Parameter Estimation Procedures for Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement

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

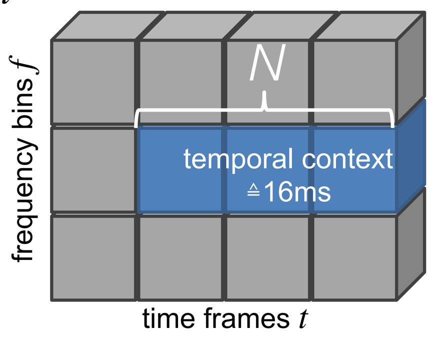
Problem Statement

- multi-frame speech enhancement algorithms provide good noise reduction and low speech distortion
- multi-frame filters can be estimated using deep neural networks (DNNs) with or without imposing structure on the filter coefficients
  → **multi-frame minimum variance distortionless response (MVDR) filter**

this poster: different procedures to estimate the parameters required by the multi-frame MVDR filter

Signal Model

- noisy STFT-domain vector $\boldsymbol{y}_t = \left[y_t \; \dots \; y_{t-N+1}\right]^T = \boldsymbol{x}_t + \boldsymbol{n}_t$
- apply complex-valued **multi-frame filter** $\mathbf{w}_t$ to N frames:
  $\hat{x}_t = \mathbf{w}_t^H \mathbf{y}_t \quad \mathbf{w}_t = \left[w_{t,1} \; \cdots \; w_{t,N}\right]^T$



- **assumptions**:
  1. → decompose speech vector $\mathbf{x}_t$ into correlated and uncorrelated components [1]:
  $$\mathbf{x}_t = \left[x_t \; \dots \; x_{t-N+1}\right]^T = \boldsymbol{\gamma}_{x,t} x_t + \mathbf{x}_t'$$
  2. uncorrelated component is considered interference: $\mathbf{i}_t = \mathbf{x}_t' + \mathbf{n}_t$
  3. independent components: $\boldsymbol{\Phi}_{y,t} = E\{\mathbf{y}_t \mathbf{y}_t^H\} = \boldsymbol{\Phi}_{x,t} + \boldsymbol{\Phi}_{i,t}$
- **speech inter-frame correlation (IFC) vector** describes correlation between current and N most recent time frames:

$$\boldsymbol{\gamma}_{x,t} = \frac{E\{\mathbf{x}_t x_t^*\}}{\phi_{x,t}}, \quad \phi_{x,t} = E\{\,|x_t|^2\,\}$$

Multi-Frame MVDR Filter

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

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

- $\boldsymbol{\gamma}_{x,t}$ is **highly time-varying** and **difficult to estimate** → rewrite using more accessible noisy & interference covariance matrices and a-priori SNR ξ_t:

$$\boldsymbol{\gamma}_{x,t} = \frac{1+\xi_t}{\xi_t} \frac{\boldsymbol{\Phi}_{y,t}\mathbf{e}}{\mathbf{e}^T \boldsymbol{\Phi}_{y,t} \mathbf{e}} - \frac{1}{\xi_t} \frac{\boldsymbol{\Phi}_{i,t}\mathbf{e}}{\mathbf{e}^T \boldsymbol{\Phi}_{i,t} \mathbf{e}}, \quad \xi_t = \frac{\phi_{x,t}}{\phi_{i,t}}, \quad \mathbf{e} = \left[1 \; 0 \; \dots \; 0\right]^T$$

main objective: estimate $\boldsymbol{\Phi}_{y,t}$, $\boldsymbol{\Phi}_{i,t}$, and ξ_t

Dataset

based on deep noise suppression (DNS) challenge dataset [2]:

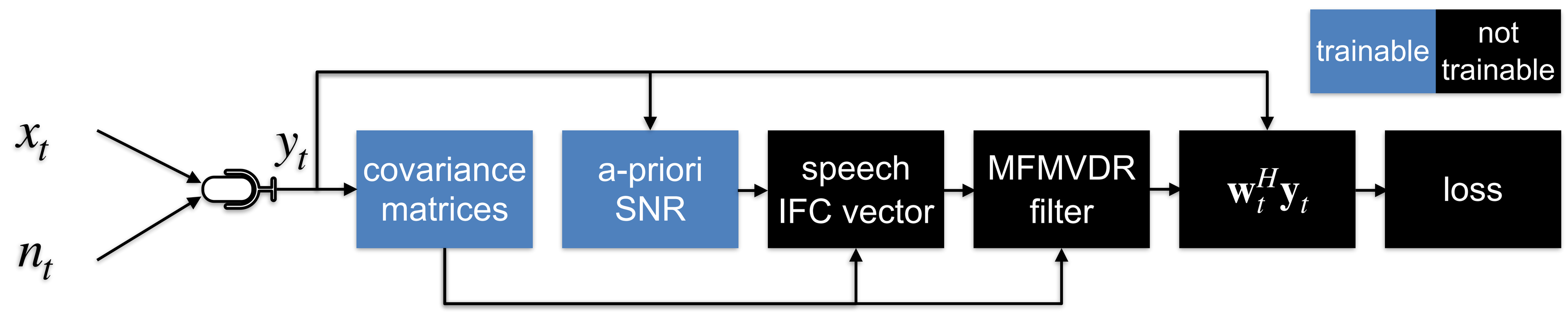
training & validation	evaluation
anechoic English speech (LibriSpeech)	anechoic English speech (Uni Graz)
noise: Audioset, Freesound, DEMAND	noise: Freesound
SNRs from 0 dB to 19 dB	
100 h	150 utterances

