LCMV Beamforming with Subspace Projection for Multi-Speaker Speech Enhancement

Amin Hassani, Alexander Bertrand, Marc Moonen

KU Leuven, Dept. of Electrical Engineering-ESAT

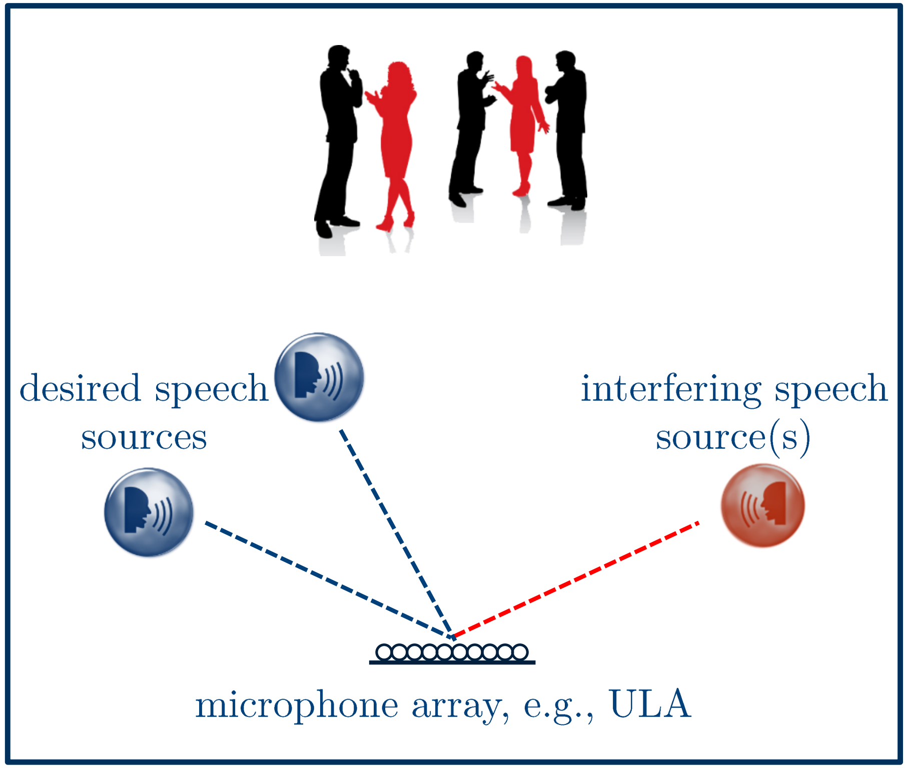
ICASSP 2016

Presentation outline

- Motivation and Problem Statement
- LCMV Beamformer
- Proposed Subspace Projection-based Approach
- Simulation Results
- Conclusion

Motivation

- Microphone arrays for audio and speech enhancement
- **Problem:** extracting desired speech signals from microphone signals, polluted by other interfering speech signals and noise components
- **Approach:** Linearly Constrained Minimum Variance (LCMV) beamformer
- **Main Goal:** using subspace projection-based approach to improve the performance of the LCMV beamformer when <u>insufficient relevant samples are available</u>



LCMV beamformer (1/2)

- Data model of microphone signals (STFT):

$$\boxed{\begin{aligned}\mathbf{y} &= \mathbf{A}_d\mathbf{s}_d + \mathbf{A}_i\mathbf{s}_i + \mathbf{n}\\ &\triangleq \mathbf{d} + \mathbf{i} + \mathbf{n}\end{aligned}}$$

$\mathbf{y}$: contains M microphone signals
$\mathbf{s}_d$: contains N_d desired speech sources
$\mathbf{s}_i$: contains N_i interfering speech sources
$\mathbf{A}_d$: $M \times N_d$ desired steering matrix
$\mathbf{A}_i$: $M \times N_i$ interfering steering matrix

