Deep Learning Based Localization of Near-Field Sources with Exact Spherical Wavefront Model

Wenyi Liu[†‡], Jingmin Xin[†‡], Weiliang Zuo[†‡], Jie Li[†‡], Nanning Zheng[†‡], Akira Sano[§]
[†] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China
[‡] National Engineering Laboratory for Visual Information Processing and Applications,
Xi'an Jiaotong University, Xi'an 710049, China
[§] Department of System Design Engineering, Keio University, Yokohama 223-8522, Japan

***Abstract*—Source localization for near-field narrowband signal is an important topic in array signal processing. Deep neural network (DNN) based methods are data-driven and free of pre-assumptions about data model and are expected to learn the intricate nonlinear structure in large data sets. This paper proposes a framework of DNN where a regression layer is utilized to address the problem of near-field source localization. Unlike previous studies in which DOA estimation is modeled as a classification problem and have a relatively low resolution, we exploit a regression model and aim to improve the estimation accuracy. In the training stage, we propose a novel form of feature representation to take full advantage of the convolution networks. In addition, the architecture of deep neural networks is well designed taking in to consideration the trade-off between the expression ability and under-training risks. The simulation results show that the proposed approach has a rather high validation accuracy with a high resolution, and also outperforms some conventional methods in adverse environments such as low signal to noise ratio (SNR) or small number of snapshots.**

***Index Terms*—Source localization, deep neural network (DNN), near-field signal, regression model**

I. Introduction

Source localization is a widely studied problem in various areas such as radar, sonar, speech recognition and wireless communications (e.g., [1]–[5], [20]). Many high-resolution methods like MUSIC [2] and ESPRIT [3] have been proposed to estimate the direction-of-arrival (DOA) of far-field signals. When a signal source is localized in the Fresnel region of the array aperture, the wave impinging on the array has a spherical wavefront and thus must be characterized by both DOA and range. As a result, most aforementioned high-resolution methods with the far-field assumption are not applicable for near-field source localization problems. Many methods were then developed using second-order Taylor expansion to approximate the spherical wavefront [4]. Other popular methods include high-order statistics (HOS), maximum likelihood estimation (MLE) methods [5] and the generalized ESPRIT based method [20]. In fact, those parametric methods mentioned above not only include strict limitations on signal/noise models, they also rely heavily on the consistency of the forward mappings from signal direction to array outputs. Consequently, the performance of these methods may suffer from various imperfections in practical systems.

In comparison, data-driven deep learning techniques have the advantage of reconstructing complicated propagation models via training processes. They generally involve an extraction procedure of features, one of the main reasons for the success of deep learning, such as generalized cross correlation (GCC) vectors [10] or the phase component of the short-time Fourier transform (STFT) of the received microphone signals [11]. While it has succeeded in very demanding scenarios such as dynamic acoustic and broadband signals [10]–[12], [15], synthesized noise signals [8], [11] and reverberant multi-room environments [10], [14], it can hardly be employed directly in general source localization because of limited information on the features [16]. In terms of the selection of the training neural networks, the majority of the deep learning methods choose classification models rather than regression ones [10]–[15], which can be seen as a compromise on the estimation accuracy. However, the regression model provides a more reasonable explanation. More recently, Chakrabarty et al. [8] proposed a framework that yields the best localization performance with $M-1$ convolution layers, given an array of M microphone. While it provides insight into the modification of a network parameter due to the change in the number of microphones, it could lead to a high computational cost as the number of the microphones increase. Furthermore, most of the deep learning methods deal with DOA estimation exclusively with a resolution of 5° [8], [13] or even 10° [12], [15], which is too low a resolution to be put into practice in most general near-field source localization applications.

