REAL-TIME MULTICHANNEL SPEECH SEPARATION AND ENHANCEMENT USING A BEAMSPACE-DOMAIN-BASED LIGHTWEIGHT CNN

Marco Olivieri[⋆]*, Luca Comanducci*[⋆]*, Mirco Pezzoli*[⋆]*, Davide Balsarri*[†]*, Luca Menescardi*[†]*, Michele Buccoli*[†]*, Simone Pecorino*[†]*, Antonio Grosso*[†]*, Fabio Antonacci*[⋆]*, Augusto Sarti*[⋆]

[⋆] Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano, Milan, Italy
[†]BdSound S.r.l., Milan, Italy

ABSTRACT

The problems of speech separation and enhancement concern the extraction of the speech emitted by a target speaker when placed in a scenario where multiple interfering speakers or noise are present, respectively. A plethora of practical applications such as home assistants and teleconferencing require some sort of speech separation and enhancement pre-processing before applying Automatic Speech Recognition (ASR) systems. In the recent years, most techniques have focused on the application of deep learning to either time-frequency or time-domain representations of the input audio signals. In this paper we propose a real-time multichannel speech separation and enhancement technique, which is based on the combination of a directional representation of the sound field, denoted as beamspace, with a lightweight Convolutional Neural Network (CNN). We consider the case where the Direction-Of-Arrival (DOA) of the target speaker is approximately known, a scenario where the power of the beamspace-based representation can be fully exploited, while we make no assumption regarding the identity of the talker. We present experiments where the model is trained on simulated data and tested on real recordings and we compare the proposed method with a similar state-of-the-art technique.

***Index Terms*—** Multichannel Speech Separation, Speech Enhancement, Neural Beamformer

1. INTRODUCTION

Speech separation and enhancement refer to the tasks of suppressing background noises or interfering speakers, respectively, while leaving intact the target speech generated by the desired speaker, also denoted as Signal Of Interest (SOI). These techniques are of the object of a consistent amount of research since they are usually applied during pre-processing steps of the now ubiquitous Automatic Speech Recognition (ASR) models [1].

In recent years, deep learning-based models for speech separation and enhancement have rapidly beaten previous state-of-the-art techniques [2]. Impressive results have been obtained by applying different deep learning techniques in the field of monaural speech separation [3, 4, 5, 6] becoming the state-of-the art approach. In the multichannel scenario, instead, the more commonly used methods are still based on spatial filtering, i.e. beamforming [7]. Recently, however, several methods combine deep learning with multichannel signal processing exploiting deep neural networks in order to compute the beamformer weights [8, 9, 10, 11]. Other solutions aim at directly using the neural network in order to perform spatial processing, without explicitly modeling beamformer-like filters [12, 13, 14, 15].

More recently, another line of research follows the combination of the beamspace model of the sound field with the power of deep learning [16]. The beamspace transform, proposed in [17], projects the multichannel input signal into a set of steered directions, through a beamforming operation. The obtained feature is then used as input to a deep learning model, which is able to take advantage of the directional representation of the acoustic scene. Moreover, the main advantage of the beamspace is that when applied to different microphone array configurations, it generates representations retaining the same dimensionality, which effectively enables the training of models that are agnostic from a geometry point of view.

In this paper we propose a real-time model for speech separation and enhancement, following a similar approach to the one proposed in [17]. It is important to note that our model is extremely more lightweight with respect to the one proposed in [17], containing 33 times less parameters and being 187 times computationally cheaper. The system is able to separate the signal of the SOI from the one of other speech interferers and from both additive and diffuse noises. We consider the Direction-Of-Arrival (DOA) of the speaker to be approximately known a priori, which gives us the possibility of fully exploiting the directional representation given by the beamspace. More specifically, we propose a convolutional neural network that, taking as input few frames of the beamspace-transformed signals of a Uniform Linear microphone Array (ULA) is able to retrieve the Ideal Ratio Mask (IRM) [18]. The SOI is therefore estimated by applying the obtained IRM on the beamspace signal associated to the source DOA.

