ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096047

GEOMETRY-AWARE DOA ESTIMATION USING A DEEP NEURAL NETWORK WITH MIXED-DATA INPUT FEATURES

Ulrik Kowalk[1]*, Simon Doclo*[2]*, and Joerg Bitzer*[1]

[1]Jade University of Applied Sciences, Institute for Hearing Technology and Audiology, Oldenburg, Germany
[2]University of Oldenburg, Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Oldenburg, Germany

ABSTRACT

Unlike model-based direction of arrival (DoA) estimation algorithms, supervised learning-based DoA estimation algorithms based on deep neural networks (DNNs) are usually trained for one specific microphone array geometry, resulting in poor performance when applied to a different array geometry. In this paper we illustrate the fundamental difference between supervised learning-based and model-based algorithms leading to this sensitivity. Aiming at designing a supervised learning-based DoA estimation algorithm that generalizes well to different array geometries, in this paper we propose a geometry-aware DoA estimation algorithm. The algorithm uses a fully connected DNN and takes mixed data as input features, namely the time lags maximizing the generalized cross-correlation with phase transform and the microphone coordinates, which are assumed to be known. Experimental results in a reverberant scenario demonstrate the flexibility of the proposed algorithm towards different array geometries and show that the proposed algorithm outperforms model-based algorithms such as steered response power with phase transform.

***Index Terms*—** DoA estimation, deep neural network, microphone array processing

1. INTRODUCTION

Estimating the direction of arrival (DoA) of an acoustic source is required in many speech communication applications, e.g. to steer a beamformer in hearing aids or to steer a camera in a teleconferencing system [1, 2]. Most DoA estimation techniques use multiple microphone signals that are recorded by a microphone array with a known geometry [3, 4]. Apart from classical model-based approaches that exploit time differences at the microphones directly, e.g., the generalized cross-correlation with phase transform (GCC-PHAT) [5, 6], steered response power with phase transform (SRP-PHAT) [7], and subspace-based methods such as multiple signal classification (MUSIC) [8], in the last years many supervised learning-based algorithms based on deep neural networks (DNNs) have been proposed for DoA estimation [9–14].

While model-based DoA estimation algorithms exploit different signal properties, such as time difference of arrival [5] or the covariance matrix estimated from the microphone signals [8], at their core is the computation of an analytic function that incorporates the microphone array geometry. For example in SRP-PHAT [7, 15] the acoustic power is sampled at candidate DoAs by an acoustic beamformer whose steering vector depends on the microphone array geometry. Similarly, the MUSIC pseudo-spectrum [8] is based on the orthogonal projection of the noise subspace of the covariance matrix of the microphone signals with an array geometry-dependent steering vector.

The preceding observations do not apply to supervised learning-based DoA estimation algorithms, where a DNN learns the relationship between input features and the DoA. These data-driven systems are trained using either microphone array recordings (simulated or real-world) or features that are extracted from such recordings. As a consequence, all training data is implicitly based on the underlying array geometry [16], of which the DNN compiles some form of internal representation. Without explicit information on the array geometry, the DNN is only able to learn the relationship between input features and DoA provided that all data originates from the same geometry. This holds for fully connected deep neural networks (FC-DNNs) as well as for convolutional neural networks (CNNs), in which the first layers perform a convolution operation on the input features. If this geometry assumption is not satisfied at inference, e.g., a DNN trained for one array geometry is applied to another array geometry, the DoA estimation performance may be substantially degraded unless a retraining step is performed, requiring a completely new data set [17] besides processing time and power. Fig. 1 illustrates the general dependency of model-based and supervised learning-based DoA estimation algorithms on the array geometry, i.e. model-based algorithms require knowledge of the microphone array geometry to compute an analytic function, whereas supervised learning-based techniques operate using an internal geometry representation.