In this paper, we propose a deep neural network with a regression model to address the issue of near-field source localization. Different from previous methods which mainly focus on DOA estimation in far-field scenarios or acoustic signals in a reverberant environment, we concentrate on localization of general DOA and range parameters for near-field source. Rather than inputting the feature into the networks in the form of a vector, we reshape the feature information into a matrix to better utilize the structural advantages of the networks. Additionally, we design a regression layer to provide more reasonable explanations with a relatively high resolution.

This work was supported in part by the National Natural Science Foundation of China under Grants 61671373 and 61790563, and the Programme of Introducing Talents of Discipline to Universities under Grant B13043. This work was supported by China Scholarship Council.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

II. Data Model

Consider K near-field noncoherent signals $\{s_k(n)\}$ impinging on a uniform linear array (ULA) consisting of M sensors with spacing d. The received noisy signal $x_m(n)$ at the m-th sensor can be expressed as

$$x_m(n) = \sum_{k=1}^{K} s_k(n) e^{j\tau_{m,k}} + w_m(n) \tag{1}$$

where $m = 1, \cdots, M$, $w_m(n)$ is the additive noise, and τ_{mk} is the phase delay due to the time delay between the reference sensor and the m-th sensor for the k-th signal, which is given by [7]

$$\tau_{m,k} = \frac{2\pi}{\lambda}\left(\sqrt{r_k^2 + (md)^2 - 2r_k md \sin\theta_k} - r_k\right) \tag{2}$$

where θ_k and r_k are the DOA and range of the kth signal, and λ is the wavelength. When the k-th signal is in the Fresnel region (i.e., $r_k \in (0.62(D^3/\lambda)^{1/2}, 2D^2/\lambda)$, where D is the array aperture [9], we can rewrite (1) as

$$\begin{aligned} \boldsymbol{x}(n) &= [x_1(n), x_2(n), \cdots, x_M(n)]^T \qquad (3)\\ &= \sum_{k=1}^{K} \boldsymbol{a}(\theta_k, r_k) s_k(n) + \boldsymbol{w}(n) \qquad (4)\\ &= \boldsymbol{A}\boldsymbol{s}(n) + \boldsymbol{w}(n) \qquad (5) \end{aligned}$$

where $(\,\cdot\,)^T$ denotes transpose, $\boldsymbol{s}(n)$ and $\boldsymbol{w}(n)$ are the vectors of incident signals and additive noises given by

$$\begin{aligned} \boldsymbol{s}(n) &= [s_1(n), s_2(n), \cdots, s_K(n)]^T \qquad (6)\\ \boldsymbol{w}(n) &= [w_1(n), w_2(n), \cdots, w_M(n)]^T. \qquad (7) \end{aligned}$$

while $\boldsymbol{A}$ is the steering matrix of the calibrated ULA defined by $\boldsymbol{A} \triangleq [\boldsymbol{a}(\theta_1, r_1), \boldsymbol{a}(\theta_2, r_2), \cdots, \boldsymbol{a}(\theta_K, r_K)]$, and $\boldsymbol{a}(\theta_k, r_k)$ is the array steering vectors which can be expressed as

$$\boldsymbol{a}(\theta_k, r_k) = [e^{j\tau_{1,k}}, e^{j\tau_{2,k}}, \cdots, e^{j\tau_{M,k}}]^T. \tag{8}$$

In this paper, we assume that the incident signals $\{s_k(n)\}$ are zero-mean stationary random processes, while the additive noises $\{w_m(n)\}$ are uncorrelated with the incident signals and are temporally and spatially complex white Gaussian random processes with zero-mean and variance σ^2.

III. Method

We are interested in generating a system that can localize the near-field source using the deep neural network technique. Most previous methods have formulated the estimation as an N-class classification problem, where each class corresponds to a possible value [10]–[15]. We nonetheless design a deep neural network featuring a regression layer, for it not only makes more sense in physical interpretation, but it also generates a higher estimation precision. The feature extraction that precedes the network also has a sophisticated design that takes full advantage of the convolution layer.

