



UNIVERSITÀ  
DEGLI STUDI  
DI UDINE

## Università degli studi di Udine

### Exploiting CNNs for Improving Acoustic Source Localization in Noisy and Reverberant Conditions

*Original*

*Availability:*

This version is available <http://hdl.handle.net/11390/1129835> since 2021-03-27T13:03:21Z

*Publisher:*

*Published*

DOI:10.1109/TETCI.2017.2775237

*Terms of use:*

The institutional repository of the University of Udine (<http://air.uniud.it>) is provided by ARIC services. The aim is to enable open access to all the world.

*Publisher copyright*

(Article begins on next page)

# Exploiting CNNs for Improving Acoustic Source Localization in Noisy and Reverberant Conditions

Daniele Salvati, Carlo Drioli, *Member, IEEE*, and Gian Luca Foresti, *Senior Member, IEEE*

**Abstract**—The paper discusses the application of convolutional neural networks (CNNs) to minimum variance distortionless response (MVDR) localization schemes. We investigate the direction of arrival (DOA) estimation problem in noisy and reverberant conditions using an uniform linear array (ULA). CNNs are used to process the multichannel data from the ULA and to improve the data fusion scheme which is performed in the steered response power (SRP) computation. CNNs improve the incoherent frequency fusion of the narrowband response power by weighting the components, reducing the deleterious effects of those components affected by artifacts due to noise and reverberation. The use of CNNs avoids the necessity of previously encoding the multichannel data into selected acoustic cues with the advantage to exploit its ability in recognizing geometrical pattern similarity. Experiments with both simulated and real acoustic data demonstrate the superior localization performance of the proposed SRP beamformer with respect to other state-of-the-art techniques.

**Index Terms**—Convolutional neural networks, source localization, direction of arrival estimation, broadband steered response power, acoustic analysis, microphone array.

## I. INTRODUCTION

Multichannel audio processing techniques have been broadly investigated in teleconferencing systems, audio surveillance, autonomous robots, human-computer interaction, and have a central role in a number of applications related to the acoustic analysis and speech technology area. Within the research on acoustic sensor arrays, spatial localization of acoustic sources and active speakers has certainly received large attention, and baseline techniques are now available that offer appreciable performances in a wide number of real-world conditions, including indoor/outdoor scenarios, reverberant and noisy environment, near-field/far-field monitoring [1]–[8]. In general, the localization can be performed by indirect and direct methods. The indirect (two-step) approach computes a set of time difference of arrivals (TDOAs) using measurements across various combinations of microphones [9], [10], and then estimates the source position using geometric considerations [11], [12]. Direct methods are based on the steered response power (SRP) beamformers [1], [13], [14], on subspace algorithms [15]–[17], or on maximum-likelihood estimators [18]–[20].

Most of the aforementioned methods can be designed to act selectively on a limited frequency range (narrowband

beamformer), while their broadband frequency range version can be obtained by fusing the narrowband components in an opportune manner. In this paper, we propose a SRP scheme which employs convolutional neural networks (CNNs) [21], [22] to refine the frequency-domain multichannel fusion operation of the minimum variance distortionless response (MVDR) beamformer [23] by learning how to opportunistically weight the narrowband components. It is shown that this approach improves the localization of acoustic sources and speakers in noisy and reverberant conditions. The novel convolutional-based scheme also contributes to better investigate the structure of acoustic cues from the multichannel spectral densities. In the SRP-weighted MVDR (SRP-WMVDR) schemes proposed previously in [24]–[26], these computations relied on a pre-processing stage in which the multichannel acoustic input was transformed into a set of cues serving as input to the machine learning component. The idea of selecting or weighting the MVDR components was proposed in [24], where a radial basis function network (RBFN) was used as narrowband frequency components classifier, using the marginal distribution of the narrowband components as input. The approach in [24] was extended in [25], [26], in which a support vector machine (SVM) learning component was used. This scheme, which used a different set of input features based on marginal distributions of the acoustic data, proved to outperform the RBFN-based one.

We extend here the hybrid beamforming-plus-machine learning approach to the use of CNNs, whose principal advantage is to avoid the explicit selection and computation of a set of acoustic cues since these are effectively computed by the convolutional layers of the network. With respect to previous research in the field, the paper addresses for the first time the exploitation of convolutional features in the context of multichannel audio processing for acoustic source localization purposes. We study the integration of CNNs in the signal processing chain on which the acoustic localization problem is based and present a new algorithm, referred in the following to as SRP-WMVDR-CNN.

The algorithm is presented in two variants, the first one based on a classification CNN, and the second one based on a regression CNN. In the classification-oriented scheme, the CNN is trained to classify the narrowband SRPs into two classes: constructively contributing SRPs vs. disruptively contributing ones. In the information fusion step, which sums up the contribution of each narrowband SRP, the latter are discarded. In the regression-oriented scheme, the CNN is trained to provide the weighting coefficients of an improved SRP fusion function, which weights the contribution of each narrowband SRP while adding it to the sum of contributions.

Copyright (c) 2016 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

D. Salvati, C. Drioli, and G.L. Foresti are with the Department of Mathematics, Computer Science and Physics, University of Udine, Udine 33100, Italy, e-mail: daniele.salvati@uniud.it, carlo.drioli@uniud.it, gianluca.foresti@uniud.it.

As for the acoustic setting, we consider the far-field direction of arrival estimation (DOA) problem of a single source in noisy and reverberant conditions, using a uniform linear array (ULA).

Applications of this scenario include videoconferencing systems [27], in which the estimation of sound coordinates can be used to automatically steer a videocamera towards an active speaker; human-computer interaction systems [28], in which localization and beamforming are used to enhance the signal and improve audio recognition; or even multimedia interactive systems for performing arts, in which acoustic source localization can be integrated into digital musical interfaces and used for performance control [29]. The method can be extended in principle to a multiple-source scenario, which would require to improve the routine devoted to peak searching in the acoustic response power.

## II. RELATED WORKS

### A. Conventional methods for DOA estimation

The DOA estimation problem concerns the processing of acoustic data collected by a microphone array with the aim of obtaining information on the direction from which the acoustic source signal originates. At today, the methods for DOA estimation can be broadly classified in two classes: TDOA-based indirect methods, and direct methods. The indirect methods aim at estimating the time difference of the acoustic wavefront arrivals between microphone pairs and then the DOA using geometric considerations [10], [30]–[33]. The generalized cross-correlation (GCC) [9] is considered a baseline practical method for TDOA estimation, but often improved versions are used in practice. The multichannel cross-correlation coefficient (MCCC) [34], for example, is based on TDOAs estimation obtained by the GCC paired with a prediction of the spatial error to provide a more robust estimate of the DOA. Direct methods, on the other hand, estimate the DOA of an acoustic source in a single step by exploiting some power density function representing the spatially-relevant information distribution, and they are considered in general more robust under noisy and reverberation conditions if compared to the TDOA-based methods. The SRP localization involves computing the output power of a beamformer steered towards each DOA of interest. The conventional SRP is performed with the delay and sum beamformer [35], which consists in the synchronization of the array signals to steer the array in a certain direction, and of summing the signals to estimate the power of the spatial filter. The SRP phase transform (SRP-PHAT) [13] is a widely used filtered SRP beamforming. The PHAT filter [9] assigns equal importance to each frequency by dividing the spectrum by its magnitude. The SRP-PHAT can be efficiently computed by the global coherent field (GCF) [36] approach, that coherently sums the GCC-PHAT from the microphone pairs for each possible point of interest. Among conventional beamformers, the MVDR [23] filter is a well-known data-dependent beamformer that provides better resolution if compared to the conventional beamformer. Both MVDR and SRP localization have been described as maximum-likelihood problems in [18]–[20]. Yet another class of high resolution methods is based on subspace

analysis and decomposition. The multiple signal classification (MUSIC) method [15] exploits the subspace orthogonality property to build the spatial spectrum and to localize the DOA sources. The estimation of signal parameters via rotational invariance techniques (ESPRIT) is also based on subspace decomposition exploiting the rotational invariance [16], [37].

### B. Machine learning methods for multichannel processing

Since many decades, machine learning and neural network methods have been successfully employed in a wide range of speech and audio processing applications, such as automatic speech recognition (ASR) [38]–[41], audio forensic [42], music information retrieval [43], [44], sound classification [45]. However, their use for the improvement or the new design of multichannel processing localization schemes has been explored only recently [25], [26], [46], [47]. Moreover, since the new computational and performance advances brought by the recent developments in the field of deep neural networks (DNNs) research, their use is now being investigated in a variety of acoustic and speech oriented applications involving multichannel processing, including in a few cases the specific problem of acoustic source localization. To date, the application of DNNs to multichannel processing problems has focused principally on ASR [28], [48], speech enhancement [49], acoustic source separation [50], and acoustic source localization [51]. In [28], a DNN-based feature enhancement method using multichannel inputs is proposed for robust ASR. The multichannel information is used in the pre-enhanced spectral features that are obtained by DOA-constrained independent component analysis. In [49], multichannel speech enhancement is addressed, and beamforming based enhancement is achieved by time-frequency (T-F) masking. The algorithm combines single- and multi-microphone processing, in which a DNN is trained to map the spectral features to a T-F mask, which is used in turn to calculate the noise covariance matrix and the steering vectors related to the speaker position. The steering vectors are then used to enhance the speech signal coming from the speaker position through an MVDR beamformer. Based on these steps, the method iterates masking and beamforming, and its application to ASR shows improved performance over state-of-the-art recognition. Note that T-F masking beamforming has been previously addressed by supervised and unsupervised machine learning methods [46], [47]. In [46], a mask is obtained by an unsupervised spatial vector clustering. A speech spectral model based on a complex Gaussian mixture model is designed to estimate the T-F masks and the steering vectors related to the speaker position.