We present results, where we show that the proposed model, trained on simulated data, is able to generalize and perform separation and enhancement over real recordings. We train and test the model on multiple array geometries at the same time, showing the independence of the proposed method from the chosen setup, an extremely important characteristic for the application of deep learning-based models to real-world scenarios. We compare the obtained results with the model proposed in [16] and with the well-known superdirective beamformer [19] used to compute the beamspace.

The rest of this paper is structured as follows. In Sec. 2 we present the signal model and the necessary beamspace background. Sec. 3 describes the proposed speech separation technique, while Sec. 4 presents the results. Finally, in Sec. 5 we draw some conclusions.

2. SIGNAL MODEL AND BACKGROUND

Let us consider a ULA with I microphones having inter-sensor spacing d, acquiring the acoustic scene of a noisy and reverberant environment where J speakers are present. Then, the Short-Time Fourier

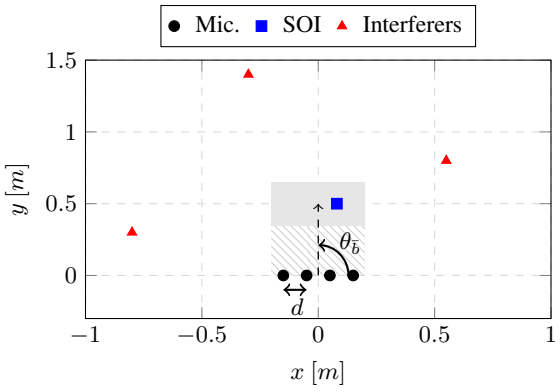

Fig. 1. Setup of the system. The ULA is composed by I microphone with distance d. The SOI is placed in a predefined region in front of the array (shown in filled grey), thus $\theta_{\bar{b}} = 90°$. Other interferers are present in the room (outside the SOI region and the one with gray lines) along with diffuse noise components.

Transform (STFT) representation of the signal acquired by the i^{th} microphone can be written as:

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

where $t = 1,\dots,T$ and $f = 1,\dots,F$ are the STFT time and frequency indexes, respectively, y_i is the signal acquired at microphone i, $h_{j,i}$ is the acoustic transfer function from source j to the i^{th} sensor, s_j is the signal emitted by the j^{th} speaker, while γ_i and v_i are the diffuse and additive noise components, respectively, measured at microphone i and $x_{j,i}$ is the reverberant speech emitted by speaker j and measured at microphone i. The overall setup of the system is depicted in Fig. 1. Notice that the SOI is placed in a region of interest in front of the array. Therefore the DOA of the SOI is approximately known in advanced as $\theta_{\bar{b}} = 90°$. Moreover, multiple interfering talkers are present in the reverberant environment along with diffuse noise components.

3. PROPOSED METHOD

In this section we describe the proposed lightweight speech separation and enhancement technique. Given that in the proposed method, we consider the DOA of the SOI to be approximately known a-priori, we exploit this information by applying to the mixture STFT the beamspace transform [17], which is based on a plane-wave decomposition of the signal in B directions $\theta_b, b = 1,\dots,B$ [20]. More specifically, the beamspace transform used in the proposed method is based on the application of a superdirective beamformer [19].

Let us define $\mathbf{Y} \in \mathbb{C}^{T\times F\times I}$ as the 3D tensor created by stacking together the STFTs of the signals acquired by the I microphones. If we define $\mathbf{W} \in \mathbb{C}^{I\times B}$ as the beamspace transform matrix, we can then compute, for each STFT frame t, the beamspace $\tilde{\mathbf{Y}}$ of the signal acquired by the microphones

$$\tilde{\mathbf{Y}}_t = \mathbf{Y}_t\mathbf{W}. \tag{2}$$

Given the desired speaker $\bar{j}$, whose DOA from the region of interest corresponds to the beamspace channel $\bar{b}$, the desired output of the network then consists of the Ideal Ratio Mask (IRM) [18] $\mathbf{M} \in \mathbb{R}^{T\times F}$. The mask is computed by considering as the target signal the one that corresponds to direction $\theta_{\bar{b}}$ of the beamspace of the matrix $\mathbf{X}_{\bar{j}}$, built by stacking together the STFTs of the signal emitted by speaker $\bar{j}$, acquired by the I microphones.

