Imperial College London

Speech and Audio Processing Lab

FRAME-BASED SPACE-TIME COVARIANCE MATRIX ESTIMATION FOR POLYNOMIAL EIGENVALUE DECOMPOSITION-BASED SPEECH ENHANCEMENT

Emilie d'Olne, Vincent W. Neo, Patrick A. Naylor
Department of Electrical and Electronic Engineering, Imperial College London, U.K.

Summary

- Polynomial eigenvalue decomposition (**PEVD**) has been proposed for speech enhancement in [1]
- Current algorithms rely on long-term (**batch**) estimate of statistics
- This work shows first steps towards **frame-based implementation**

The space-time covariance matrix

Reverberant microphone signal

$$x_m(n) = \mathbf{h}_m^T \mathbf{s}(n) + v_m(n)$$

For M microphones

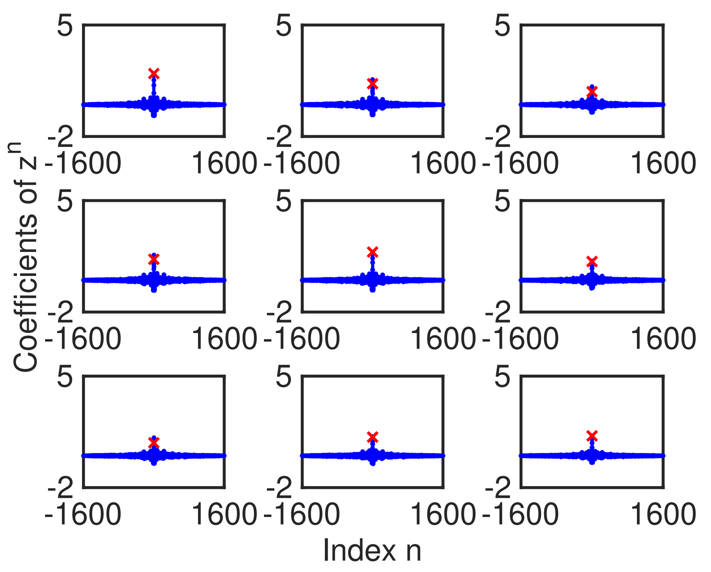
$$\mathbf{x}(n) = [x_1(n), x_2(n)\ldots, x_M(n)]^T$$

Space-time Covariance Matrix (STCOV)

$$\mathbf{R}_{\mathbf{xx}}(\tau) = \mathbb{E}[\mathbf{x}(n)\mathbf{x}^H(n-\tau)]$$

Para-Hermitian Polynomial Matrix

$$\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) = \sum_{\tau=-W}^{W} \mathbf{R}_{\mathbf{xx}}(\tau) z^{-\tau}$$



PEVD-based enhancement

The PEVD of $\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z)$ is [2]

$$\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) \approx \boldsymbol{\mathcal{U}}^P(z)\boldsymbol{\mathit{\Lambda}}(z)\boldsymbol{\mathcal{U}}(z)$$

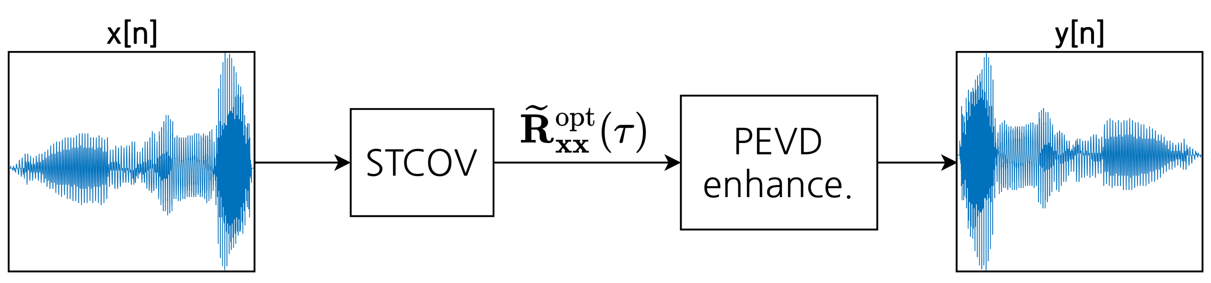
$$= \left[\ \boldsymbol{\mathcal{U}}^P_{\tilde{s}}(z) \mid \boldsymbol{\mathcal{U}}^P_{\tilde{v}}(z)\ \right] \left[\begin{array}{c|c} \boldsymbol{\mathit{\Lambda}}_{\tilde{s}}(z) & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\mathit{\Lambda}}_{\tilde{v}}(z) \end{array}\right] \left[\begin{array}{c} \boldsymbol{\mathcal{U}}_{\tilde{s}}(z) \\ \hline \boldsymbol{\mathcal{U}}_{\tilde{v}}(z) \end{array}\right]$$