While in the aforementioned cases source localization is subordinate to other signal processing tasks, such as ASR or speech enhancement, the research in [51] especially addresses the localization problem of a single sound source. This approach is based on a discriminative machine learning to compute the location estimator in the frequency domain, in which a DNN encodes the steering vectors by applying the orthogonality principle used in the MUSIC method [15]. The eigenvectors of power spectral density matrices are treated as the input vector by constructing directional image activators,

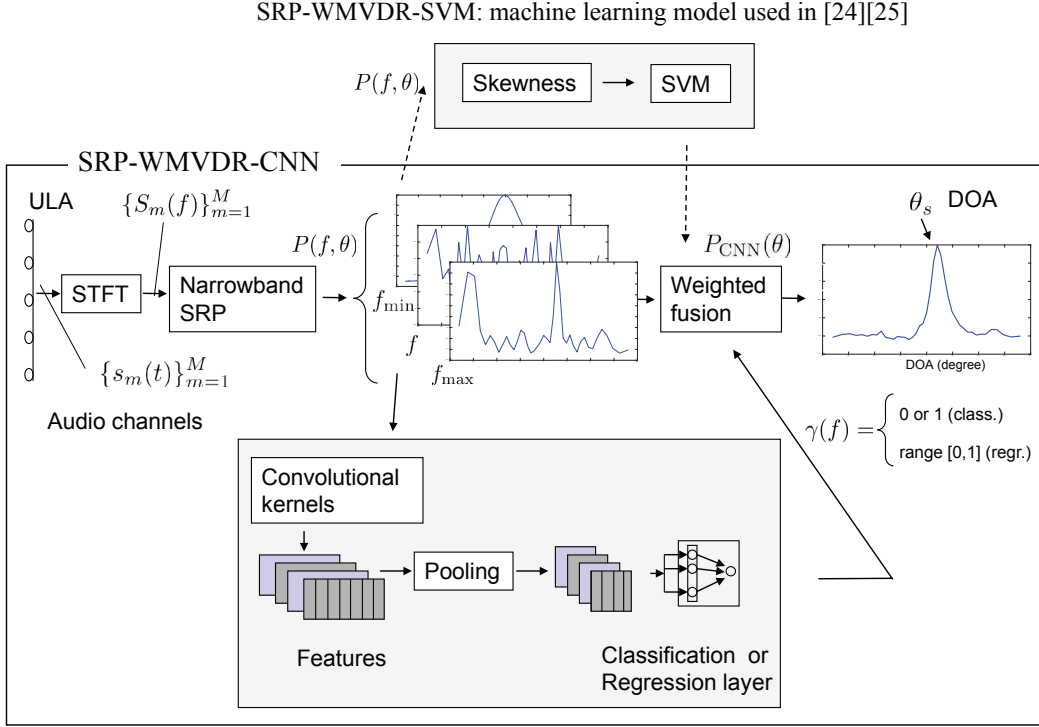


Fig. 1. System overview: signal processing and information flow.

whose relationships with the source DOAs are learned in turn by a DNN. Unfortunately, the authors state that their DNN-based method resulted ineffective in noisy and reverberant conditions, and did not result in significant localization performance improvements.

Recently, we have discussed a scheme which employs a machine learning component to refine the multichannel fusion scheme and improves the localization of acoustic sources and speakers in near-field noisy and reverberant conditions [24], [25] and far-field noisy condition [26]. These investigations underline the importance of the way in which broadband fusion of narrowband components is performed, and the usefulness of exploiting the knowledge on which components contribute constructively to the localization and which do not. These computation schemes rely, however, on a preprocessing stage in which the multichannel acoustic input is transformed into a set of cues serving as input to the machine learning component (i.e., the skewness, the kurtosis, the crest factor, and the marginal distribution of the acoustic input data were used). We extend here the hybrid beamforming-plus-machine learning approach to the use of CNNs, whose principal advantage is to exploit its ability in recognize geometrical pattern similarity and to avoid the explicit selection and computation of a set of acoustic cues since these are effectively computed in the CNN layers.

### III. ACOUSTIC SOURCE LOCALIZATION BASED ON CNNs

The signal processing pipeline structure is illustrated in Figure 1. The middle-part of the scheme describes the acoustics based processing steps, including the short-time Fourier

transform (STFT) of the multichannel input  $s_m(t)$  ( $m = 1, 2, \dots, M$ , where  $M$  is the number of microphones), the frequency bin-dependent narrowband SRP,  $P(f, \theta)$  (where  $f$  is the frequency bin and  $\theta$  is the DOA), and the enhanced fusion step used to build the final broadband acoustic map,  $P_{\text{CNN}}(\theta)$ , by exploiting the weighting information provided by the CNN output. The input to the CNNs is provided by the narrowband SRP components. The lower part of the scheme contains a convolutional layer and a pooling layer, followed by an output layer, i.e. a classification or regression fully connected NN layer. The upper processing path in Figure 1 shows the SVM based scheme proposed in [25], [26], and used here only for comparison. Note that the upper SVM processing path is sketched with dashed lines to emphasize the fact that it is not part of the presently proposed algorithm, and it does not used in conjunction with the lower CNN based processing path.

#### A. Acoustic Localization Elements

Beamforming methods search for the maximum of the SRP functions computed from the output of the sensor array. Straightforward calculation can be achieved through a delay-and-sum procedure in the time domain [35]. However, for computational efficiency the broadband SRP is typically computed in the frequency-domain by calculating the power spectral density (PSD) matrix and the narrowband SRP on each frequency bin, and by finally fusing these narrowband responses. The PSD at frequency  $f$  for the looking direction



$\theta^1$  can be written as

$$P(f, \theta) = E[|\mathbf{w}^H \mathbf{s}(f)|^2] = \mathbf{w}^H(f, \theta) \Phi(f) \mathbf{w}(f, \theta), \quad (1)$$

where  $\mathbf{w}(f, \theta)$  is the weighting and steering vector,  $\mathbf{s}(f) = [S_1(f), S_2(f), \dots, S_M(f)]$  is the sensor array output in the frequency domain,  $H$  denote the conjugate transpose, and  $\Phi(f) = E[\mathbf{s}(f)\mathbf{s}^H(f)]$  is the symmetric and positive definite PSD matrix ( $E[\cdot]$  denotes here the mathematical expectation).

Throughout this paper, we will make use of the specific class of MVDR beamformers [23], whose narrowband PSD has the following form:

$$P(f, \theta) = \frac{1}{\mathbf{a}^H(f, \theta) \Phi^{-1}(f) \mathbf{a}(f, \theta)}, \quad (2)$$

where  $\mathbf{a}(f, \theta)$  is the steering vector, i.e. the set of phase delays affecting a plane wave when it reaches each sensor in the array. In the far-field, the array steering vector is defined for the ULA as

$$\mathbf{a}(f, \theta) = [1, e^{-j\frac{2\pi f \tau(\theta)}{L}}, \dots, e^{-j\frac{2\pi f (N-1)\tau(\theta)}{L}}]^T, \quad (3)$$

where  $L$  is the size of the DTFT,  $j$  is the imaginary unit, and  $(n-1)\tau(\theta)$  is time difference of arrival (TDOA) between the  $n$ th and reference microphone. The relationship between the TDOA  $\tau(\theta)$  and the DOA  $\theta$  is given by

$$\tau(\theta) = \frac{d \sin(\theta)}{c}, \quad (4)$$

where  $c$  is the speed of sound and  $d$  is the inter-microphone distance of the ULA. The fusion of these narrowband PSDs to obtain the SRP-MVDR beamformer is then computed as the sum of all the frequency bin components, i.e.  $P(\theta) = \sum_{f=0}^{f=L-1} P(f, \theta)$ . Usually, however, some sort of normalization is operated on the components before the fusion, since the normalization has the beneficial effect of increasing the spatial resolution of the beamformer. Example of such beamformers are the delay-and-sum SRP phase transform (SRP-PHAT) [13], in which the normalization is achieved discarding the magnitude and only keeps the phase of the PSD matrix, or the SRP normalized MVDR (SRP-NMVDR) [52], where each PSD component is normalized by the maximum value of the PSDs for that frequency with respect to all possible DOAs. The normalization is known to improve the spatial resolution of the beamformer, however it also emphasizes the noise at those frequency components with low signal-to-noise ratio (SNR), causing localization errors and performance degradation.