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, cos and sin of phase of noisy microphone signals
- **DNN architecture**: causal temporal convolutional networks (TCNs) [3]
  - 2 stacks of 4 layers; hidden dimensions chosen to yield similar number of parameters across compared algorithms (ca. 5 M)
  - temporal receptive field size: 128 ms
- scale-invariant signal-to-distortion ratio **(SI-SDR) loss function**
- trained using AdamW optimizer for ≤ 150 epochs (with early stopping)
- minimum gain of -17 dB during evaluation
- diagonal loading applied to estimated covariance matrices before inversion
- **baseline algorithms**: direct estimation of mask or multi-frame filter [4]

Deep Multi-Frame MVDR Filter

- integrate multi-frame MVDR filter into end-to-end supervised learning framework [5]: $\boldsymbol{\Phi}_{y,t}$, $\boldsymbol{\Phi}_{i,t}$, and ξ_t **estimated using DNNs**:



i) Recursive Smoothing (RS)

y_t → DNN → $\lambda_t \in \mathbb{R}$ →

$$\widehat{\boldsymbol{\Phi}}_t^{\text{RS}} = \lambda_t \widehat{\boldsymbol{\Phi}}_{t-1}^{\text{RS}} + (1-\lambda_t)\mathbf{y}_t\mathbf{y}_t^H$$

ii) Cholesky Decomposition (CD)

y_t → DNN → $\mathbf{a}_t^{\Re}, \mathbf{a}_t^{\Im} \in \mathbb{R}^{N^2/2}$ →

$$\widehat{\mathbf{L}}_t = \begin{bmatrix} a_{t,1}^{\Re} + ja_{t,1}^{\Im} & 0+0 & \dots & 0+j0 \\ a_{t,2}^{\Re} + ja_{t,2}^{\Im} & a_{t,N+1}^{\Re} + ja_{t,N+1}^{\Im} & \dots & 0+j0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{t,N}^{\Re} + ja_{t,N}^{\Im} & a_{t,2N-1}^{\Re} + ja_{t,2N-1}^{\Im} & \dots & a_{t,N^2/2}^{\Re} + ja_{t,N^2/2}^{\Im} \end{bmatrix}$$

imposes **positive-definite** structure

$$\widehat{\boldsymbol{\Phi}}_t^{\text{CD}} = \mathbf{L}_t \mathbf{L}_t^H$$

iii) Vandermonde Factorization (PDT)

y_t → DNN → $\boldsymbol{\theta}_t, \mathbf{a}_t \in \mathbb{R}^N$ →

$$\widehat{\mathbf{V}}_t = \begin{bmatrix} 1 & \exp(1j\theta_{t,1}) & \dots & \exp(Nj\theta_{t,1}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \exp(1j\theta_{t,N}) & \dots & \exp(Nj\theta_{t,N}) \end{bmatrix}$$

imposes **positive-definite Toeplitz** structure (stationarity over N frames)

$$\widehat{\mathbf{D}}_t = \begin{bmatrix} a_{t,1} & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & a_{t,N} \end{bmatrix} \qquad \widehat{\boldsymbol{\Phi}}_t^{\text{PDT}} = \mathbf{V}_t \mathbf{D}_t \mathbf{V}_t^H$$

iv) Rank-1 (R1)