- LCMV minimizes the total **output variance**, under a set of linear constraints (generalization of Minimum Variance Distortionless Response (MVDR)):

$$\boxed{\begin{aligned}&\min_{\mathbf{w}} E\{|\mathbf{w}^H\mathbf{y}|^2\}\\ &\text{s.t. } \mathbf{A}^H\mathbf{w} = \mathbf{f}\end{aligned}} \Rightarrow \boxed{\mathbf{w} = \mathbf{R}_{yy}^{-1}\mathbf{A}(\mathbf{A}^H\mathbf{R}_{yy}^{-1}\mathbf{A})^{-1}\mathbf{f}} \Rightarrow \boxed{\bar{d} = \mathbf{w}^H\mathbf{y}}$$

$\mathbf{A} = [\mathbf{A}_d \ \mathbf{A}_i]$
$\mathbf{f} = [\underbrace{1\ldots1}_{N_d} \ \underbrace{0\ldots0}_{N_i}]^T$ is the vector of desired responses

KU LEUVEN

LCMV beamformer (2/2)

- Two main classes of LCMV beamformer:

  I. All Acoustic Transfer Functions (ATFs) are **known** ➔ LCMV output contains mixture of desired source signals (mixture of *dry* speech signals)

  II. **Unknown** ATFs: '*blind beamforming*' requires subspace estimation ➔ LCMV output contains mixture of the desired source signals as observed by a *reference microphone* (mixture of *wet* speech signals)

- If ATFs (class I) or subspaces (class II) are not accurately estimated, the LCMV beamformer that minimizes the *output variance* delivers severe speech distortion [1]

[1] Harry L. Van Trees," Detection, Estimation, and Modulation Theory, Optimum Array Processing", 2004.

Blind LCMV beamformer (1/2)

- '*desired-sources-only*' correlation matrix: $\mathbf{R}^{d}_{yy} = \mathbf{A}_d \mathbf{\Pi}_d \mathbf{A}_d^H + \mathbf{R}_{nn}$
- '*interfering-sources-only*' correlation matrix: $\mathbf{R}^{i}_{yy} = \mathbf{A}_i \mathbf{\Pi}_i \mathbf{A}_i^H + \mathbf{R}_{nn}$
- '*noise-only*' correlation matrix: $\mathbf{R}_{nn}$

- Estimating $\mathbf{R}^{d}_{yy}$ and $\mathbf{R}^{i}_{yy}$ via sample averaging (e.g., as in [2])

- Subspace estimation via Generalized EigenValue Decomposition (GEVD): better suited for scenarios with spatially correlated (e.g., localize noise sources) and/or nonstationary noise (e.g., interfering speakers)

[2] S. Markovich Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), USA, Mar. 2010, pp. 201 –204.

Blind LCMV beamformer (2/2)

- Compute $\mathbb{Q}_d$: $M \times N_d$ subspace of desired speech

$$\text{GEVD } (\mathbf{R}_{yy}^{d}, \mathbf{R}_{nn}) \Rightarrow \mathbb{Q}_d$$

- Compute $\mathbb{Q}_i$: $M \times N_i$ subspace of interfering speech

$$\text{GEVD } (\mathbf{R}_{yy}^{i}, \mathbf{R}_{nn}) \Rightarrow \mathbb{Q}_i$$

- With modified constrain set $\mathbb{Q} \triangleq [\mathbb{Q}_d \ \mathbb{Q}_i]$, LCMV becomes [2]

$$\boxed{\begin{aligned} \min_{\mathbf{w}} \ & E\{|\mathbf{w}^H \mathbf{y}|^2\} \\ \text{s.t. } & \mathbb{Q}^H \mathbf{w} = \mathbf{f} \end{aligned}} \Rightarrow \boxed{\mathbf{w} = \mathbf{R}_{yy}^{-1} \mathbb{Q} (\mathbb{Q}^H \mathbf{R}_{yy}^{-1} \mathbb{Q})^{-1} \mathbf{f}} \Rightarrow \boxed{\bar{d} = \mathbf{w}^H \mathbf{y}}$$

$\mathbf{f} = [\mathbb{q}_d^T \ \underbrace{0 \ldots 0}_{N_i}]^T$, $\mathbb{q}_d$ is the r-th (reference) column of $\mathbb{Q}_d^H$

[2] S. Markovich Golan, S. Gannot, and I. Cohen, “Subspace tracking of multiple sources and its application to speakers extraction,” in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), USA, Mar. 2010, pp. 201 –204.