To avoid to use the disruptive information provided by such components, especially for localization in reverberant and noisy environments, the narrowband SRP components are weighted in the fusion process [24]–[26]. In order to improve the localization performance, we further extend here this weighting concept, and define the SRP output as a weighted sum of narrowband components:

$$P_{\text{CNN}}(\theta) = \sum_{f=f_{\min}}^{f_{\max}} \gamma(f) \frac{P(f, \theta)}{\max_{\theta} [P(f, \theta)]} = \sum_{f=f_{\min}}^{f_{\max}} \gamma(f) \bar{P}(f, \theta), \quad (5)$$

<sup>1</sup>We address here the far-field localization problem, where the searched information is the DOA of the acoustic source. The discussion can be extended to the near-field case by substituting to the DOA  $\theta$  a position  $\mathbf{r}$  in the search space.

where  $f_{\min}$  and  $f_{\max}$  denote the frequency range of the broadband source,  $\bar{P}(f, \theta)$  are the normalized narrowband SRPs,  $\gamma(f)$  are weighting factors provided by the output of the CNN component. They might assume the values 0 or 1 in the classifier-CNN configuration, or might assume values in the range  $[0, 1]$  in the regression-CNN configuration.

The DOA estimation of the acoustic source is computed by a maximum search procedure on the  $P_{\text{CNN}}(\theta)$  function, i.e.

$$\hat{\theta}_s = \underset{\theta}{\operatorname{argmax}} [P_{\text{CNN}}(\theta)]. \quad (6)$$

In the new scheme based on CNNs, the input features to the classification or regression layer are computed by the convolutional components, thus avoiding the problem of searching for the best feature class, and providing a new class of features especially suited for the specific task. Moreover, in the regression-CNN configuration, the information provided by a given frequency component is weighted in order to prevent the use of incorrect information taking into account the errors of machine learning component, in particularly when the training and testing conditions considerably change.

## B. CNN-based Component

The overall structure of the CNN component is made by a convolution-pooling hidden layer, followed by a fully-connected layer. The input to the CNN is provided by the low-level narrowband normalized SRP  $\bar{P}(f, \theta)$ , which is encoded as a b/w image  $\mathbf{V}$  as follows exploiting the ability of CNNs to recognize geometrical similarity patterns without being affected by their position nor by small distortions of their shapes [21], [22]. For a given inter-microphone distance  $d$ , the set of distinct discrete TDOA of  $\tau(\theta)$  values will have cardinality  $T = 2 \lfloor \frac{df_s}{c} \rfloor + 1$ , where  $\lfloor \cdot \rfloor$  denotes the floor function that maps a real number to the largest previous integer and  $f_s$  is the sampling frequency. Therefore, we have that the input matrix  $\mathbf{V}$  will have dimension  $T \times T$ , its element  $v_{ij}$ ,  $i = 1, 2, \dots, T$ , will be set to 255 if  $j = \lfloor \bar{P}(f, i)T \rfloor$ , otherwise it will be set to 0. This operation allows to transform the mono-dimensional output power of the ULA into a two-dimensional input, encoding an image-like representation of the SRP function, thus emphasizing the shape-oriented nature of the processing which occurs in the subsequent CNN layers. Note that by using the mono-dimensional input we cannot identify the shape of SRPs. Figure 2 shows some examples of the input  $\mathbf{V}$ , each one representing a narrowband SRP at a different frequency, for an ULA with inter-microphone distance of 0.15 m and a sample frequency of 48 kHz. The frequencies were chosen arbitrarily among those classified positively (upper plots) and those classified negatively (lower plots) for that frame.

This input raw data undergoes a filtering and activation detection step operated through the convolutional layer kernel  $\mathbf{W}$ , as

$$h = \sigma(\mathbf{W} * \mathbf{V} + b), \quad (7)$$

where  $\mathbf{W}$  is a trained kernel,  $b$  is a bias parameter, and  $\sigma(\cdot)$  is the activation function. We use here the rectified linear unit (ReLU) [53] for generating the output of the convolutional

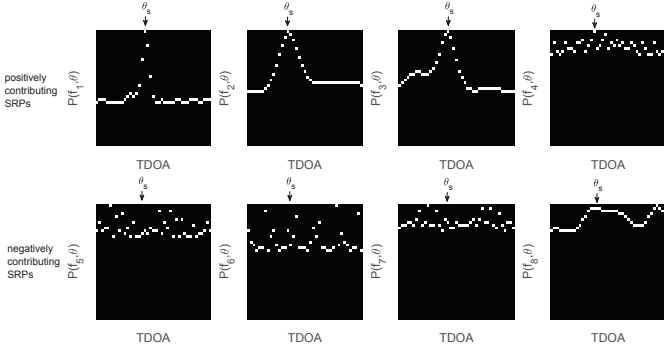


Fig. 2. Examples of the input matrix  $\mathbf{V}$  encoding the narrowband SRPs. The plots are related to a signal frame from a sound source impinging the array with a DOA  $\theta_s = -11.72$  degree, simulated with an ULA with inter-microphone distance of 0.15 m and a sample frequency of 48 kHz. Each plot represents a different frequency, chosen arbitrarily among those classified positively (upper plots) and those classified negatively (lower plots).

layer. The bias guarantees that every node has a trainable constant value. The kernels are computed through a stochastic gradient descent method [54], which minimizes a loss function measuring the discrepancy between the CNN prediction and the target. The loss function for classification is cross entropy [55] and for regression is mean squared error. Next, the pooling layer operates a dimensionality reduction through an averaging or maximizing operation with respect to the two dimensions of the feature. In this work, we adopt a max-pooling layer [56]. The output of the convolutional-pooling layer is then used as the input of the final fully connected layer, in which each neuron is connected to all neurons of the previous layer.

The CNN must be trained using a supervised procedure, based on a set of known target  $\theta_t$  DOAs. This step is achieved by computing the contribution of each frequency component to the global localization error. If

$$\hat{\theta}(f) = \underset{\theta}{\operatorname{argmax}}[P(f, \theta)] \quad (8)$$

is the source DOA estimate based only on the component related to frequency  $f$ , the contribution of this frequency to the localization error is

$$\Omega(f, \theta_t) = |\theta_t - \hat{\theta}(f)|. \quad (9)$$

The localization error is then used to build the output training values of the CNN model, as follows:

1) *Classifier-based Configuration:* In the classifier-based configuration, the last fully connected layer combines convolutional features to classify the input as 0 or 1. The activation function used in fully connected classification layer is the softmax function [57]. The classifier is trained to remove those narrowband components which contribute negatively to the localization. Namely, consider the  $i$ -th input  $\mathbf{V}_i$ , the  $i$ -th training set output  $\bar{\gamma}_i^c(f)$  is set as

$$\bar{\gamma}_i^c(f) = \begin{cases} 0, & \text{if } \Omega(f, \theta_t) > \eta, \\ 1, & \text{otherwise,} \end{cases} \quad (10)$$

where  $\eta$  is a given threshold.

2) *Regression-based Configuration:* In the regression-based configuration, the output variable is continuous in the range  $[0, 1]$  and the  $i$ -th training set output  $\bar{\gamma}_i^r(f)$  is set as

$$\bar{\gamma}_i^r(f) = \begin{cases} 0, & \text{if } \Omega(f, \theta_t) > \eta, \\ \max\left[1 - \frac{\Omega(f, \theta_t)^2}{\eta}, 0\right], & \text{otherwise.} \end{cases} \quad (11)$$

Hence, we have that the contribution of positively narrow-band SRP is weighted as a quadratic function of narrowband localization error. The activation function used for the fully connected regression layer is the mean squared error.

The choice of the parameter  $\eta$  is crucial for a good training. In general, we aim at selecting a value that allows a balanced number of positively and negatively contributing maps on the whole training set. A very small value has the effect of providing a small number of positively contributing maps. On the other hand, large values of  $\eta$  may have the effect of allowing some disruptive narrowband components to take part in the fusion [25], [26]. In [25], it has been demonstrated that a value in the range 0.3-0.6 m is a good choice for the near-field. In [26], it was successfully used a value of 3 degrees for a far-field noisy condition. In this work, we have empirically found that a value of 5 degrees provides satisfactory results for the far-field noisy and reverberant case.

