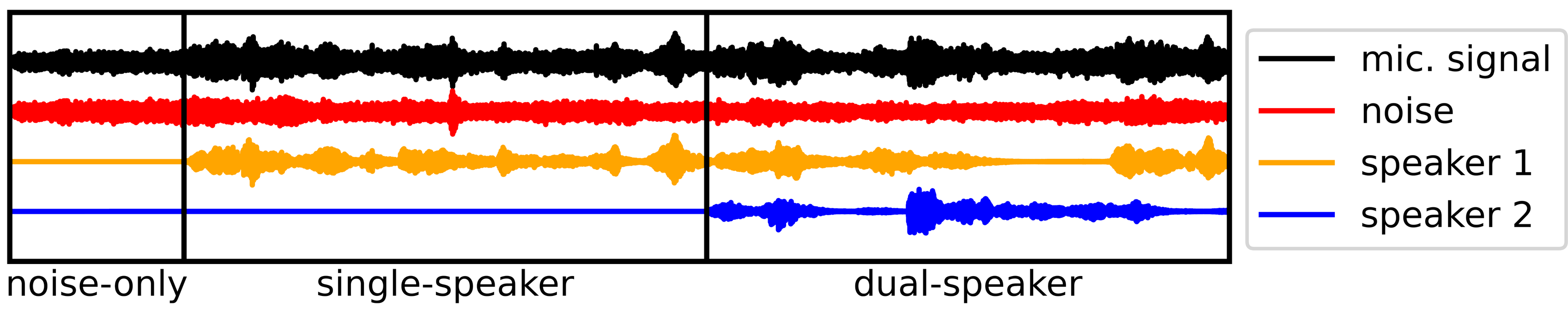
Covariance Blocking and Whitening Method for Successive Relative Transfer Function Vector Estimation in Multi-Speaker Scenarios

Henri Gode, Simon Doclo

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

Introduction

- Common multi-microphone noise reduction methods, e.g., linearly constrained minimum variance (LCMV) beamforming, rely on estimates of the relative transfer function (RTF) vectors of all speakers
- In this work: acoustic scenario with two successively activating speakers



- **Objective: Estimate RTF vector of second speaker during overlapping speech segments**

MAIN IDEAS

Covariance Blocking and Whitening (CBW) method for estimating the RTF vector of the second speaker

1. **Block the initial speaker** to isolate information about the second speaker
2. **Whiten the noise** to minimize its influence

Problem Statement

Signal Model in STFT-domain

- Two speakers and noise recorded with M microphones
- Noisy covariance matrix

$$\mathbf{R}_\mathbf{y} = \overbrace{\mathbf{h}\phi_x\mathbf{h}^\mathrm{H}}^{\text{speaker 2}} + \underbrace{\overbrace{\mathbf{g}\phi_u\mathbf{g}^\mathrm{H}}^{\text{speaker 1}} + \mathbf{R}_\mathbf{n}}_{\mathbf{R}_\mathbf{v}} \in \mathbb{C}^{M\times M}$$

- Noise covariance matrix $\mathbf{R}_\mathbf{n}$ and RTF vector $\mathbf{g}$ of speaker 1 can be estimated in noise-only and single-speaker segments
- Power spectral densities ϕ_u and ϕ_x of are unknown and time-varying

GOAL

Estimate RTF vector $\mathbf{h}$ of speaker 2 in dual-speaker segment using noisy covariance matrix $\mathbf{R}_\mathbf{y}$ and estimates of $\mathbf{R}_\mathbf{n}$ and $\mathbf{g}$

Conventional Methods

1.) Covariance Whitening (CWu) [1]

- **Jointly whiten speaker 1 and noise** with undesired covariance matrix $\mathbf{R}_\mathbf{v}$

$$\widehat{\mathbf{h}}^{(\mathrm{CW})} = \tilde{\mathbf{h}}/\mathbf{e}_r^\mathrm{T}\tilde{\mathbf{h}} \quad \text{with} \quad \tilde{\mathbf{h}} = \mathbf{R}_\mathbf{v}^{\mathrm{H}/2}\mathcal{P}\{\mathbf{R}_\mathbf{v}^{-\mathrm{H}/2}\mathbf{R}_\mathbf{y}\mathbf{R}_\mathbf{v}^{-1/2}\}$$

with $\mathcal{P}\{.\}$ denoting principal eigenvector

2.) Blind Oblique Projection (BOP) [2]