In order for the network to be both lightweight and able to operate in real-time we output the $\mathbf{M}_t \in \mathbb{R}^{1\times F}$ mask relative to one STFT frame t at a time, using as input T_{ctx} frames of $\mathbf{Y}$. More formally, the function $\mathcal{U}(\cdot)$ modeled by the proposed network can be written as

$$\hat{\mathbf{M}}_t = \mathcal{U}(\tilde{\mathbf{Y}}_{t-T_{\text{ctx}}/2:t+T_{\text{ctx}}/2}), \tag{3}$$

where $\hat{\mathbf{M}}_t$ is an estimate of the ground truth IRM mask $\mathbf{M}_t$ at frame t. Finally, an estimate $\hat{\mathbf{X}}_{\bar{j},t}$ of the desired signal $\mathbf{X}_{\bar{j},t}$ at frame t can be simply obtained through

$$\hat{\mathbf{X}}_{\bar{j},t} = \hat{\mathbf{M}}_t \odot \tilde{\mathbf{Y}}_{\bar{b},t}, \tag{4}$$

where $\odot$ denotes the Hadamard product and $\tilde{\mathbf{Y}}_{\bar{b},t}$ corresponds to channel $\bar{b}$ of the beamspace of the acquired signals pointing at $\theta_{\bar{b}} = 90°$ at frame t. The network pipeline is depicted in Fig. 2.

3.1. Network Architecture

The proposed network architecture is defined as follows. We first compute the Power Spectral Density of $\tilde{\mathbf{Y}}$ and then convert it into a 64-bands log-mel spectrogram representation, before feeding it into three convolutional blocks, each consisting of two convolutional layers. The three blocks compute the following number of feature maps: i) 16, ii) 32, iii) 64. While the first convolutional layer in each block has stride $(1,1)$, the second one has an asymmetric stride $(1,2)$ in order to compress the representation along the frequency axis. We apply no padding along the time axis, while we adopt a symmetric padding for what concerns the frequency axis, in order to avoid spurious artefacts. All convolutional layers have kernel size (3×3), and are followed by Batch Normalization and a ReLU activation. The output of the convolutional blocks is averaged along the time axis and is then flattened before being fed into a fully connected layer with 64 neurons, followed by Batch normalization and ReLU. Finally the output $\mathbf{M}_t$ is obtained through a fully connected layer with F neurons, followed by a sigmoid activation.

3.2. Training Procedure

During the training phase, the beamspace-transformed mixture $\tilde{\mathbf{Y}}$ is computed along with the ground truth IRM mask $\mathbf{M}$. Then for a STFT frame t the corresponding T_{ctx} frames are extracted. Finally, the estimated mask $\hat{\mathbf{M}}_t$ is used in the loss computation at frame t as

$$\mathcal{L}(t) = \frac{1}{F}\sum_{f=1}^{F}(\mathbf{M}_{t,f} - \hat{\mathbf{M}}_{t,f})^2, \tag{5}$$

where the batch index is omitted for simplicity.

