Poster ID #5066

Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement

Marvin Tammen, Simon Doclo

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

IEEE ICASSP 2021

Hearing4all

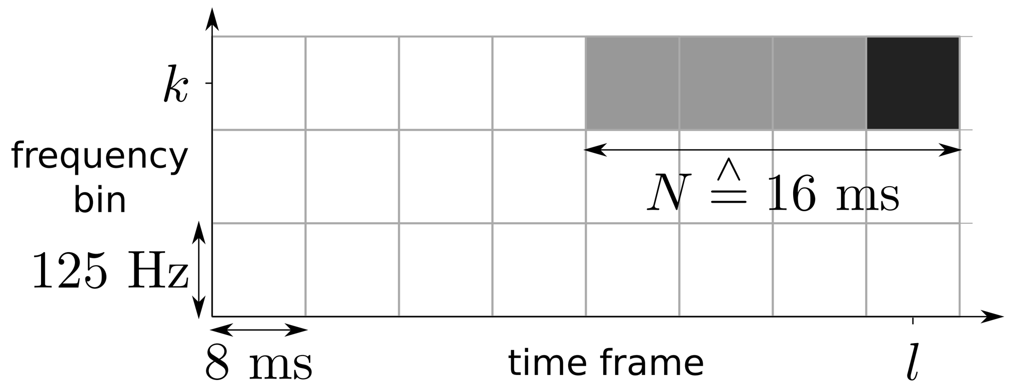
Carl von Ossietzky Universität Oldenburg

Problem Statement

- **microphone signal degraded by** (potentially highly time-varying) **ambient noise**
- multi-frame minimum-variance-distortionless-response **(MFMVDR) filter can yield good noise reduction and low speech distortions**
- **MFMVDR filter requires accurate estimates** of interference covariance matrix and speech interframe correlation (IFC) vector
- **embed MFMVDR filter within deep learning framework**

Signal Model

- noisy STFT coefficients: $Y_l = X_l + N_l$
- **multi-frame vector**: $\mathbf{y}_l = [Y_l \; Y_{l-1} \; \dots \; Y_{l-N+1}]^\top$



- assumptions:
  1. **independent speech and noise** components:
  $$\boldsymbol{\Phi}_{\mathbf{y},l} = \mathcal{E}\{\mathbf{y}_l\mathbf{y}_l^H\} = \boldsymbol{\Phi}_{\mathbf{x},l} + \boldsymbol{\Phi}_{\mathbf{n},l} \in \mathbb{C}^{N\times N}$$
  2. **decompose speech** into **correlated and uncorrelated component**:
  $$\mathbf{x}_l = \boldsymbol{\gamma}_{\mathbf{x},l} X_l + \mathbf{x}'_l, \quad \boldsymbol{\gamma}_{\mathbf{x},l} = \frac{\mathcal{E}\{\mathbf{x}_l X_l^*\}}{\mathcal{E}\{|X_l|^2\}} = \frac{\boldsymbol{\Phi}_{\mathbf{x},l}\mathbf{e}}{\mathbf{e}^\top\boldsymbol{\Phi}_{\mathbf{x},l}\mathbf{e}}$$
  3. **uncorrelated speech** $\triangleq$ **interference**: $\mathbf{u}_l := \mathbf{n}_l + \mathbf{x}'_l$
  $$\Rightarrow \boldsymbol{\Phi}_{\mathbf{y},l} = \phi_{X,l}\boldsymbol{\gamma}_{\mathbf{x},l}\boldsymbol{\gamma}_{\mathbf{x},l}^H + \boldsymbol{\Phi}_{\mathbf{u},l}$$

