Carl von Ossietzky
Universität
Oldenburg
Signal Processing Group

DNN-based speech enhancement for hearing devices

Prof. Dr. Simon Doclo

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

Hearing
4all

Outline

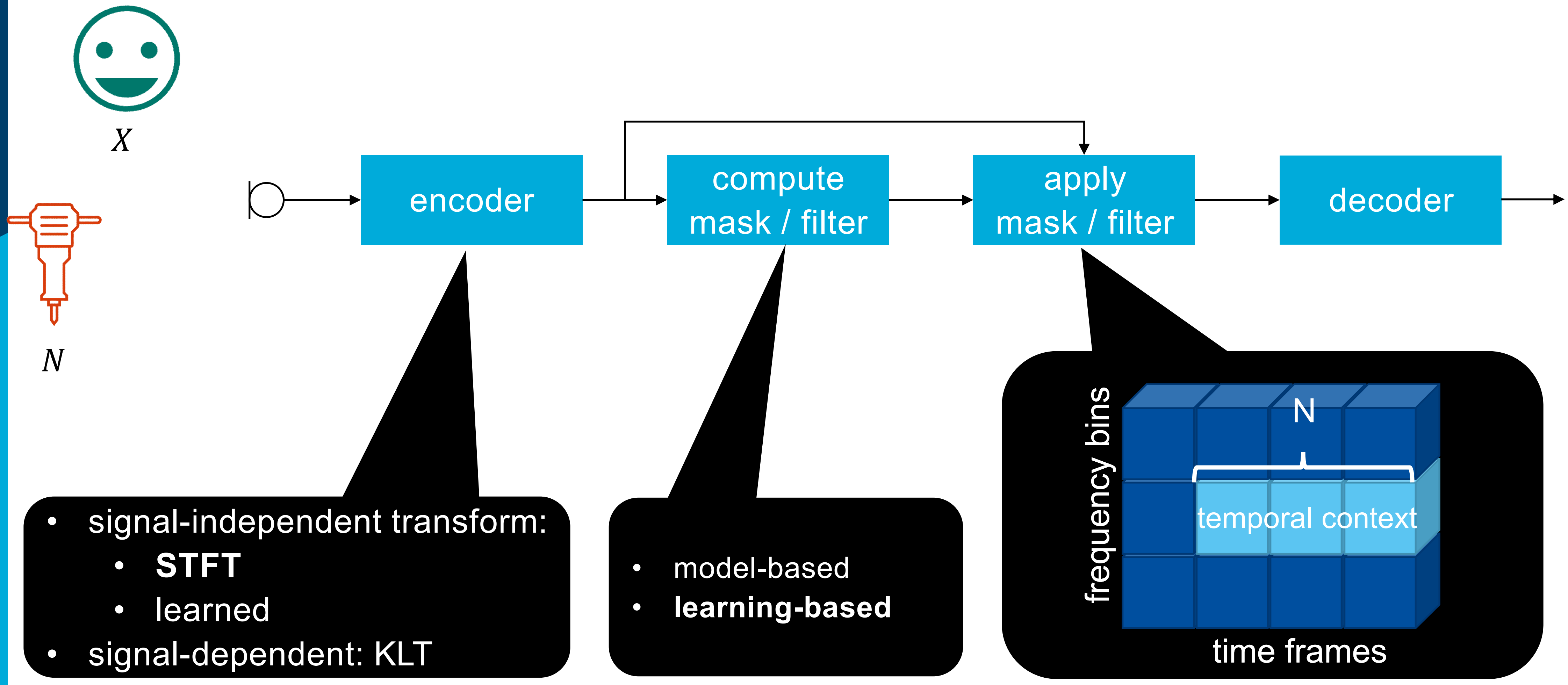
- **Deep multi-frame MVDR-based noise reduction**
  - combination of model-based and learning-based approach
  - single-microphone processing + extension towards binaural processing
- **Low-complexity single-channel noise reduction based on Skip-GRUs**
- **DNN-based own voice extraction using model-based data augmentation**

Disclaimer: all presented algorithms are on-line but complexity not always low enough for hearing devices

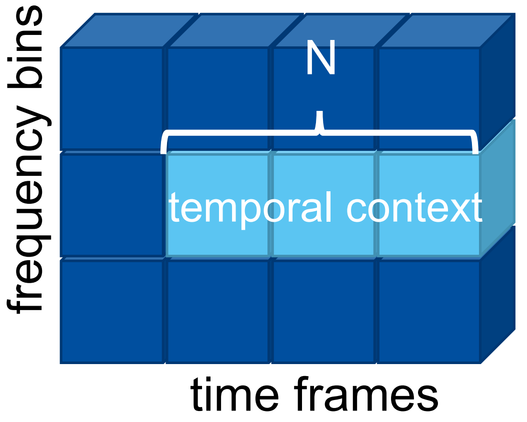
Carl von Ossietzky Universität Oldenburg
Signal Processing Group