4. RESULTS

In this Section, we present results obtained with real recordings in order to show the speech separation capabilities of the proposed model and we compare them with the Neural Beamspace-Domain Filter (NBDF) [1] method proposed in [16] and with the beamformer used

[1] https://github.com/lucacoma/NeuralBeamspaceDomainFilter

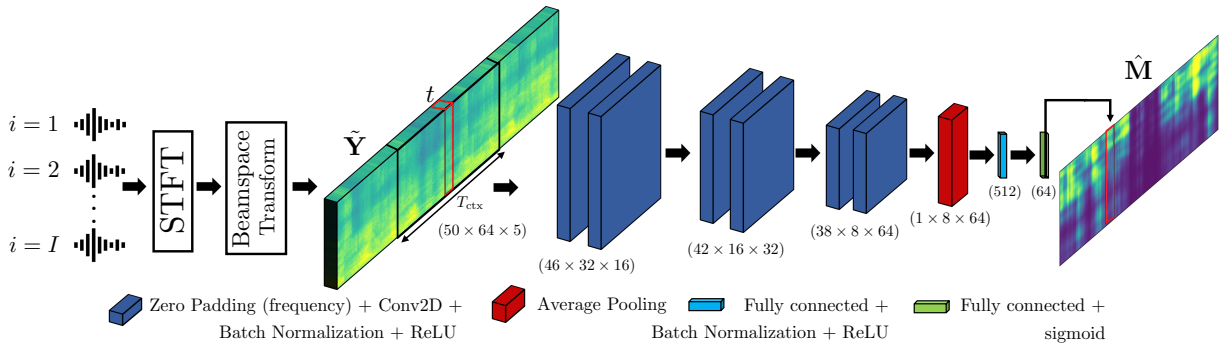

Fig. 2. Schematic representation of the proposed speech separation and enhancement pipeline operating at frame t.

to compute the beamspace fed as input to the proposed model. Audio examples are contained in the accompanying website [2].

4.1. Training Setup

The signal model reported in Eq. (1) considers a recorded sound scene consisting of J different speakers. Among all possible configurations of the J speakers we also consider two possibilities, namely having either a SOI only or no SOI at all. When the SOI is present in the acoustic scene, the number of sources considered in Eq. (1) is $J = R + 1$, and the objective is to extract the single SOI from the R interfering speakers, where $R \in \{0, \dots, 4\}$. Instead, if no SOI is active in the scene, then $J = R$ and the estimated IRM should completely remove all the R active interferers. We considered a wide range of acoustic conditions in which the overall system is required to operate. The training set has been generated through an extensive simulation campaign by sampling with a uniform distribution the operational ranges of each room parameter and considering five different array configurations, for a total of 10 000 different simulated rooms. We used gpuRIR [21] to compute the Room Impulse Responses (RIRs) of rectangular rooms with dimensions $L_x \in [3,\ 8]$m, $L_y \in [3,\ 8]$m, $L_z \in [2.6,\ 4]$m, and reverberation time $\mathrm{T}_{60} \in [0.2,\ 1.4]$s. We considered different ULA configurations by changing I and d parameters, while keeping fixed the elevation along the $z-$axis to 1.3 m. In particular, we used the following ULA configurations: $I = 3$ and $d = 20\,\text{mm}$ or $30\,\text{mm}$, $I = 4$ and $d = 20\,\text{mm}$ or $d = 30\,\text{mm}$ and $I = 4$ and $d = 26\,\text{mm}$. In order to take into account the calibration error related to the microphone positioning, we applied a random error by randomly sampling uniform distributions defined in the ranges $\pm 3\,\text{mm}$, $\pm 0.5\,\text{mm}$, and $\pm 1\,\text{mm}$ for the x, y, z coordinates of the sensors, respectively. The SOI was randomly placed in the region of interest as shown in Fig. 1. We defined the area of this region as $x_{\text{SOI}} \in [-0.2,\ 0.2]$m, $y_{\text{SOI}} \in [0.35,\ 0.65]$m, and $z_{\text{SOI}} \in [1.3,\ 1.9]$m. The microphone signals are then computed as the convolution between the RIRs and clean speech signals extracted from the LibriSpeech corpus [22] both for SOI and interferers. In the 80 % of the generated scenarios the SOI is present while in the remaining 20 % there are only interferers. When both SOI and interferers are active we applied different Signal-to-Interferences Ratio (SIR), with $\text{SIR} \in [-3,\ 3]$dB, to simulate the SOI loudness with respect to the acoustic energy of all the interferers. We simulated diffuse sound field components of different noises, e.g., babble speech and canteen noise, with the algorithm presented in [23] by varying the Signal-to-Diffuse Ratio (SDR) [24], with $\text{SDR} \in [-3,\ 60]$dB. Finally, in order to simulate different microphone arrays we varied the Signal-to-Noise Ratio (SNR) of the sensors, $\text{SNR} \in [30,\ 70]$dB, and the array gain $\text{G} \in [-40,\ -1]$dB, which serves the purpose of modeling the possible different dynamics of the array recordings. The beamspace transform matrix $\mathbf{W}$ used in (2) is composed by $B = 5$ predefined directions. Given the center of the ULA as reference, we steered the beamformer to $\theta_b = \{0°, 45°, 90°, 135°, 180°\}$. Therefore, $\bar{b} = 3$ is the desired beamspace channel corresponding to the SOI at $\theta_{\bar{b}} = 90°$. We applied a 512 point STFT resulting in $F = 257$, with 16 ms of Hamming window, 50 % overlap and 16 kHz of sampling frequency. The input of the network consisted of slices of length $T_{\text{ctx}} = 50$ frames, resulting in tensors of dimensionality $50 \times 257 \times 5$, corresponding to a total of 400 ms splitted into 200 ms for both the look-ahead and past frames. The proposed network was trained for approximately 300 epochs, using the Adam optimizer [25], with a learning rate of $1e-3$, following an exponential decay schedule.