A. Feature Extraction

Feature extraction is essential to the learning process. Efficient and sufficient information is needed in the feature representation for the source localization task. While most methods reformulate the covariance matrix of the received signal into a vector [22], we keep the covariance matrix in the matrix form, since the convolution layer in the deep learning network is initially designed to handle image data. In fact, experiment shows that feature represented in a matrix form generated slightly higher accuracy than vector form in the case of using a regression model.

Under the basic assumption and data model, the covariance matrix $\boldsymbol{R}$ of the array output is given by

$$\boldsymbol{R} = E\{\boldsymbol{x}(n)\boldsymbol{x}(n)^H\} = \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H + \sigma^2\boldsymbol{I}_M \tag{9}$$

where $\boldsymbol{R}_s \triangleq E\{\boldsymbol{s}(n)\boldsymbol{s}(n)^H\}$. In general $\boldsymbol{R}$ can be approximated by averaging over the time index n. Since $\boldsymbol{R}$ is a Hermitian matrix, there are redundant information in the upper and lower triangular parts of the matrix. And considering that we need to decompose the complex value of each entry of $\boldsymbol{R}$, except the diagonal elements, into separate real values, we thus propose a form of feature representation $\boldsymbol{r}$, given by

$$\boldsymbol{r} = \begin{bmatrix} r_{1,1} & \Re(r_{1,2}) & \cdots & \Re(r_{1,M}) \\ \Im(r_{2,1}) & r_{2,2} & \cdots & \Re(r_{2,M}) \\ \vdots & & \ddots & \\ \Im(r_{M,1}) & \Im(r_{M,2}) & \cdots & r_{M,M} \end{bmatrix} \tag{10}$$

where $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary parts of a complex value respectively. The dimension of the matrix $\boldsymbol{r}$, i.e., the size of the inputs remain the same as the covariance matrix.

B. Deep Neural Network Framework

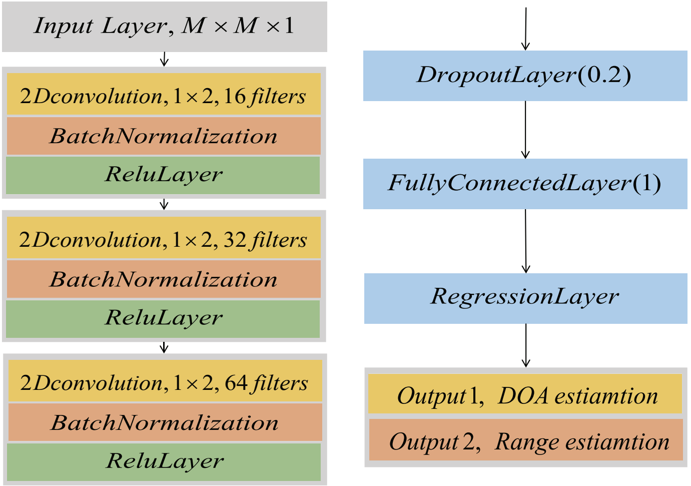

Fig. 1. Deep learning neural network framework for near-field source localization

The proposed DNN-based near-field source localization framework is depicted in Fig. 1. With the feature-enhanced covariance matrix $\boldsymbol{r}$ as the input of the network, we mainly employ 3 convolution layers to learn the local features of the input data. In this paper, small filters of size 1×2 are applied to learn the correlations between the neighboring microphones with "same" padding. This is in contrast to [11], [12], [15], where square filters of size such as 2×2 or filters of larger size such as 5×7 were used to learn the features. In fact, a smaller filter size not only capture more detailed and complex features in the input, it also benefits from weight sharing and reduction in computational costs. The convolution layer is followed by a batch normalization layer, which normalizes the activations and gradients propagating through a network, making network training an easier optimization problem [25]. Additionally, it is noted that using batch normalization layers between convolution layers and non-linearities, such as ReLU, can speed up network training and reduce the sensitivity to network initialization. There are no max pooling layers to perform down sampling, because it causes performance degradation in our network. Furthermore, we use a dropout layer with a probability of 0.2 to prevent from overfitting [23]. Finally, since we define the source localization as a regression problem, a fully connected layer must precede the regression layer at the end of the network.

