Joint Multi-Channel Dereverberation and Noise Reduction Using a Unified Convolutional Beamformer With Sparse Priors

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

Introduction

- **Reverberation** and **noise degrades speech quality**.
- The weighted power minimization distortionless response **(WPD) beamformer unifies** weighted prediction error **(WPE) dereverberation** and the minimum power distortionless response **(MPDR) beamformer** [1, 2].
- In the STFT-domain reverberation and noise lead to a **less sparse representation** of a signal compared to its clean version.

IN THIS POSTER

- Similar as in [3] for WPE we introduce a shape parameter p to **control the sparsity** of the cost function for WPD.
- Additionally we investigate the effect of **single- and multi-channel initialization** of the iterative optimization.

Convolutional Beamformer

Convolutional Signal Model (multi-frame) in STFT-domain

$$\mathbf{y}_t = \sum_{l=0}^{L_a-1} \mathbf{a}_l s_{t-l} + \mathbf{n}_t = \underbrace{\sum_{l=0}^{\tau-1} \mathbf{a}_l s_{t-l}}_{\text{direct/early } \mathbf{d}_t} + \underbrace{\sum_{l=\tau}^{L_a-1} \mathbf{a}_l s_{t-l}}_{\text{late reverb } \mathbf{r}_t} + \underbrace{\mathbf{n}_t}_{\text{noise}}$$

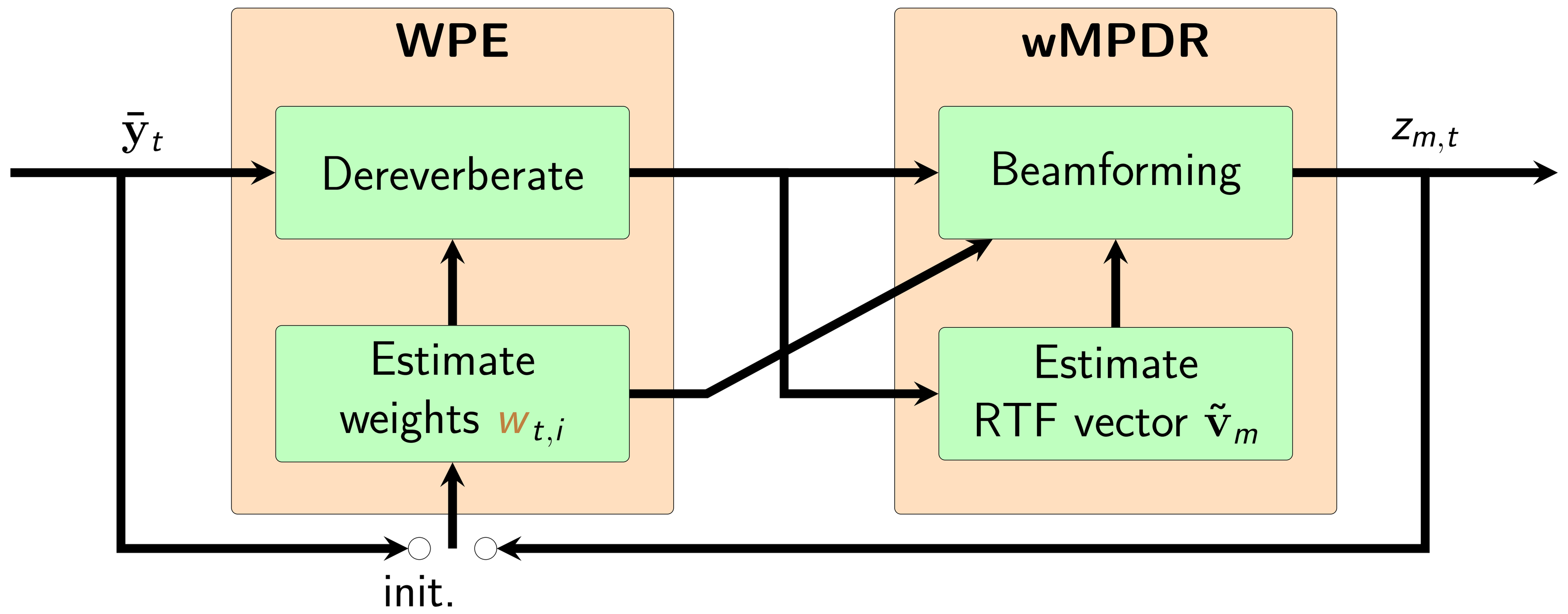
- Prediction delay τ separates early reflections and late reverberation
- Desired component $\mathbf{d}_t$ can be approximated by: $\mathbf{d}_t \approx \mathbf{v} s_t = \tilde{\mathbf{v}}_m d_{m,t}$ with the **relative transfer function (RTF)** vector $\tilde{\mathbf{v}}_m$

