End-to-End Signal-Aware Direction-of-Arrival Estimation Using Weighted Steered-Response Power

Julian Wechsler, Wolfgang Mack, Emanuël A. P. Habets
International Audio Laboratories Erlangen*, Am Wolfsmantel 33, 91058 Erlangen, Germany
{julian.wechsler,wolgang.mack,emanuel.habets}@audiolabs-erlangen.de

***Abstract*—The direction-of-arrival (DOA) of acoustic sources is an important parameter used in multichannel acoustic signal processing to perform, e.g., source extraction. Deep learning-based time-frequency masking has been widely used to make DOA estimators signal-aware, i.e., to localize only the sources of interest (SOIs) and disregard other sources. The mask is applied to feature representations of the microphone signals. DOA estimators can either be model-based or deep learning-based, such that the combination with the deep learning-based masking estimator can either be hybrid or fully data-driven. Although fully data-driven systems can be trained end-to-end, existing training losses for hybrid systems like weighted steered-response power require ground-truth microphone signals, i.e., signals containing only the SOIs. In this work, we propose a loss function that enables training hybrid DOA estimation systems end-to-end using the noisy microphone signals and the ground-truth DOAs of the SOIs, and hence does not dependent on the ground-truth signals. We show that weighted steered-response power trained using the proposed loss performs on par with weighted steered-response power trained using an existing loss that depends on the ground-truth microphone signals. End-to-end training yields consistent performance irrespective of the explicit application of phase transform weighting.**

***Index Terms*—Direction-of-Arrival, Steered-Response Power, Deep Learning, Time-Frequency Masking, End-to-End**

I. Introduction

In an acoustic environment, multiple sound sources can be simultaneously active. Consider, e.g., a scenario where a desired speaker and undesired directional sound are captured using a microphone array. To perform multichannel speech enhancement (SE) or steer a camera towards the desired speaker, the direction-of-arrival (DOA) of the source of interest (SOI) with respect to the microphone array could be used. Reverberation, noise, and interference render estimating the DOA challenging. The task of estimating the DOAs of the SOIs is commonly referred to as signal-aware DOA estimation.

DOA estimation is a well studied subject [1]–[3]. Signal processing methods are usually based on time differences of arrival between the microphone signals of an array, which depend on the DOAs. As natural sounds typically exhibit structure in frequency, a frequency transform, e.g., the short-time Fourier transform (STFT), is usually applied to the microphone signals. The time differences of arrival translate to phase differences per time-frequency (TF) bin. Consequently, DOA estimation can be performed "narrowband" or "broadband", i.e., based on single or all frequency bands, respectively. Signal processing methods for DOA estimation can rely on i) inter-microphone cross-correlations [4], [5], where the dependency between the DOAs and the phase difference is directly exploited, ii) beamformers [6]–[9], where spatial sampling is performed, and the DOAs are obtained by peak-picking, or iii) signal or noise subspaces [9]–[14] that are constructed from the power spectral density matrix.

Progress in deep learning has enabled DOA estimation also in more adverse and challenging scenarios. Commonly, a feature representation in the STFT domain is fed into a deep neural network (DNN) where the objective is to i) estimate the DOA directly [15]–[20] by mapping the features to a representation of the DOAs, or ii) aid a further DNN or a signal processing method that estimates the DOA by masking the features [20]–[24] or weighting the averaging of the narrowband DOA estimates [25], [26]. Including DNNs can robustify an estimator against reverberation and noise or make it signal-aware [23], [24]. Estimation systems can be fully data-driven or hybrid, i.e., consist solely of DNNs, or of DNNs and signal processing methods, respectively. Hybrid systems were found to be more flexible at lower computational complexity and comparable performance [24]. We focus on hybrid systems where the masking introduces signal-awareness.

When combining the masking and the DOA estimators, training (if applicable) can be performed separately or jointly. Possible training objectives are i) SE objectives, e.g., a phase-sensitive mask (PSM) [23], ii) objectives based on the spatial pseudo-spectrum (SPS) [24], or iii) end-to-end objectives [27], [28]. The former two require the ground-truth (GT) microphone signals, i.e., signals containing only the SOIs, the latter only the GT DOA. As only the GT DOA may be available in real-world scenarios, end-to-end training is a desirable option. To the best of the authors' knowledge, there are no end-to-end training strategies for hybrid signal-aware DOA estimation (HYSADE) systems. This work proposes an interpretable loss function for end-to-end training of HYSADE systems. Here, we demonstrate the applicability of this loss function for deep learning-based weighted steered-response power with phase transform (SRP-PHAT). We show that the estimation performance is comparable to a state-of-the-art method that requires the GT signals.