The regression computes the half-mean-squared-error loss of the training process, which is given by

$$loss = \frac{1}{2}\sum_{i=1}^{N}\frac{(\boldsymbol{t}_i - \boldsymbol{y}_i)^2}{N} \tag{11}$$

where N is the number of responses, $\boldsymbol{t}_i$ is the target output, and $\boldsymbol{y}_i$ is the network's prediction for the response variable, i.e., the DOA and range estimation of the network.

After defining the network structure, we selected the best configuration of the parameters in the network based on fine tuning. The network is trained using stochastic gradient descent with momentum (SGDM) with an initial learning rate of 0.001. To prevent the training from reaching a sub-optimal result or diverging, we drop the learnin by multiplying it with 0.1 every 5 epochs. Additionally we set the maximum number of epochs for training to 20, and the size of the mini-batch to use for each training iteration is 128. We shuffle the training data before each training epoch and the validation data before each network validation.

IV. Simulations and Analyses

A. Simulation Settings

The proposed method is evaluated on a 9-element uniform linear array, with inter-element spacing of the ULA being a quarter of the wavelength. Theoretically, multi-sources can be trained simultaneously to output their DOA and range estimations using our DNN framework, while for the sake of simplicity, we take one source as an example. And the location of the source is set as $(r, \theta) = (1.7\lambda, 12^\circ)$, which is in the Fresnel region of the array aperture $(0.62\lambda < r < 2\lambda)$, according to the definition in Section II. To generate the training data set, the source's DOA impinges from the spatial scope of $[-60^\circ, 60^\circ)$ from the center of the array within the range of 0.70λ to 1.90λ deliberately. For simulations on the DOA of the source, the spatial spectrum is discreted at 1° intervals, thus there are $I = 120$ grids in total with $\theta_1 = -60^\circ$, $\theta_2 = -59^\circ, \cdots, \theta_I = 59^\circ$. As for the simulations of the range of the source, the plane wave is discreted at 0.01λ intervals from the spatial scope of $[0.70, 1.89)$. Similarly, there are also $P = 120$ grids in total. 1000 examples are generated to sample each DOA and range entry. Hence there are 120000 samples in total, and we set aside 12000 samples for network validation. The training set and validation set for each number of snapshots are of the same size as for SNR. The results are compared with the GESPRIT method in [20] and the Cramer-Rao Bound (CRB) in [21].