Deep Multi-Frame MVDR-based Noise Reduction

Deep Multi-Frame Noise Reduction



Multi-Frame 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:

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

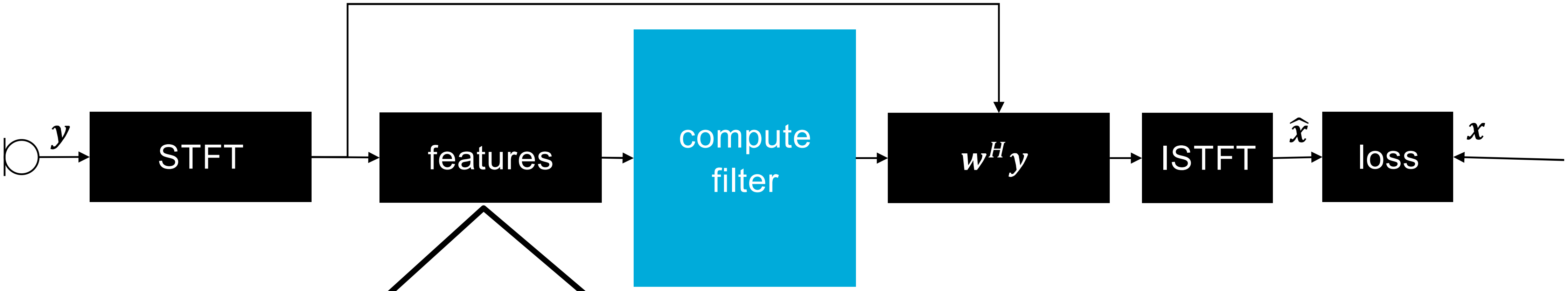
- solved by multi-frame MVDR (MFMVDR) filter:

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

  - 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

not trainable | trainable

Supervised Learning-Based Parameter Estimation



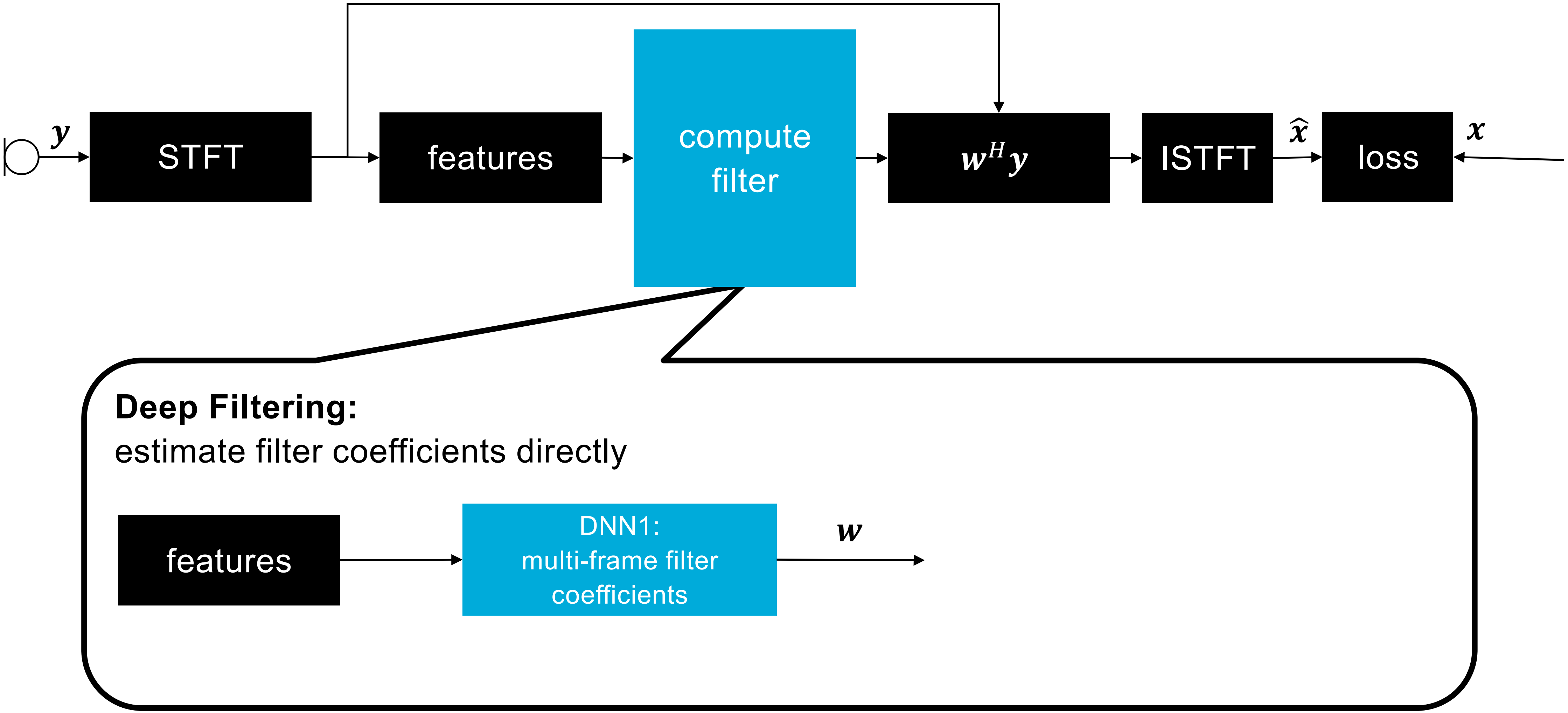
Features: concatenation of

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

[Tammen & Doclo, IEEE/ACM TASLP, in press]

not trainable | trainable

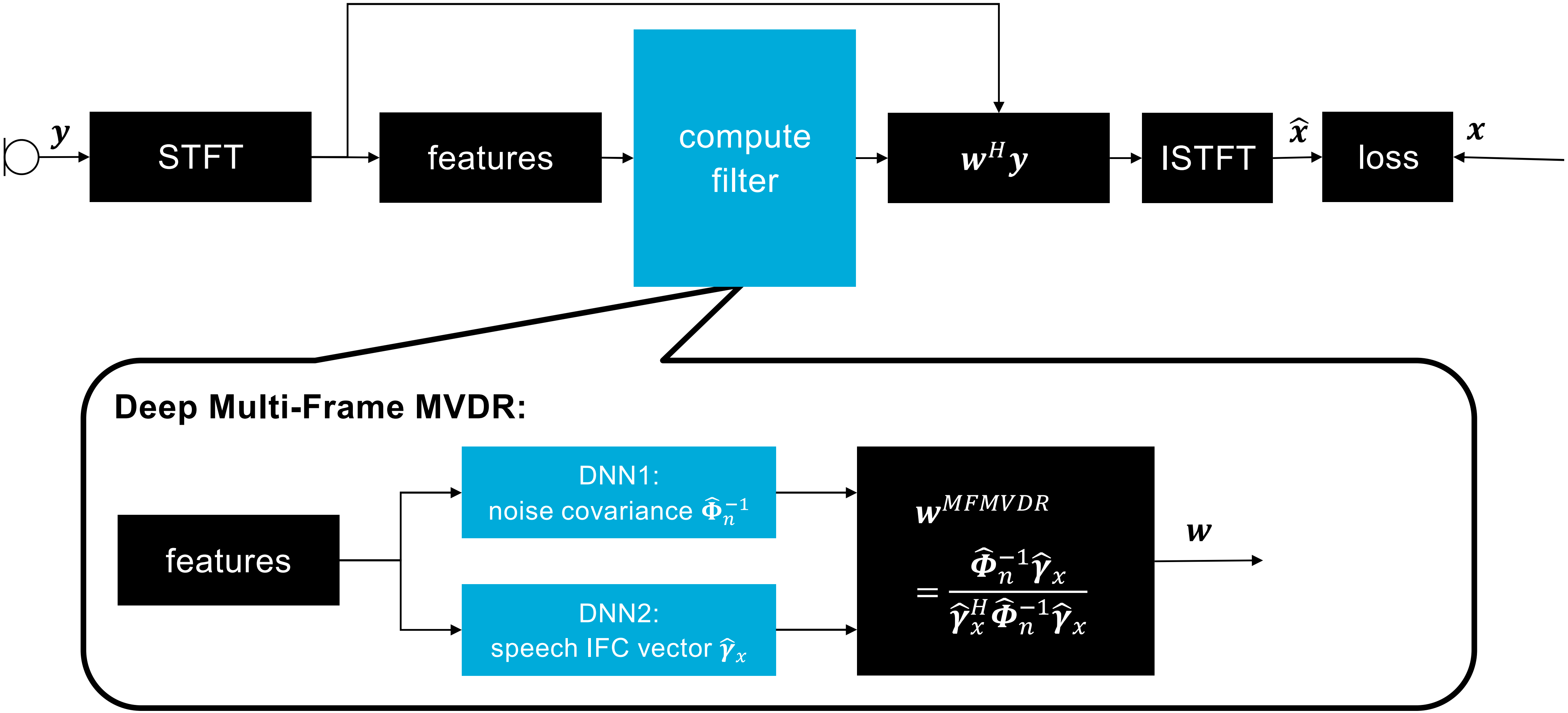
Supervised Learning-Based Parameter Estimation