In this paper we present a feasibility study on geometry-aware supervised learning-based DoA estimation. We propose a deep neural network that takes as input two separate types of independent data, namely the time lags maximizing the GCC-PHAT and the microphone array geometry, which is assumed to be perfectly known. Based on two experiments we demonstrate that the proposed geometry-aware DNN is flexible towards different array geometries and outperforms state-of-the-art model-based algorithms.

2. MODEL-BASED DOA ESTIMATION

We assume an array of M omnidirectional microphones with a known geometry $\mathbf{r}$ capturing a single static acoustic source in a reverberant environment where some background noise is present. In the frequency domain, the signal at the m-th microphone is composed of

$$Y_m(\omega) = S(\omega)H_m(\omega) + V_m(\omega), \tag{1}$$

This work was funded by the German Federal Ministry of Education and Research under the funding program "Forschung an Fachhochschulen", Project ID: 13FH666IB6.

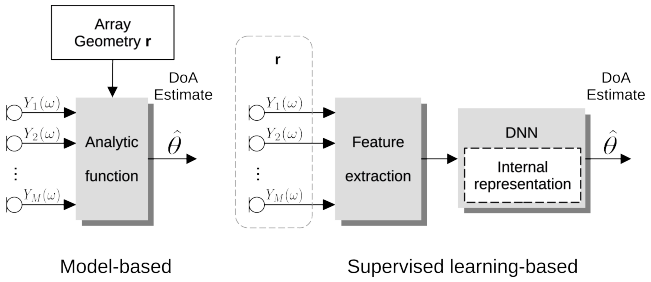

Fig. 1. Model-based DoA estimation approaches: Computation of an analytic function based on the array geometry $\mathbf{r}$. Supervised learning-based approaches: DoA estimation using an internal geometry representation, which depends on the array geometry assumed during training

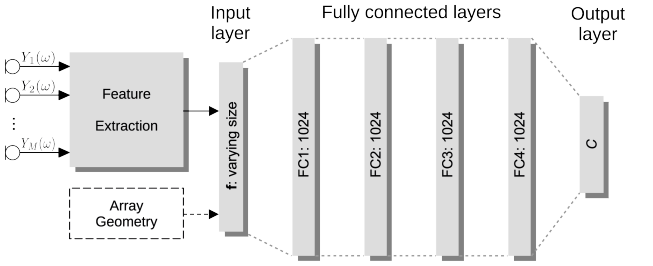

Fig. 2. Universal FC-DNN architecture comprising an input layer of varying size, depending on the algorithm-specific input features, four fully connected (FC) layers, and an output layer with C neurons.

where $S(\omega)$ denotes the source signal, $H_m(\omega)$ denotes the acoustic transfer function between the acoustic source and the m-th microphone, and $V_m(\omega)$ denotes the additive noise component at the m-th microphone.

Based on this signal model, model-based algorithms estimate the DoA of the acoustic source by computing an analytic function incorporating knowledge of the microphone array geometry. For example, SRP-PHAT computes an acoustic power map for candidate DoAs θ using an acoustic beamformer whose steering vector is based directly on the microphone array geometry $\mathbf{r}$ as

$$P(\theta, \mathbf{r}) = 2\pi \sum_{k=1}^{M} \sum_{l=1}^{M} \int_{-\infty}^{\infty} \Gamma_{k,l}(\omega) e^{j\omega\tau_{k,l}(\theta)} d\omega, \tag{2}$$

where $\mathbf{\Gamma}_{k,l}$ denotes the frequency-domain representation of the GCC-PHAT, $\boldsymbol{\gamma}_{k,l}$, between microphones k, l, which is defined as

$$\boldsymbol{\gamma}_{k,l} = \mathcal{F}^{-1} \left\{ \frac{Y_k(\omega) \cdot Y_l^*(\omega)}{|Y_k(\omega) \cdot Y_l^*(\omega)|} \right\}, \tag{3}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform and $(\cdot)^*$ denotes complex conjugation. The DoA-specific time delay $\tau_{k,l}(\theta)$ between microphones k, l is determined by their coordinates, i.e.

