Real-time Multichannel Speech Separation and Enhancement using a Beamspace-domain-based Lightweight CNN

Marco Olivieri[⋆], L. Comanducci[⋆], M. Pezzoli[⋆], D. Balsarri[†], L. Menescardi[†], M. Buccoli[†], S. Pecorino[†], A. Grosso[†], F. Antonacci[⋆], A. Sarti[⋆]

[⋆] *Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano, Milan, Italy*

[†] *BdSound S.r.l., Milan, Italy*

POLITECNICO MILANO 1863

BdSound

CONTEXT

Noisy and reverberant environment

CONTEXT

Extract the talker in front of the array

case 1

CONTEXT

Discard all the interferers

case 2

SETUP and GOALS

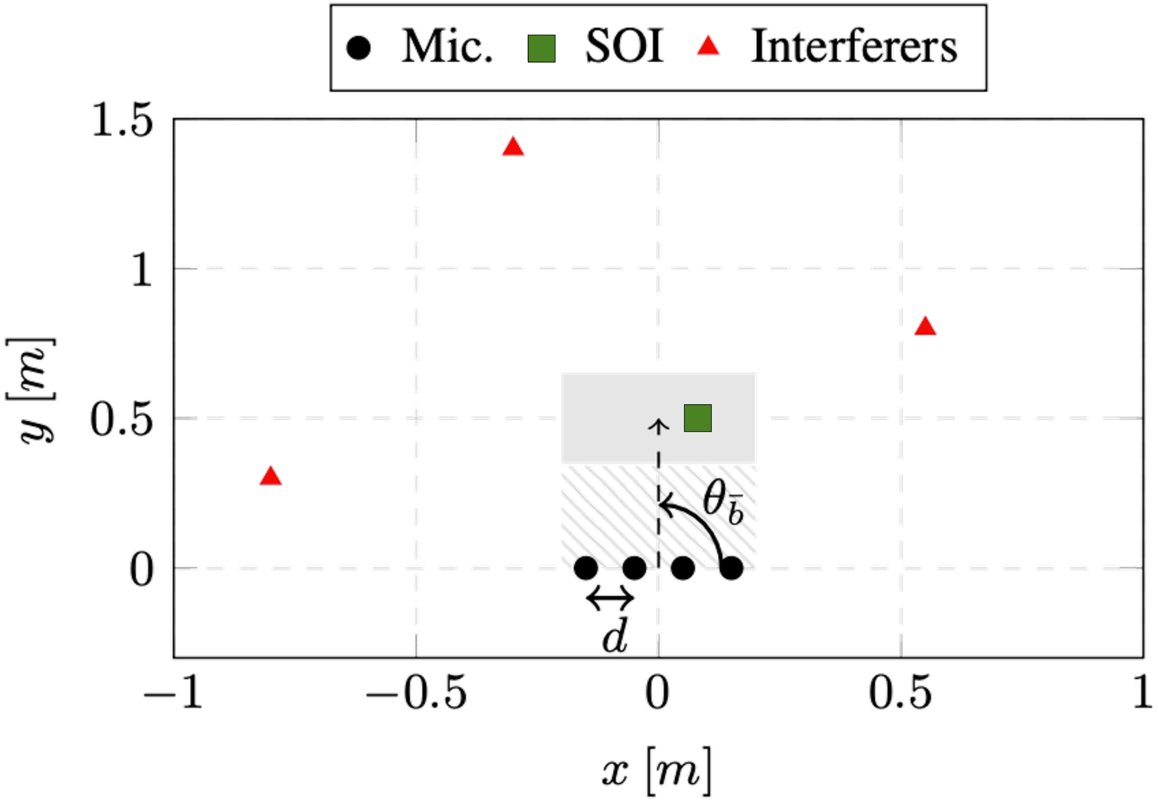
Uniform Linear Array (**ULA**)

1 Signal Of Interest (**SOI**)

- within a region in front of the array
- DOA $\approx \theta_{\bar{b}} = 90^\circ$

