Joint Estimation of RETF Vector and PSDs for Speech Enhancement Based on Alternating Least Squares

Marvin Tammen* Ina Kodrasi† Simon Doclo*

* University of Oldenburg, Dept. Medical Physics and Acoustics and Cluster of Excellence Hearing4All, Germany
† Idiap Research Institute, Speech and Audio Processing Group, Switzerland

Problem Statement

- microphone signals corrupted by **reverberation, diffuse noise + residual noise**, e.g., sensor noise
- **multichannel Wiener filter** (MWF) requires estimates of **relative early transfer functions** (RETFs) of target source and **target, diffuse and residual power spectral densities** (PSDs)
- goal: develop an **online joint estimator for RETFs and noise PSDs**

Signal Model

- signal model in STFT-domain (frame index l):

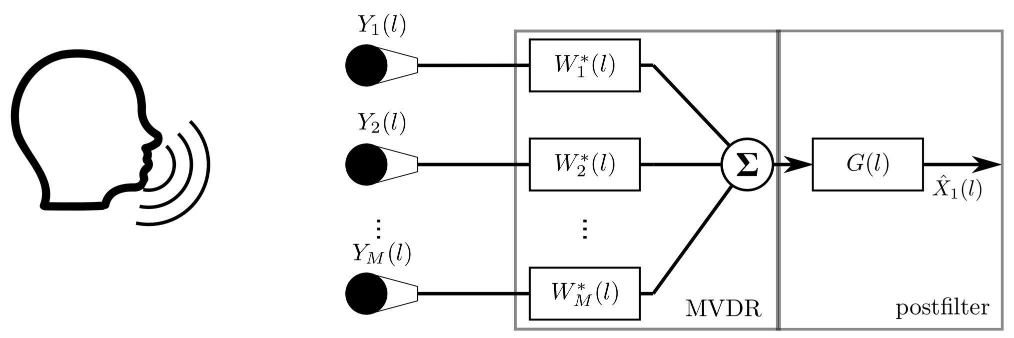
$$\mathbf{y}(l) = \mathbf{x}(l) + \mathbf{d}(l) + \mathbf{v}(l)$$

  - $\mathbf{x}(l)$: direct and early speech component (target)
  - $\mathbf{d}(l)$: diffuse noise and late reverberation
  - $\mathbf{v}(l)$: residual noise (e.g., sensor noise)
- assumptions:
  - $\mathbf{x}(l)$, $\mathbf{d}(l)$ and $\mathbf{v}(l)$ uncorrelated:
  - diffuse and residual noise: **stationary (known) spatial coherence and time-varying PSD**

$$\mathbf{\Phi_y}(l) = \mathbb{E}\{\mathbf{y}(l)\mathbf{y}^H(l)\} = \phi_x(l)\mathbf{a}(l)\mathbf{a}^H(l) + \phi_d(l)\mathbf{\Gamma} + \phi_v(l)\mathbf{\Psi}$$

- $\phi_x(l) = \mathcal{E}\{|X_1(l)|^2\}$: target PSD
- $\phi_d(l)$, $\phi_v(l)$: diffuse and residual PSD
- $\mathbf{a}(l)$: RETF vector (possibly time-varying)
- $\mathbf{\Gamma}$: diffuse coherence matrix, $\mathbf{\Psi}$: residual coherence matrix

Multichannel Wiener Filter



- target estimate: $\hat{X}_1(l) = G(l)\mathbf{w}^H_{\mathrm{MVDR}}(l)\mathbf{y}(l)$
- **requires** estimates of **RETFs** $\mathbf{a}(l)$ **and PSDs** $[\phi_x(l), \phi_d(l), \phi_v(l)]^T =: \boldsymbol{\phi}(l)$

$$\mathbf{w}_{\mathrm{MVDR}}(l) = \frac{[\phi_d(l)\mathbf{\Gamma} + \phi_v(l)\mathbf{\Psi}]^{-1}\mathbf{a}(l)}{\mathbf{a}^H(l)\,[\phi_d(l)\mathbf{\Gamma} + \phi_v(l)\mathbf{\Psi}]^{-1}\mathbf{a}(l)}$$

