Complexity Reduction of EVD-Based Diffuse Power Spectral Density Estimators using the Power Method

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 and noise**
- multi-channel Wiener filter (MWF) requires PSD estimates
- recently proposed **diffuse PSD estimator based on eigenvalue decomposition (EVD)**, which does not require relative early transfer function (RETF) vector of desired speech source [1]
- **goal**: reduce **computational cost** of EVD-based PSD estimator

Signal Model

- microphone signal model in STFT-domain, independent processing in each subband

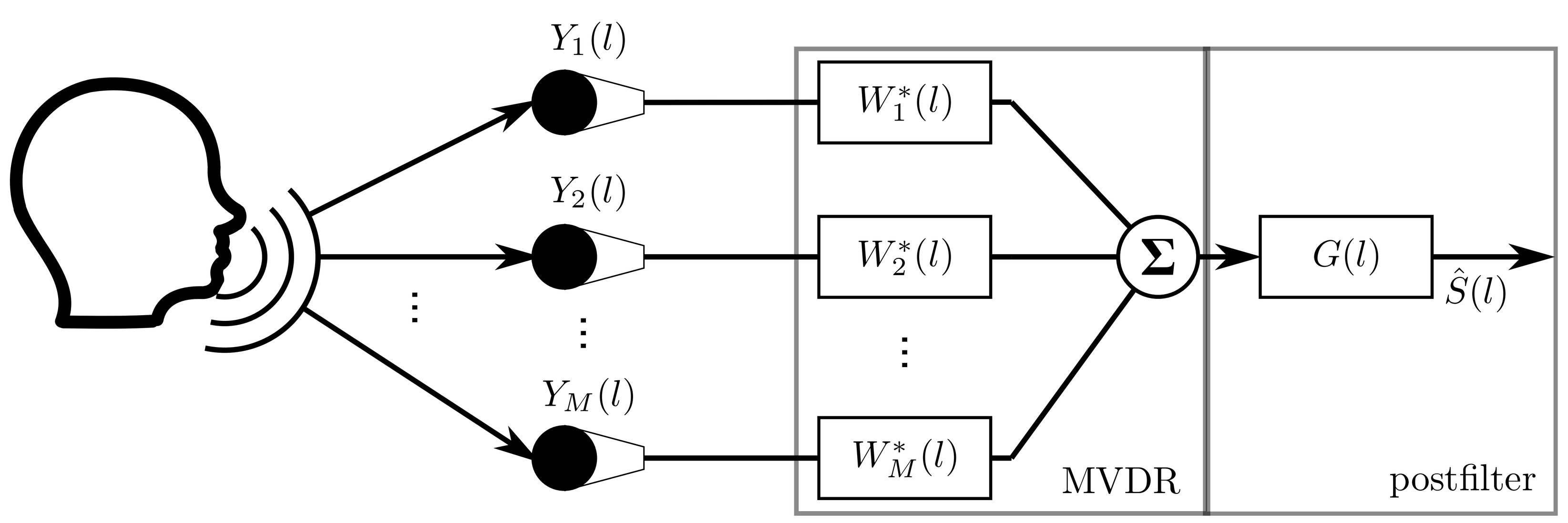
$$\mathbf{y}(l) = \mathbf{x}(l) + \mathbf{d}(l) \text{ with } \mathbf{y}(l) = \begin{bmatrix} Y_1(l) \\ \vdots \\ Y_M(l) \end{bmatrix}$$

  - M microphones, time frame l
  - $\mathbf{x}(l)$: direct and early speech component
  - $\mathbf{d}(l)$: diffuse noise and reverberation
- microphone PSD matrix

$$\mathbf{\Phi}_{\mathbf{y}}(l) = \mathcal{E}\{\mathbf{y}(l)\mathbf{y}^H(l)\} = \Phi_s(l)\mathbf{a}(l)\mathbf{a}^H(l) + \Phi_d(l)\mathbf{\Gamma}$$

  - $\mathbf{a}(l)$: RETF vector, $\mathbf{\Gamma}$: diffuse coherence matrix
  - Φ_s, Φ_d: speech and diffuse PSD

Multichannel Wiener Filter



- stage I: MVDR beamformer $\mathbf{w}(l) = \frac{\mathbf{\Gamma}^{-1}\mathbf{a}(l)}{\mathbf{a}^H(l)\mathbf{\Gamma}^{-1}\mathbf{a}(l)}$
- stage II: spectro-temporal postfilter $G(l) = \frac{\hat{\Phi}_s(l)}{\hat{\Phi}_s(l) + \hat{\Phi}_d(l)/(\mathbf{a}^H(l)\mathbf{\Gamma}^{-1}\mathbf{a}(l))}$

  requires estimate of speech PSD $\Phi_s(l)$ and diffuse PSD $\Phi_d(l)$

EVD-Based Diffuse PSD Estimator

- exploit PSD matrix structure
  - prewhitening based on Cholesky decomposition: $\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^H$

$$\mathbf{\Phi}_{\mathbf{y}}^w(l) = \mathbf{L}^{-1}\mathbf{\Phi}_{\mathbf{y}}(l)\mathbf{L}^{-H} = \Phi_s(l)\underbrace{\mathbf{L}^{-1}\mathbf{a}(l)}_{\mathbf{b}(l)}\underbrace{\mathbf{a}^H(l)\mathbf{L}^{-H}}_{\mathbf{b}^H(l)} + \Phi_d(l)\mathbf{I}_M$$

