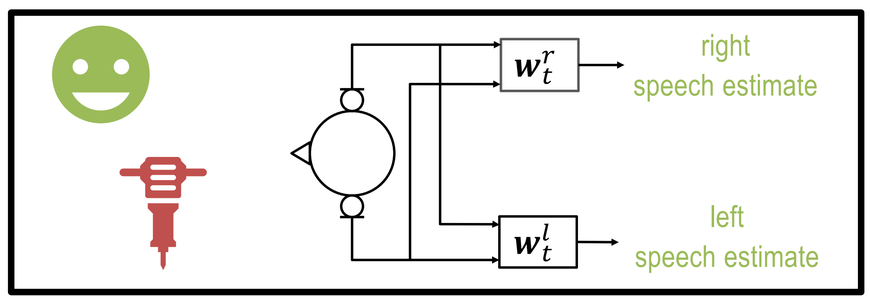
Deep Multi-Frame MVDR Filtering for Binaural Noise Reduction

Marvin Tammen, Simon Doclo
Dept. Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany

Problem Statement

- microphone signals on hearing aids degraded by ambient noise
  → decreased speech quality and intelligibility
- **multi-frame speech enhancement algorithms** have been shown to yield good noise reduction and low speech distortion [1]
- **multi-frame filter coefficients are difficult to estimate** using traditional approaches

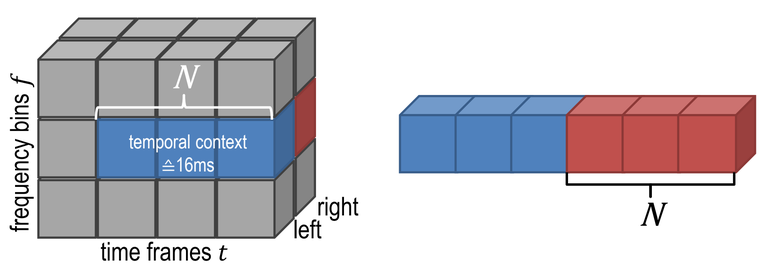
this poster: supervised learning-based approaches to estimate binaural **multi-frame filter** coefficients: **should filter structure be imposed**?



Signal Model

- apply complex-valued **multi-frame filter** to N frames in STFT domain:

$$\boldsymbol{y}_{t,f} = \left[Y_{t,f}^l \quad \dots \quad Y_{t-N+1,f}^l \quad Y_{t,f}^r \quad \dots \quad Y_{t-N+1,f}^r\right]^T = \boldsymbol{x}_{t,f} + \boldsymbol{n}_{t,f}$$



- **assumptions**:
  1. independent speech and noise components:
  $$\boldsymbol{\Phi}_{y,t} = \mathcal{E}\{\boldsymbol{y}_t \boldsymbol{y}_t^H\} = \boldsymbol{\Phi}_{x,t} + \boldsymbol{\Phi}_{n,t} \in \mathbb{C}^{2N \times 2N}$$
  2. decompose speech into spatio-temporally correlated and uncorrelated components [1]:
  $$\boldsymbol{x}_t = \boldsymbol{\gamma}_{x,t}^l X_t^l + \boldsymbol{x}_t'^l = \boldsymbol{\gamma}_{x,t}^r X_t^r + \boldsymbol{x}_t'^r$$
  with $\boldsymbol{\gamma}_{x,t}^{\{l,r\}}$ the **speech correlation vectors**

- relation to beamforming approaches:

binaural multi-frame filtering	binaural beamforming
exploits spatio-temporal correlations	exploits spatial correlations
spatio-temporal correlation vector: **signal-dependent** → highly time-varying	steering vector: **room- and geometry-dependent** → quasi-stationary

Multi-Frame MVDR Filter

- minimizes output noise power spectral density while leaving correlated speech component undistorted:

$$\mathbf{w}_t = \underset{\tilde{\boldsymbol{w}}_t}{\operatorname{argmin}}\ \tilde{\boldsymbol{w}}_t^H \boldsymbol{\Phi}_{n,t} \tilde{\boldsymbol{w}}_t \ \text{ s.t. } \ \tilde{\boldsymbol{w}}_t^H \boldsymbol{\gamma}_{x,t} = 1$$