#### IV. EXPERIMENTS AND RESULTS

In this section, the performance of the CNN-based localization schemes (classification and regression) is assessed by addressing a 2D source localization task in the far-field scenario (DOA estimation). The multichannel noisy and reverberant acoustic data used in the first experimental setup were obtained by numerical simulation of the room acoustics, whereas the data used in the second experimental setup are actual multichannel recordings of an acoustic source located in reverberant environments. The performance of the proposed SRP-WMVDR-CNN methods is assessed in terms of the localization accuracy rate (AR) for a threshold error of 5 degrees and the root mean square error (RMSE), and compared with the SRP-WMVDR-SVM [25], [26], the SRP-NMVDR [52], and the SRP-PHAT [13]. In the SRP-WMVDR-SVM beamformer scheme, the weighting factors of narrowband MVDR response are estimated with an SVM supervised model defined as

$$\gamma(f) = \operatorname{sgn}\left(\sum_{i=0}^Q \alpha_i \bar{\gamma}_i \psi(\bar{\zeta}_i, \zeta(f)) + b\right), \quad (12)$$

where  $Q$  is the training sample size,  $\psi(\bar{\zeta}_i, \zeta(f))$  is the inner-product kernel for the  $i$ -th training sample input  $\bar{\zeta}_i$  and the sample input  $\zeta(f)$  for the narrowband PSD at frequency  $f$ ,  $\bar{\gamma}_i$  is the  $i$ -th target value so that it takes values  $\{1, -1\}$ ,  $\alpha_i \geq 0$ , and  $b$  is a real constant. The parameters  $\alpha_i$  are found as usual by solving a convex maximization quadratic programming problem. The skewness of the normalized narrowband PSDs is taken as input to the classifier. The radial basis function kernel was adopted for the SRP-WMVDR-SVM by setting  $\lambda = 1$  and  $\sigma = 1$  using a cross-validation in accordance to [25], [26]. The sample frequency was 48 kHz and the window size  $L$  was 2048

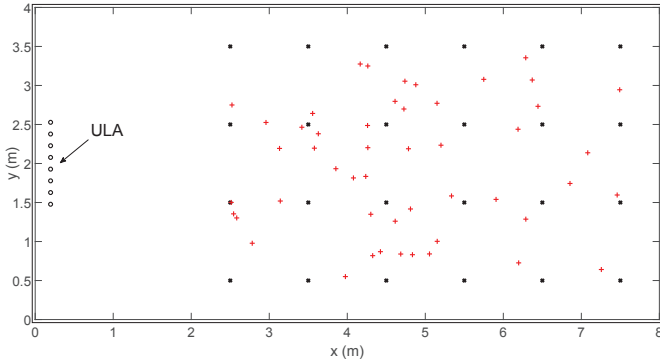


Fig. 3. The simulated room setup with the positions of the 8-microphone ULA, the 24 training positions (black circle) and the 50 testing positions (red plus).

samples. We have set  $f_{\min}$  and  $f_{\max}$  to 50 Hz and 15000 Hz, respectively. A diagonal loading regularization [58] was used for the narrowband MVDR filter to improve the robustness of the SRP. The PSD matrix is estimated using an averaging of 10 snapshots in all methods. The inter-microphone distance of the ULA is 0.15 m, resulting in an angular discretization of  $N=41$  samples. Hence, the input matrix  $\mathbf{V}$  has dimension  $41 \times 41$ . In our CNN configuration, we used 20 convolutional kernels with a size of  $5 \times 5$ , since it allows a simple structure balancing the recognition accuracy and the overfitting problem. We adopted a max-pooling layer with size  $2 \times 2$ . Thus, the feature size is reduced by a factor of four. The parameter  $\eta$  was set to 5 degrees since we have empirically found that it provides satisfactory results for the far-field noisy and reverberant case [25], [26]. The CNN and SVM have been implemented using the Matlab R2017a Neural Network Toolbox and Statistics and Machine Learning Toolbox. We used our own implementation for the MVDR filter.

We investigate the generalization properties of the proposed method with respect to three characteristics: 1) the source position (training and testing positions are different in all experiments); 2) the acoustic source nature (training is performed with an USASI signal [59], whereas the testing is performed with speech, impulsive or narrowband signals); 3) the environment characteristics (training and testing are performed in the same room and in different rooms evaluating the localization performance with different noise and reverberant conditions).

#### A. Simulations

Simulations of reverberant environments were obtained with the image-source method (ISM) [60], implemented using the improved algorithm reported in [61]. The simulations were conducted with different SNR levels, obtained by adding mutually independent white Gaussian noise to each channel.

In the first set of simulations, an ULA of 8 microphones was used. A localization task in a room of  $8 \text{ m} \times 4 \text{ m} \times 3 \text{ m}$  was considered. The sources and microphones were considered omnidirectional. The room setup is shown in Figure 3, in which we can see the 24 source positions used in the training phase. The training was performed using an USASI signal with a  $RT_{60}$  of 0.6 s and a SNR of 20 dB. The same

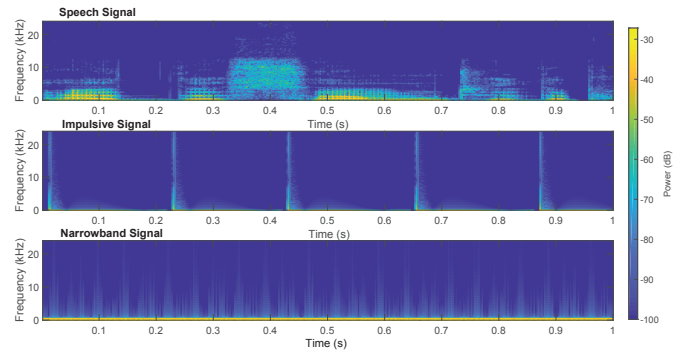


Fig. 4. STFT of the three signals used in simulations.

training setup was used for the classification CNN, regression CNN, and SVM. The resulting training set has a size of 16368 input matrices  $\mathbf{V}$ . We compare the localization performance of three different acoustic sources, i.e. a speech signal, an impulsive signal (starter pistol), and a narrowband signal (range [100,750] Hz). The three signals are depicted in Figure 4 (excerpts of one second each). We consider 50 random source positions (different from the training positions) with a distance from the array in the range 2.5-7.5 m. The test positions are shown in Figure 3. Table I shows the results in different noisy and reverberant conditions. As we can observe, the classification CNN provides in general the best performance, although in some noisy condition (-20 dB) the regression CNN outperforms it. In such situation, the binary classification error may discard some useful information and include incorrect ones, while the regression output allows a weighting of narrowband components resulting a more robust localization accuracy. Both proposed CNN methods perform better than all of the other algorithms and provide good generalization performances in all noisy and reverberation conditions with respect to the source position (training and test positions are different) and to the acoustic source nature (training is performed with USASI noise and testing is performed with speech, impulsive and narrowband signals). We can also note that the SVM-based classifier become ineffective in low noise conditions and in some narrowband source cases, confirming the results in [25] and in [26]. Next, we can observe that the localization performance of CNN methods in term of AR and RMSE is better with the narrowband signal if compared to that SRP-NMVD, since the most of the spectrum of the signal is affected by noise. This performance difference is reduced with the speech and the impulsive signals. Both classification and regression CNN-based methods demonstrate a good robustness when reverberation is increased, as we can note in Table I. When  $RT_{60}$  is increased, in general the performance gap between CNN and NMVD increases.

Next, to assess the generalization characteristics with respect to room dimensions, the system trained on data from this room geometry was tested on a set of acoustic data obtained with room dimensions of  $7 \text{ m} \times 11 \text{ m} \times 4 \text{ m}$ , in which the 8-microphone ULA is positioned. Table II shows the results for the speech signal. We can see that the regression CNN provides the best performance in this case. This result



suggests that, when the training and testing room are different, the classification CNN performance is affected by a larger number of narrowband SRP classification errors, since the room response is changed. In this case, the regression CNN seems to be less sensitive to room geometry differences and provides a better classification performance.

In the second set of simulations, a small configuration ULA of 3 microphones was used. The room setup is shown in Figure 6. We consider 20 training positions, resulting 13640 input matrices  $\mathbf{V}$ . The training was performed using an USASI signal with a  $RT_{60}$  of 0.3 s and a SNR of 20 dB. Two localization tests were performed using the same room configuration of the training (5 m  $\times$  4 m  $\times$  3 m) and a different room configuration with a size of 6.37 m  $\times$  2.98 m  $\times$  3.6 m. We consider 50 random source positions (different from the training positions) with a distance from the array in the range 1-3 m. The results with a speech signal are reported in Table III. As we can observe, the simulation confirms the efficiency of CNN-based SRP methods. Specifically, the classification output has a better performance when the room setup is equal to that of the training. On the other hand, the regression output provides a better localization accuracy in a different room setup. In this case, the SVM does not provide any improvement, and the classification CNN tends to provide lower performance when the SNR decreases.

Next, we evaluated the localization performance using a speech signal corrupted by two different types of noise: babble noise (i.e., background noise originating from a large number of simultaneously talking people, as it is typically observed in a cocktail party) and diffuse noise field [62]. The room setup used for this evaluation is depicted in Figure 3. The  $RT_{60}$  was set to 0.3 s and the SNR was set to 20 dB. Table IV shows how the classification and regression SRP-WMVDR-CNN methods perform better than the other SRP-based methods.

Afterwards, we compared the CNN-based approach with the DNN method described in [51], in which the noise eigenvectors of the power spectral density matrices are used as input to the neural network that, in turn, outputs an estimate of the DOA. The DNN structure is composed by a directional image activators layer, a partially integrated layer and an integrated layer. The DNN was trained using the same setup used for the CNN and shown in Figure 3. Table V shows the results for two noisy and reverberant conditions, when the training and testing data are organized as in the first experiment. As we can observe, both CNN-based approaches outperform the DNN-based scheme.