$R \in \{0, \ldots, 4\}$ **interferers**

- in the noisy and reverberant room



OBJECTIVES

- ❑ **Real-time model** for the SOI separation and enhancement
- ❑ Evaluation on **real recordings** whereas training on simulated data
- ❑ **Robust system** with respect to multiple array geometries and acoustic conditions

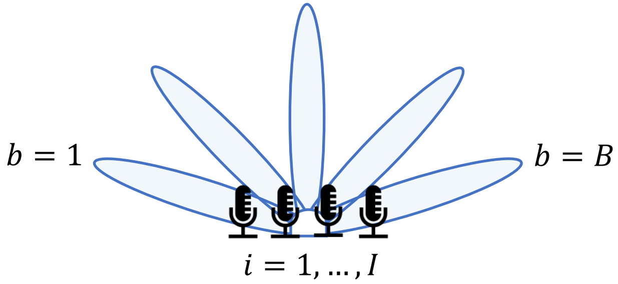
SIGNAL MODEL and BACKGROUND

- I **microphones** with d **inter-sensor spacing**
- J **speakers**
- γ **diffuse noise** component
- ν **additive noise** component

$$y_i[t,f] = \sum_{j=1}^{J} h_{j,i}[t,f] s_j[t,f] + \gamma_i[t,f] + v_i[t,f]$$
$$= \sum_{j=1}^{J} x_{j,i}[t,f] + \gamma_i[t,f] + v_i[t,f],$$

STFT representation of the signal acquired by the i^{th} microphone

BEAMSPACE REPRESENTATION:



$$\mathbf{Y} \in \mathbb{C}^{T \times F \times I}$$

3D tensor of **stacked STFT** of the signals acquired by the I microphones

$$\mathbf{W} \in \mathbb{C}^{I \times B}$$

beamspace transform matrix of a plane-wave decomposition in B directions $\theta_b, b = 1, \dots, B$

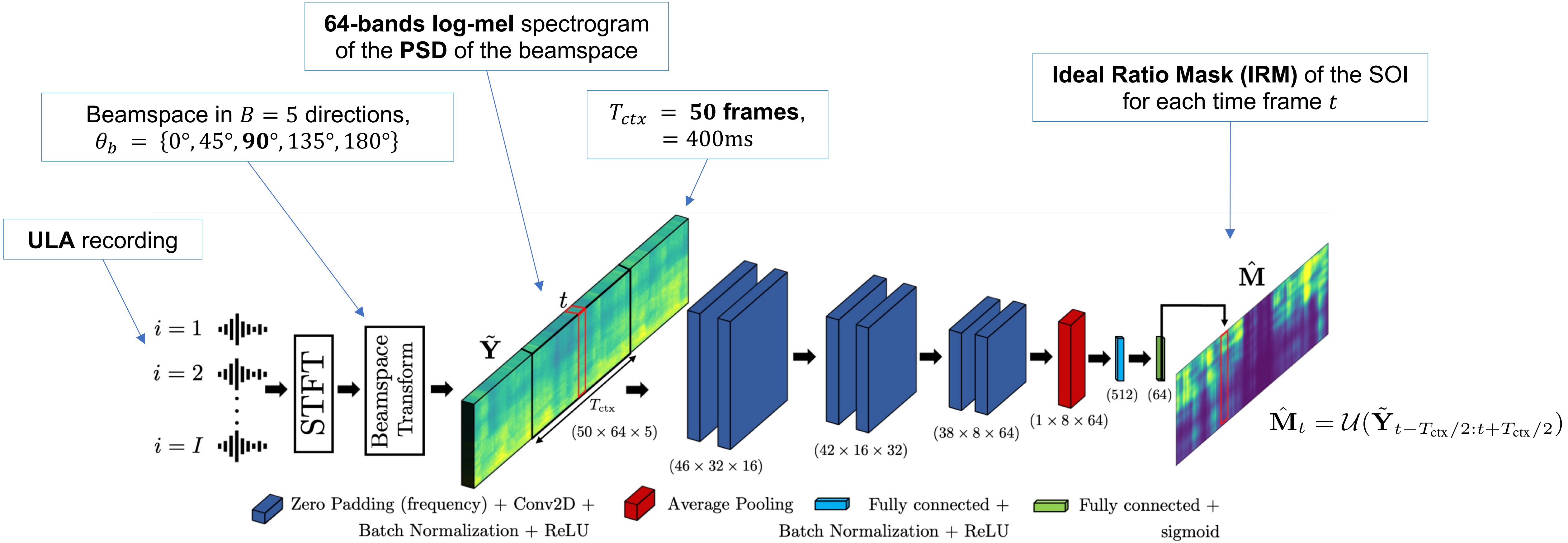
$$\tilde{\mathbf{Y}}_t = \mathbf{Y}_t \mathbf{W}$$