4.1.1. Baseline and computational complexity

We used the same audio pipeline for training the NBDF baseline, which outputs an STFT of the same dimensionality of the input. The NBDF baseline was trained for 120 epochs (60 for each module) using the Adam optimizer with a learning rate of $5e-4$, which was halved after two consecutive epochs with no validation loss improvement, as in [16]. The proposed network consists of roughly $120,000$ parameters, and requires 1.06 millions of MACs per frame, while NBDF consists of $4,006,236$ parameters and requires 198.5 millions of MACs per frame. Since we have 125 frames per second, NBDF requires 24.81 billions of MACs per second, while the proposed network requires 132.5 million of MACs per second, about 187 times computationally cheaper.

4.2. Metrics

The following metrics have been used in order to evaluate the separation and enhancement performances, namely: Signal-to-Interferences Ratio (SIR), Signal-to-Artifacts Ratio (SAR), Signal-to-Distortion Ratio (SDR) [26], Perceptual Evaluation of Speech Quality (PESQ) [27], and Extended Short-Time Objective Intelligibility (ESTOI) [28]. We also defined the rejection of the signal energy in decibel as R_{soi} and $\mathrm{R}_{\text{interf}}$ for the cases when only the SOI or only interferers are present in the room, respectively. More specifically, let us denote E_{soi} and E_{interf} as the signal energies coming from the beamformer pointing to $\theta_{\bar{b}} = 90°$ when only the SOI or only interferers and noise are present, respectively, and $\hat{E}$ the energy of the signal estimate coming from network, then

$$\mathrm{R}_{\text{soi}} = 10\,\log_{10}\left(\frac{\hat{E}}{E_{\text{soi}}}\right), \mathrm{R}_{\text{interf}} = 10\,\log_{10}\left(\frac{\hat{E}}{E_{\text{interf}}}\right). \quad (6)$$

[2] https://polimi-ispl.github.io/beamspace_cnn_speech_separation.github.io/

ULA setups	$I = 4$, $d = 26\,\text{mm}$			$I = 3$, $d = 52\,\text{mm}$			$I = 4$, $d = 52\,\text{mm}$			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.

Notice that, although the optimal R_{soi} corresponds to $0\,\text{dB}$, this condition cannot be practically achieved due to the presence of sensor noise in E_{soi}. On the contrary, R_{interf} should be as low as possible, ideally $-\infty$ in the optimum working condition, since the network should completely attenuate all interfering sources and noises.