- **Noise is neglected** by assuming a sufficiently high SNR
- **Block speaker 2** using parameterized oblique projection matrix $\mathbf{P}_{\mathbf{g}\boldsymbol{\theta}}^{\angle}$ while **keeping speaker 1 distortionless** and minimizing the power

$$\widehat{\mathbf{h}}^{(\mathrm{BOP})} = \tilde{\mathbf{h}}/\mathbf{e}_r^\mathrm{T}\tilde{\mathbf{h}} \quad \text{with} \quad \tilde{\mathbf{h}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left(\mathrm{Tr}\{\mathbf{P}_{\mathbf{g}\boldsymbol{\theta}}^{\angle}\mathbf{R}_\mathbf{y}\mathbf{P}_{\mathbf{g}\boldsymbol{\theta}}^{\angle\mathrm{H}}\}\right) \quad \text{with} \quad \mathbf{P}_{\mathbf{g}\boldsymbol{\theta}}^{\angle} = \mathbf{g}\left(\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{g}\right)^{+}$$

Proposed Method

Covariance Blocking and Whitening (CBW)

1.) Block speaker 1 using orthogonal projection matrix $\mathbf{P}_\mathbf{g}^{\perp} = \mathbf{I}_M - \frac{\mathbf{g}\mathbf{g}^\mathrm{H}}{(\mathbf{g}^\mathrm{H}\mathbf{g})}$

- Noise whitening requires full column rank → remove one column

$$\mathbf{R}_\mathbf{y}\mathbf{P}_{\mathbf{g},\mathrm{r}}^{\perp} = \mathbf{h}\phi_x\mathbf{h}^\mathrm{H}\mathbf{P}_{\mathbf{g},\mathrm{r}}^{\perp} + \mathbf{R}_\mathbf{n}\mathbf{P}_{\mathbf{g},\mathrm{r}}^{\perp} \in \mathbb{C}^{M\times M-1}$$

2.) Whiten the noise using pseudo-inverse of blocked noise covariance matrix

$$\left(\mathbf{R}_\mathbf{n}\mathbf{P}_{\mathbf{g},\mathrm{r}}^{\perp}\right)^{+}\mathbf{R}_\mathbf{y}\mathbf{P}_{\mathbf{g},\mathrm{r}}^{\perp} - \mathbf{I}_{M-1} = \underbrace{\left(\mathbf{R}_\mathbf{n}\mathbf{P}_{\mathbf{g},\mathrm{r}}^{\perp}\right)^{+}\mathbf{h}}_{\propto\,\mathbf{q}_\mathrm{L}}\,\phi_x\,\underbrace{\mathbf{h}^\mathrm{H}\mathbf{P}_{\mathbf{g},\mathrm{r}}^{\perp}}_{\propto\,\mathbf{q}_\mathrm{R}^\mathrm{H}} \in \mathbb{C}^{M-1\times M-1}$$

Covariance Blocking and Whitening (CBW)

3.) Set up non-linear equation system using left & right principal singular vectors $\mathbf{q}_\mathrm{L}$ and $\mathbf{q}_\mathrm{R}$ and unknown scaling factor α

$$\begin{bmatrix}\mathbf{q}_\mathrm{L}\\ \mathbf{q}_\mathrm{R}\alpha\end{bmatrix} \overset{!}{=} \mathbf{B}\tilde{\mathbf{h}} \quad \text{with} \quad \mathbf{B} = \begin{bmatrix}\left(\mathbf{R}_\mathbf{n}\mathbf{P}_{\mathbf{g},\mathrm{r}}^{\perp}\right)^{+}\\ \left(\mathbf{P}_{\mathbf{g},\mathrm{r}}^{\perp}\right)^{\mathrm{H}}\end{bmatrix}$$