[Tammen & Doclo, IEEE/ACM TASLP, in press]

not trainable | trainable

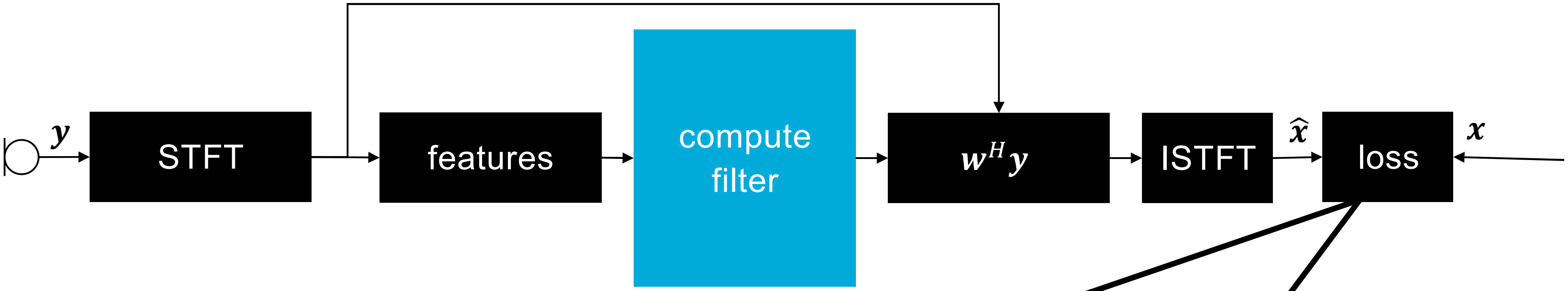
Supervised Learning-Based Parameter Estimation



[Tammen & Doclo, IEEE/ACM TASLP, in press]

not trainable | trainable

Supervised Learning-Based Parameter Estimation



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

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

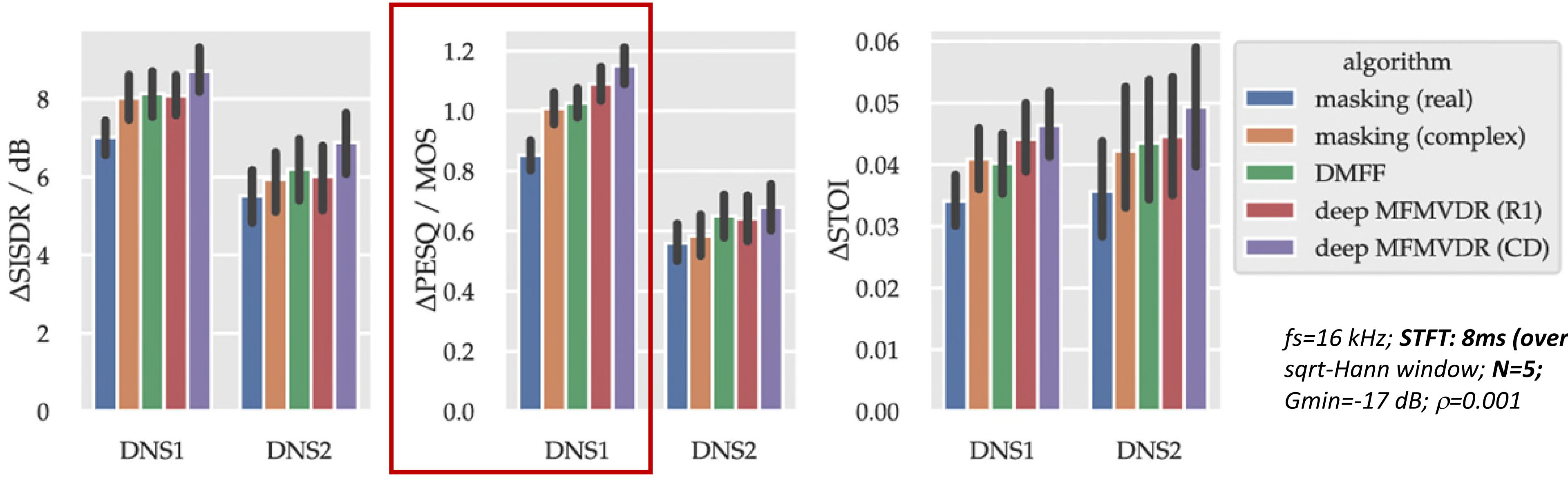
- [J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, in *Proc. 2019 ICASSP]*
- popular time-domain loss for speech enhancement and separation algorithms