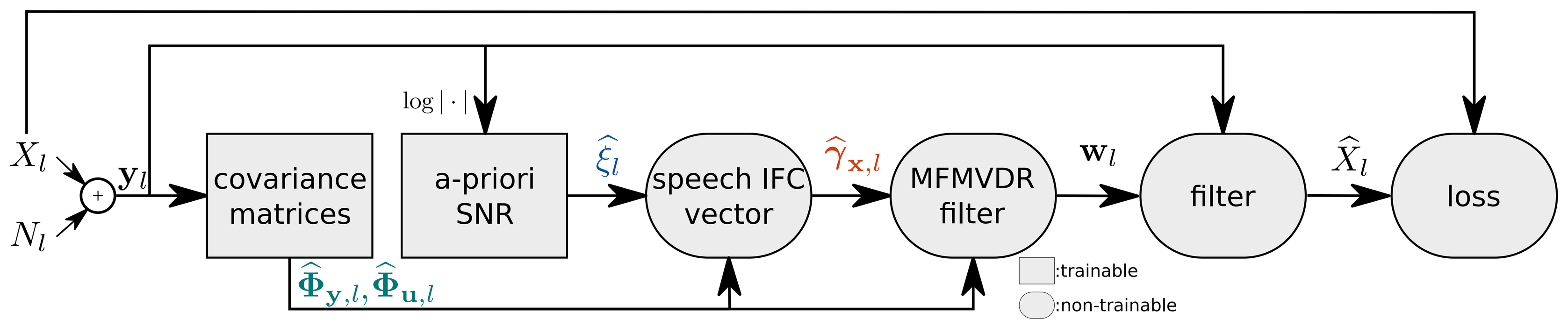
Multi-Frame MVDR Filter

- **minimize output interference power** while **preserving correlated speech component**
$$\underset{\mathbf{w}\in\mathbb{C}^N}{\text{argmin}} \; \mathbf{w}^H\boldsymbol{\Phi}_{\mathbf{u},l}\mathbf{w}, \quad \text{s.t.} \quad \mathbf{w}^H\boldsymbol{\gamma}_{\mathbf{x},l} = 1$$
- solved by MFMVDR filter
$$\boxed{\mathbf{w}_l = \frac{\boldsymbol{\Phi}_{\mathbf{u},l}^{-1}\boldsymbol{\gamma}_{\mathbf{x},l}}{\boldsymbol{\gamma}_{\mathbf{x},l}^H\boldsymbol{\Phi}_{\mathbf{u},l}^{-1}\boldsymbol{\gamma}_{\mathbf{x},l}}}$$
- **requires** estimates of highly time-varying **interference covariance matrix** $\boldsymbol{\Phi}_{\mathbf{u},l}$ **and speech interframe correlation (IFC) vector** $\boldsymbol{\gamma}_{\mathbf{x},l}$

Speech IFC Vector

- linear combination of noisy and interference IFC vectors
$$\boldsymbol{\gamma}_{\mathbf{x},l} = \frac{1+\xi_l}{\xi_l}\boldsymbol{\gamma}_{\mathbf{y},l} - \frac{1}{\xi_l}\boldsymbol{\gamma}_{\mathbf{u},l} = \frac{1+\xi_l}{\xi_l}\frac{\boldsymbol{\Phi}_{\mathbf{y},l}\mathbf{e}}{\mathbf{e}^\top\boldsymbol{\Phi}_{\mathbf{y},l}\mathbf{e}} - \frac{1}{\xi_l}\frac{\boldsymbol{\Phi}_{\mathbf{u},l}\mathbf{e}}{\mathbf{e}^\top\boldsymbol{\Phi}_{\mathbf{u},l}\mathbf{e}}$$
with a-priori SNR $\xi_l = \frac{\mathbf{e}^\top\boldsymbol{\Phi}_{\mathbf{x},l}\mathbf{e}}{\mathbf{e}^\top\boldsymbol{\Phi}_{\mathbf{u},l}\mathbf{e}}$
- $\boldsymbol{\gamma}_{\mathbf{y},l}$ and $\boldsymbol{\gamma}_{\mathbf{u},l}$ can be assumed to be less time-varying than $\boldsymbol{\gamma}_{\mathbf{x},l}$
  $\Rightarrow$ **estimate** ξ_l, $\boldsymbol{\Phi}_{\mathbf{y},l}$, and $\boldsymbol{\Phi}_{\mathbf{u},l}$

Deep MFMVDR Filter



- **supervised learning-based approach** to estimate ξ_l, $\boldsymbol{\Phi}_{\mathbf{y},l}$, and $\boldsymbol{\Phi}_{\mathbf{u},l}$
1. **covariance matrices** $\widehat{\boldsymbol{\Phi}}_{\mathbf{y},l}$ and $\widehat{\boldsymbol{\Phi}}_{\mathbf{u},l}$:
   - Hermitian positive-semidefinite matrices
     $\Rightarrow N^2$ real-valued coefficients $\boldsymbol{h}_{\mathbf{y},l}$ and $\boldsymbol{h}_{\mathbf{u},l}$
   - DNN inputs: concatenated real and imaginary STFT components
   - DNN outputs: coefficients $\boldsymbol{h}_{\mathbf{y},l}$ and $\boldsymbol{h}_{\mathbf{u},l}$; linear activation
   - construct matrices:
   $$\widehat{\boldsymbol{\Phi}}_{\mathbf{y},l} = \boldsymbol{H}_{\mathbf{y},l}\boldsymbol{H}_{\mathbf{y},l}^H, \quad \boldsymbol{H}_{\mathbf{y},l} = \text{Hermitian}\{\boldsymbol{h}_{\mathbf{y},l}\}$$
   $$\widehat{\boldsymbol{\Phi}}_{\mathbf{u},l} = \boldsymbol{H}_{\mathbf{u},l}\boldsymbol{H}_{\mathbf{u},l}^H, \quad \boldsymbol{H}_{\mathbf{u},l} = \text{Hermitian}\{\boldsymbol{h}_{\mathbf{u},l}\}$$
2. **a-priori SNR estimate** $\widehat{\xi}_l$
   - DNN inputs: logarithm of noisy STFT magnitude
   - DNN output: $\widehat{\xi}_l$; softplus activation to ensure $\widehat{\xi}_l > 0$
- all DNNs are **trained with speech enhancement-related loss**
  $\Rightarrow$ **no target** covariance matrices or a-priori SNRs **required**