LCMV Beamforming with Subspace Projection

- The estimation of $\mathbb{Q}_d$ and $\mathbb{Q}_i$ may yield poor results when insufficient available *'desired-sources-only'* and/or *'interfering-sources-only'* samples

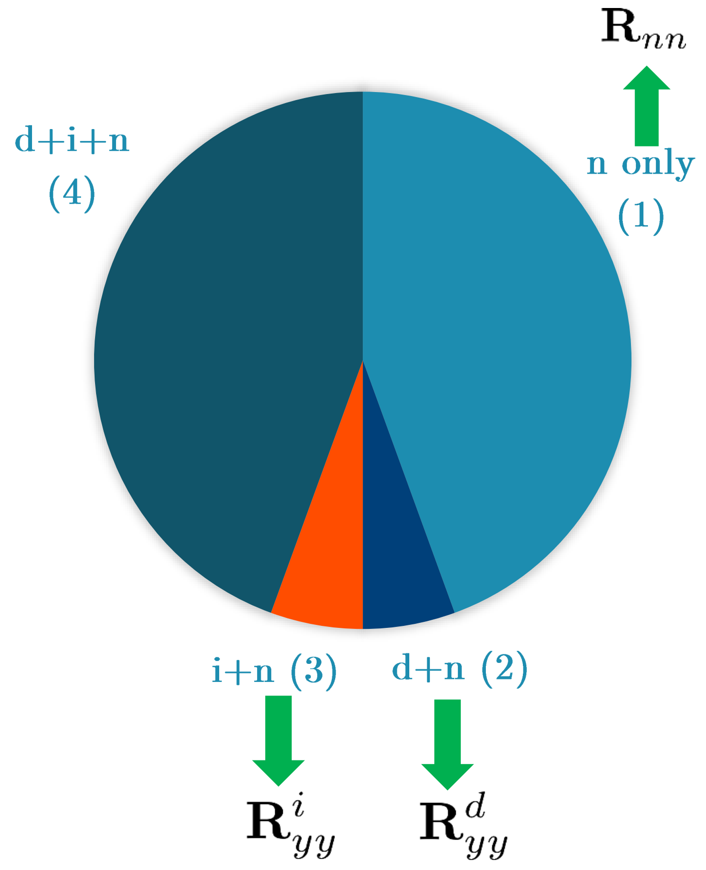
- *'desired+interfering'* segments were not exploited for the estimation of $\mathbb{Q}_d$ and $\mathbb{Q}_i$

- Only excluding the samples of *'noise-only'* segments

$$\mathbf{R}_{yy}^{d,i} = \mathbf{A}_d \mathbf{\Pi}_d \mathbf{A}_d^H + \mathbf{A}_i \mathbf{\Pi}_i \mathbf{A}_i^H + \mathbf{R}_{nn}$$

- Compute $\mathbb{Q}_{d,i}$: $M \times (N_d + N_i)$ joint subspace of desired and interfering speech

$$\text{GEVD}\ (\mathbf{R}_{yy}^{d,i}, \mathbf{R}_{nn}) \Rightarrow \mathbb{Q}_{d,i}$$



KU LEUVEN

LCMV Beamforming with Subspace Projection

- In theory: Col $(\mathbb{Q}_{d,i})=$ Col $([\mathbb{Q}_d \;\mathbb{Q}_i])$
- In practice: Col $(\mathbb{Q}_{d,i}) \neq$ Col $([\mathbb{Q}_d \;\mathbb{Q}_i])$ due to different data segments
- Correction via projection:

$$\boxed{\begin{aligned}\mathbb{Q}_d^{\text{proj}} &\triangleq \mathbb{Q}_{d,i}(\mathbb{Q}_{d,i}^T\mathbb{Q}_{d,i})^{-1}\mathbb{Q}_{d,i}^T\mathbb{Q}_d \\ \mathbb{Q}_i^{\text{proj}} &\triangleq \mathbb{Q}_{d,i}(\mathbb{Q}_{d,i}^T\mathbb{Q}_{d,i})^{-1}\mathbb{Q}_{d,i}^T\mathbb{Q}_i\end{aligned}}$$