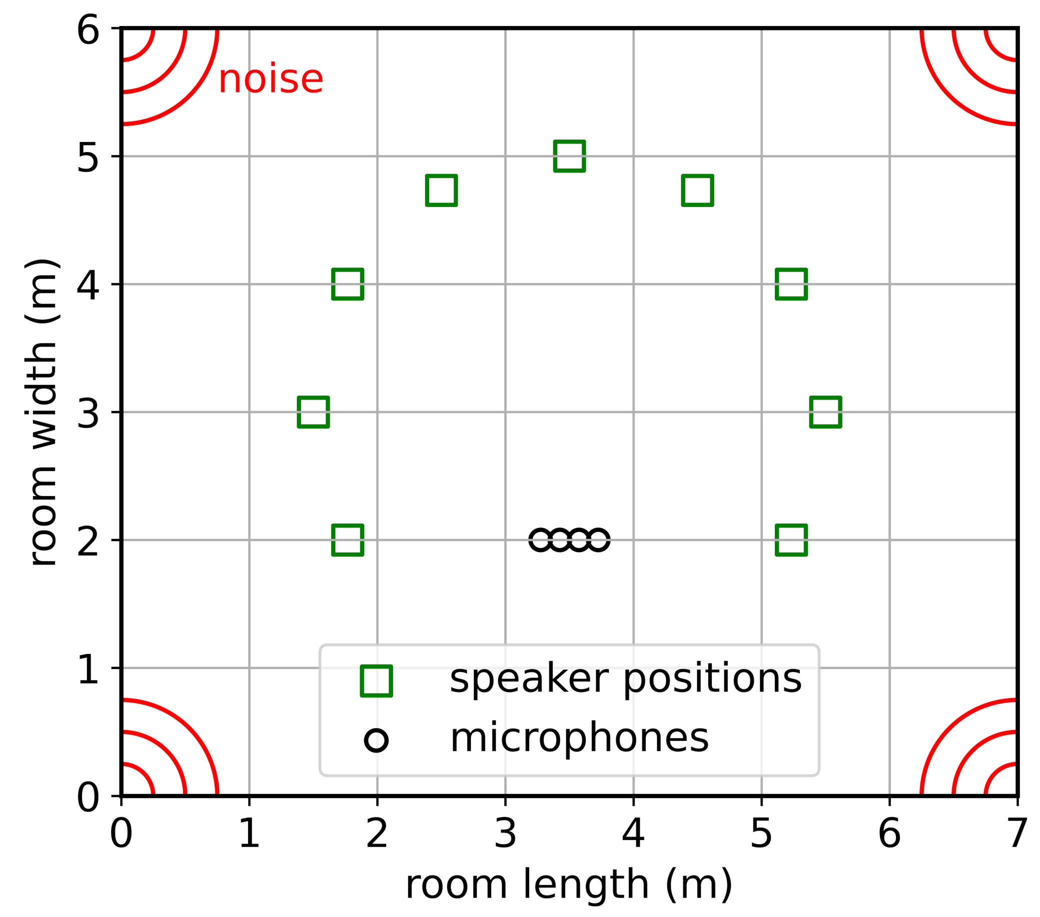
$$\boxed{\widehat{\mathbf{h}}^{(\mathrm{CBW})} = \tilde{\mathbf{h}}/\mathbf{e}_r^\mathrm{T}\tilde{\mathbf{h}} \quad \text{with} \quad \tilde{\mathbf{h}} = \mathbf{B}^{+}\begin{bmatrix}\mathbf{q}_\mathrm{L}\\ -\mathbf{q}_\mathrm{R}\left(\mathbf{P}_\mathbf{B}^{\perp,\mathrm{R}}\mathbf{q}_\mathrm{R}\right)^{+}\mathbf{P}_\mathbf{B}^{\perp,\mathrm{L}}\mathbf{q}_\mathrm{L}\end{bmatrix}}$$

Method Overview

	CWu [1]	BOP [2]	CBW (prop.)
1.) Blocking	X	speaker 2	speaker 1
2.) Whitening	speaker 1 & noise	X	noise
required estimates	$\mathbf{R}_\mathbf{v}$	$\mathbf{g}$	$\mathbf{R}_\mathbf{n}$ & $\mathbf{g}$