II. Background

In this section, we formalize HYSADE, describe the signal model, and review TF masks from [23], [24], which are later used as baselines.

*A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

A. Problem Formulation

We assume a uniform linear array with Q microphones[1] in a reverberant room. The sound field comprises S directional sound sources with index $s \in \mathcal{S} = \{1, 2, \ldots, S\}$, where $\mathcal{I} \subseteq \mathcal{S}$ is the set of SOIs. The objective is to localize the SOIs. The STFTs of the microphone signals are denoted as $Y \in \mathbb{C}^{K\times N\times Q}$, where K and N denote the total number of frequency bands and time frames, respectively. The set of complex numbers is represented by $\mathbb{C}$. The microphone signals consist of an additive superposition of the direct and reverberant sound components of the individual sources, X_s^{dir} and X_s^{rev}, respectively, and microphone self-noise V, i.e.,

$$Y = \sum_{s=1}^{S} \left(X_s^{\mathrm{dir}} + X_s^{\mathrm{rev}} \right) + V. \tag{1}$$

Using the SRP method [8], the DOAs can be estimated from Y via spatial sampling in C directions (i.e., DOA candidates), power computation (yielding the SPS), and peak-picking; the estimated DOAs are the angles corresponding to the chosen peaks.

It has been shown that applying a frequency-dependent weighting to Y can significantly improve the DOA estimation when using the SRP method [8]. We consider different real-valued weights $M \in \mathbb{R}^{K\times N\times Q}$,

$$\widetilde{Y} = M \odot Y, \tag{2}$$

where $\odot$ denotes element-wise multiplication. The set of real numbers is represented by $\mathbb{R}$. A weighting known to robustify SRP against adverse acoustic conditions [1], [8] is PHAT weighting, i.e., magnitude normalization of the signals,

$$W_{\mathrm{PHAT}} = 1^{K\times N\times Q} \oslash \left(|Y| + \rho \, 1^{K\times N\times Q} \right), \tag{3}$$

where $1^{K\times N\times Q}$ is an all-ones matrix of dimension $K \times N \times Q$, $\oslash$ denotes element-wise division, $|\cdot|$ denotes the element-wise absolute value, and $\rho \in \mathbb{R}^{+}$ is a small regularization constant.

The weighting can also be devised to only respond to certain source signals, e.g., the SOIs. If the sources in $\mathcal{I}$ are to be localized, the information contained in $X_{\text{SOIs}}^{\mathrm{dir}} = \sum_{i\in\mathcal{I}} X_i^{\mathrm{dir}}$ can be exploited as it contains the direct-path sound from the sources to the array. In an anechoic mixture of speech signals, each TF bin can be assumed to be dominated by a single source [29]. This property is violated, e.g., in the presence of reverberation and broadband sources. To focus on the TF bins supporting the DOA estimates of the sources in $\mathcal{I}$, we additionally apply a TF mask $M_{\mathcal{I}} \in [0, 1]^{K\times N\times Q}$ to the microphone signals Y. In general, the weighting in (2) is a concatenation of all employed individual weightings, i.e.,

$$M = W_{\mathrm{PHAT}} \odot M_{\mathcal{I}}. \tag{4}$$

Finally, the SRPs are computed from $\widetilde{Y}$ after (2). Note that W_{PHAT} is always channel dependent, whereas $M_{\mathcal{I}}$ can also be devised to be channel independent, cf. Section IV-B.

B. Baseline Methods

HYSADE systems have been realized using different acoustic signal processing methods, e.g., multiple signal classification [30] and SRP [22], [23]. Hereinafter, we briefly review the training strategies as adopted in [23], [24].