[Tammen & Doclo, IEEE/ACM TASLP, in press]

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);***
*sqrt-Hann window; **N=5;***
Gmin=-17 dB; ρ=0.001

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

[Tammen & Doclo, IEEE/ACM TASLP, in press]

Simulation Results

- **Network size, complexity and real-time factor (RTF)**

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

[Tammen & Doclo, IEEE/ACM TASLP, in press]

Simulation Results - Audio examples

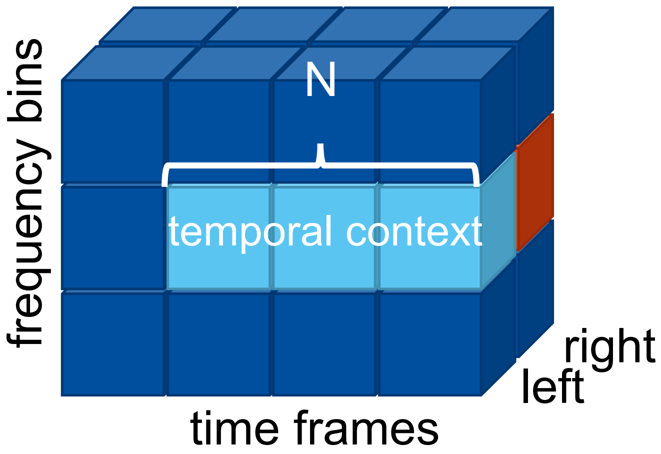
noisy		
single-frame mask, complex		
multi-frame filter, direct estimation		
multi-frame filter, MFMVDR structure		

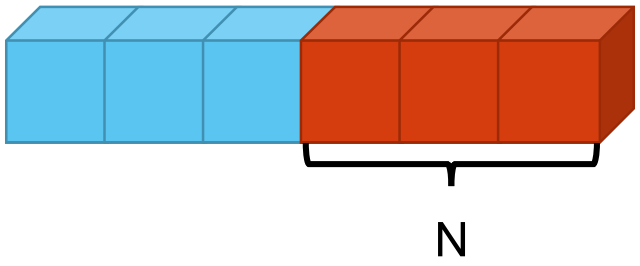
[Tammen & Doclo, IEEE/ACM TASLP, in press]

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

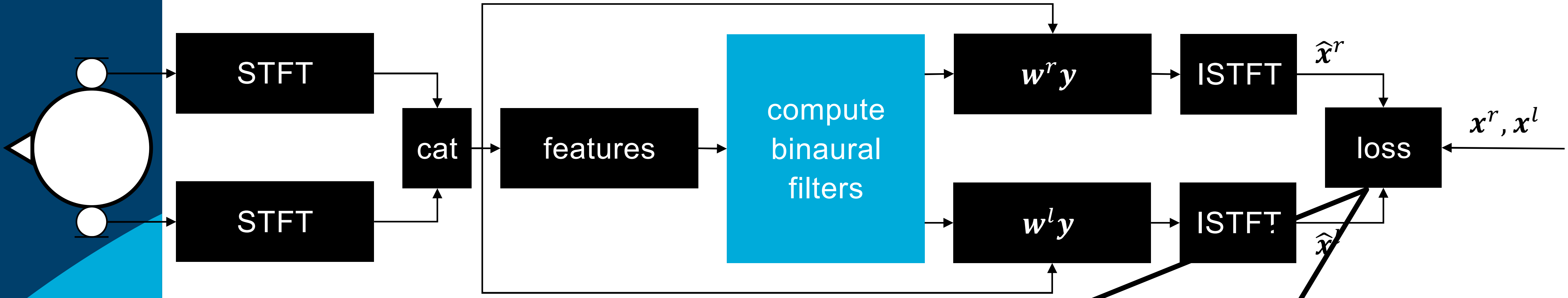




[Tammen & Doclo, Proc. IWAENC 2022]