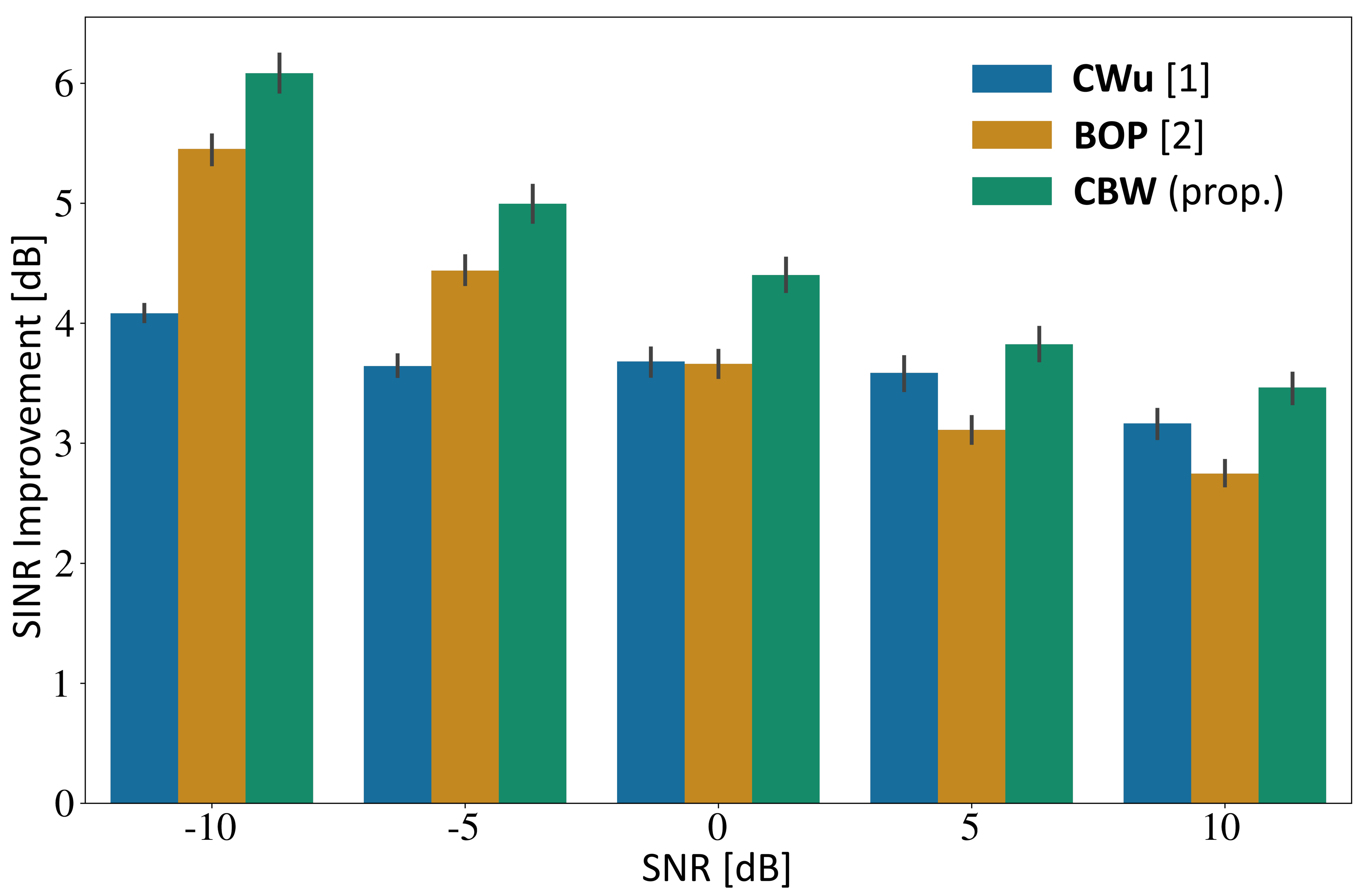
Evaluation

- Linear array with $M = 4$ microphones ($d = 2$ cm)
- Clean signals convolved with measured room impulse responses ($T_{60} \approx 500$ ms)
- 72 combinations of dual-speaker positions
- Quasi-diffuse babble noise with Signal-to-noise ratio (SNR): $-10:5:10$ dB
- Signal-to-interferer ratio (SIR): $-10:5:10$ dB
- $f_s = 16$ kHz
- STFT framework: frame length 200 ms, 75% overlap
- segment borders are assumed to be known



Results

- Signal-to-interferer-and-noise ratio (SINR) improvement of LCMV beamformer using estimated noise covariance matrix and RTF vectors of both speakers



Conclusions

- ✓ **The proposed CBW method combines blocking of the initial speaker and whitening of the noise to estimate the RTF vector of the second speaker**
- ✓ **In terms of SINR improvement, the proposed CBW method outperforms conventional RTF vector estimation methods**

References

[1] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," IEEE Trans. Audio, Speech, and Language Processing, vol. 15, no. 5, pp. 1529–1539, 2007.

[2] D. Cherkassky and S. Gannot, "Successive Relative Transfer Function Identification Using Blind Oblique Projection," IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 28, pp. 474–486, 2020.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 390895286 – EXC 2177/1