$$\tau_{k,l}(\theta) = \frac{|\mathbf{r}_k - \mathbf{r}_l| \sin\theta}{c}, \tag{4}$$

where c is the speed of sound, indicating the direct dependence of P on $\mathbf{r}$. The MUSIC algorithm [8] computes the noise subspace from the covariance matrix of the microphone signals and establishes the so-called MUSIC pseudo-spectrum $P(\theta, \mathbf{r})$ as

$$P(\theta, \mathbf{r}) = \frac{1}{||\mathbf{a}^H(\theta, \mathbf{r}) \mathbf{E}_N \mathbf{E}_N^H \mathbf{a}(\theta, \mathbf{r})||}, \tag{5}$$

where $\mathbf{E}_N$ denotes the noise subspace and $\mathbf{a}(\theta, \mathbf{r})$ is an array geometry-dependent steering vector, again indicating a direct dependence of P on $\mathbf{r}$ and showing the flexibility of model-based algorithms towards different microphone array geometries. Please refer to [8] for more detailed information on MUSIC.

3. SUPERVISED LEARNING FOR DOA ESTIMATION

In this section we discuss three supervised learning-based DoA estimation algorithms, which share the same FC-DNN architecture (Section 3.1). After describing the conventional GCC-PHAT input features in Section 3.2, in Section 3.3 we introduce a reduced set of GCC-PHAT-based input features. In Section 3.4, we propose a geometry-aware DoA estimation algorithm, which uses the microphone geometry as additional input features.

3.1. DNN Architecture

In this study, the DoA estimation problem is formulated as a multi-class classification task, with C DoA classes $\boldsymbol{\theta}$ that span the 360° azimuth range. For each signal frame the class maximizing the posterior probability map $\boldsymbol{P}$ in the output layer is determined by

$$I = \underset{i \in C}{\arg\max}\, P_i, \tag{6}$$

with the DoA estimate given by $\hat{\theta} = \theta_I$. Because in most cases the posterior probability is not concentrated in a single class, a better estimate is obtained through parabolic interpolation [18] incorporating the three DoA classes centered around the maximum. For signals consisting of multiple frames, the global DoA estimate $\tilde{\theta}$ is calculated as the median value over all frames.

For all supervised learning-based algorithms we consider an FC-DNN architecture (see Fig. 2) consisting of an input layer, four fully connected hidden layers with 1024 neurons each, followed by an output layer with C neurons. Every fully connected layer comprises a 20 % dropout stage and is activated by a ReLU function. Because each of the FC-DNNs uses different input features, the sizes of their input layers are different, as indicated in Fig. 2.

3.2. FC-DNN with GCC-PHAT (FC$_{full}$)

The first FC-DNN uses a discrete time version of the GCC-PHAT in (3) as input feature. GCC-PHAT is a representation of the time differences between microphone signals and has been shown to be robust against reverberation and noise due to its phase transform property [19]. For practical reasons we constrain the discrete time lag τ^δ to the interval $[-\tau^\delta_{max}, \tau^\delta_{max}-1]$, where τ^δ_{max} corresponds to the largest possible inter-microphone time delay measured in samples, i.e.

$$\tau^\delta_{max} = \left\lceil \frac{r_{max} \cdot f_s}{c} \right\rceil + \eta, \tag{7}$$

where r_{max} denotes the largest inter-microphone distance within the array, f_s denotes the sampling rate, $\lceil \cdot \rceil$ indicates the operation of rounding up, and η denotes an additional margin. By concatenating the discrete GCC-PHAT vectors for all non-redundant microphone pairs, the final feature vector is constructed as

$$\boxed{\mathbf{f}_{full} = [\boldsymbol{\gamma}^\delta_{1,2}, \boldsymbol{\gamma}^\delta_{1,3}, \ldots, \boldsymbol{\gamma}^\delta_{M-1,M}]} \tag{8}$$

Because this FC-DNN uses the full GCC-PHAT feature vector in (8) as input feature, it is referred to as FC$_{full}$. The size of the input layer is equal to $M(M-1)/2 \cdot 2\tau^{\delta}_{max}$, i.e., the number of non-redundant microphone pairs multiplied by the width of each GCC-PHAT.