In [23], the authors propose to leverage a SE mask, the PSM, for the objective of DOA estimation with SRP. The mask is estimated by a DNN trained separately from the DOA estimation task to minimize the mean squared error (MSE) between the network output and the oracle PSM extracting $X_{\text{SOIs}}^{\mathrm{dir}}$ from the microphone signals. In [24], the MSE between the reference SPS based on $X_{\text{SOIs}}^{\mathrm{dir}}$ and the respective SPS based on $\widetilde{Y}$ is used as an optimization criterion. Training of the DNN that computes the mask is performed implicitly within the DOA estimation framework. The localization performances using SRP and either one of the two previously introduced masks were found to be comparable for the task of localizing a talker in the presence of a non-speech interferer and microphone self-noise [24]. As the PSM enables simultaneous SE and accurate localization, it is used as a baseline in this paper.

The training strategies in [23], [24] require $X_{\text{SOIs}}^{\mathrm{dir}}$. As DNNs adapt to their training data, learning from measurements is highly desirable as it is likely to improve the performance over training with simulated data. In particular, when considering measured data, the GT microphone signals are unobservable while the GT DOA can be obtained using, e.g., an optical tracking system [31]. Consequently, a loss function that only requires the GT DOA as training reference is highly desirable. In [24], [27], [28], objectives for purely data-driven signal-aware DOA estimation are proposed that do not require the GT signals. To the best of the authors' knowledge, no such training strategy for HYSADE systems has been proposed yet.

III. Proposed Training Strategy

We propose the power minimization loss (PML), a solely DOA-based loss for HYSADE systems, to train the mask-estimation DNN. Hereinafter, we restrict ourselves to the case of a single SOI. The PML is based on two components, i) the SPS obtained from $\widetilde{Y}$ for time frame $n \in \{1, 2, \ldots, N\}$ and DOA candidate $c \in \{1, 2, \ldots, C\}$, denoted by $\mathrm{SPS}(\widetilde{Y})[c, n]$, and ii) a mapping function $f : \mathbb{R}^{C} \mapsto \mathbb{R}$. The time-averaged SPS is normalized using the time-averaged SPS value associated with the DOA of the SOI, c_{SOI}, and then processed by f, i.e.,

$$\mathrm{PML} = f\left(\frac{\sum_{n=1}^{N} \mathrm{SPS}\left(\widetilde{Y}\right)[c, n]}{\sum_{n=1}^{N} \mathrm{SPS}\left(\widetilde{Y}\right)[c_{\text{SOI}}, n]} \right). \tag{5}$$

In this work, we use the arithmetic mean over c as the mapping function. By minimizing the mean normalized SPS, the loss function mimics the minimum power distortionless response beamformer objective, i.e., minimizing the overall output power while retaining it from one direction. As the normalization guarantees unity of the SPS in the desired direction, training cannot yield an all-zero mask.

[1] Linear arrays sample sound fields along a line in three-dimensional space. As the beam pattern is rotationally symmetric, only DOAs in a half-plane can be distinguished and are identified with the azimuth angle $\vartheta \in [0, \pi]$.

TABLE I
PARAMETERS OF THE RIRS FOR THE THREE DATA SETS. (FROM [24].)

	Training	Validation	Test [37]
T_{60} [s]	$\{.2, .3, .4, .6, .8\}$	$\{.45, .6, .75\}$	$\{.16, .36, .61\}$
SMD [m]	$\{1, 2\}$	$\{1.2, 2.3\}$	$\{1, 2\}$
ϑ [°]	$\{0, 5, \ldots, 180\}$	$\{0, 5, \ldots, 180\}$	$\{0, 15, \ldots, 180\}$
# files	$11.1 \cdot 10^4$	$1.3 \cdot 10^4$	1560

IV. EXPERIMENTAL SETUP

This section briefly describes the generation of the required data sets and elaborates on the model details.