Baseline Algorithms

1. **complex-valued direct filtering** (multi-frame)
2. **complex-valued masking** (single-frame)
3. **real-valued masking** (single-frame)
4. **ConvTasNet** [2] (causal implementation)
- all compared algorithms with **same architecture and similar number of parameters** ($\approx$ 5 M)

Simulations – Dataset

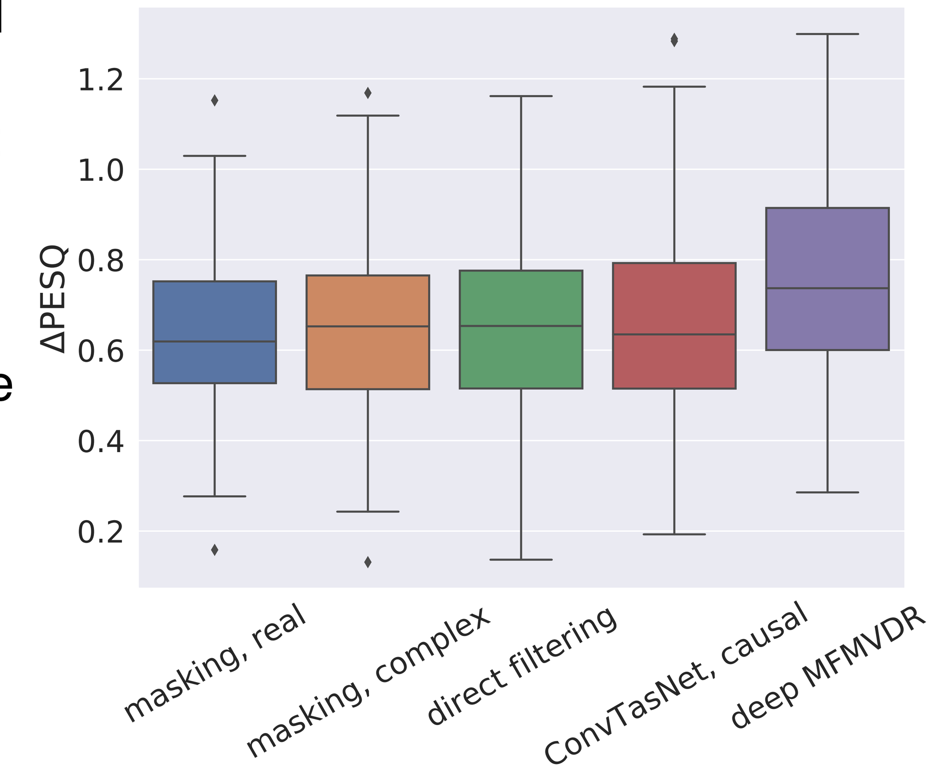
- based on **DNS challenge dataset** [3]
- training and validation
  - clean: from Librivox (anechoic)
  - noise: from Audioset, Freesound, and DEMAND
  - SNR $\in$ [0, 20) dB
  - 4 s utterances, **dataset size 50 h**
- testing
  - clean: from U Graz dataset (anechoic)
  - noise: 15 clips each from 12 classes, Freesound
  - SNR $\in$ [0, 25) dB
- disjoint training, validation, and test sets

Simulations – Settings

- STFT: 8 ms frame length, 2 ms shift, Hann window
- multi-frame algorithms: $N = 5$ frames (16 ms temporal context)
- deep MFMVDR: diagonal loading applied to estimated covariance matrices with constant 10^{-3}
- **temporal convolutional network architecture** [4]; hidden dimension size varied to obtain $\approx$ 5 M parameters per algorithm
- time-domain **scale-invariant signal-to-distortion ratio (SI-SDR) loss**
- Adam optimizer with learning rate $3 * 10^{-4}$ and scheduling, batch size 6, max. 50 epochs, early stopping, gradient norm clipped to 5

Simulations – Results

- all compared algorithms yield high PESQ improvement
- **deep MFMVDR** with **highest performance**
- **complex masking** slightly **better than real masking**
- complex masking comparable to direct filtering
- **deep MFMVDR better than direct filtering**
- STOI improvement shows similar tendencies



Guiding multi-frame **filter estimation by relying on MFMVDR structure** is **beneficial** compared to estimating the filter directly.

[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*, 20(4):1256–1269, May 2012.

[2] Y. Luo and N. Mesgarani. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 27(8):1256–1266, August 2019.

[3] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke. The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results. *arXiv:2005.13981 [cs, eess]*, May 2020.

[4] S. Bai, J. Z. Kolter, and V. Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv:1803.01271 [cs]*, March 2018.