3.3. FC-DNN with GCC-PHAT maximum locations (FC$_{max}$)

The second FC-DNN uses a reduced feature set of GCC-PHAT-based features as input feature, where we propose to only use the location of the maximum in each GCC-PHAT vector. This is performed for three reasons. First, the size of the feature set is drastically reduced. Second, this excludes local maxima in the GCC-PHAT, e.g. arising from room reflections. Third, it can be argued that the location of the maximum is the most important piece of information in GCC-PHAT, where we would like the DNN to focus on. We first determine the discrete time lag that maximizes the GCC-PHAT using

$$d_{k,l} = \arg\max_{\tau^{\delta}} \boldsymbol{\gamma}^{\delta}_{k,l}. \tag{9}$$

Since the true maximum is most likely situated between two discrete time lags, we then determine the interpolated time lag $\tilde{d}_{k,l}$ by applying parabolic interpolation between $d_{k,l}$-1 and $d_{k,l}$+1. The final feature vector is constructed by concatenating all estimated time lags as

$$\boxed{\mathbf{f}_{max} = [\tilde{d}_{1,2}, \tilde{d}_{1,3}, \ldots, \tilde{d}_{M-1,M}]} \tag{10}$$

Because this DNN uses the locations of GCC-PHAT maxima, it is referred to as FC$_{max}$. The size of the input layer is equal to $M(M-1)/2$, i.e., the number of non-redundant microphone pairs.

3.4. Geometry-aware FC-DNN (FC$_{GA}$)

Whereas the FC-DNNs in the previous sections only use signal-based input features, in this section we propose a geometry-aware FC-DNN, which uses a combination of two types of independent data as input feature: the GCC-PHAT maximum locations, $\mathbf{f}_{max}$ in (10), and the microphone coordinates $\mathbf{f}_r$, separated into their x and y components, i.e.

$$\mathbf{f}_r = [x_1, \ldots, x_M, y_1, \ldots, y_M] \tag{11}$$

Both parts, maximum locations and array coordinates, are then concatenated to form the final feature set:

$$\boxed{\mathbf{f} = [\mathbf{f}_{max}, \mathbf{f}_r]} \tag{12}$$

It should be noted that experiments have shown that the combination of GCC-PHAT maximum locations and microphone coordinates produces more accurate and robust estimates than the combination of the full GCC-PHAT vector and microphone coordinates. Since this DNN is geometry-aware, it is referred to as FC$_{GA}$. The size of the input layer is equal to $M(M-1)/2 + 2M$, i.e., the number of non-redundant microphone pairs plus the x- and y-coordinates of the microphones.

4. EXPERIMENTAL EVALUATION

In this section the performance of the proposed geometry-aware DNN is validated and compared to several baseline algorithms under different acoustic conditions. In Section 4.1 the considered acoustic setup is presented, while Section 4.2 describes the training procedure. The performance metrics are introduced in Section 4.3, and the evaluation results are presented in Section 4.4. In addition to FC$_{full}$ and FC$_{max}$ as baseline algorithm we consider the CNN-based algorithm from [11], which uses raw signal phases as input feature. We further consider SRP-PHAT and MUSIC, both implemented in [20], as baseline model-based algorithms as well.

Room dimensions:	[9.0, 5.0, 3.0] m $\pm$ [1.0, 1.0, 0.5] m
Array position:	[4.5, 2.5, 1.5] m $\pm$ [0.5, 0.5, 0.5] m
Source distance:	1.0 - 3.0 m [within boundaries]
Source direction:	0° : 5° : 355°
T_{60}:	0.13 s - 1.0 s
SNR:	0 - 30 dB