beamspace of the STFT signal at frame t

POLITECNICO MILANO 1863

BdSound

PROPOSED METHOD



$$\mathcal{L}(t) = \frac{1}{F}\sum_{f=1}^{F}(\mathbf{M}_{t,f} - \hat{\mathbf{M}}_{t,f})^2$$

Loss function

$$\hat{\mathbf{X}}_{\bar{j},t} = \hat{\mathbf{M}}_t \odot \tilde{\mathbf{Y}}_{\bar{b},t}$$

Final estimate of the **desired signal** $\bar{j}$ at frame t

DATASET GENERATION

Extensive **simulation campaign** by sampling with a uniform distribution the operational ranges

RIRs computed with **gpuRIR** [1]

operational ranges

ULA setup	$I = 3/4,\ d = 20/30\text{mm}$
Room dimensions	$L_x \in [3, 8]\text{m}$; $L_y \in [3, 8]\text{m}$; $L_z \in [2.6, 4]\text{m}$
T60	$[0.2, 1.4]$ s
SOI presence	$80/20$ % of rooms with / without SOI
R number of interferers	from 0 to 4
SIR (loudness simulation)	$[-3, 3]$ dB
SDR (babble noise)	$[-3, 60]$ dB
SNR (microphone noise)	$[30, 70]$ dB
Array Gain (signal dynamic)	$[-40, -1]$ dB
LibriSpeech dataset	5 sec signals
Total training rooms	$\mathbf{250,000}$

[1] D. Diaz-Guerra, et al. "gpuRIR: A python library for room impulse response simulation with gpu acceleration," Multimedia Tools and Applications, 2021.

EVALUATION

Evaluation on **real recordings in ETSI room** with different **ULA unseen** during training:

- **Varying number of sensors $\boldsymbol{I}$**
- **Varying inter-sensor distance $\boldsymbol{d}$**

→ $I = 3/4, d = 20/26mm$

→ $I = 5, d = 52mm$

Comparison wrt

- **NBDF** method [2]
- $\tilde{\mathbf{Y}}_{90^\circ}$ **input beamformer** steering to the SOI region

[2] W. Liu, et al., "A neural beamspace-domain filter for real-time multi-channel speech enhancement," Symmetry, 2022