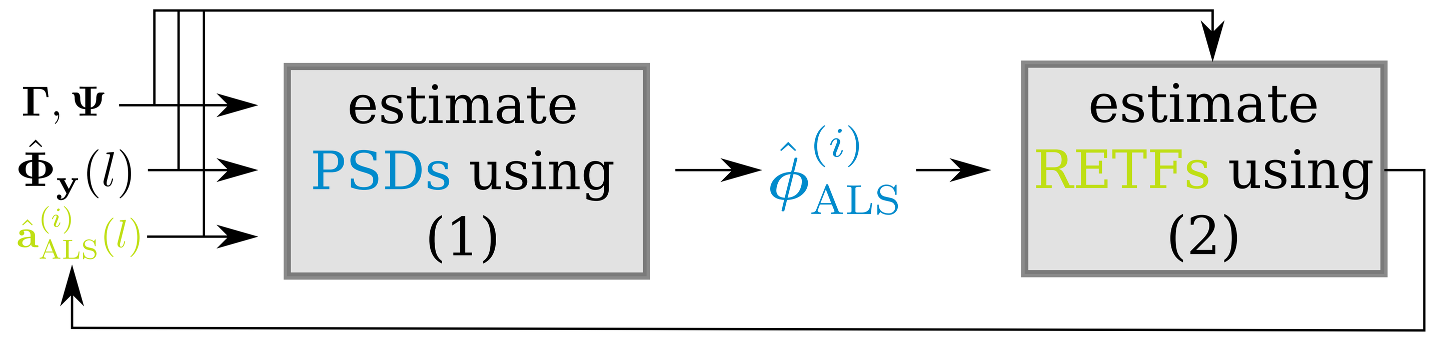
$$G(l) = \frac{\xi(l)}{1+\xi(l)}, \text{ with } \xi(l) = \frac{\phi_x(l)}{\mathbf{w}^H_{\mathrm{MVDR}}(l)\,[\phi_d(l)\mathbf{\Gamma} + \phi_v(l)\mathbf{\Psi}]\,\mathbf{w}_{\mathrm{MVDR}}(l)}$$

Proposed Method

- cost function to **jointly estimate PSDs and RETFs**:

$$\left(\hat{\mathbf{a}}_{\mathrm{ALS}}, \hat{\boldsymbol{\phi}}_{\mathrm{ALS}}\right) = \underset{\mathbf{a},\boldsymbol{\phi}}{\operatorname{argmin}} \left\| \hat{\mathbf{\Phi}}_{\mathbf{y}} - \left(\phi_x\mathbf{a}\mathbf{a}^H + \phi_d\mathbf{\Gamma} + \phi_v\mathbf{\Psi}\right)\right\|_F^2$$

- no closed-form solution ⇒ **alternating least-squares** (ALS) **approach**



(i) initialize RETFs $\hat{\mathbf{a}}^{(0)}_{\mathrm{ALS}}$ (e.g., using DOA); set $i=1$
(ii) estimate PSDs $\hat{\boldsymbol{\phi}}^{(i)}_{\mathrm{ALS}}$ using (1), assuming $\mathbf{a} = \hat{\mathbf{a}}^{(i-1)}_{\mathrm{ALS}}$
(iii) estimate RETFs $\hat{\mathbf{a}}^{(i)}_{\mathrm{ALS}}$ using (2), assuming $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}^{(i)}_{\mathrm{ALS}}$; set $i = i+1$
(iv) repeat steps (ii) — (iii) until convergence

PSDs and RETFs Estimation

- **PSDs**
  - assume fixed $\mathbf{a} = \hat{\mathbf{a}}^{(i-1)}_{\mathrm{ALS}}$
  - estimate PSDs as [1]

$$\hat{\boldsymbol{\phi}}^{(i)}_{\mathrm{ALS}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left\| \hat{\mathbf{\Phi}}_{\mathbf{y}} - \left(\hat{\phi}^{(i)}_{x,\mathrm{ALS}}\mathbf{a}\mathbf{a}^H + \hat{\phi}^{(i)}_{d,\mathrm{ALS}}\mathbf{\Gamma} + \hat{\phi}^{(i)}_{v,\mathrm{ALS}}\mathbf{\Psi}\right)\right\|_F^2 = \mathbf{A}^{-1}\mathbf{b},$$