- We define the new constraint matrix $\mathbb{Q}_{\text{proj}} \triangleq [\mathbb{Q}_d^{\text{proj}} \;\mathbb{Q}_i^{\text{proj}}]$:

$$\boxed{\mathbf{w}_{\text{proj}} = (\mathbf{R}_{yy}^{d,i})^{-1}\mathbb{Q}_{\text{proj}}(\mathbb{Q}_{\text{proj}}^H(\mathbf{R}_{yy}^{d,i})^{-1}\mathbb{Q}_{\text{proj}})^{-1}\mathbf{f}_{\text{proj}}} \quad \boxed{\bar{d}_{\text{proj}} = \mathbf{w}_{\text{proj}}^H\mathbf{y}}$$

$\mathbf{f}_{\text{proj}} = [(\mathbb{q}_d^{\text{proj}})^T \; \underbrace{0\ldots0}_{N_i}]^T$, $(\mathbb{q}_d^{\text{proj}})$ is the r-th (reference) column of $(\mathbb{Q}_d^{\text{proj}})^H$

Simulations

- Two scenarios:
  I. Monte Carlo (MC) simulations with narrowband source signals (multiple desired + multiple interfering sources)
  II. multi-talker speech enhancement in a simulated cubic room

- Performance measure 1: output Signal to Interference plus Noise Ratio (oSINR):

$$\text{oSINR} = 10\log_{10}\frac{E\{|\mathbf{w}^H\mathbf{d}|^2\}}{E\{|\mathbf{w}^H\mathbf{i}|^2\} + E\{|\mathbf{w}^H\mathbf{n}|^2\}}$$

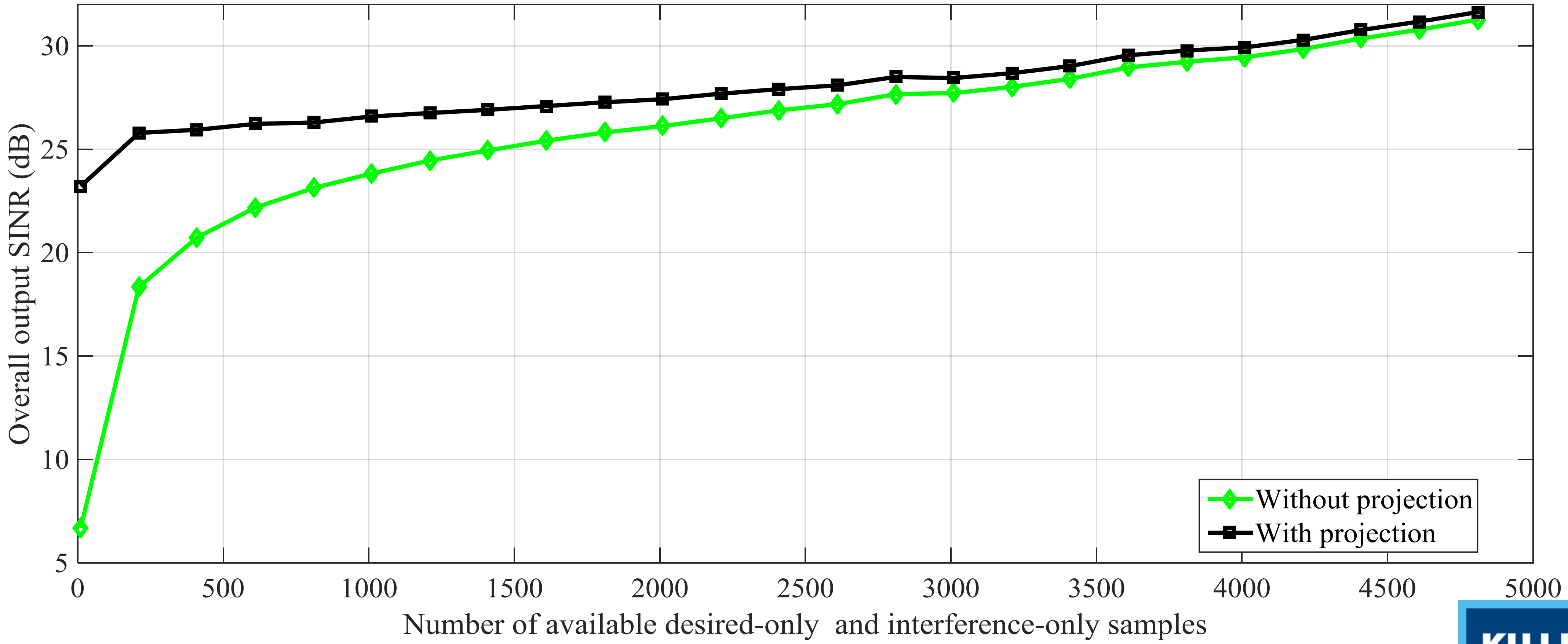
- Performance measure 2: output Signal to Distortion Ratio (oSDR):