with orthogonal signal, $\{\cdot\}_{\tilde{s}}$ and noise subspaces, $\{\cdot\}_{\tilde{v}}$.
For a single source, the enhanced signal is

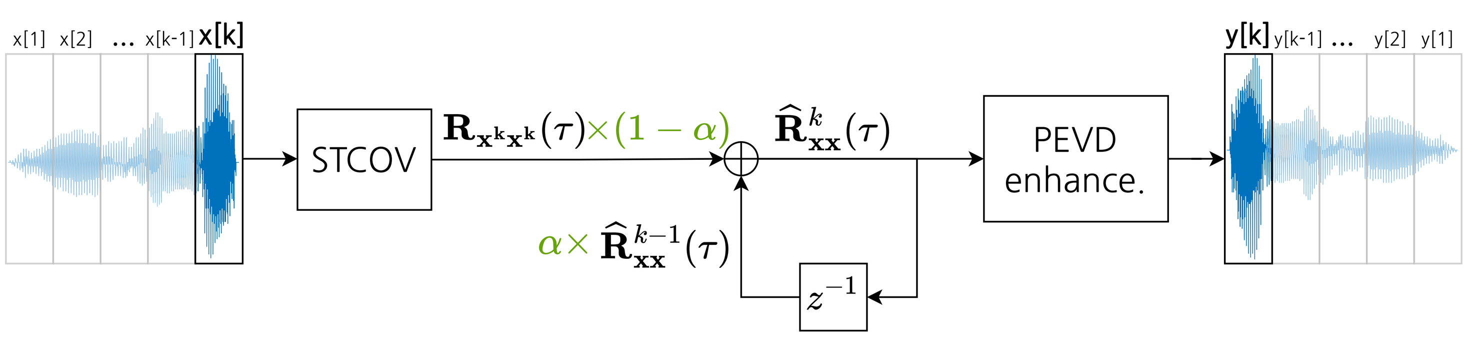
$$y(z) = \boldsymbol{\mathcal{U}}^P_{\tilde{s}}(z)\mathbf{x}(z)$$