A. Data Sets

We generated training, validation, and test sets comprising signals with a duration of $1.6\,\mathrm{s}$, a sampling frequency of $16\,\mathrm{kHz}$, $S = 2$ sources, and spatiotemporally white noise for a uniform linear array with $Q = 4$ microphones and $8\,\mathrm{cm}$ inter-microphone distance. The (single) SOI is always a speech source from LibriSpeech [32], and the undesired source is a non-speech interference from FSDnoisy18k [33], [34]. Reverberation was simulated by convolving the sources with room impulse responses (RIRs) of different reverberation times T_{60} and of different source microphone-center distances (SMDs). For training and validation, RIRs were simulated using the image method [35], [36], whereas measured RIRs [37] were used for the test (RIRs of the central four microphones of the $8\,\mathrm{cm}$ configuration). All RIR parameters are summarized in Table I and correspond to rooms with sizes in meters ($[\text{length}, \text{width}, \text{height}]$) as follows: $\{[6, 6, 2.7], [5, 4, 2.7], [10, 6, 2.7], [8, 3, 2.7], [8, 5, 2.7]\}$ for training, $\{[9, 11, 2.7], [10, 10, 2.7], [9, 5, 2.7]\}$ for validation, $\{[6, 6, 2.4]\}$ for testing. Note that the angular separation of both sources was larger than $10°$. We steered SRP into $C = 37$ directions, resulting in an angular resolution of $5°$. The angular resolution should be determined depending on the employed array architecture and the expected SMD. We simulated signal-to-interference ratios (SIRs) in the range $[-6, 6]\,\mathrm{dB}$ and signal-to-noise ratios in the range $[20, 30]\,\mathrm{dB}$. All signals were transformed into the STFT-domain with a window length of $32\,\mathrm{ms}$ and a hop size of $16\,\mathrm{ms}$, yielding $K = 257$ frequency bands and $N = 100$ time frames.

B. Models

The considered weighted SRP method combines traditional DOA estimation and deep learning-based mask estimation. We adapted the DNN from [24] as the mask estimator and trained it with the baseline [23] as well as the proposed loss functions. The DNN consists of two long short-term memory layers (hidden dim. = 512) followed by a feed-forward layer with sigmoid activation to ensure mask values in the range $[0, 1]$. The input of the DNN per time frame is the magnitude STFT of one microphone, and the output per time frame is a mask of output dim. = input dim. = K = 257. This mask is then applied to all microphones. This choice was motivated by [24], where no performance difference between using a single PSM applied to all microphones or using multiple (channel-dependent) PSMs was observed.

V. PERFORMANCE EVALUATION

In this section, we evaluate and compare the performance of the proposed end-to-end PML training with the baseline MSE training yielding a PSM [23] for the objective of signal-aware DOA estimation using weighted SRP[2]. One instance of the DNN was trained with the baseline objective, as it is independent of the DOA estimation. Two instances of the DNN were trained with the proposed end-to-end objective, as it depends on the DOA estimation, and the training and outcome might be influenced by including PHAT weighting.

A. Performance Metrics

We use two measures [23], [24], [28] to assess the performance of the methods, whereby all DOA estimates correspond to the maximum of the time-averaged SPS. The mean absolute error (MAE) for a method is given in degrees,

$$\text{MAE} = \frac{1}{F} \cdot \sum_{f} \left(|\vartheta_f - \widehat{\vartheta}_f| \right), \tag{6}$$

where F is the total number of files in the data set, $f \in \{1, 2, \ldots, F\}$ is the file index, ϑ_f is the true DOA of the SOI in file f, and $\widehat{\vartheta}_f$ is the corresponding estimate. Let $F^{(\text{deg})}$ denote the number of files where the absolute error (AE) fulfils $\text{AE} \in [0°, \text{deg}°)$. The accuracy, i.e., the detection rate for an admissible $\text{AE} \in [0°, \text{deg}°)$ (deg-ACC), is given in percent,

$$\text{deg-ACC} = \frac{F^{(\text{deg})}}{F} \cdot 100\%. \tag{7}$$

Using $C = 37$ DOA candidates, the 5-ACC states how often the correct DOA candidate is detected, i.e., how often the AE is below 5 degrees. The 10-ACC states how often the correct DOA candidate or one of its neighboring candidates is detected, i.e., how often the AE is lower than 10 degrees.

B. Performance Analysis

All results are summarized in Table II and are broken down gradually; illustrations of SPSs and masks resulting from the baseline [23] as well as the proposed training can be found in Figures 1 and 2, respectively. The reported performance refers to the test set as described in Section IV-A.