not trainable | trainable

Supervised Learning-Based Parameter Estimation



Loss: Combined Mean Absolute Spectral Error

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

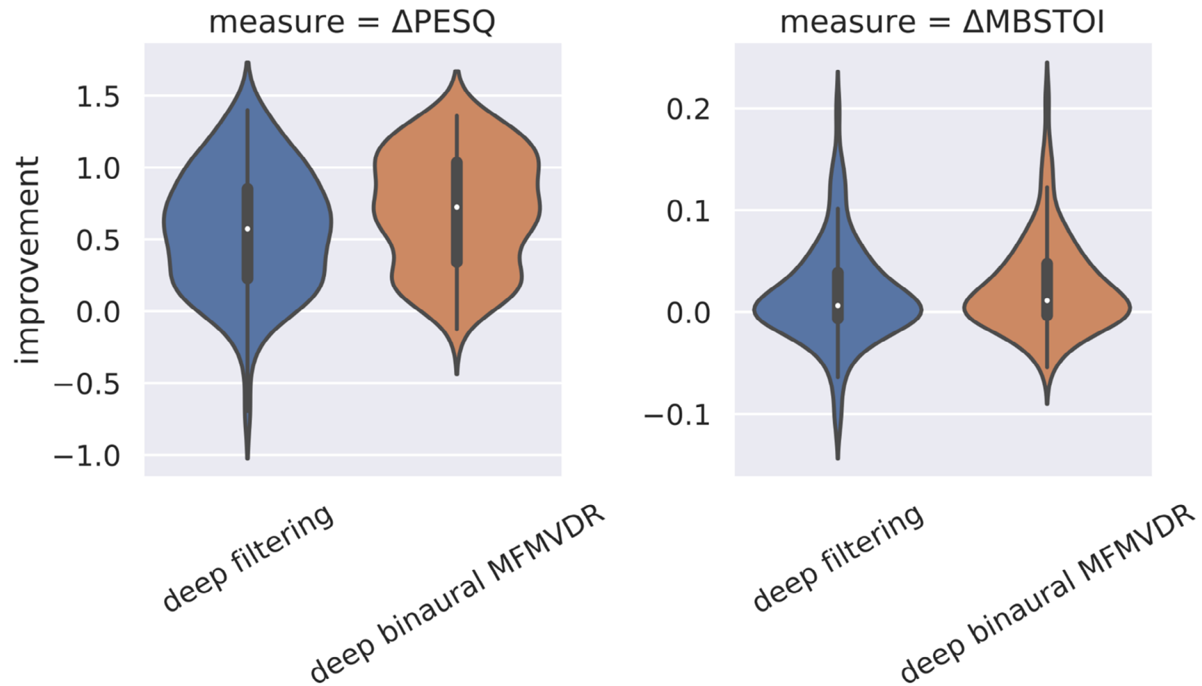
- more robust against reverberation than SI-SDR
- [Z.-Q. Wang, P. Wang, and D. Wang, *IEEE/ACM TASLP*, 2020]

[Tammen & Doclo, Proc. IWAENC 2022]

Carl von Ossietzky
Universität
Oldenburg
Signal Processing Group

Simulation Results

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



[Tammen & Doclo, Proc. IWAENC 2022]

Simulation Results – Audio Examples

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

[Tammen & Doclo, Proc. IWAENC 2022]

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

Carl von Ossietzky
Universität
Oldenburg
Signal Processing Group

Conclusions

- **Considerable monaural and binaural noise reduction** performance using supervised learning-based approaches
- Consistent **benefit by imposing multi-frame MVDR structure**
- Co

Deep Multi-Frame Filtering for Hearing Aids

Hendrik Schröter[1], *Tobias Rosenkranz*[2], *Alberto N. Escalante-B.*[2], *Andreas Maier*[1]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, Pattern Recognition Lab
[2]WS Audiology, Research and Development, Erlangen, Germany

hendrik.m.schroeter@fau.de

- **Cu**
  - Joint noi … acoustic scene us

arXiv > eess > arXiv:2305.08225

Electrical Engineering and Systems Science > Audio and Speech Processing

[Submitted on 14 May 2023]

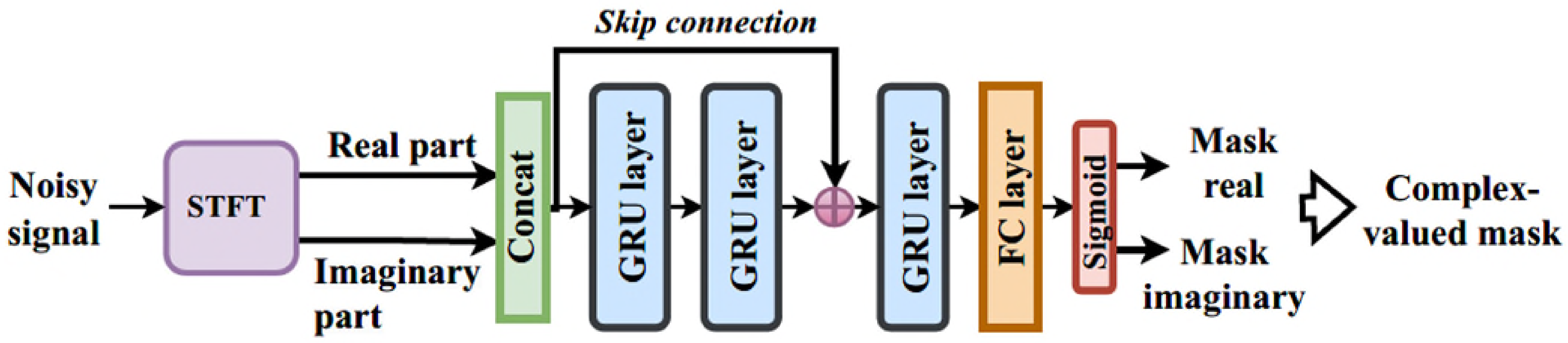
Carl von Ossietzky Universität Oldenburg
Signal Processing Group