  - rank-1 term $\Phi_s(l)\mathbf{b}(l)\mathbf{b}^H(l)$ has one non-zero eigenvalue $\sigma(l)$
  - term $\Phi_d(l)\mathbf{I}_M$ adds offset $\Phi_d(l)$ to eigenvalues

$$\lambda_1\{\mathbf{\Phi}_{\mathbf{y}}^w(l)\} = \sigma(l) + \Phi_d(l), \qquad \lambda_i\{\mathbf{\Phi}_{\mathbf{y}}^w(l)\} = \Phi_d(l), \quad i \in \{2, ..., M\}$$

  - in practice, **model not perfect** $\rightarrow \lambda_i \neq \lambda_j, \quad i, j \in \{2, \ldots, M\}, \; i \neq j$
- estimate $\Phi_d(l)$ using either second eigenvalue or mean of smallest $M-1$ eigenvalues

$$\hat{\Phi}_{d,\mathrm{EIG2}}(l) = \lambda_2\{\mathbf{\Phi}_{\mathbf{y}}^{\mathrm{w}}(l)\}$$
$$\hat{\Phi}_{d,\mathrm{EIG1}}(l) = \frac{1}{M-1}\left(\mathrm{trace}\left\{\mathbf{\Phi}_{\mathbf{y}}^{\mathrm{w}}(l)\right\} - \lambda_1\left\{\mathbf{\Phi}_{\mathbf{y}}^{\mathrm{w}}(l)\right\}\right)$$

- eigenvalue decomposition of $M \times M$ matrix required for each STFT bin

high performance high complexity

Power Method

- since only first or second eigenvalue of $\mathbf{\Phi}_{\mathbf{y}}^{\mathrm{w}}(l)$ are required, computational cost can be reduced using **power method**
- power method iteratively estimates **dominant eigenvalue** provided that

$$|\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_M| \quad ✓$$

  and can also estimate **additional eigenvalues** using rank reduction

- **initialization**: random or estimate from previous frame
- **convergence speed**: depends on $|\lambda_1|/|\lambda_2|$
- **complexity** (flops):
  - λ_1: $N(16M^2 + 2M - 2)$
  - λ_2: $2N(16M^2 + 2M - 2) + 5M^2$
  - full EVD using Hessenberg QR algorithm: $4/3M^3 + \mathcal{O}(M^2)$

In: $\mathbf{\Phi}_{\mathbf{y}}^w(l) \in \mathbb{C}^{M \times M}$, number of iterations N
Out: estimates $\hat{\lambda}_1\{\mathbf{\Phi}_{\mathbf{y}}^w(l)\}$, $\hat{\lambda}_2\{\mathbf{\Phi}_{\mathbf{y}}^w(l)\}$
for $m = 1$ **to** 2 **do**
  initialize $\mathbf{u}_m^{(0)} \in \mathbb{C}^M$;
  for $n = 1$ **to** N **do**
    $\mathbf{t} = \mathbf{\Phi}_{\mathbf{y}}^w(l)\mathbf{u}_m^{(n-1)}$; `// power iteration`
    $\mathbf{u}_m^{(n)} = \mathbf{t}/||\mathbf{t}||_2$;
    $\lambda_m^{(n)} = \mathbf{u}_m^{(n),H}\mathbf{\Phi}_{\mathbf{y}}^w(l)\mathbf{u}_m^{(n)}$; `//RAYLEIGH quotient`
  end
  $\hat{\lambda}_m\{\mathbf{\Phi}_{\mathbf{y}}^w(l)\} = \lambda_m^{(N)}$;
  `// matrix rank reduction`
  $\mathbf{\Phi}_{\mathbf{y}}^w(l) = \mathbf{\Phi}_{\mathbf{y}}^w(l) - \hat{\lambda}_m\{\mathbf{\Phi}_{\mathbf{y}}^w(l)\}\mathbf{u}_m^{(N)}\mathbf{u}_m^{(N),H}$;
end