4.3. Evaluation

The test set was built by measuring RIRs in an acoustically treated room, according to the standard ETSI ES 202 396-1, with dimensions $L_x = 6\,\text{m}$, $L_y = 4.8\,\text{m}$, $L_z = 2.6\,\text{m}$, and $\text{T}_{60} = 0.8\,\text{s}$. We evaluate the performances on three test sets with different array configurations. We consider an array configuration with $I = 4$ and $d = 26\,\text{mm}$, seen also during training. Moreover, to assess the system robustness we consider two array configurations unseen during training, with $I = 3$, $I = 4$ and $d = 52\,\text{mm}$. The audio pipeline is the same as the one used for the training procedure. Therefore, we compare the results of the proposed method with the NBDF approach [16] and the beamformer steered to $\theta_{\bar{b}} = 90^\circ$, denoted as $\tilde{\mathbf{Y}}_{90^\circ}$. Notice that $\tilde{\mathbf{Y}}_{90^\circ}$ corresponds to comparing the solution with the beamformer used for computing the input beamspace filter. In Table 1 we report the average metrics obtained by applying the proposed method and the baselines to the three different datasets. In general, the proposed approach is able to extract the SOI speech in all the test sets, hence proving that the network is independent from the array configuration used to record the acoustic scene. By inspecting Table 1, we can notice that the proposed method outperforms the beamformer $\tilde{\mathbf{Y}}_{90^\circ}$ when applied to all the test sets. In particular, when using the proposed method, the SIR and SDR increase of more than $5.5\,\text{dB}$ and $1.3\,\text{dB}$, respectively, and in average by $7.13\,\text{dB}$ and $2.62\,\text{dB}$, with respect to $\tilde{\mathbf{Y}}_{90^\circ}$. SAR values for $\tilde{\mathbf{Y}}_{90^\circ}$ have not been reported, since they tend to infinity. In fact the beamspace processing applied to the mixture and the desired target is the same, so no additional artifacts are added. When applied to the test sets unseen during training, NBDF reaches higher SIR than the proposed method, with an increment limited to $1.98\,\text{dB}$ and $4.37\,\text{dB}$ for the datasets with $I = 3$ and $I = 4$, respectively. However, this comes at the cost of more distortions in the estimates, that reduce the SAR value by $4\,\text{dB}$ and the SDR value by $2.2\,\text{dB}$, on average. In Fig. 3 we report the average SDR as a function of the number of interferers R. It is worth noticing that SDR incorporates both SIR and SAR [26]. As expected, the performances of all methods decrease as the number of interferers increases. However, the SDR of the proposed method is steadily above NBDF and $\tilde{\mathbf{Y}}_{90^\circ}$. Inspecting the PESQ and ESTOI values in Table 1 we can notice similar results for all the three approaches and for the different test sets. In general both PESQ and ESTOI present moderate results due to the fact that we tested on real data, while the networks have been trained on simulations. Notice that, $\tilde{\mathbf{Y}}_{90^\circ}$ reaches the best PESQ results due to the limited filtering process responsible of signal distortions. However, the general perceptual intelligibility of the devised solution outperforms both NBDF and the input beamformer $\tilde{\mathbf{Y}}_{90^\circ}$. Indeed, although the use of the beamspace is similar to [16], the network architectures are very different in terms of layers and overall dimensions. In fact the proposed method has less than 30 times the number of parameters used in NBDF. Thus, we obtain better generalization results using an extremely more lightweight model, which is a highly sought after characteristic for the application of such models to real-world hardware devices. As far as the scenarios when only the SOI is active, results show that the system is able to let unfiltered the SOI signal with no distortion while attenuating the background noises. As a matter of fact, we achieve an average $\text{R}_{\text{soi}} = -4.67\,\text{dB}$ for the three test sets. On the other hand, when only interferers are present in the rooms, the network correctly suppresses the energy when filtering the recordings, obtaining an estimate close to a signal consisting of silence. For these cases, we achieved a mean R_{interf} grater than $-14.31\,\text{dB}$ for all the test sets. Notice that R_{soi} and R_{interf} cannot be computed for $\tilde{\mathbf{Y}}_{90^\circ}$ since their values are used as reference in the definition (6).