$$\text{oSDR} = 10\log_{10}\frac{E\{|d_{\text{ref}}|^2\}}{E\{|d_{\text{ref}} - \mathbf{w}^H\mathbf{d}|^2\}}$$

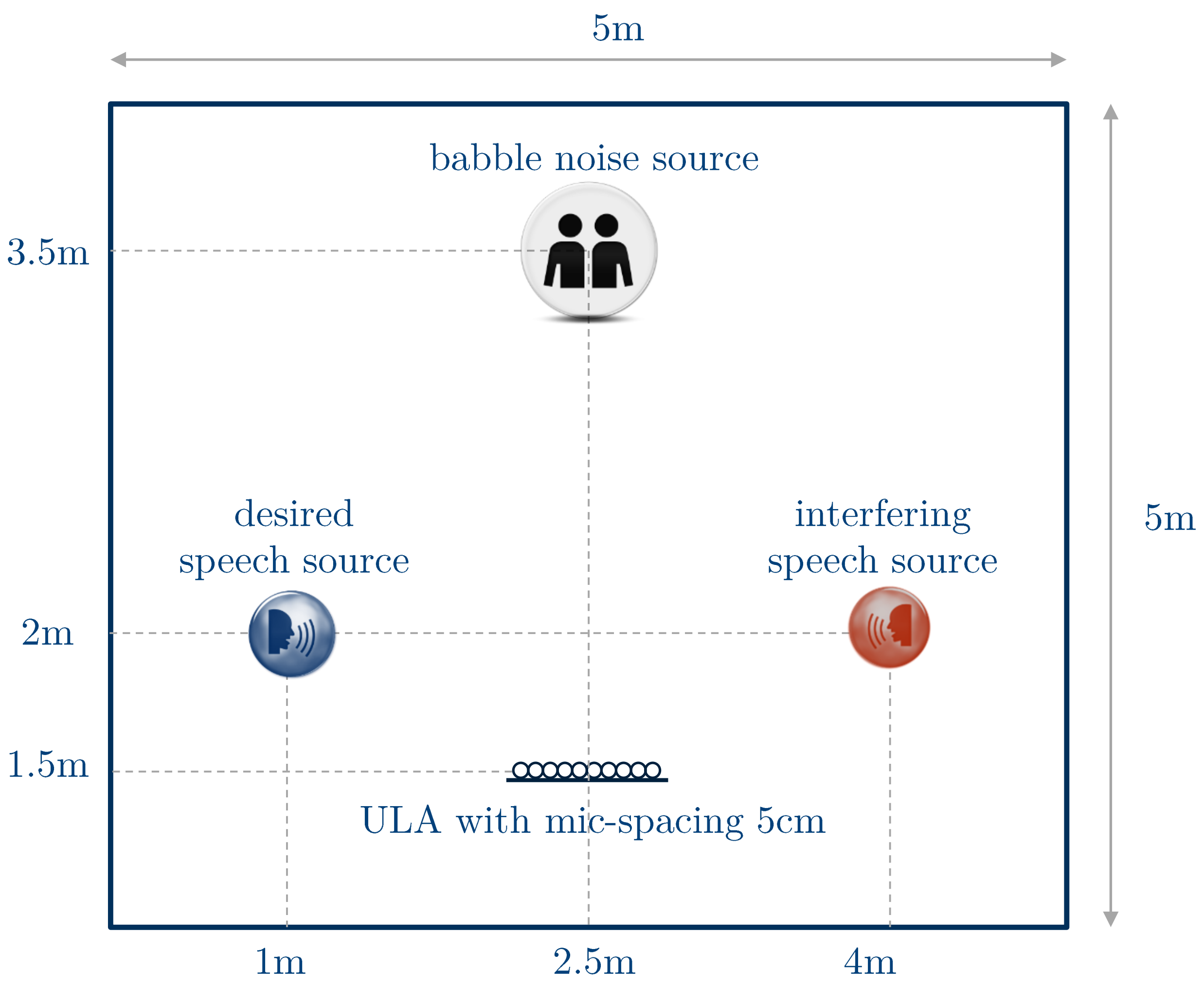
KU LEUVEN

Narrowband simulations (MC=1000)

$M = 10$, $N_d = 2$ (power P), $N_i = 3$ (power P), 2 localized noise(power $0.5P$)
total # of samples= 20000
# of samples in which both desired and interfering sources are active= 7000
increasing Nb_{only} (number of desired/interfering-only samples)