Simulation Setup

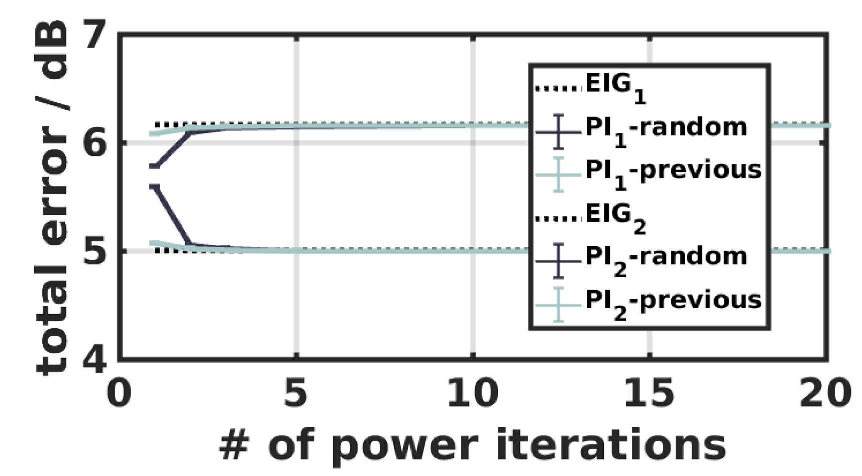
- $M = 4$ microphones, $f_s = 16\,\mathrm{kHz}$
- diffuse babble noise [2] at $\mathrm{SNR}_{\mathrm{in}} = \{10, \ldots, 40\}\,\mathrm{dB}$, no sensor noise

	array geometry	mic. distance	θ	T_{60}
AS_1	linear	$d = 8\,\mathrm{cm}$	45°	$0.61\,\mathrm{s}$
AS_2	circular	$r = 10\,\mathrm{cm}$	45°	$0.73\,\mathrm{s}$
AS_3	linear	$d = 6\,\mathrm{cm}$	-15°	$1.25\,\mathrm{s}$

- STFT: 64 ms frame length ($N_{\mathrm{FFT}} = 1024$), 16 ms shift
- $\mathbf{\Phi}_{\mathbf{y}}(l)$ estimated using recursive averaging, 40 ms smoothing constant
- speech PSD $\Phi_s(l)$ estimated using decision-directed approach[3]
- performance measures:
  - PSD estimation error (averaged over frames and frequencies)
  - speech quality of processed signal $\hat{S}(l)$ using fwsSNR and PESQ

Results

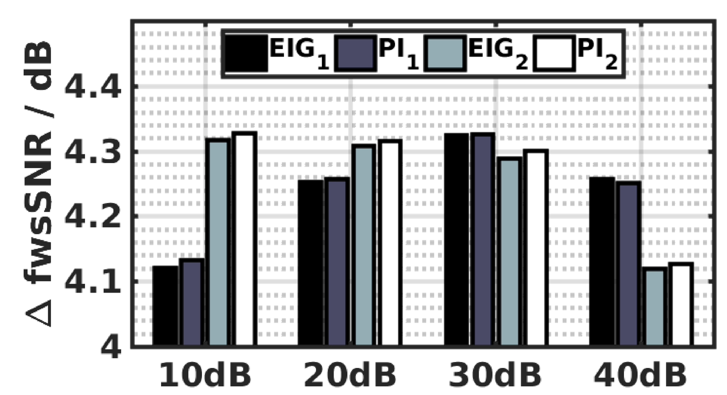
- **PSD estimation accuracy**: different initialization, $\mathrm{SNR}_{\mathrm{in}} = 10\,\mathrm{dB}$, AS_1
  - convergence to full EVD after few iterations
  - best initialization utilizing estimate of previous frame

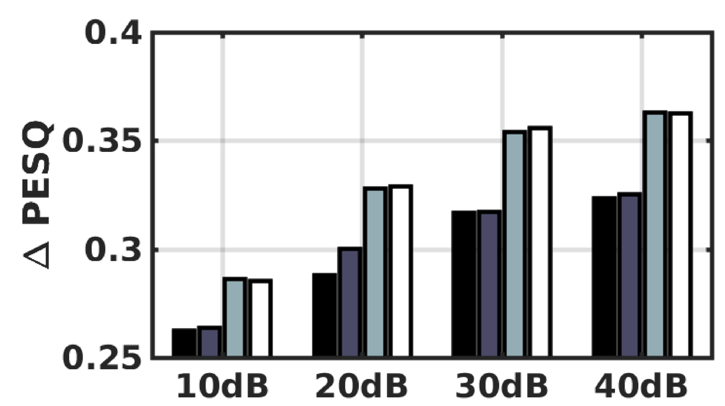


- **computational complexity**: non-optimized runtimes, $N = 2$
  - power method significantly faster
  - no noticeable difference in accuracy

Method	Total Error / dB	Av. Duration / 10^{-5} s
Power Method, λ_1	6.09	0.35
QR Method, λ_1	6.11	1.03
QZ Method (MATLAB), λ_1	6.16	0.55

- **speech quality** of processed signal using different PSD estimates (average over AS_{1-3}, $N = 2$)





  - no significant difference between power method and full EVD

References

[1] I. Kodrasi and S. Doclo. Analysis of Eigenvalue Decomposition-Based Late Reverberation Power Spectral Density Estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2018. in press.

[2] E. A. P. Habets, I. Cohen, and S. Gannot. Generating nonstationary multisensor signals under a spatial coherence constraint. *Journal of the Acoustical Society of America*, 124(5):2911–2917, November 2008.

[3] 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.