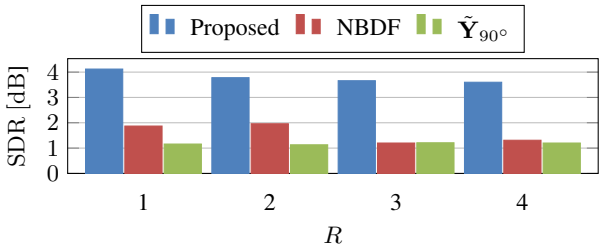

Fig. 3. Average SDR as function of the number of interferers R.

5. CONCLUSIONS

In this paper we proposed a lightweight CNN architecture for speech separation and enhancement of a main talker placed in a region of interest in noisy and reverberant environments. The system is able to work in real time and is independent from the geometry of the array, in terms of number of microphones and inter-sensor distance, thanks to the adoption of the beamspace representation of the sound field. To prove the effectiveness of the proposed method, we present results where the network is trained on simulated data generated with different array configurations and tested on real data. We compared the proposed approach with respect to the beamformer used to compute the input beamspace and with a recently proposed approach based on the beamspace domain. Results show the superiority of the devised approach and its ability to generalize to different setups.

6. REFERENCES

[1] Z.-Q. Wang and D. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust asr," in *2018 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 5709–5713, IEEE, 2018.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Int. Conf. on Acoust., speech and signal Process. (ICASSP)*, pp. 31–35, IEEE, 2016.

[4] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 241–245, IEEE, 2017.

[5] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700, IEEE, 2018.

[6] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 21–25, IEEE, 2021.

[7] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[8] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 196–200, IEEE, 2016.

[9] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks.," in *Proc. Interspeech*, pp. 1981–1985, 2016.

[10] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[11] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," in *Intl. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, pp. 6089–6093, IEEE, 2021.

[12] R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 7319–7323, IEEE, 2020.

[13] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 6389–6393, IEEE, 2020.

[14] D. Markovic, A. Defossez, and A. Richard, "Implicit Neural Spatial Filtering for Multichannel Source Separation in the Waveform Domain," in *Proc. Interspeech*, pp. 1806–1810, 2022.

[15] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 6487–6491, IEEE, 2022.

[16] W. Liu, A. Li, X. Wang, M. Yuan, Y. Chen, C. Zheng, and X. Li, "A neural beamspace-domain filter for real-time multi-channel speech enhancement," *Symmetry*, vol. 14, no. 6, p. 1081, 2022.

[17] S. Lee, S. H. Park, and K.-M. Sung, "Beamspace-domain multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Process. Letters*, vol. 19, no. 1, pp. 43–46, 2011.

[18] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. on audio, speech, and language Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[19] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 35, no. 10, pp. 1365–1376, 1987.

[20] M. Pezzoli, J. J. Carabias-Orti, M. Cobos, F. Antonacci, and A. Sarti, "Ray-space-based multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Process. Letters*, vol. 28, pp. 369–373, 2021.

[21] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE Int. Conf. on Acoust., Speech and signal Process. (ICASSP)*, pp. 5206–5210, IEEE, 2015.

[23] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, 2008.

[24] O. Thiergart, T. Ascherl, and E. A. Habets, "Power-based signal-to-diffuse ratio estimation using noisy directional microphones," in *Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 7440–7444, IEEE, 2014.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on audio, speech, and language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE Int. Conf. on Acoust., Speech, and Signal Process. Proc.*, vol. 2, pp. 749–752, IEEE, 2001.

[28] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.