Fraunhofer IDMT

Low-complexity single-channel noise reduction

Low-complexity single-channel noise reduction

- **Compact system which is resource-efficient during inference**
- **Skip-GRU architecture:** smooth flow of information without increasing complexity



[Sinha, Rollwage & Doclo, Proc. ITG Speech Comm. 2023]

Carl von Ossietzky Universität Oldenburg
Signal Processing Group

Fraunhofer IDMT

Low-complexity single-channel noise reduction

- **Evaluation on DNS challenge dataset**
- **Latency** for all algorithms **32 ms** (lower latency possible)

Systems	SI-SDR (dB)	WB-PESQ	STOI	MACs (G/s)	#Param
Input Noisy Signals	9.2	1.8	0.87	-	-
FullSubNet+ (ret) [14]	13.5	2.6	0.89	31.81	8.7 M
DEMUCS-48 (ret) [9]	14.5	2.5	0.90	1.51	18.9 M
DTLN (ret) [11]	12.2	2.1	0.88	0.12	1.0 M
GRU (real)	13.2	2.2	0.89	0.21	1.8 M
Skip-GRU (real)	13.9	2.3	0.90	0.21	1.8 M
GRU (complex)	14.1	2.4	0.90	0.39	4.4 M
Skip-GRU (complex)	**14.4**	**2.4**	**0.90**	0.39	4.4 M

Proposed Skip-GRU system achieves similar performance as best SOA system with about 4 times lower complexity

[Sinha, Rollwage & Doclo, Proc. ITG Speech Comm. 2023]

Carl von Ossietzky Universität Oldenburg
Signal Processing Group

Fraunhofer IDMT

Own voice extraction

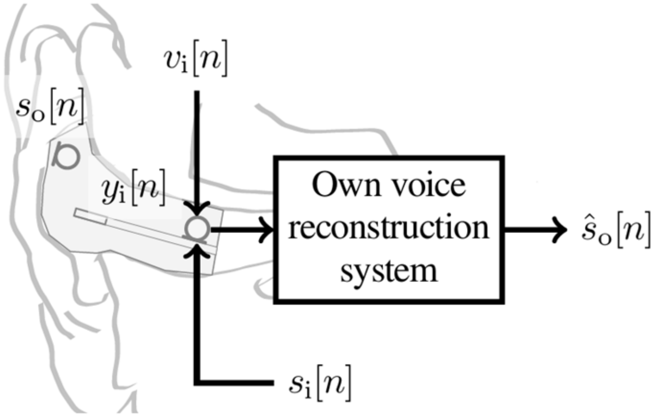
Carl von Ossietzky
Universität
Oldenburg
Signal Processing Group

Fraunhofer
IDMT

Own voice extraction

- **Aim:** enhance own voice of user wearing earpiece in noisy acoustic environment (e.g. industrial workplace)
- **Different characteristics** for own voice and external noise at in-ear and outer microphones
  - **in-ear microphone:** bandlimited own voice, high SNR (external noise), body noise
  - **outer microphone:** full bandwidth, low SNR (external noise)

- **Objectives of algorithm:** estimate clean speech signal at outer microphone from
  - in-ear microphone: **combined bandwidth extension, equalization and noise reduction** (body + external noise)
  - in-ear and outer microphone