Without masking, the localization of the speech source succeeded only poorly with an MAE of about 40 degrees, irrespective of the application of PHAT weighting. Note that a clear improvement in the deg-ACCs can be observed by including PHAT weighting (e.g., PHAT weighting improved the 5-ACC by 16 %). At constant MAE, this corresponds to sharper peaks of the SPS, as shown in Figure 1 for some typical SPSs (the estimates in the bottom row with PHAT are peakier than those in the top row without PHAT). The low accuracy and the high MAE without masking are expected, as we always selected the highest peak of the SPS to correspond to the estimated DOA of the SOI. Without masking, the

[2] Note that typical end-to-end loss functions like the categorical cross-entropy cannot be employed for weighted SRP due to the broad lobes of the SPS; the loss cannot be effectively minimized in the vicinity of the desired DOA. We reimplemented a standard version of SRP [8], [38] in PyTorch to enable backpropagation of the gradients over SRP.

SPS is expected to have two peaks corresponding to the two directional sources; picking the highest peak can be thought of as localizing either of the two present sources. Assuming that masking improves the performance of localizing the SOI, this result constitutes a lower bound on the performance for our experiments.

For the considered scenario, the inclusion of either signal-aware mask substantially improved the performance. Employing PHAT weighting and the baseline masking [23] achieved the top performance in our experiments, improving the results by 33 degrees in MAE, 31 % in 5-ACC, and 38 % in 10-ACC, with respect to no masking. The proposed method combined with PHAT weighting performed comparably, i.e., only 1 degree worse in MAE, 2 % in 5-ACC, and 1 % in 10-ACC. Training without the GT signals is also successfully possible for HYSADE systems, and the proposed end-to-end loss function yields state-of-the-art performance. Observe how the application of either mask lead to a successful localization of the SOI instead of the interferer in Figure 1.

In Figure 1, the SPSs based on $\widetilde{Y}$ resemble the SPSs based on $X^{\mathrm{dir}}_{\mathrm{SOIs}}$ for the cases "PSM excl. PHAT", "PSM incl. PHAT", and "End-to-End Mask incl. PHAT". Interestingly, for the case "End-to-End Mask excl. PHAT", the SPS based on $\widetilde{Y}$ is peakier than the SPS based on $X^{\mathrm{dir}}_{\mathrm{SOIs}}$, even though no channel-dependent weighting was included. It is comparable to the SPS based on $\widetilde{Y}$ for the case "End-to-End Mask incl. PHAT". This observation is generally reflected in comparable performance with and without PHAT weighting for the proposed training method. We hypothesize that the proposed end-to-end loss enabled the DNN to learn the necessary normalization.

Lastly, we illustrate $\log_{10}(\mathrm{SIR})$ and typical TF masks, i.e., $\log_{10}(M_{\mathcal{I}})$, as output by the DNN in Figure 2. Note that the input to the DNN was the magnitude STFT of the noisy microphone signals $|Y|$ and that the PSM remains unchanged irrespective of the inclusion of PHAT weighting. Within the PSM, the magnitude structure of speech in the TF domain can be observed as it is an SE mask. This structure is not present in the end-to-end masks; however, comparable onsets in time can be observed in the PSM and the end-to-end masks. The end-to-end masks cannot be used for SE purposes[3]. In the lower frequency bands with a lot of speech energy, the end-to-end masks are generally sparser. This finding can be explained by the broader beamformer lobes for low frequencies, which provide little additional information for accurate DOA estimation. The end-to-end training was able to adapt to that property. Other than a slight sparsification by the inclusion of PHAT weighting, the structure of the two end-to-end masks is similar. Note that the end-to-end masks put less emphasis on the low-SIR regions, whereas the PSM also extracts the low-SIR TF bins corresponding to the SOI.

VI. Conclusion

We proposed the PML, an end-to-end training objective for deep learning-based weighted SRP as an example of an

[3] For a demonstration, please visit https://www.audiolabs-erlangen.de/resources/2022-EUSIPCO-HYSADE.

TABLE II
Results using the baseline training from [23] and the proposed end-to-end training, evaluated on the test set as described in Section IV-A. The best results are printed in bold face.