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

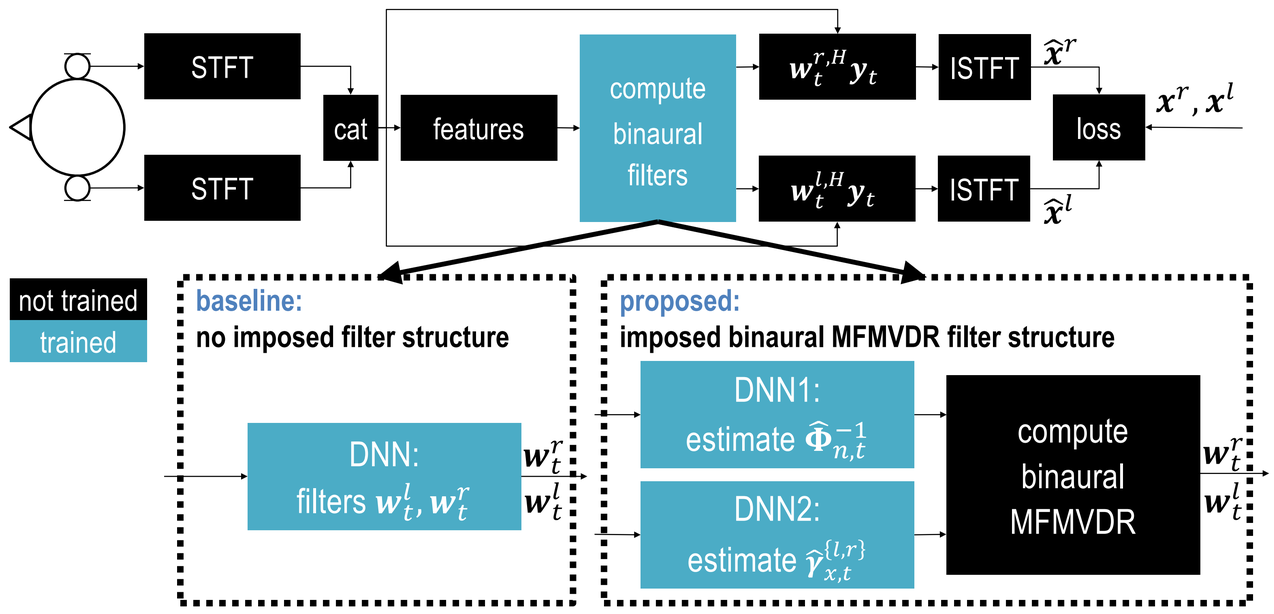
➢ **requires estimate** of
- inverse noise covariance matrix $\boldsymbol{\Phi}_{n,t}^{-1}$
- speech correlation vectors $\boldsymbol{\gamma}_{x,t}$ (left / right, **highly time-varying**)

Dataset

	training	evaluation
hearing aids	BTE hearing aids	
binaural room impulse responses (speech and noise)	**simulated**; from Clarity Enhancement Challenge [2]	**measured**; daily-life communication scenarios [3]
clean speech	deep noise suppression (DNS) challenge 1 training set	DNS challenge 3 test set
noise		
SNR	$[0, 15]$ dB	$[-5, 20]$ dB
speech azimuth	$[-30, 30]°$	$[-180, 180)°$

Supervised Learning-Based Filter Estimation

- integrate **fully-differentiable binaural MFMVDR filter** into **end-to-end supervised learning framework** [4]
  - all parameters estimated using DNNs
  - training is guided using binaural speech enhancement loss function