Table 1. Acoustic simulation parameters

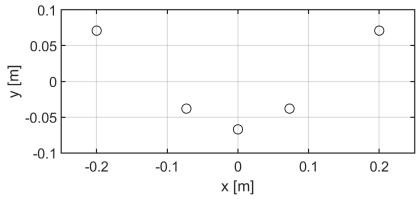

Fig. 3. 2-dimensional arc-shaped microphone array geometry

4.1. Acoustic Setup and Algorithm Parameters

For the evaluation we consider a 2-dimensional arc-shaped microphone array comprising M=5 microphones with a width of 0.4 m and a depth of about 0.15 m, as shown in Fig. 3. The microphone array is situated inside a rectangular room and records the signal from a single acoustic source and noise. For both training and evaluation we utilize data consisting of 50% speech and 50% white noise as suggested in [14]. Speech data is taken from the "clean" section of the *LibriSpeech* corpus [21]. Directional cues are simulated by convolving monophonic signals with room impulse responses (RIRs) generated using the image source method implemented in *pyroomacoustics* [20]. The additive noise consists of diffuse-like stationary babble noise generated using [22, 23].

In the output layer of the FC-DNNs we consider C=72 DoA classes, leading to an angular resolution of 5°. Since for the considered array geometry τ^{δ}_{max}=14, including a margin, the input layer size for FC$_{full}$, FC$_{max}$, and FC$_{GA}$ is equal to 280, 10 and 20, respectively. Training as well as evaluation are performed on non-overlapping Hann windowed 32 ms signal frames at a sampling rate of f_s=8 kHz, yielding a frame size of 256 samples.

4.2. Training

Every training sample consists of a single frame containing a single acoustic source as well as noise. We introduce variability in the acoustic parameters defining the training data aiming to achieve a robust algorithm that generalizes well to unmatched acoustic conditions. To this end, for every training sample a rectangular room with different dimensions and acoustic properties and an acoustic source with different directions and distances according to Table 1 is considered. FC$_{full}$, FC$_{max}$ and the CNN are trained using the microphone array geometry in Fig. 3. The proposed geometry-aware FC$_{GA}$ is trained using a 2-dimensional array of randomly positioned microphones with a width and depth of 0.4 m, whose coordinates are perfectly known, where a completely different array is considered for every training sample. We employ mini batches of 32 samples, the Adam optimizer with a learning rate of 10^{-4}, and a cross entropy loss function. Training is concluded when no reduction of the validation loss can be observed for 10 epochs.

4.3. Performance Metrics

Two performance metrics are used to quantify the performance of the considered algorithms: mean absolute error (*MAE*) and *Accuracy*. In order to account for circular wrapping, first the absolute angular error, δ_n, is calculated for every estimate n as

$$\delta_n = \left| \arg\left\{ e^{i2\pi \cdot (\tilde{\theta}_n - \theta_n) \,/\, 360^\circ} \right\} \right| \cdot \frac{360^\circ}{2\pi}, \qquad (13)$$

where $\tilde{\theta}_n$ is the estimated global DoA defined in Section 3.1 and θ_n is the ground truth DoA. Based on δ_n both performance metrics are then calculated as

$$MAE\,[^\circ] \quad = \quad \frac{1}{N} \sum_{n=1}^{N} \delta_n, \qquad (14)$$

$$Accuracy\,[\%] \quad = \quad \frac{1}{N} \sum_{n=1}^{N} \Theta\left(\varepsilon - \delta_n\right) \times 100, \qquad (15)$$

where N is the overall number of trials, Θ represents the Heaviside step function and ε denotes the margin of tolerance – here we use a margin of one DoA class, i.e. 5°.

4.4. Evaluation Results

This section presents the results of two experiments employing signals of length 5 s with moderate reverberation and low noise.

First experiment: Coordinates deviating from trained geometry

The first experiment investigates the sensitivity of the baseline supervised learning-based algorithms (CNN, FC$_{full}$, FC$_{max}$) to coordinates deviating from the trained array geometry compared to the sensitivity of the proposed geometry-aware algorithm (FC_{GA}) and the performance of the baseline model-based algorithms. In this experiment, random rooms and source positions are simulated according to Table 1, but with a fixed reverberation time of T_{60}=0.5 s and SNR=20 dB. Starting from the microphone array in Fig. 3, we consider microphone coordinate deviations in all directions. For every deviation in steps of 10^{-2} m, all microphone coordinates deviate by the same amount, but in separate, random directions. Each deviation step is simulated 10^4 times.