$$\mathbf{A} = \begin{bmatrix} (\mathbf{a}^H\mathbf{a})^2 & \mathbf{a}^H\mathbf{\Gamma}\mathbf{a} & \mathbf{a}^H\mathbf{\Psi}\mathbf{a} \\ \mathbf{a}^H\mathbf{\Gamma}\mathbf{a} & \mathrm{trace}\{\mathbf{\Gamma}^H\mathbf{\Gamma}\} & \mathrm{trace}\{\mathbf{\Gamma}^H\mathbf{\Psi}\} \\ \mathbf{a}^H\mathbf{\Psi}\mathbf{a} & \mathrm{trace}\{\mathbf{\Psi}^H\mathbf{\Gamma}\} & \mathrm{trace}\{\mathbf{\Psi}^H\mathbf{\Psi}\} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathrm{Re}\left\{\mathbf{a}^H\hat{\mathbf{\Phi}}_{\mathbf{y}}\mathbf{a}\right\} \\ \mathrm{Re}\{\mathrm{trace}\{\hat{\mathbf{\Phi}}_{\mathbf{y}}\mathbf{\Gamma}^H\}\} \\ \mathrm{Re}\{\mathrm{trace}\{\hat{\mathbf{\Phi}}_{\mathbf{y}}\mathbf{\Psi}^H\}\} \end{bmatrix} \quad (1)$$

  - constrain PSDs to sensible range: $\mathrm{eps} \leq \{\hat{\phi}^{(i)}_{x,\mathrm{ALS}}, \hat{\phi}^{(i)}_{d,\mathrm{ALS}}, \hat{\phi}^{(i)}_{v,\mathrm{ALS}}\} \leq \frac{1}{M}\mathbf{y}^H\mathbf{y}$
- **RETFs**
  - assume fixed $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}^{(i)}_{\mathrm{ALS}}$
  - define $\hat{\mathbf{\Phi}}^{(i)}_{\mathbf{x}} = \hat{\mathbf{\Phi}}_{\mathbf{y}} - \left(\hat{\phi}^{(i)}_{d,\mathrm{ALS}}\mathbf{\Gamma} + \hat{\phi}^{(i)}_{v,\mathrm{ALS}}\mathbf{\Psi}\right)$

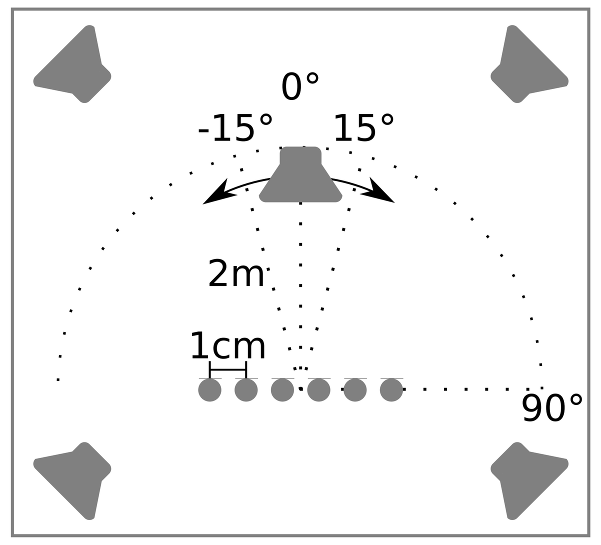
$$\Rightarrow \hat{\mathbf{a}}^{(i)}_{\mathrm{ALS}} = \underset{\mathbf{a}}{\operatorname{argmin}} \left\| \hat{\mathbf{\Phi}}^{(i)}_{\mathbf{x}} - \hat{\phi}^{(i)}_{x,\mathrm{ALS}}\mathbf{a}\mathbf{a}^H \right\|_F^2 \triangleq \text{rank-1 approximation of } \hat{\mathbf{\Phi}}^{(i)}_{\mathbf{x}}$$

$$= \sqrt{\frac{\lambda_1}{\hat{\phi}^{(i)}_{x,\mathrm{ALS}}}}\mathbf{u}_1, \text{ with } \lambda_1, \mathbf{u}_1 \text{ the principal eigenvalue/vector of } \hat{\mathbf{\Phi}}^{(i)}_{\mathbf{x}} \quad (2)$$

  - scaling of RETFs affects PSDs estimate $\hat{\boldsymbol{\phi}}^{(i)}_{\mathrm{ALS}}$
  - after convergence: normalize with reference component (first mic.)