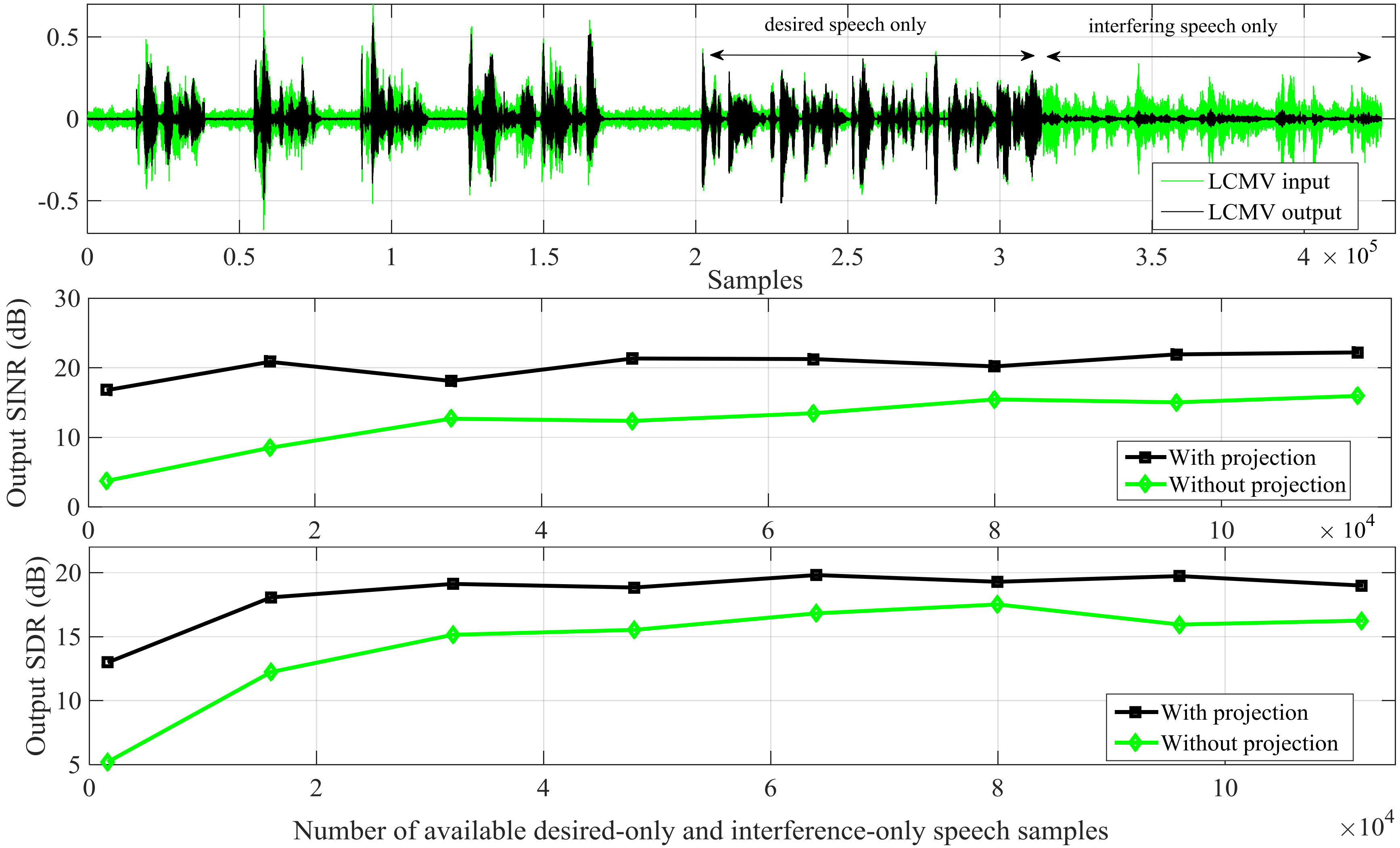
Multi-talker speech simulations



$M = 10$
$F_s = 16kHz, \text{DFTsize} = 512$
desired and interfering sources power $P_s = P_i$
babble noise power $0.5P_s$