SRP	No PHAT			PHAT		
	MAE	5-ACC	10-ACC	MAE	5-ACC	10-ACC
No Masking	39.9°	20 %	33 %	39.8°	36 %	43 %
Baseline [23]	12.0°	30 %	58 %	**6.8°**	**67 %**	**81 %**
Proposed	8.1°	64 %	79 %	7.8°	65 %	80 %

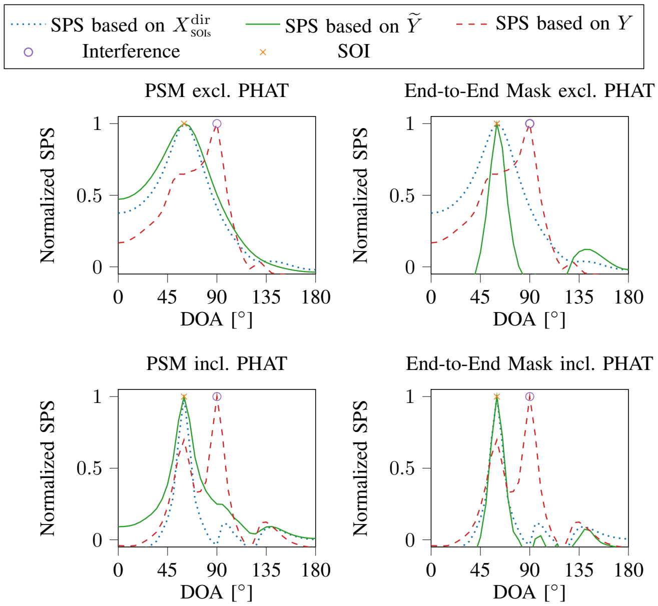

Fig. 1. Illustration of the normalized SPSs incl./excl. PHAT weighting and using the two training objectives for the DNN, the baseline objective after [23] (left column) and the proposed objective (right column). Observe how PHAT weighting improved the localization of the SOI (at $\vartheta = 60°$), while disregarding the interferer (at $\vartheta = 90°$) only for the baseline training.

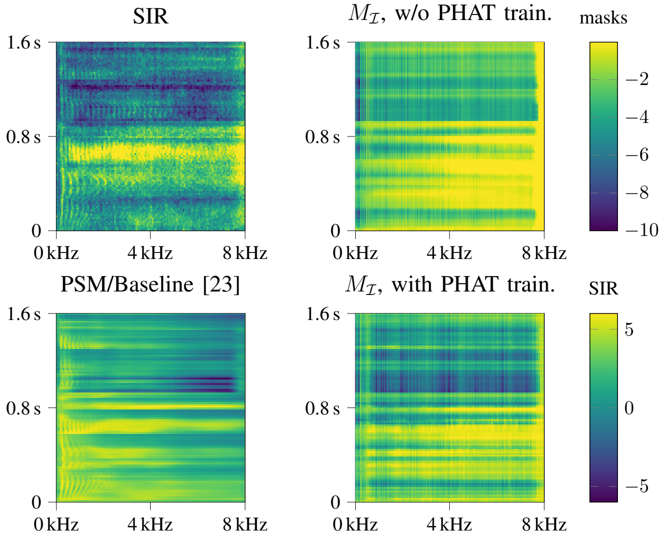

Fig. 2. Illustration of $\log_{10}(\mathrm{SIR})$ and the TF masks ($\log_{10}(M_{\mathcal{I}})$) corresponding to the SPSs in Figure 1, estimated from $|Y|$. The PSM is identical regardless of the application of PHAT as the accompanying DNN training is performed separately from the DOA estimation, whereas the two end-to-end masks depend on the incl./excl. of the PHAT weighting during training.

HYSADE system. As opposed to existing training strategies for HYSADE systems requiring microphone signals containing only the SOI, the proposed PML only requires the GT DOA as training reference, which can be obtained using, e.g., an optical tracking system. Using a test set generated using measured RIRs, we showed that the proposed method achieves state-of-the-art performance, exemplified by the objective of localizing a speech source in the presence of a directional non-speech interferer and spatiotemporally white noise. This finding holds without the use of (channel-dependent) PHAT weighting. Future work involves testing the PML with i) multiple SOIs where a suitable strategy for the choice of the normalizing must be devised (e.g., using the lowest SPS value associated with the SOIs) and ii) different mapping functions.

Acknowledgment

The authors would like to thank the Erlangen Regional Computing Center (RRZE) for providing computing resources and support.

References

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin-Heidelberg: Springer-Verlag, 2001.

[2] E. Tuncer and B. Friedlander, Eds., *Classical and Modern Direction-of-Arrival Estimation*. Burlington, MA: Academic Press, 2009.