Convolutional Filter (multi-frame) $\bar{\mathbf{h}}_m$

$$z_{m,t} = \bar{\mathbf{h}}_m^{\mathrm{H}} \bar{\mathbf{y}}_t \quad \text{with} \quad \bar{\mathbf{y}}_t = \left[\mathbf{y}_t^{\mathrm{T}} \mid \mathbf{y}_{t-\tau}^{\mathrm{T}} \cdots \mathbf{y}_{t-L_h+1}^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{M(L_h-\tau+1)\times 1}$$

- Stacked signal vector $\bar{\mathbf{y}}_t$ contains the current frame and the past frames corresponding to late reverberation $\rightarrow$ **Gap of $\tau - 1$ frames** to keep early reflections

Iterative WPD algorithm unifying WPE dereverberation and MPDR beamformer [2]



Filter Optimization

Distortionless Constraint

Distortionless constraint for desired component using RTF vector

$$\bar{\mathbf{h}}_m^{\mathrm{H}} \bar{\mathbf{v}}_m = 1 \quad \text{with} \quad \bar{\mathbf{v}}_m = \left[\tilde{\mathbf{v}}_m^{\mathrm{T}} \ \mathbf{0}^{\mathrm{T}}\right]^{\mathrm{T}}$$

- **Conventional WPD** uses time-varying circular Gaussian model for STFT coefficients as objective function
- **Proposed generalization**: ℓ_p-norm reformulation of the generalized Gaussian prior from [3], which introduces a shape parameter p.

Conventional Objective Function

Time-varying Gaussian model

- Variances $\lambda_t = \mathbb{E}\left[|d_{m,t}|^2\right]$
- $\boldsymbol{\Lambda} = \mathrm{diag}\left(\left[\lambda_1 \cdots \lambda_T\right]^{\mathrm{T}}\right)$

$$\mathcal{L}\left(\bar{\mathbf{h}}_m, \boldsymbol{\Lambda}\right) \propto \left(\mathrm{tr}\left(\ln \boldsymbol{\Lambda}\right) + \mathbf{z}_m \boldsymbol{\Lambda}^{-1} \mathbf{z}_m^{\mathrm{H}}\right)$$

$\rightarrow$ iterative optimization scheme

Proposed Objective Function

ℓ_p-norm cost function

$$\mathcal{L}\left(\bar{\mathbf{h}}_m\right) \propto \|\mathbf{z}_m\|_p^p \propto \sum_{t=1}^{T} |z_{m,t}|^p$$

$\rightarrow$ iterative reweighted optimization scheme using ℓ_2-norm

$$\mathcal{L}\left(\bar{\mathbf{h}}_{m,i}, \mathbf{W}_i\right) \propto \sum_{t=1}^{T} w_{t,i} |z_{m,t}|^2 \propto \mathbf{z}_{m,i} \mathbf{W}_i \mathbf{z}_{m,i}^{\mathrm{H}}$$

Iterative Algorithms

1. Optimize and update the beamformer $\bar{\mathbf{h}}_{m,i}$ (iteration index i)
2. Optimize and update the variances $\boldsymbol{\Lambda}_i$ and weights $\mathbf{W}_i$

Conventional

$$\bar{\mathbf{h}}_{m,i}^{\mathrm{opt}} = \frac{\bar{\mathbf{R}}_{y,i}^{-1} \bar{\mathbf{v}}_m}{\bar{\mathbf{v}}_m^{\mathrm{H}} \bar{\mathbf{R}}_{y,i}^{-1} \bar{\mathbf{v}}_m}$$

with noisy power-weighted multi-frame covariance matrix

$$\bar{\mathbf{R}}_{y,i} = {}^{1}/_{T} \bar{\mathbf{Y}} \boldsymbol{\Lambda}_i^{-1} \bar{\mathbf{Y}}^{\mathrm{H}}$$