B. DOA and Range Estimation Performance

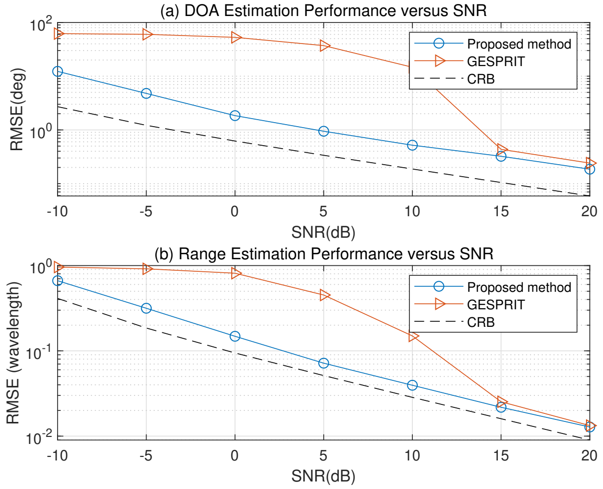

Fig. 2. RMSEs of the (a) DOA and (b) range estimates versus SNR

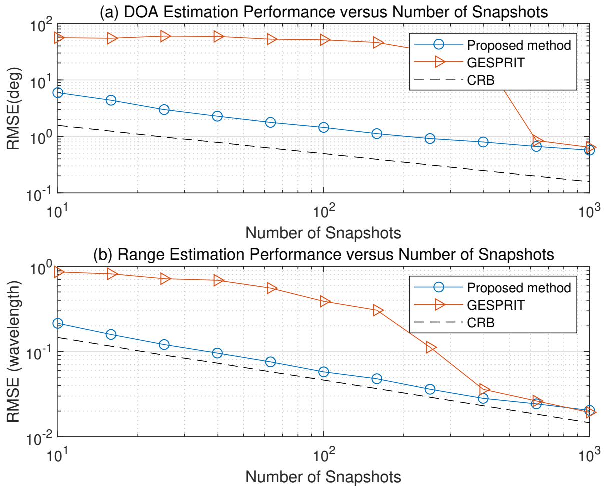

Fig. 3. RMSEs of the (a) DOA and (b) range estimates versus the number of snapshots

TABLE I
VALIDATION ACCURACY PERFORMANCE OF DOA AND RANGE VERSUS SNR

SNR	-10	-5	0	5	10	15	20	25	30	35
DOA	0.1077	0.2928	0.5112	0.7822	**0.9222**	**0.9702**	**0.9765**	**0.9825**	**0.9842**	**0.9800**
Range	0.1390	0.2917	0.5037	0.7032	0.7787	0.8365	0.8670	0.8817	0.8865	**0.9070**

TABLE II
VALIDATION ACCURACY PERFORMANCE OF DOA AND RANGE VERSUS NUMBER OF SNAPSHOTS

Snapshots	$10^{1.0}$	$10^{1.2}$	$10^{1.4}$	$10^{1.6}$	$10^{1.8}$	$10^{2.0}$	$10^{2.2}$	$10^{2.4}$	$10^{2.6}$	$10^{2.8}$
DOA	0.3315	0.3932	0.4755	0.5642	0.6310	0.7098	0.7938	0.8560	0.8813	**0.9407**
Range	0.3037	0.3428	0.4273	0.4862	0.5610	0.6357	0.7112	0.7938	0.7820	**0.9010**

Fig. 2 illustrates the RMSEs performance of the DOA and range estimation versus SNR, respectively. The number of snapshots is set as 64, and SNR varies from -10dB to 20 dB. It can be seen that the RMSEs of both DOA and range decrease with the increasing SNR. Additionally, it shows that the proposed method outperforms GESPRIT method in adverse conditions such as low SNR, which can be better adapted to practical applications.

Similarly, Fig. 3 displays the RMSEs performance of the DOA and range estimation versus the number of snapshots, respectively. In this simulation, SNR is set as 4dB, and the number of snapshots vary from 10 to 1000. It shows that the RMSEs of DOA estimation and range estimation roughly observe a monotonic decreasing pattern with the increasing number of snapshots. Corresponding to the estimation performance versus SNR, the proposed method generally performs better than GESPRIT, especially when the number of snapshots is relatively small. It is noted that most conventional methods suffer from demanding tasks like estimation in environments with low SNR or small number of snapshots, while the proposed method shows potential capability of complementary use of such DNNs with other techniques.