Batch vs frame-based approaches

Batch processing



Frame-based processing



Proposed method

Recursive estimation of the space-time covariance matrix

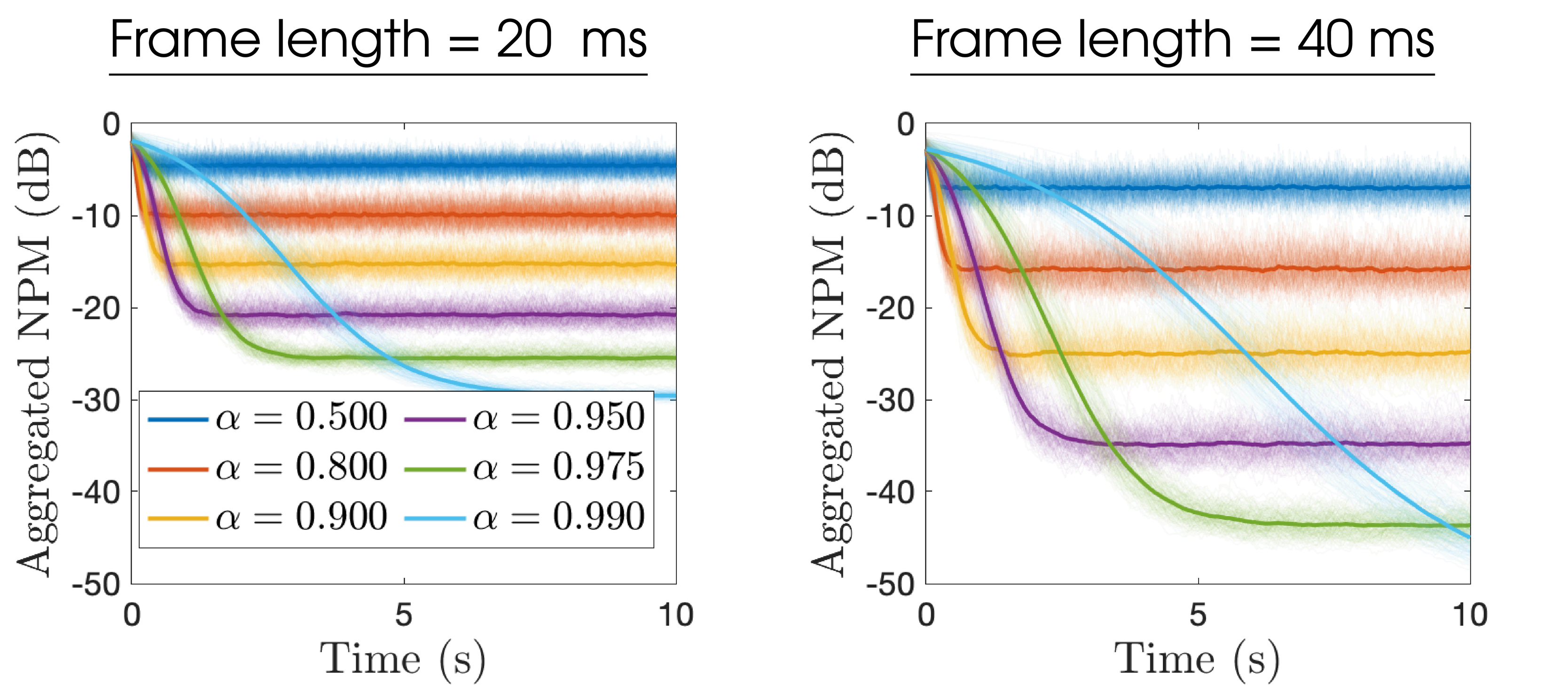
$$\widehat{\mathbf{R}}^k_{\mathbf{xx}}(\tau) = \alpha\widehat{\mathbf{R}}^{k-1}_{\mathbf{xx}}(\tau) + (1-\alpha)\mathbf{R}_{\mathbf{x}^k\mathbf{x}^k}(\tau)\,,$$

Simulation setup

- Using ULA with **3 microphones**, separated by 5 cm
- Target source is located at 90° azimuth
- Reverberant room (T60=400 ms) simulated using image-source method

Experiment 1: Estimation accuracy

- When $\mathbf{s}(n)$ is white Gaussian, the ground-truth STCOV is given by acoustic impulse responses
- Use **aggregated** normalised projection misalignement (**NPM**) between estimate $\widehat{\mathbf{R}}_{\mathbf{xx}}(\tau)$ and ground-truth $\mathbf{R}_{\mathbf{xx}}(\tau)$