Variance update

$$\lambda_{t,i+1} = |z_{m,t,i}|^2 = \left|\bar{\mathbf{h}}_{m,i}^{\mathrm{opt,H}} \bar{\mathbf{y}}_t\right|^2$$

Proposed

$$\bar{\mathbf{h}}_{m,i}^{\mathrm{opt}} = \frac{\left(\bar{\mathbf{R}}_{y,i}^{\mathbf{W}}\right)^{-1} \bar{\mathbf{v}}_m}{\bar{\mathbf{v}}_m^{\mathrm{H}} \left(\bar{\mathbf{R}}_{y,i}^{\mathbf{W}}\right)^{-1} \bar{\mathbf{v}}_m}$$

with noisy **reweighted** multi-frame covariance matrix

$$\bar{\mathbf{R}}_{y,i}^{\mathbf{W}} = {}^{1}/_{T} \bar{\mathbf{Y}} \mathbf{W}_i \bar{\mathbf{Y}}^{\mathrm{H}}$$

Weight update

$$w_{t,i+1}^{-1} = |z_{m,t,i}|^{2-p} = \left|\bar{\mathbf{h}}_{m,i}^{\mathrm{opt,H}} \bar{\mathbf{y}}_t\right|^{2-p}$$

$\longrightarrow$ **Proposed method equals conventional method for $p = 0$**

Initialization of Variances/Weights

No output signal power available in first iteration $\rightarrow$ Use input signal power

Single-channel initialization: $\lambda_{t,1} = |y_{m,t}|^2$ and $w_{t,1}^{-1} = |y_{m,t}|^{2-p}$

Multi-channel initialization: $\lambda_{t,1} = \frac{\|\mathbf{y}_t\|_2^2}{M}$ and $w_{t,1}^{-1} = \frac{\|\mathbf{y}_t\|_2^{2-p}}{M}$

Experimental Evaluation

Reverb Challenge development dataset [4]

- Reverberation time $T_{60} \in \{0.3\,\mathrm{s}, 0.6\,\mathrm{s}, 0.7\,\mathrm{s}\}$, SNR of about 20 dB
- Circular array with 8 microphone channels (speaker-to-mic distance 50 cm or 200 cm)

