardt

Wiebke Middelberg



Marvin Tammen

Reza Varzandeh

- **Funding:**
  - Cluster of Excellence Hearing4all (DFG)
  - Collaborative Research Centre Hearing Acoustics (DFG)
  - Joint Lower-Saxony Israel Project "Acoustic scene aware speech enhancement for binaural hearing aids"
  - Marie Skłodowska-Curie Actions European Training Network SOUNDS

Questions ?

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

YouTube Signal Processing Uni Oldenburg