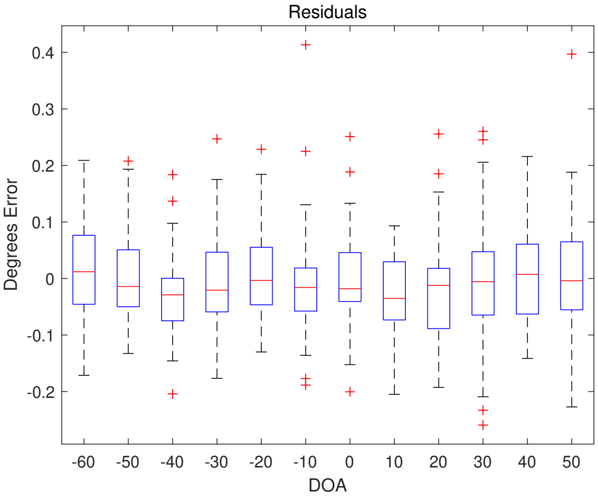

Fig. 4. Box plot of degrees error for some DOA estimations

Further more, we want to explore the parameters estimation performance in details, such as the distribution of degrees error. A box plot is thus proposed like Fig. 4 for this purpose. In this experiment, we take the DOA estimation as an example, where the number of snapshots and SNR are set as 500 and 10dB, respectively. We select 12 directions as representatives of the whole spatial scope. Through the quartiles information depicted in the figure, we can see that the DOAs have a mean close to zero and little variability outside the upper and lower quartiles. Besides this, only a small number of outliers appear as individual points, which demonstrates the robustness of the proposed method.

Table. 1 and Table. 2 show the validation accuracy performance of DOA and Range versus SNR and the number of snapshots, where the simulation settings are the same as the experiments above. Different from most classification model based methods, which consider the DOA estimation is correct as long as the deviation of the estimated direction is within 1°, or even 10°, we set our evaluation resolution to be 0.5° for DOA estimation and 0.05λ for range estimation. The simulation results are rather satisfactory nonetheless. For DOA estimation versus SNR, the validation accuracy approaches to above 90 percent when SNR is above 10 dB. Besides, the proposed method also gave reliable performance when the number of snapshots was sufficiently large. In fact, if the deviation is set as 1° for DOA estimation and/or 0.1λ for range estimation, the validation accuracy will be much higher than the results showed in the tables.

C. *Impact of Convolution Layers*

In addition to the evaluation about the parameters of interests, we also carried on experiments to investigate the impacts of the convolution layers to the whole network, which could provide insight into the improvements of the deep neural networks. The different networks are evaluated with the number of snapshots equals to 64. We compared the DOA estimation performance versus SNR with number of convolution layers varying from 1 to 4, separately. According to Fig. 5, it can be seen that the estimation performance improves ideally with the number of convolution layers increasing from 1 to 3, which is reasonable because a deeper network comes with stronger

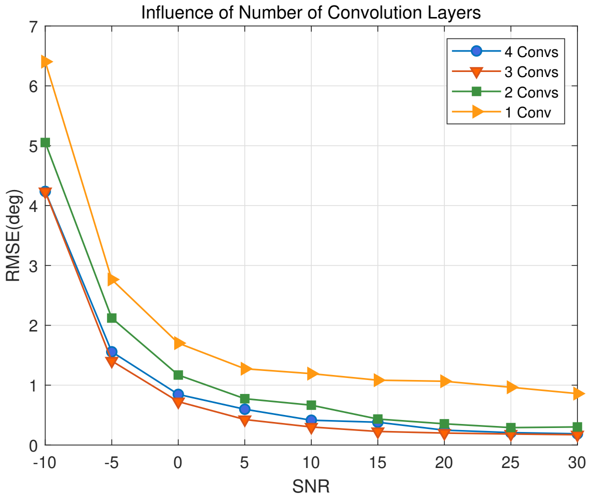

Fig. 5. Convolution layers' impact on RMSE of DOA estimates versus SNR

expression ability, theoretically [24]. However, it is noted that the DNNs consisting of three convolution layers outperformed those of four layers. This can be explained by the increasing under-training risk caused by excessive parameters [25]. In fact, too many convolution layers can also lead to overfitting and the degeneration of the networks, not to mention the increasing computing costs. As a result, we choose three convolution layers as a trade-off between the expression power and various risks.