We are firmly convinced that the effectiveness of the proposed method lies in the hybrid nature of the processing scheme. To provide a comparison of this solution with a simpler one, in which the localization is based solely on the convolutional neural network component, we have implemented an end-to-end scheme in which the phase of the STFT is encoded directly as the input to the CNN, while the CNN output is used to directly encode the DOA value. Thus, a regression configuration was assumed. The system overview of the CNN-based end-to-end configuration is shown in Figure 5. We have performed a simulation with a speech signal and an ULA of 8 microphones in the room setup depicted in Figure 3 with

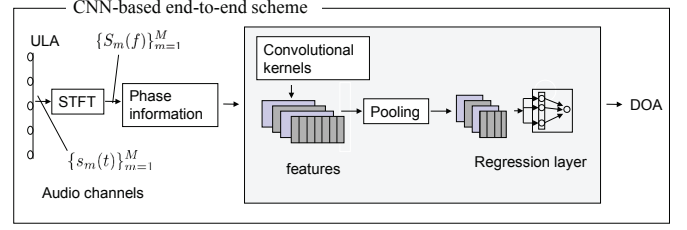


Fig. 5. System overview of the CNN-based end-to-end configuration.

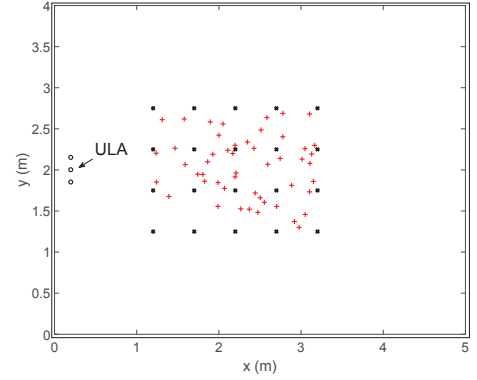


Fig. 6. The simulated room setup with the positions of the 3-microphone ULA, the 20 training positions (black circle) and the 50 testing positions (red plus).

$RT_{60}=0.3$  s and  $SNR=20$  dB. The localization performance is  $AR=25.78$  % and  $RMSE=11.488$  degree. When training the model using training data and conditions comparable to the one used for the proposed schemes, the performances of the model resulted extremely poor. Apparently in this processing chain choice, the localization-related information is so overwhelmed by unrelated information, that the amount of training data and training time required to provide the same performances as an hybrid scheme, would be both much larger.

The proposed CNN methods improve the localization performance in noisy and reverberant conditions, compared to other state-of-the-art methods. Specifically, SRP-WMVDR-CNN methods prove to effectively generalize with respect to the source position, the acoustic source nature and the environment characteristics. For the latter, the classifier-based configuration performs better than the regression-based configuration when the training and the test environments are the same. On the other hand, the regression-based configuration proved to be more robust in our tests when the environment characteristics used in the test procedure were different from those used during training.

### B. Analysis of the convolutional layer features

To gain further insight into the features learned by the CNNs, we report the high-level features at the output of the fully connected layer. Figures 7 and 8 show the feature maps that strongly activate the two channels for the 8-microphone and 3-microphone ULA, respectively. It can be seen that the features are characterized by a similar pattern (i.e., energy concentrated in the upper central region for the positively contributing features), for both array configurations.



TABLE I

AR (%) AND RMSE (degree) OF DOA ESTIMATION FOR SIMULATED DATA WITH A ULA OF 8 MICROPHONES IN A ROOM OF 8 m  $\times$  4 m  $\times$  3 m. THE TRAINING AND TESTING ROOMS ARE THE SAME. THE BEST PERFORMANCE IS SHOWN IN BOLD TEXT.

speech signal							
RT (s)	SNR (dB)		SRP-NMVDR	SRP-WMVDR-CNN (class.)	SRP-WMVDR-CNN (regr.)	SRP-WMVDR-SVM	SRP-PHAT
0.3	20	AR	95.56	<b>99.33</b>	97.56	97.56	93.11
		RMSE	2.418	<b>1.909</b>	2.122	2.248	2.832
	0	AR	90.55	<b>92.44</b>	92.22	80.33	88.12
		RMSE	4.129	<b>3.022</b>	3.296	6.049	5.38
	-20	AR	42.89	<b>52.00</b>	51.33	35.11	35.56
		RMSE	19.252	<b>9.237</b>	11.757	28.324	21.509
0.6	20	AR	79.11	<b>88.00</b>	83.56	82.67	73.78
		RMSE	4.422	<b>3.437</b>	3.837	4.194	4.888
	0	AR	74.44	<b>83.33</b>	83.12	72.66	70.00
		RMSE	6.793	<b>4.896</b>	4.965	8.158	9.391
	-20	AR	31.11	<b>42.56</b>	42.22	20.00	27.56
		RMSE	22.747	<b>11.057</b>	13.026	30.646	25.510
0.9	20	AR	64.89	<b>74.22</b>	66.44	68.67	60.89
		RMSE	5.824	<b>4.942</b>	5.603	5.342	6.109
	0	AR	56.00	<b>63.11</b>	60.44	48.44	52.00
		RMSE	6.008	<b>4.917</b>	4.982	16.745	6.641
	-20	AR	27.78	38.22	<b>41.11</b>	17.56	22.67
		RMSE	25.163	12.763	<b>12.187</b>	31.946	26.878
impulsive signal							
RT (s)	SNR (dB)		SRP-NMVDR	SRP-WMVDR-CNN (class.)	SRP-WMVDR-CNN (regr.)	SRP-WMVDR-SVM	SRP-PHAT
0.3	20	AR	87.33	<b>90.44</b>	88.67	86.89	84.67
		RMSE	3.595	<b>3.259</b>	3.412	3.666	3.930
	0	AR	81.33	<b>87.44</b>	85.24	75.88	79.55
		RMSE	4.882	<b>3.571</b>	3.775	6.344	5.323
	-20	AR	28.44	34.22	<b>39.56</b>	10.22	25.33
		RMSE	21.538	<b>12.452</b>	13.329	36.295	23.766
0.6	20	AR	75.11	<b>78.52</b>	78.44	76.89	72.00
		RMSE	5.532	<b>4.375</b>	4.872	6.429	6.560
	0	AR	70.888	<b>79.11</b>	75.11	66.88	68.77
		RMSE	5.897	<b>4.938</b>	5.162	11.475	7.417
	-20	AR	20.44	31.11	<b>33.33</b>	11.33	19.56
		RMSE	25.696	14.904	<b>13.227</b>	34.681	26.800
0.9	20	AR	65.56	<b>72.44</b>	67.78	66.67	63.56
		RMSE	5.991	<b>5.044</b>	5.741	8.226	6.239
	0	AR	60.88	<b>68.88</b>	69.88	52.66	57.55
		RMSE	7.180	<b>5.202</b>	5.159	16.498	8.457
	-20	AR	18.44	<b>29.56</b>	27.78	11.33	15.33
		RMSE	28.866	<b>14.916</b>	15.688	36.608	30.188
narrowband signal							
RT (s)	SNR (dB)		SRP-NMVDR	SRP-WMVDR-CNN (class.)	SRP-WMVDR-CNN (regr.)	SRP-WMVDR-SVM	SRP-PHAT
0.3	20	AR	61.78	<b>69.11</b>	68.00	66.67	43.33
		RMSE	11.044	<b>4.924</b>	8.832	13.155	13.163
	0	AR	53.55	<b>66.00</b>	64.22	60.44	35.77
		RMSE	8.282	<b>5.476</b>	6.8794	13.7261	15.7299
	-20	AR	29.33	<b>50.22</b>	45.56	25.33	24.00
		RMSE	21.143	<b>9.766</b>	10.612	32.070	25.305
0.6	20	AR	44.22	<b>59.56</b>	53.78	42.89	35.11
		RMSE	14.781	<b>7.701</b>	8.603	14.428	17.047
	0	AR	42.22	<b>53.77</b>	49.11	42.44	29.11
		RMSE	13.33	<b>8.55</b>	9.13	18.45	20.17
	-20	AR	22.22	<b>39.11</b>	38.89	20.00	15.56
		RMSE	23.270	<b>10.649</b>	11.299	32.382	28.461
0.9	20	AR	34.22	<b>48.00</b>	44.67	30.67	24.22
		RMSE	21.363	<b>10.777</b>	14.749	27.004	24.810
	0	AR	29.11	<b>44.33</b>	40.77	30.44	25.55
		RMSE	18.668	<b>11.603</b>	13.305	25.543	22.981
	-20	AR	16.89	<b>39.11</b>	30.44	14.00	15.11
		RMSE	28.203	<b>12.211</b>	19.534	33.793	29.853

Then, we reported the average recognition accuracy of positively contributing maps (1-value classification) and negatively contributing maps (0-value classification) for the classification CNN and the SVM using skewness. The results showed

in Table VI confirm the better recognition accuracy of the classification CNN.