Figs. 4 and 5 illustrate the effect of microphone coordinate deviation on all considered DoA estimation algorithms in terms of MAE and Accuracy, respectively. First, it can be observed that the baseline supervised learning-based algorithms (CNN, FC$_{full}$, FC$_{max}$) outperform the baseline model-based algorithms (SRP-PHAT, MUSIC) when no coordinate deviations occur. Second, up to a deviation of about 0.01 m, no considerable decrease in performance can be observed for the supervised learning-based baseline systems. Larger deviations, however, lead to a substantial decrease in performance. Third, it can be clearly observed that the proposed geometry-aware algorithm (FC$_{GA}$) is much more robust to coordinate deviations than the baseline supervised learning-based algorithms.

Second experiment: Fully randomized geometry

We conducted 10^4 simulations, each using a different 2-dimensional array geometry with uniformly distributed coordinates, with a maximum width and depth of 0.4 m. For all considered algorithms in this experiment, we assume that the microphone coordinates are perfectly known. All other acoustic parameters are the same as in the first experiment, again with T_{60}=0.5 s and SNR=20 dB.

Table 2 shows the results in terms of MAE and Accuracy averaged over all simulations. These results show that the proposed geometry-aware algorithm (FC$_{GA}$) not only generalizes to unseen array geometries, but also outperforms the model-based baseline algorithms (SRP-PHAT, MUSIC), both in terms of MAE and Accuracy.

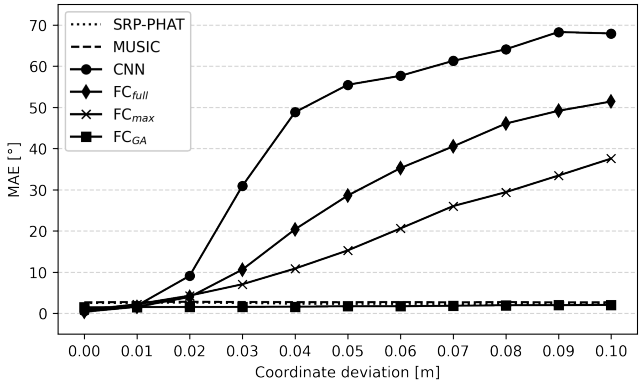

Fig. 4. MAE averaged over all simulations of model-based (SRP-PHAT, MUSIC) and supervised learning-based (CNN, FC$_{full}$, FC$_{max}$, FC$_{GA}$) algorithms with respect to deviating array coordinates (T_{60}=0.5 s, SNR=20 dB)

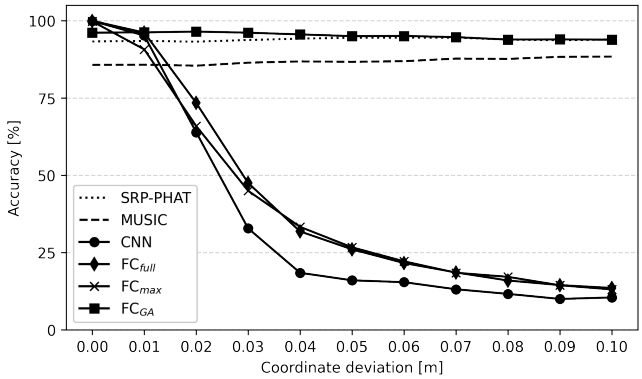

Fig. 5. Accuracy of model-based (SRP-PHAT, MUSIC) and supervised learning-based (CNN, FC$_{full}$, FC$_{max}$, FC$_{GA}$) algorithms with respect to deviating array coordinates (T_{60}=0.5 s, SNR=20 dB)