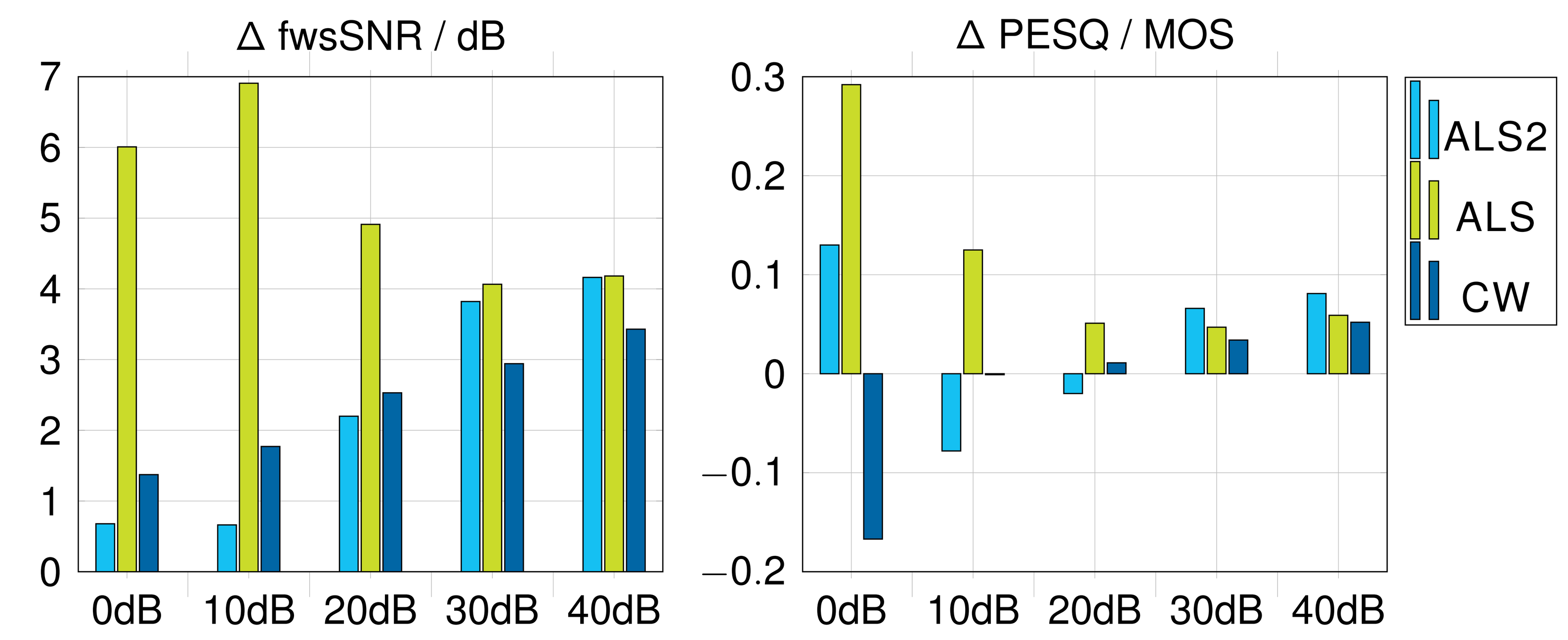
Experimental Results

- $M = 6$ microphones, $T_{60} = 0.35$ s
- **dynamic scenarios**:
  (i) slowly moving from 0° to 90°
  (ii) moving from 0° to 90°, then stationary
  (iii) moving between -15° and 15°
- **STFT settings**: f_s = 16 kHz, $N_{\mathrm{FFT}} = 512$ (frame length 32 ms), 75 % overlap, Hamming window



- pseudo-diffuse babble noise (SDR 10 dB)
- white noise at different diffuse-to-noise-ratios (DNRs)
- $\hat{\mathbf{\Phi}}_{\mathbf{y}}$ obtained via **recursive smoothing** (≈ 20 ms smoothing constant)
- **spatial coherence matrices**:
  - $\mathbf{\Gamma}$: spherically diffuse based on microphone geometry + regularization
  - $\mathbf{\Psi}$: identity matrix (matched condition)
- **DDA** [2] to estimate a-priori SNR ξ for postfilter G
- **baseline methods**:
  - ALS method ignoring residual noise ($\mathbf{\Psi} = \mathbf{0}$ in model) (ALS2) [3]
  - state-of-the-art covariance whitening method [4] (CW)

objective performance vs. DNR



- consistent improvement compared to covariance whitening method
- difference larger for lower DNRs

By considering noise with **multiple spatial coherence matrices**, the proposed method has **increased modeling capacity**, **outperforming state-of-the-art covariance whitening**.

References

[1] O. Schwartz, S. Gannot, and E. A. P. Habets. Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm. In *Proc. European Signal Processing Conference*, pages 1123–1127, Budapest, Hungary, September 2016. IEEE.

[2] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(6):1109–1121, December 1984.

[3] M. Tammen, I. Kodrasi, and S. Doclo. Iterative Alternating Least-Squares Approach to Jointly Estimate the RETFs and the Diffuse PSD. In *Proc. ITG Conference on Speech Communication*, pages 221–225, Oldenburg, Germany, October 2018. VDE.

[4] S. Markovich-Golan, S. Gannot, and I. Cohen. Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1071–1086, August 2009.