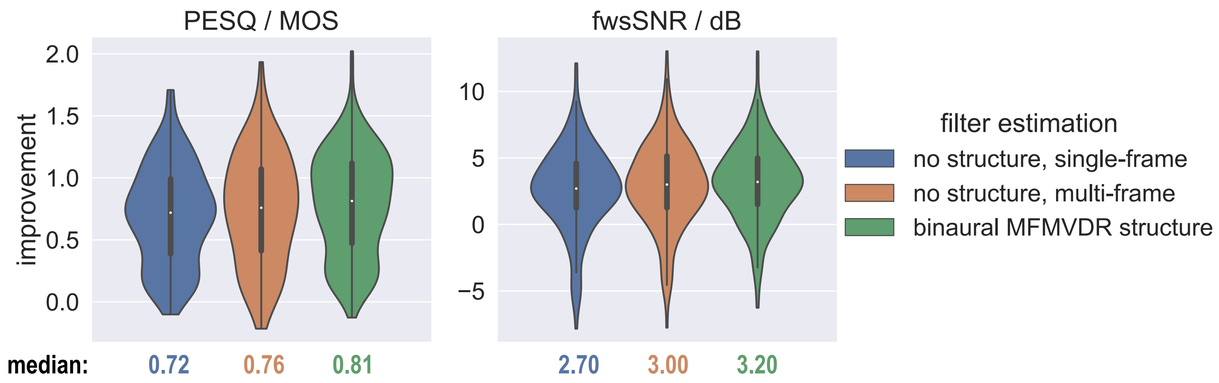
Simulations

Settings

- $f_s = 16$ kHz; **STFT**: $\sqrt{\text{Hann}}$ window, 8 ms frame length, 75 % overlap
- filter length $N = 5$ → temporal context of 16 ms
- **features**: log-magnitude, cosine and sine of phase of microphone signals
- **DNN architecture**: causal temporal convolutional networks [5]
  - 2 stacks of 6 layers; hidden dimensions chosen to yield similar number of parameters across compared algorithms
  - temporal receptive field size: 512 ms
- **loss function**: combined complex and magnitude absolute spectral error [6]
- trained using **AdamW optimizer** for ≤ 150 epochs (with early stopping)
- **minimum gain** of -20 dB during evaluation

Results

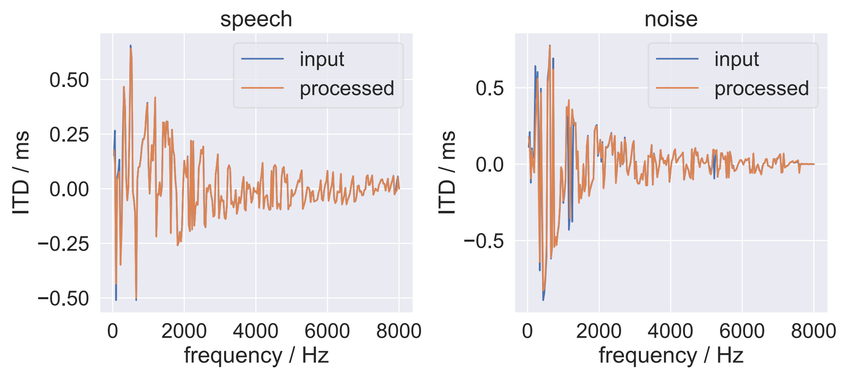
1. **Speech Enhancement Performance** (averaged across left / right):
   - perceptual evaluation of speech quality (PESQ)
   - frequency-weighted segmental SNR (fwsSNR)



   - all algorithms yield **considerable speech quality improvements**
   - multi-frame filters outperform single-frame filter

imposing binaural MFMVDR structure yields **highest improvements**

2. **Preservation of Interaural Time Differences:**



both speech and noise cues are well preserved

Outlook

- **reduce complexity** of binaural MFMVDR filter by assuming, e.g., slowly varying *spatial* correlations, with final goal of **real-time implementation**
- evaluation of binaural MFMVDR filters using **listening test**

check out audio demos:

References
[1] Y. A. Huang and J. Benesty, "A Multi-Frame Approach to the Frequency-Domain Single-Channel Noise Reduction Problem," IEEE Trans. Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1256–1269, May 2012.
[2] S. Graetzer et al., "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in Proc. Interspeech, Brno, Czech Republic, Aug. 2021, pp. 686–690.
[3] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses," EURASIP J. Adv. Signal Process., vol. 2009, no. 1, Dec. 2009.
[4] M. Tammen and S. Doclo, "Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, Jun. 2021, pp. 8443–8447.
[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[6] Z.-Q. Wang, P. Wang, and D. Wang, "Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR," IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 28, pp. 1778–1787, Jan. 2020.
This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 - Project ID 390895286.