Algorithm parameters

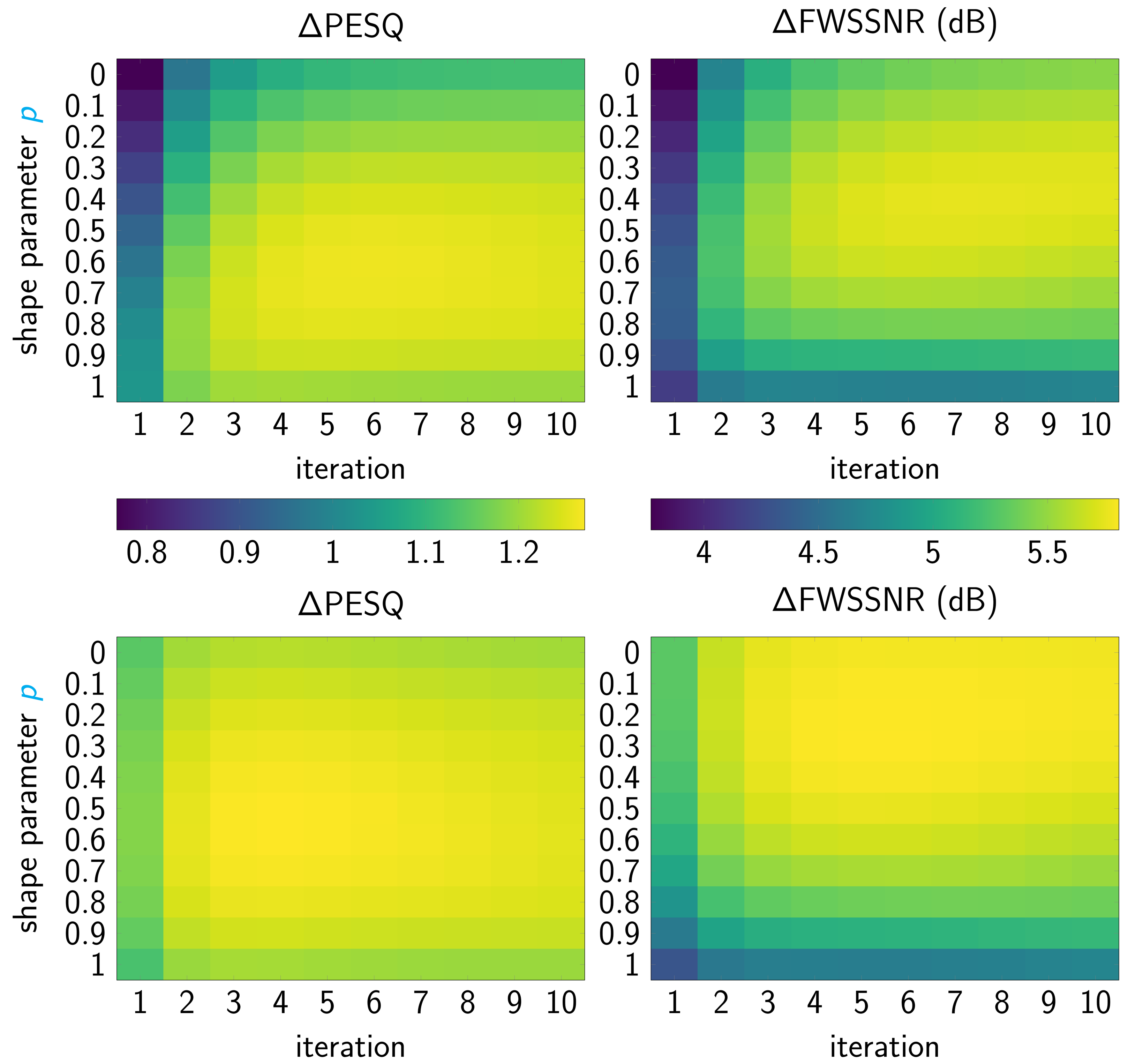
- STFT: 32 ms frame length with 25% overlap and square-root-Hann window
- Prediction delay $\tau = 4$ and prediction filter length $L_h = 12$

RTF estimation

- RTF vector $\tilde{\mathbf{v}}_m$ is estimated using covariance whitening
- Assuming only-noise period in the first 225 ms $\rightarrow$ noise covariance matrix

Results

Influence of shape parameter p, initialization and number of iterations



- ✓ **Multi-channel initialization outperforms single-channel initialization (higher performance and faster convergence)**
- ✓ **Proposed beamformer with sparse priors outperforms conventional WPD beamformer (only slightly if multi-channel initialization is used)**
- ✓ **We extended the proposed method towards a weighted LCMP beamformer and participated in the Clarity Challenge** [5, 6]

References

[1] T. Nakatani and K. Kinoshita, "A Unified Convolutional Beamformer for Simultaneous Denoising and Dereverberation," IEEE Signal Processing Letters, vol. 26, no. 6, pp. 903–907, Jun. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8691481/

[2] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly Optimal Dereverberation and Beamforming," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, May 2020, pp. 216–220. [Online]. Available: https://ieeexplore.ieee.org/document/9054393/

[3] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-Channel Linear Prediction-Based Speech Dereverberation With Sparse Priors," IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 23, no. 9, pp. 1509–1520, Sep. 2015.

[4] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz NY, USA, Oct. 2013, pp. 1–4.

[5] A. Aroudi, M. Delcroix, T. Nakatani, K. Kinoshita, S. Araki, and S. Doclo, "Cognitive-Driven Convolutional Beamforming Using EEG-Based Auditory Attention Decoding," in Proc. IEEE International Workshop on Machine Learning for Signal Processing. Espoo, Finland: IEEE, Sep. 2020, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9231657/

[6] S. Graetzer and T. Cox, "Clarity challenge."