[3] Z. Chen, G. Gokeda, and Y. Yu, *Introduction to Direction-of-Arrival Estimation*. Norwood, MA: Artech House, 2010.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Ac., Speech, Sig. Proc.*, vol. 24, no. 4, pp. 320–327, 1976.

[5] J. Chen, Y. Huang, and J. Benesty, "Time delay estimation via multichannel cross-correlation [audio signal processing applications]," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, vol. 3, Apr. 2005, pp. 49–53.

[6] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[7] M. Al-Nuaimi, R. Shubair, and K. Al-Midfa, "Direction of arrival estimation in wireless mobile communications using minimum variance distortionless response," in *The Second International Conference on Innovations in Information Technology (IIT'05)*, Sep. 2005, pp. 1–5.

[8] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University Providence, RI, May 2000.

[9] A. M. Johansson, G. Cook, and S. Nordholm, "Acoustic direction of arrival estimation, a comparison between root-MUSIC and SRP-PHAT," in *IEEE Reg. 10 Conf., TENCON*, vol. B, Nov. 2004, pp. 629–632.

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

[11] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: Opportunities and challenges for multiple source localization," in *Proc. IEEE W. on Appl. of Sig. Proc. to Aud. and Ac. (WASPAA)*, 2007, pp. 18–21.

[12] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Ac., Speech, Sig. Proc.*, vol. 37, no. 7, pp. 984–995, 1989.

[13] J. Bermudez, R. C. Chin, P. Davoodian, A. T. Y. Lok, Z. Aliyazicioglu, and H. K. Hwang, "Simulation study on DOA estimation using ESPRIT algorithm," in *Proc. World Congress on Engineering and Computer Science (WCECS)*, vol. 1, Oct. 2009, pp. 431–436.

[14] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, vol. 3, Mar. 2005, pp. iii/89–iii/92.

[15] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2444–2453, 2017.

[16] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, Apr. 2015, pp. 2814–2818.

[17] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, pp. 37–48, 2017.

[18] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Oct. 2017.

[19] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE W. on Appl. of Sig. Proc. to Aud. and Ac. (WASPAA)*, Oct. 2017.

[20] W. Mack, U. Bharadwaj, S. Chakrabarty, and E. A. P. Habets, "Signal-aware broadband DOA estimation using attention mechanisms," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, May 2020, pp. 4930–4934.

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

[22] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2017, pp. 6125–6129.

[23] Z. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Aud., Sp., Lang. Proc.*, vol. 27, no. 1, pp. 178–188, 2019.

[24] W. Mack, J. Wechsler, and E. A. P. Habets, "End-to-end signal-aware direction-of-arrival estimation using attention mechanisms," *Computer Speech & Language*, vol. 75, p. 101363, 2022. [Online]. Available: https://doi.org/10.1016/j.csl.2022.101363

[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] ——, "Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 2, no. 2, pp. 103–116, 2018.

[27] P. Pertilä and M. Parviainen, "Time difference of arrival estimation of speech signals using deep neural networks with integrated time-frequency masking," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2019, pp. 436–440.

[28] J. Wang, X. Qian, Z. Pan, M. Zhang, and H. Li, "GCC-PHAT with speech-oriented attention for robotic sound source localization," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 5876–5883.

[29] S. Rickard and Ö. Yılmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, vol. 1, 2002, pp. I–529–I–532.

[30] C. Xu, X. Xiao, S. Sun, W. Rao, E. S. Chng, and H. Li, "Weighted spatial covariance matrix estimation for MUSIC based TDOA estimation of speech source," in *Proc. Interspeech Conf.*, Aug. 2017, pp. 1894–1898.

[31] (2021, Sep.) OptiTrack. [Online]. Available: https://optitrack.com

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2015, pp. 5206–5210.

[33] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, Mar. 2017, pp. 776–780.

[34] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, May 2019, pp. 21–25.

[35] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Ac. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[36] E. A. P. Habets. (2021, Sep.) Room impulse response (RIR) generator. [Online]. Available: https://github.com/ehabets/RIR-Generator

[37] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Intl. W. Ac. Sig. Enh. (IWAENC)*, Sep. 2014, pp. 313–317.

[38] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, May 2018, pp. 351–355.