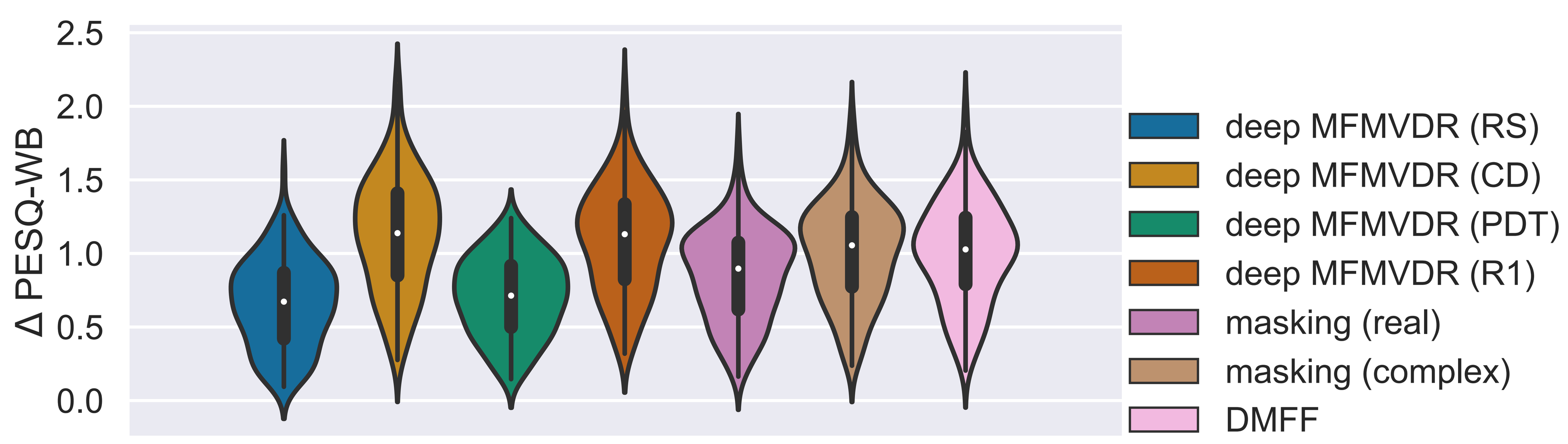
y_t → DNN → $\mathbf{a}_t^{\Re}, \mathbf{a}_t^{\Im} \in \mathbb{R}^N$ →

$$\widehat{\boldsymbol{\Phi}}_t^{\text{R1}} = \left(\mathbf{a}_t^{\Re} + j\mathbf{a}_t^{\Im}\right)\left(\mathbf{a}_t^{\Re} + j\mathbf{a}_t^{\Im}\right)^H$$

- imposes **dominant principal subspace**
- circumvent explicit matrix inversion → **lower computational complexity**

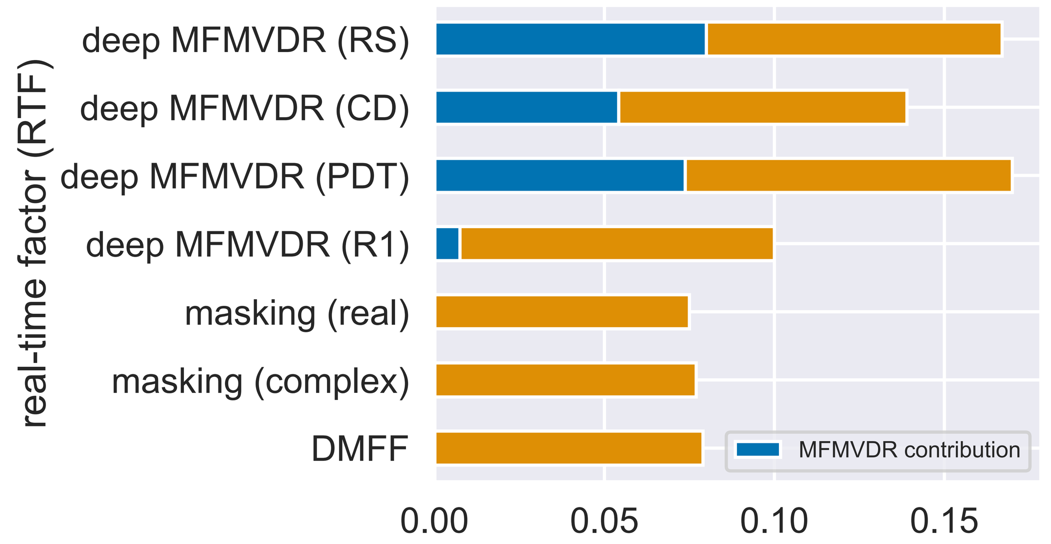
Results

1. Speech Enhancement Performance



- deep MFMVDR employing positive semi-definite matrix structure (CD) and rank-1 matrix structure (R1) yield highest performance
- **baseline algorithms are outperformed**: direct estimation of real-valued mask, complex-valued mask, or multi-frame filter (DMFF) [4]
- recursive smoothing (RS) and positive semi-definite Toeplitz structure (PDT) yield much worse performance

2. Computational Complexity



- **RTF**: ratio between processing time and signal duration
- **all RTFs < 1**
- deep MFMVDR filters more complex than baseline algorithms, primarily due to additional linear algebra operations in MFMVDR filter

rank-1 matrix structure yields good trade-off between speech enhancement / complexity

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] C. K. A. Reddy et al., “The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results,” in Proc. Interspeech, Shanghai, China, Oct. 2020, pp. 2492–2496.
[3] 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.
[4] W. Mack and E. A. P. Habets, “Deep Filtering: Signal Extraction and Reconstruction Using Complex Time-Frequency Filters,” IEEE Signal Processing Letters, vol. 27, pp. 61–65, Nov. 2020.
[5] 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, vol. 31, pp. 3237-3248, 2023.

Acknowledgement: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2177/1 - Project ID 390895286.