To further compare the effectiveness of the convolutional layer features and other specific features such as the skewness

TABLE II  
AR (%) AND RMSE (degree) OF DOA ESTIMATION FOR SIMULATED DATA WITH A ULA OF 8 MICROPHONES IN A ROOM OF 7 m × 11 m × 4 m DIFFERENT. THE TRAINING AND TESTING ROOMS ARE DIFFERENT. THE BEST PERFORMANCE IS SHOWN IN BOLD TEXT.

speech signal							
RT (s)	SNR (dB)		SRP-NMVDR	SRP-WMVDR-CNN (class.)	SRP-WMVDR-CNN (regr.)	SRP-WMVDR-SVM	SRP-PHAT
0.3	20	AR	92.86	98.32	<b>99.16</b>	98.16	90.34
		RMSE	2.709	2.252	<b>2.156</b>	2.449	3.073
	0	AR	87.77	82.66	<b>95.11</b>	82.33	84.88
		RMSE	5.124	9.338	<b>3.758</b>	5.947	6.381
	-20	AR	48.00	48.00	<b>56.00</b>	36.00	32.00
		RMSE	18.172	8.088	<b>6.302</b>	21.073	18.288

TABLE III  
AR (%) AND RMSE (degree) OF DOA ESTIMATION FOR SIMULATED DATA WITH A ULA OF 3 MICROPHONES WITH A SPEECH SIGNAL. THE UPPER PART OF THE TABLE REPORT THE PERFORMANCE IN CASE OF SAME ROOM FOR TRAINING AND TESTING. THE LOWER PART OF THE TABLE REPORT THE PERFORMANCE IN CASE OF DIFFERENT ROOMS FOR TRAINING AND TESTING. THE BEST PERFORMANCE IS SHOWN IN BOLD TEXT.

room of 5 m × 4 m × 3 m (the same of the training)							
RT (s)	SNR (dB)		SRP-NMVDR	SRP-WMVDR-CNN (class.)	SRP-WMVDR-CNN (regr.)	SRP-WMVDR-SVM	SRP-PHAT
0.6	30	AR	80.22	<b>89.33</b>	88.44	81.33	77.11
		RMSE	4.078	<b>3.286</b>	3.305	6.644	4.563
	15	AR	78.67	<b>88.67</b>	88.00	79.56	73.78
		RMSE	4.267	<b>3.482</b>	3.580	6.159	6.956
	0	AR	67.33	<b>70.22</b>	68.67	57.33	52.00
		RMSE	10.378	<b>6.279</b>	6.490	17.437	18.409
room of 6.37 m × 2.98 m × 3.6 m							
RT (s)	SNR (dB)		SRP-NMVDR	SRP-WMVDR-CNN (class.)	SRP-WMVDR-CNN (regr.)	SRP-WMVDR-SVM	SRP-PHAT
0.6	30	AR	81.91	<b>85.00</b>	<b>85.00</b>	77.78	80.89
		RMSE	4.006	<b>3.605</b>	<b>3.605</b>	5.691	4.614
	15	AR	81.11	84.89	<b>85.78</b>	78.89	79.33
		RMSE	4.052	3.895	<b>3.618</b>	6.883	6.052
	0	AR	63.11	62.44	<b>64.89</b>	52.00	51.78
		RMSE	9.388	8.055	<b>6.991</b>	18.054	16.085

TABLE IV  
AR (%) AND RMSE (degree) OF DOA ESTIMATION FOR SIMULATED DATA IN DIFFERENT NOISE TYPE CONDITIONS WITH A ULA OF 8 MICROPHONES IN A ROOM OF 8 m × 4 m × 3 m. THE RT<sub>60</sub> WAS SET TO 0.3 s AND THE SNR WAS SET TO 20 dB. THE BEST PERFORMANCE IS SHOWN IN BOLD TEXT.

speech signal						
		SRP-NMVDR	SRP-WMVDR-CNN (class.)	SRP-WMVDR-CNN (regr.)	SRP-WMVDR-SVM	SRP-PHAT
babble noise	AR	95.33	<b>98.89</b>	97.33	95.77	91.33
	RMSE	2.586	<b>1.969</b>	2.336	2.527	3.054
diffuse noise	AR	95.55	<b>99.11</b>	98.22	97.11	92.00
	RMSE	2.428	<b>1.920</b>	2.023	2.244	2.914

TABLE V  
COMPARISON BETWEEN CNN AND DNN: AR (%) AND RMSE (degree) OF DOA ESTIMATION FOR SIMULATED DATA WITH A ULA OF 8 MICROPHONES IN A ROOM OF 8 m × 4 m × 3 m.

RT (s)	SNR (dB)		SRP-WMVDR-CNN (class.)	SRP-WMVDR-CNN (regr.)	DNN
0.3	30	AR	99.33	97.78	77.11
		RMSE	1.897	2.142	5.765
0.5	10	AR	92.89	92.22	76.00
		RMSE	2.895	2.956	6.001

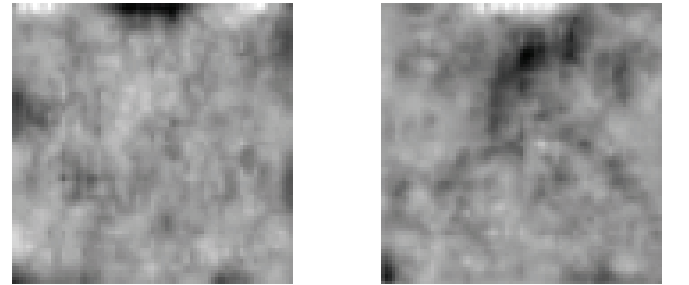


Fig. 7. Features computed by the convolutional layer of the 8-microphone ULA. Left: feature mapped to a 0-valued classification label; Right: feature mapped to a 1-valued classification label

mentioned in this section, we also report in Table VII the Fisher's discriminant ratio (FDR) average values for the two choices, the average being referred to all positions for the 3-microphones and for the 8-microphones case. The FDR is defined as the ratio of the between-class scatter matrix to the within-class scatter matrix, and can be employed to quantify the discriminatory power of individual features between

classes [63]. A common problem in the computation of the FDR is the high dimensionality of the features with respect

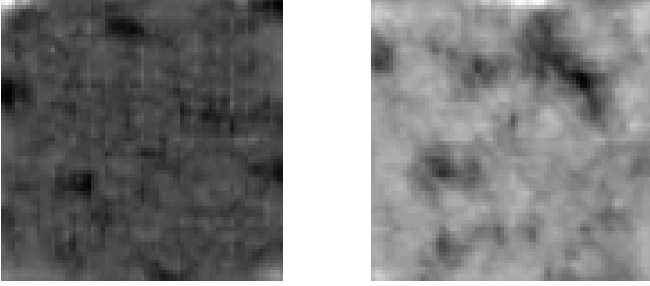


Fig. 8. Features computed by the convolutional layer for the 3-microphone ULA. Left: feature mapped to a 0-valued classification label; Right: feature mapped to a 1-valued classification label

TABLE VI  
THE RECOGNITION ACCURACY (%) OF NARROWBAND SRPs FOR THE CLASSIFICATION CNN AND THE SVM USING SKEWNESS.

ULA	Method	1-value class.	0-value class.
3 mic.	CNN (class.)	48.87	85.18
	SVM (ske.)	15.29	94.62
8 mic.	CNN (class.)	39.81	93.20
	SVM (ske.)	36.44	86.17

to the observation data, which leads to poorly conditioned within-class scatter matrices. To face this issue, we perform a preprocessing data dimensionality reduction based on principal component analysis, which is a commonly accepted solution [64], [65]. The numerical results reported confirm the substantial increment of the discriminatory power of convolutional features if compared to skewness.

Last, we further investigated how the CNN-based method improves robustness to noise, by using an ad-hoc low-pass random noise signal as source, with energy up to 1 kHz, corrupted by white uncorrelated noise. The uncorrelated noise energy was gradually increased so that the related SNR spanned the range [30 dB, -30 dB]. We considered a ULA of 8 microphones in the room setup depicted in Figure 3 with  $RT_{60}=0.3$  s, and a source impinges the array with  $\theta_s=10.72$  degree. Table VIII reports the average of the weighting factors  $\gamma(f)$  with different noise levels for the two frequency intervals [50 Hz, 1 kHz] (in which the source and the noise are both present), and [1 kHz, 16 kHz] (in which only the uncorrelated noise is present). As we can note, when the SNR is 30 dB and 15 dB all the narrowband SRP components in the range [50 Hz, 1 kHz] are classified as positively contributing, and assigned to class 1. When the SNR is lower the classification in the range [50 Hz, 1 kHz] reduces the average of the weighting factors, meaning that some SRP components are classified as contributing negatively. We can observe that when SNR=-30 dB, the classification operated by the CNN becomes ineffective. In all SNR conditions, in the range of [1 kHz, 16 kHz] the CNN correctly classified the most of the components as contributing negatively since only uncorrelated noise and no source signal energy is present in that frequency range. Figure 9 shows the SRP function for two specific frequencies: 500 Hz, falling in the range of the source signal spectrum, and 2500 Hz, falling in the range where only uncorrelated noise is present. As we can see, when the SNR level is 30 dB, 0 dB

TABLE VII  
FEATURE DISCRIMINABILITY POWER COMPARISON.

ULA	Feature	Fisher's discriminant ratio
3 mic.	Convolutional	18.138
	Skewness	0.016
8 mic.	Convolutional	13.339
	Skewness	0.109

TABLE VIII  
THE AVERAGE OF THE WEIGHTING FACTORS FOR TWO RANGES OF FREQUENCIES USING A LOW-PASS RANDOM NOISE WITH ENERGY UP TO 1 KHz AT VARIATION OF SNR LEVEL.