Algorithm	**MAE [°]**	**Accuracy [%]**
SRP-PHAT	2.44	93.5
MUSIC	2.69	86.0
FC$_{GA}$	**1.47**	**96.1**

Table 2. MAE and accuracy of model-based algorithms (SRP-PHAT, MUSIC) and proposed geometry-aware (FC$_{GA}$) algorithm for the randomized geometry simulation with T_{60}=0.5 s and SNR=20 dB

5. CONCLUSIONS

In this paper we presented a feasibility study on geometry-aware DNN-based DoA estimation. First, we demonstrated the sensitivity of common supervised learning-based techniques towards deviations from the array geometry employed during training. The proposed geometry-aware algorithm uses as a novel feature the locations of GCC-PHAT maxima alongside the microphone coordinates. Experimental results showed that the proposed geometry-aware algorithm outperformed the evaluated model-based algorithms for different array geometries. Further studies will investigate the generalization to 3-dimensional arrays, the performance in adverse acoustic conditions as well as the robustness to inaccuracies in the assumed microphone coordinates.

6. REFERENCES

[1] Y. Huang, J. Benesty, and G. W. Elko, "Microphone arrays for video camera steering," in *Acoustic Signal Processing for Telecommunication*, pp. 239–259. Springer, 2000.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on array processing and sensor networks*, vol. 9, pp. 269–302. Wiley Hoboken, NJ, USA, 2010.

[3] Y. A. Huang, J. Benesty, and J. Chen, "Time delay estimation and source localization," in *Springer Handbook of Speech Processing*, pp. 1043–1063. Springer, 2008.

[4] N. Madhu and R. Martin, "Acoustic source localization with microphone arrays," in *Advances in Digital Speech Transmission*, pp. 135–170. Wiley UK, 2008.

[5] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[6] G. C. Carter, "Coherence and time delay estimation," *Proc. of the IEEE*, vol. 75, no. 2, pp. 236–255, 1987.

[7] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, PhD thesis, Brown University, 2000.

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[9] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[10] Z. Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178–188, 2018.

[11] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[12] W. Zhang, Y. Zhou, and Y. Qian, "Robust DOA estimation based on convolutional neural network and time-frequency masking," in *Proc. Interspeech*, Graz, AUT, 2019, pp. 2703–2707.

[13] W. Mack, J. Wechsler, and E. A. P. Habets, "Signal-aware direction-of-arrival estimation using attention mechanisms," *Computer Speech & Language*, vol. 75, 2022, article 101363.

[14] D. Krause, A. Politis, and K. Kowalczyk, "Data diversity for improving DNN-based localization of concurrent sound events," in *Proc. European Signal Processing Conference (EUSIPCO)*, Dublin, IRL, 2021, pp. 236–240.

[15] A. Johansson, N. Grbić, and S. Nordholm, "Speaker localisation using the far-field SRP-PHAT in conference telephony," in *Proc. IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Penang, MYS, 2002.

[16] S. Zhang and X. Li, "Microphone array generalization for multichannel narrowband deep speech enhancement," in *Proc. Interspeech*, Brno, CZE, 2021, pp. 666–670.

[17] T. Yoshioka, X. Wang, D. Wang, M. Tang, Z. Zhu, Z. Chen, and N. Kanda, "Vararray: Array-geometry-agnostic continuous speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, SGP, 2022, pp. 6027–6031.

[18] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proc. International Computer Music Conference (ICMC)*, Champaign/Urbana, IL, USA, 1987, International Computer Music Conference, pp. 290–297.

[19] C. Zhang, D. Florêncio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 2565–2568.

[20] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, CAN, 2018, pp. 351–355.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, 2015, pp. 5206–5210.

[22] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[23] J. A. Adrian, T. Gerkmann, S. van de Par, and J. Bitzer, "Synthesis of perceptually plausible multichannel noise signals controlled by real world statistical noise properties," *Journal of the Audio Engineering Society*, pp. 914–928, 2017.