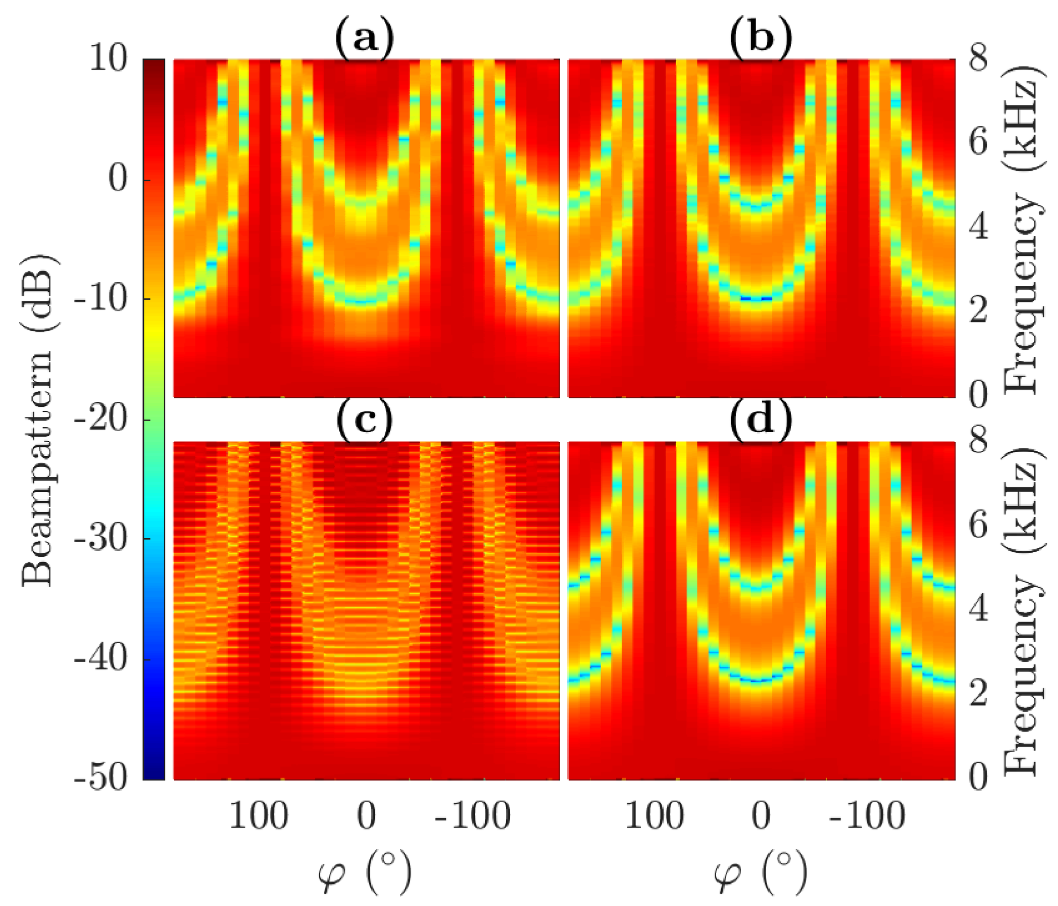
Experiment 2: Impact on speech enhancement

- **Male target speaker** (IEEE sentence) in isotropic speech-shaped noise
- No longer possible to establish ground-truth
- Instead compare performance against batch estimate using **beampatterns** and **speech enhancement metrics**

Beampattern examples

(a): Long-term estimate $\mathbf{U}^{\text{opt}}_s(\tau)$

(b), (c), (d): $\widehat{\mathbf{U}}^k_s(\tau)$ with $\alpha = 0.9$ and at $t = \{1, 1.84, 3\}$ s



SNR and STOI improvements relative to batch mode

α	0.50	0.80	0.90	0.95	0.975	0.99
ΔSNR [dB]	1.29	1.28	1.18	1.08	0.80	0.49
ΔSTOI	0.01	0.03	0.02	0.01	0.02	0.02

Conclusions

- Showed feasibility of frame-based space-time covariance estimation
- The estimate converges to ground-truth
- The proposed method performs similarly for speech enhancement than previously proposed batch method

References

[1] V. W. Neo, C. Evers, and P. A. Naylor, "Enhancement of noisy reverberant speech using polynomial matrix eigenvalue decomposition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2021.

[2] J. G. McWhirter, P. D. Baxter, T. Cooper, S. Redif, and J. Foster, "An EVD algorithm for para-Hermitian polynomial matrices," *IEEE Trans. Signal Process.*, vol. 55, no. 5, pp. 2158–2169, May 2007.