SNR (dB)	class.		regr.	
	50 Hz-1 kHz	1 kHz-16 kHz	50 Hz-1 kHz	1 kHz -16 kHz
30	1.00	0.15	0.12	0.07
15	1.00	0.10	0.12	0.06
0	0.79	0.11	0.14	0.07
-15	0.48	0.10	0.10	0.07
-30	0.17	0.12	0.08	0.07

and -15 dB, the DOA source is correctly estimated for the 500 Hz frequency and the SRP functions are correctly classified as 1. When the SNR level is -30 dB, the SRP is classified as 0, and, hence, the SRP is removed in the fusion process, since it does not contribute anymore to correctly localize the source. For the 2500 Hz case, in all noise conditions the SRP are classified as 0, and the noise components were removed in the classification scheme or removed/attenuated in the regression one.

### C. Real Data

The experiments based on multichannel recorded data were performed in an office room of 6.37 m  $\times$  2.98 m  $\times$  3.6 m with a  $RT_{60}$  of 0.6 s and in a conference room of 16 m  $\times$  7 m  $\times$  3 m with a  $RT_{60}$  of 0.9 s.

In the first experiment, an ULA of 3 microphones was used in the office room. We performed the localization with the training configuration used in the simulation with 3 microphones. The room setup is shown in Figure 10. Both microphones and the source were positioned at a distance from the floor of 0.9 m. A speech signal of 25 s duration from a male speaker was reproduced with a loudspeaker in the positions depicted in Figure 10. We estimated an average SNR of 15 dB at microphones. The SNR was computed by estimating the average speech energy vs the average noise energy (the latter is estimated from signal fragments where the speaker is not active). The results are reported in Table IX. We can note that only the regression CNN-based method improves the performance if compared to SRP-NMVDR, while both binary classification (classification CNN and SVM) fails in such case. This fact confirms that the generalization of the classification CNN is more difficult due to far-field reverberant conditions, in which the reflection components have a larger impact on SRP computation in comparison to the near-field condition [25]. This result can be compared to that reported in Table III with the same room noise and reverberant conditions (the lower part). We can note that the real results have a greater RMSE

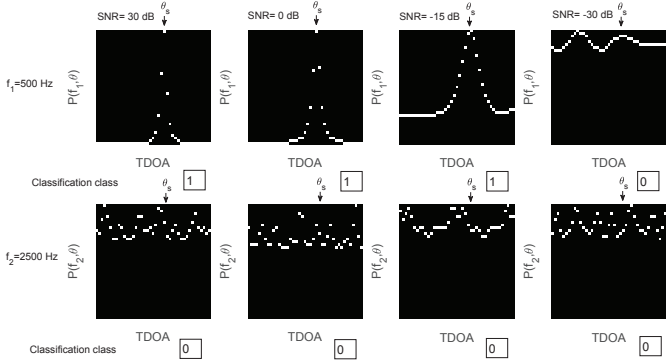


Fig. 9. The narrowband SRP input  $\mathbf{V}$  using a low-pass random noise with energy up to 1 kHz at variation of SNR level.

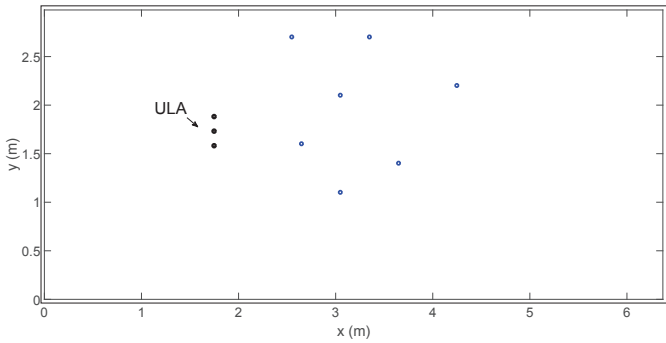


Fig. 10. The real room setup with the positions of the 3-microphone ULA and the 7 test positions.

with a better AR, due to the limited number of positions used in this experiment.

In the second experiment, an ULA of 8 microphones was used in the conference room. We performed the localization with the training configuration used in the simulation with 8 microphones. The room setup is shown in Figure 11. Both microphones and the source were positioned at a distance from the floor of 1.7 m. Three sessions were recorded using short sentences uttered by two male and one female speakers, standing up at different positions depicted in Figure 11. The results reported in Table X confirm the good performance of the regression CNN-based method. In this experiment also the classification configuration has a better performance if compared to that of the SRP-NMVDR. This fact is due to a better robustness to noise of the 8-microphone ULA in comparison to a small 3-microphone ULA.

## V. CONCLUSIONS

A WMVDR beamformer based on a CNN deep learning has been presented. It improves the localization accuracy in a single source scenario without point-source interferences. The results show that CNNs improve the incoherent frequency fusion of the narrowband response power by weighting the components in such a way as to reduce the deleterious effects of those components affected by artifacts due to noise and reverberation. The use of CNNs avoids the necessity of previously encoding the multichannel data into selected acoustic cues. We implemented the CNNs in two versions, one with

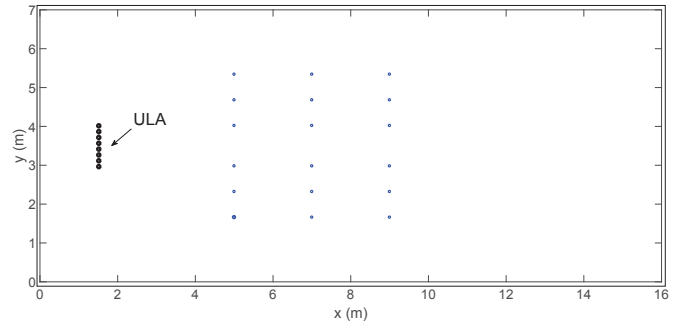


Fig. 11. The real room setup with the positions of the 8-microphone ULA and the 18 test positions.

a classification output layer, and the other with a regression output layer. Our experiments demonstrate that the CNN is robust to noise and reverberation in comparison to the state-of-the-art. Specifically, the classification CNN has a better performance when the training and test condition setup are the same (i.e., same room and array position). On the other hand, the regression CNN provides a better localization accuracy, due to its robustness against classification errors that may occur when training data and test data are referred to different acoustic conditions. The proposed method has been compared to other two possible approaches based on a neural network component. An end-to-end CNN scheme, and a DNN model proposed in the literature. In both cases, the proposed method provided superior performances. Our explanation for that is that the method exploits the hybrid nature of the processing scheme, in which the CNN component is integrated with a simple but effective information fusion model rooted on acoustic principles.

A number of issues remain to be investigated, and will be the subject of future work. In the present study, frequency components in the training set were selected as positive or negative by using a frequency independent, empirically selected threshold. This approach might be improved by investigating if different components might be more or less relevant to the localization depending on their frequency, and if this has any perceptual basis. Moreover, in future refinements of this class of signal processing paths, the machine learning components might be trained to both improve the fusion model (as done in the present case) while also contributing to recognize spectral/temporal characteristics of the acoustic sources, and distinguish for example between speech, music, or ecological sounds. In this case, they could be successfully used for effective multisource localization, or might be trained to distinguish between actual and image sources in reverberant environments.

## VI. ACKNOWLEDGEMENTS

The authors are grateful to the anonymous reviewers for their constructive comments that greatly contributed to improve the manuscript.

This work was partially supported by the "Proactive Vision for advanced UAV systems for the protection of mobile units, control of territory and environmental prevention (SUPReME)" FVG L.R. 20/2015 project.



TABLE IX

AR (%) AND RMSE (degree) OF DOA ESTIMATION FOR REAL DATA WITH A ULA OF 3 MICROPHONES IN A ROOM OF  $6.37 \text{ m} \times 2.98 \text{ m} \times 3.6 \text{ m}$  WITH A  $\text{RT}_{60}$  OF 0.6 S. THE BEST PERFORMANCE IS SHOWN IN BOLD TEXT.

	SRP-NMVDR	SRP-WMVDR-CNN (class.)	SRP-WMVDR-CNN (reg.)	SRP-WMVDR-SVM	SRP-PHAT
AR	91.13	86.45	<b>94.34</b>	74.38	85.96
RMSE	5.046	8.476	<b>4.625</b>	20.245	8.352

TABLE X

AR (%) AND RMSE (degree) OF DOA ESTIMATION FOR REAL DATA WITH A ULA OF 8 MICROPHONES IN A ROOM OF  $16 \text{ m} \times 7 \text{ m} \times 3 \text{ m}$  WITH A  $\text{RT}_{60}$  OF 0.9 S. THE BEST PERFORMANCE IS SHOWN IN BOLD TEXT.