AWGN with 5% of power of speech at the ref mic
RIR-generator, image method [3]

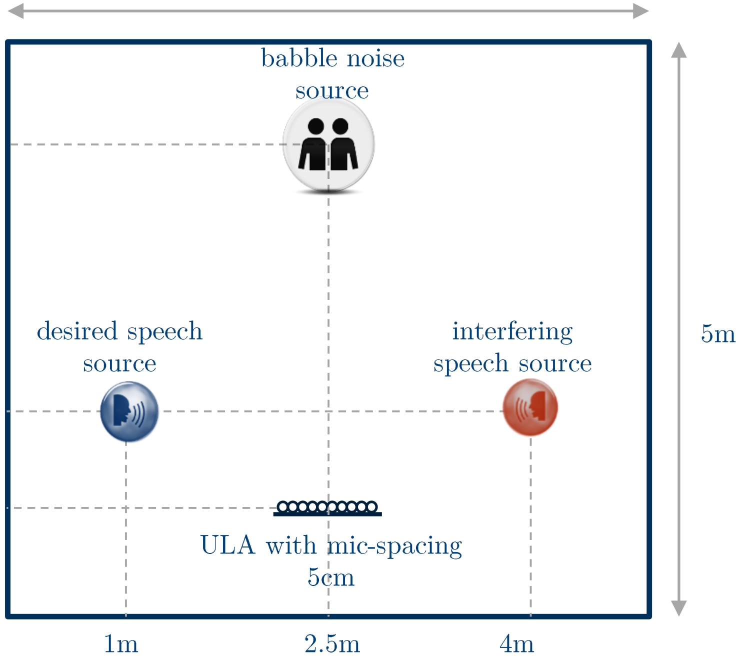
[3] E. Habets, "Room impulse response (RIR) generator," 2010.