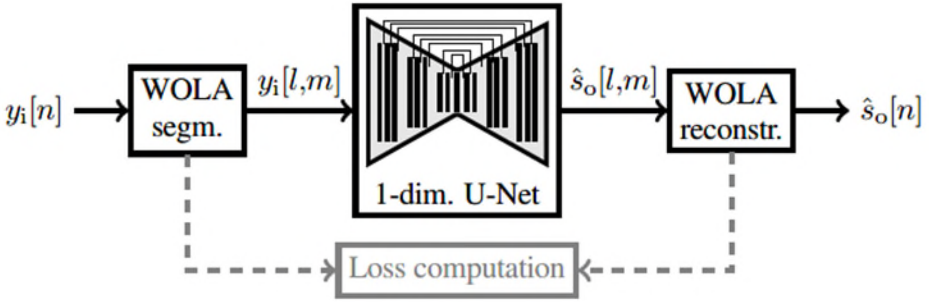


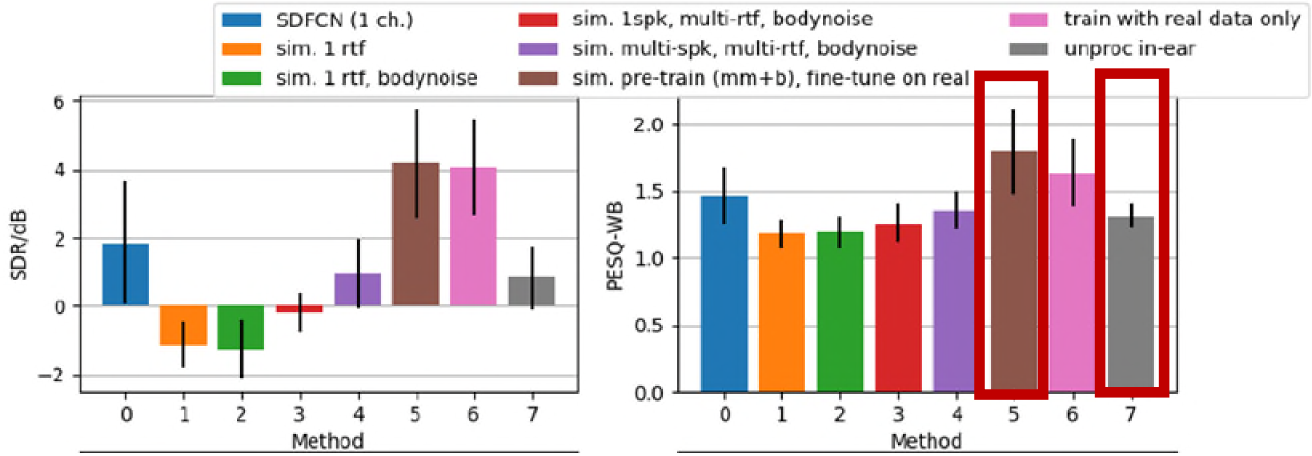
Carl von Ossietzky Universität Oldenburg
Signal Processing Group

Fraunhofer IDMT

Own voice extraction

- **Limited training data available for supervised learning-based algorithms**:
  - use **acoustic models** to generate simulated data (data augmentation):
    - Fixed relative transfer function (sp.-indep.)
    - Phoneme-dependent relative transfer function (sp.dep.)
  - **domain transfer** (train with simulated data, fine-tune with real recordings)

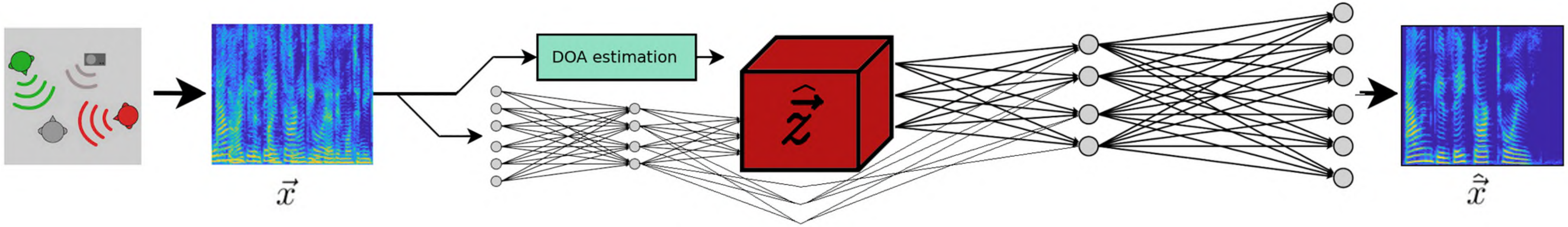




[Ohlenbusch, Rollwage & Doclo, *Proc. IWAENC 2022, Proc. Forum Acusticum 2023*]

Current / future work: challenges and opportunities…

- **Applications** to (binaural) speech enhancement, own voice extraction, DOA estimation, acoustic feedback control and active noise reduction
- Explore **trade-off between latency/complexity and performance**
- Best hybrid compromise between **model-based and learning-based approaches**
- Realistic **dynamic acoustic scenes** with moving speakers and (fast) head movements
- Integration with **individual hearing loss compensation**: 1-stage (individual) vs. 2-stage
- Explore advantages of **unsupervised/semi-supervised algorithms**



Questions ?

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

YouTube Signal Processing Uni Oldenburg