V. Conclusion

This paper proposes a regression approach with a deep neural network framework to deal with the problem of near-field source localization, so as to make up for the previous data driven methods in terms of deficiency of regression model and low estimation resolution. The proposed approach has a unique design in feature extraction form and consists mainly of three convolution layers and one regression layer in the end of the network. In spite of the simplicity of the training data set, the proposed approach is shown to have a high estimation accuracy with a reasonable high resolution. It also outperforms the conventional methods in scenarios like low SNR and/or small number of snapshots. Furthermore, one specific example shows that the degree errors have a mean close to zero and little variability.

In this work, we only discussed the single source scenario with a fixed level of noise. Future work involves testing the proposed method with different noise types and adjusting the method for multiple source localization, especially for coherent or correlated sources.

References

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67-94, 1996.

[2] R. O. Schmidt, "Multiple emitter location and signal parameters estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 267-280, 1986.

[3] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984-995, 1989.

[4] E. Grosicki, K. Meraim, and Y. Hua,"A weighted linear prediction method for near-field source localization," *IEEE Trans. Signal Process.*, vol. 53, pp. 3651-3660, 2005.

[5] N. Kabaoglu, H. A. Cirpan, E. Cekli, and S. Paker,"Deterministic maximum likelihood approach for 3-D near field source localization," *Int. J. Electron. Commun. (AE)*,vol. 57, no. 5, pp. 345-350, 2003.

[6] W. Zhi and M. Y. W. Chia, "Near-field source localization via symmetric subarrays," *IEEE Signal Process. Lett.*, vol. 14, pp. 409-412, 2007.

[7] A. L. Swindlehurst and T. Kailath,"Passive direction-of-arrival and range estimation for near-field sources," *Proc. 1988 IEEE 4th ASSP Workshop Spec. Est. Mod.*, pp. 123-128, 1988.

[8] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals" *IEEE J. Sel. Top. Signal Process.*, doi: 10.1109/JSTSP.2019.2901664, 2019.

[9] P. R. P. Hoole, *Smart antennas and signal processing for communications, biomedical and radar systems*, Southampton, U.K., WIT Press, 2001.

[10] X. Xiao et al., "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," *2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 2814-2818, 2015.

[11] S. Chakrabarty and E. A. P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," *2017 IEEE Workshop Appl. Signal Process. Audio, Acoust (WASPAA)*, pp. 136-140, 2017.

[12] Q. Li, X. Zhang and H. Li, "Online direction of arrival estimation based on deep learning," *2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 2616-2620, 2018.

[13] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," *2016 IEEE Spok. Lang. Technol. Workshop (SLT)*, pp. 603-609, 2016.

[14] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," *2016 IEEE 26th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, pp. 1-6, 2016.

[15] S. Adavanne, A. Politis and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," *2018 26th Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1462-1466, 2018.

[16] Z. Liu, C. Zhang and P. S. Yu, "Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections," *IEEE Trans. Antennas Propag.*, vol. 66, no. 12, pp. 7315-7327, Dec. 2018.

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

[18] E. M. Grais, M. U. Sen and H. Erdogan, "Deep neural networks for single channel source separation," *2014 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 3734-3738, 2014.

[19] Z. Huang, J. Xu and J. Pan, "A Regression Approach to Speech Source Localization Exploiting Deep Neural Network," *2018 IEEE 4th Int. Conf. Multimedia Big Data (BigMM)*, pp. 1-6, 2018.

[20] W. Zhi and M. Y. Chia, "Near-field source localization via symmetric subarrays," *IEEE Signal Process. Lett.*, vol. 14, no. 6, pp. 409-412, 2007.

[21] E. Grosicki, K. Abed-Meraim, and Y. Hua, "A weighted linear prediction method for near-field source localization," *IEEE Trans. Signal Process.*, vol. 53, pp. 3651–3660, 2005.

[22] Y. Kase et al., "DOA estimation of two targets with deep learning," *2018 15th Workshop Posit., Navig. and Commun (WPNC)*, pp. 1-5, 2018.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, no. 15, pp. 1929-1958, 2014.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[25] S. Ioffe, C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv: 1502.03167, 2015.