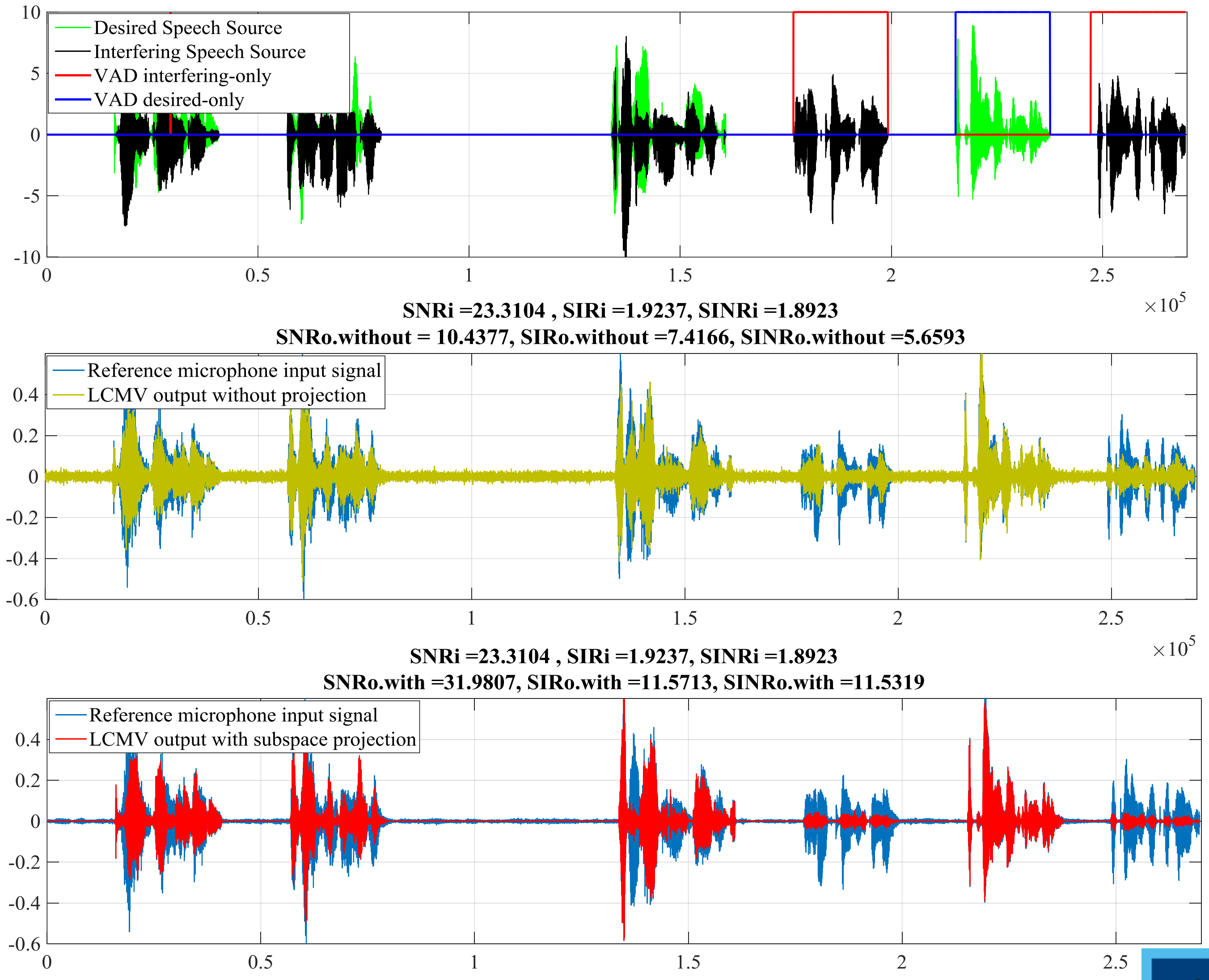
increasing Nb_{only} from $0.1F_s$ to $7F_s$



Audio demonstrations

(batch-processing)





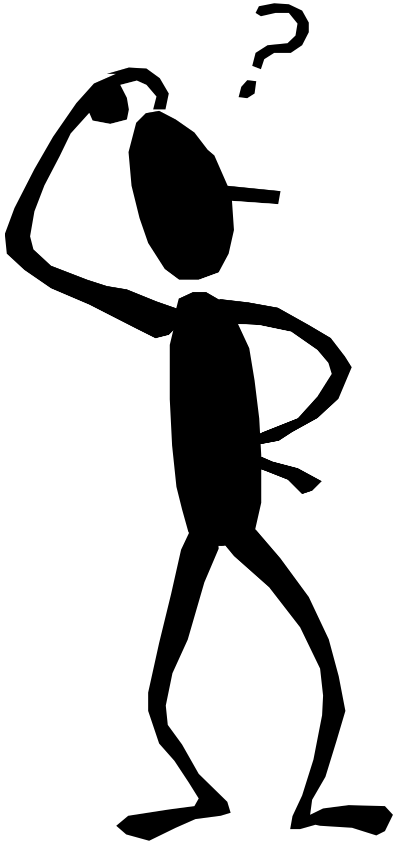
Conclusions

- We have proposed a subspace projection-based approach when insufficient relevant samples are available

- GEVD-based approach has been considered (better subspace estimation performance)

- Improvement is achieved at the cost of more complex computations, as the poorly estimated subspaces have to be projected onto the larger joint subspace
  ➔ extra GEVD

Thank you for your attention.



Discussion ...