	SRP-NMVDR	SRP-WMVDR-CNN (class.)	SRP-WMVDR-CNN (reg.)	SRP-WMVDR-SVM	SRP-PHAT
AR	71.29	74.18	<b>79.01</b>	62.04	67.05
RMSE	12.728	11.078	<b>6.898</b>	27.179	18.628

## REFERENCES

- [1] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement," *Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 586–601, 2017.
- [2] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1393–1407, 2016.
- [3] L. Kumar and R. M. Hegde, "Near-field acoustic source localization and beamforming in spherical harmonics domain," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3351–3361, 2016.
- [4] D. Yook, T. Lee, and Y. Cho, "Fast sound source localization using two-level search space clustering," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 20–26, 2016.
- [5] D. Salvati, C. Drioli, and G. L. Foresti, "Sound source and microphone localization from acoustic impulse responses," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1459–1463, 2016.
- [6] L. Petrica, "An evaluation of low-power microphone array sound source localization for deforestation detection," *Applied Acoustics*, vol. 113, pp. 162–169, 2016.
- [7] D. Salvati and S. Canazza, "Incident signal power comparison for localization of concurrent multiple acoustic sources," *The Scientific World Journal*, vol. 2014, pp. 1–13, 2014.
- [8] D. Salvati and S. Canazza, "Adaptive time delay estimation using filter length constraints for source localization in reverberant acoustic environments," *IEEE Signal Processing Letters*, vol. 20, no. 6, pp. 507–510, 2013.
- [9] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
- [10] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.
- [11] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.
- [12] P. Stoica and J. Li, "Source localization from range-difference measurements," *IEEE Signal Processing Magazine*, vol. 23, no. 3, pp. 63–66, 2006.
- [13] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001, ch. Robust localization in reverberant rooms.
- [14] J. Velasco, D. Pizarro, and J. Macias-Guarasa, "Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints," *Sensors*, vol. 12, no. 10, pp. 13 781–13 812, 2012.
- [15] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.
- [16] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.
- [17] B. D. Rao and K. V. S. Hari, "Performance analysis of root-music," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1939–1949, 1989.
- [18] K. Harmanci, J. Tabrikian, and J. L. Krolik, "Relationships between adaptive minimum variance beamforming and optimal source localization," *IEEE Transactions on Signal Processing*, vol. 48, no. 1, pp. 1–12, 2000.
- [19] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.
- [20] J. Traa, D. Wingate, N. D. Stein, and P. Smaragdis, "Robust source localization and enhancement with a probabilistic steered response power model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 493–503, 2016.
- [21] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.
- [22] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.
- [23] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.
- [24] D. Salvati, C. Drioli, and G. L. Foresti, "Frequency map selection using a RBFN-based classifier in the MVDR beamformer for speaker localization in reverberant rooms," in *Proceedings of the Conference of the International Speech Communication Association*, 2015, pp. 3298–3301.
- [25] D. Salvati, C. Drioli, and G. L. Foresti, "A weighted MVDR beamformer based on SVM learning for sound source localization," *Pattern Recognition Letters*, vol. 84, pp. 15–21, 2016.
- [26] D. Salvati, C. Drioli, and G. L. Foresti, "On the use of machine learning in microphone array beamforming for far-field sound source localization," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, 2016.
- [27] F. Ribeiro, C. Zhang, D. A. Florencio, and D. E. Ba, "Using reverberation to improve range and elevation discrimination for small array sound source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1781–1792, 2010.
- [28] H.-Y. Lee, J.-W. Cho, M. Kim, and H.-M. Park, "DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1091–1095, 2016.
- [29] D. Salvati, S. Canazza, and G. L. Foresti, "A microphone array interface for real-time interactive music performance," in *Proceedings of the International Computer Music Conference*, 2012, pp. 473–477.
- [30] W. Ma, B. Vo, S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements a random finite set approach," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.
- [31] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1479–1489, 2008.
- [32] F. Nesta, M. Omologo, and P. Svaizer, "A novel robust solution to the permutation problem based on a joint multiple TDOA estimation," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control*, 2008.

- [33] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1490–1503, 2011.
- [34] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 549–557, 2003.
- [35] B. V. Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.
- [36] M. Omologo, P. Svaizer, and R. De Mori, *Spoken Dialogue with Computers*. Academic Press, 1998, ch. Acoustic Transduction.
- [37] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005, pp. iii/89–iii/92 Vol. 3.
- [38] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.
- [39] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.
- [40] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.
- [41] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.
- [42] X. Lin, J. Liu, and X. Kang, "Audio recapture detection with convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1480–1487, 2016.
- [43] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 208–221, 2016.
- [44] S. Durand, J. P. Bello, B. David, and G. Richard, "Robust downbeat tracking using an ensemble of convolutional networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 76–89, 2017.
- [45] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.
- [46] T. Higuchi, N. Ito, T. Yoshioka, and N. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5210–5214.
- [47] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 196–200.
- [48] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 5542–5546.
- [49] X. Zhang, Z. Wang, and D. Wang, "A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 276–280.
- [50] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.
- [51] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 405–409.
- [52] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 581–585, 2014.
- [53] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceeding of the International Conference on Machine Learning*, 2010, pp. 807–814.
- [54] D. E. Rumelhart, G. E. Hinton, and R. J. Williams., "Learning internal representations by error propagation," in *Parallel distributed processing: explorations in the microstructure of cognition*, vol. 1. MIT Press, 1986, pp. 318–362.
- [55] P. T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of Operations Research*, vol. 134, no. 1, pp. 19–67, 2005.
- [56] M. Ranzato, F. J. Huang, Y. L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
- [57] J. S. Bridle, *Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition*. Springer, 1990, pp. 227–236.
- [58] J. F. Synnevag, A. Austeng, and S. Holm, "Adaptive beamforming applied to medical ultrasound imaging," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 54, no. 8, pp. 1606–1613, 2007.
- [59] NRSC, *AM Preemphasis/Deemphasis and Broadcast Audio Transmission Bandwidth Specifications*, National Association of Broadcasters - Consumer Technology Association Std., 2012.
- [60] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
- [61] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.
- [62] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.
- [63] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, *Introduction to Pattern Recognition: A Matlab Approach*. Academic Press, 2010.
- [64] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data - with application to face recognition," *Pattern Recognition*, vol. 34, no. 10, pp. 2067–2070, 2001.
- [65] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.



**Daniele Salvati** received the Laurea degree in Environmental Engineering from the Sapienza University of Rome, Italy, in 2003, the postgraduate Master degree in Sound Engineering from Tor Vergata University of Rome, Italy, in 2006, and the Ph.D. degree in Multimedia Communication from the University of Udine, Italy, in 2012, with a research on acoustic source localization using microphone arrays. Since 2013 he is a postdoctoral researcher at the Department of Mathematics, Computer Science and Physics, University of Udine, Italy. He was a system and audio consultant to many information technology companies from 2001 to 2008. His research interests are in audio and acoustic signal processing and multimedia communications.



**Carlo Drioli** is Assistant Professor at the Department of Mathematics, Computer Science and Physics of the University of Udine. He received the Laurea degree in Electronic Engineering and the Ph.D. degree in Electronic and Telecommunications Engineering from the University of Padova, in 1996 and 2003, respectively. He has been a researcher with the Centro di Sonologia Computazionale (CSC) of the University of Padova, in the field of sound and voice analysis and processing; visiting researcher at the Royal Institute of Technology (KTH), Stockholm, Sweden, with the support of the European Community through a Marie Curie Fellowship; researcher at the Department of Phonetics and Dialectology of the Institute of Cognitive Sciences and Technology of the Italian National Research Council (ISTC-CNR), where he pursued research on voice processing and emotional speech synthesis; research assistant and adjunct professor at the Department of Computer Science of the University of Verona. His current research interests are in the fields of multimedia signal processing, sound and voice coding by means of physical modeling, speech analysis and synthesis, array signal processing. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), of the Acoustical Society of America (ASA), and of the International Speech Communication Association (ISCA).

Processing, Multisensor Data and Information Fusion, Pattern Recognition and Deep Neural Networks. Prof. Foresti is author of more than 300 papers published in International Journals and International Conferences and he has been co-editor of several books in the field of Multimedia and Computer Vision. He has been Guest Editor of a Special Issue of the Proceedings of the IEEE on Video Communications, Processing and Understanding for Third Generation Surveillance Systems. In 2002, he has been awarded of best IEEE Vehicular Electronics paper, in 2010 of the Best paper Award at the International Conference on Distributed Smart Cameras (ICDSC 2010) and in 2016 of the Best Industry Related Paper Award (BIRPA) at the International Conference on Pattern Recognition (ICPR16). Prof. Foresti is Fellow member of IAPR and Senior member of IEEE.



**Gian Luca Foresti** is Full Professor of Computer Science at the University of Udine and Director of the Dept. of Mathematics, Computer Science and Physics. He is the appointed Italian member of the NATO RTO Information System Technology Panel. He was Finance Chair of the 11th IEEE Conference on Image Processing (ICIP05), General Chair of the 16th Int. Conf. on Image Analysis and Processing (ICIAP11) and of the 8th IEEE Conf. on Advanced Video and Signal Based Surveillance (AVSS11). His main interests involve Computer Vision and Image

Processing, Multisensor Data and Information Fusion, Pattern Recognition and Deep Neural Networks. Prof. Foresti is author of more than 300 papers published in International Journals and International Conferences and he has been co-editor of several books in the field of Multimedia and Computer Vision. He has been Guest Editor of a Special Issue of the Proceedings of the IEEE on Video Communications, Processing and Understanding for Third Generation Surveillance Systems. In 2002, he has been awarded of best IEEE Vehicular Electronics paper, in 2010 of the Best paper Award at the International Conference on Distributed Smart Cameras (ICDSC 2010) and in 2016 of the Best Industry Related Paper Award (BIRPA) at the International Conference on Pattern Recognition (ICPR16). Prof. Foresti is Fellow member of IAPR and Senior member of IEEE.