RESULTS

ULA setups	$I=4$, $d=26$ mm			$I=3$, $d=52$ mm			$I=4$, $d=52$ mm			Average over test sets		
Metrics	Proposed	NBDF	$\tilde{\mathbf{Y}}_{90^\circ}$	Proposed	NBDF	$\tilde{\mathbf{Y}}_{90^\circ}$	Proposed	NBDF	$\tilde{\mathbf{Y}}_{90^\circ}$	Proposed	NBDF	$\tilde{\mathbf{Y}}_{90^\circ}$
SIR	**9.46**	8.5	1.62	8.5	**10.48**	0.93	6.47	**10.84**	0.97	8.31	**10.05**	1.18
SAR	**7.73**	2.99	-	**9.34**	6.05	-	**7.58**	3.08	-	**8.29**	4.29	-
SDR	**4.15**	0.04	1.6	**4.63**	3.29	0.92	**2.28**	0.75	0.96	**3.79**	1.59	1.17
PESQ	1.66	1.19	**1.71**	**1.86**	1.39	1.79	1.66	1.24	**1.79**	1.73	1.27	**1.76**
ESTOI	0.57	0.44	**0.58**	**0.61**	0.52	0.61	0.6	0.44	**0.62**	0.59	0.46	**0.6**
R_{soi}	-5.75	**-4.7**	-	-2.61	**-1.77**	-	-6.81	**-3.81**	-	-4.67	**-3.25**	-
R_{interf}	**-17.54**	-13.47	-	-14.31	**-14.48**	-	**-15.7**	-15.2	-	**-15.65**	-14.32	-

Table 1. Comparison of the average metrics between the proposed method, the NBDF approach and the beamformer $\tilde{\mathbf{Y}}_{90^\circ}$ for the different test sets and for the average results.

COMPARISON	NBDF	Proposed solution
# parameters	4,006,236	**120,000**
MACS/frame	198.5 millions	**1.06 millions**

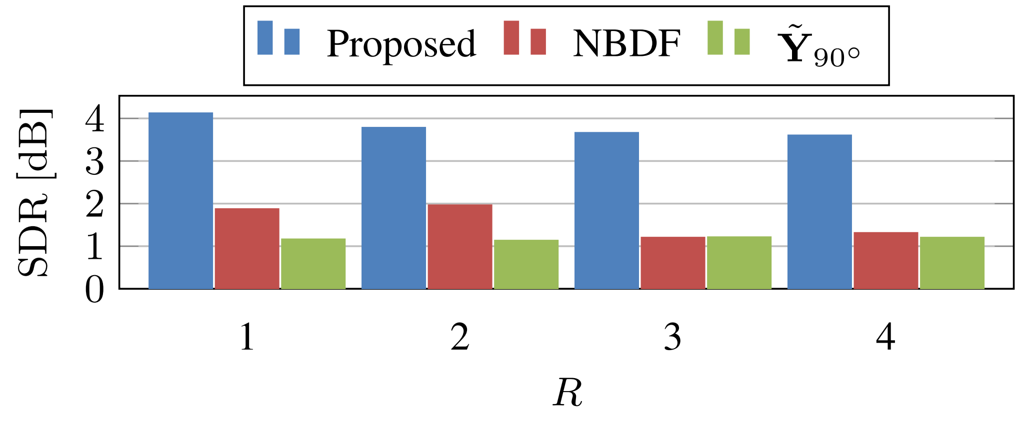


Fig. 3. Average SDR as function of the number of interferers R.

Perceptual intelligibility of the devised solution outperforms both NBDF and $\tilde{Y}_{90^\circ}$

- *https://polimi-ispl.github.io/beamspace_cnn_speech_separation.github.io/*

CONCLUSION

✅ **SOI speech extraction** and enhancement in noisy and reverberant environments

✅ Lightweight CNN architecture for **real-time computation**

✅ Robust system wrt **setup generalization**:

- number of **speakers** and **microphones**
- inter-sensor **spacing**
- **reverberation** time of the room
- **noise** components of the array and the environment

❑ Generalization wrt different array geometries

Real-time Multichannel Speech Separation and Enhancement using a Beamspace-domain-based Lightweight CNN

Thank you!

marco1.olivieri@polimi.it

M. Olivieri[*], L. Comanducci[*], M. Pezzoli[*], D. Balsarri[†], L. Menescardi[†], M. Buccoli[†], S. Pecorino[†], A. Grosso[†], F. Antonacci[*], A. Sarti[*]

[*] *Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano, Milan, Italy*

[†] *BdSound S.r.l., Milan, Italy*

POLITECNICO MILANO 1863

BdSound®