Dereverberation of Relative Harmonic Coefficients via CNNs for Acoustic Source DOA estimation

Gioele Greco[1], Silvia Messana[1], Mirco Pezzoli[1], Maximo Cobos[2], Fabio Antonacci[1]

[1]Dipartimento Elettronica Informatica e Bioingegneria, Politecnico di Milano, Italy,
[2]Departament d'Informàtica, Universitat de València, Spain

r ***Abstract*—Relative Harmonic Coefficients (RHCs) are a promising audio descriptor for Direction of Arrival (DOA) estimation but are vulnerable to noise and reverberation. We introduce RHC-ED, a convolutional encoder-decoder architecture that processes noisy and reverberant RHCs, restoring their ideal properties by suppressing unwanted artifacts. Using stacked CNNs, RHC-ED compresses and reconstructs RHCs for improved DOA estimation. Experiments across diverse acoustic conditions confirm RHC-ED's effectiveness in reducing estimation errors and outperforming recent state-of-the-art methods for source localization, especially using first-order spherical harmonics.**

***Index Terms*—Direction of Arrival, Relative Harmonic Coefficients, Convolutional Encoder-Decoder, Denoising, Dereverberation.**

I. Introduction

Audio processing is integral to many modern technologies, enabling devices to analyze, synthesize, and enhance complex audio features [1]–[5]. Understanding the Direction of Arrival (DOA) of sound sources is crucial for improving user experiences in applications like teleconferencing, smart speakers, surveillance, and spatial audio rendering [6], [7].

DOA estimation is a well-established but challenging research problem. Traditional techniques can be broadly categorized into three main approaches [1]: Steered Response Power (SRP)-based, subspace methods, and Time Difference of Arrival (TDOA)-based methods. SRP involves steering a beamformer towards candidate DOAs to maximize output power, but its precision deteriorates in noisy and reverberant environments due to multiple local maxima [8]. Subspace methods, such as MUSIC [9] and ESPRIT [10], analyze the signal covariance matrix and perform well in multi-source scenarios under specific a priori conditions. However, these methods are ineffective at low Signal-to-Noise Ratios (SNRs) [11]. Lastly, TDOA-based methods rely on signal cross-correlation matrices but also suffer precision loss in diffuse sound fields [8].

Spherical Microphone Arrays (SMAs) and Spherical Harmonics Coefficients (SHCs) have improved sound field descriptions [12]–[14], solving frequency correlation issues in DOA estimation [15]. Hu *et al.* [16] introduced Relative Harmonic Coefficients (RHCs), which depend only on DOA, offering spatial uniqueness. However, noise and reverberation reduce their effectiveness.

The authors would like to thank the Multilayered Urban Sustainability Action (MUSA) project, which is funded by the European Union, for their contributions to this work. Grant TED2021-131003B-C21 funded by MCIN/AEI/10.13039/501100011033 and by the "EU NextGeneration EU/PRTR". Grant PID2022-137048OB-C41 funded by MICIU/AEI/-10.13039/501100011033 and "ERDF A way of making Europe".

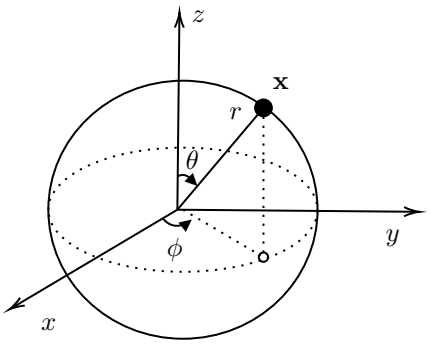

Fig. 1: Spherical coordinates system, defined by three components: azimuth ϕ, inclination θ and radius r.

RHCs have been applied in both model-based methods [17], [18] and deep learning approaches [19]. Dwivedi *et al.* proposed a hybrid Convolutional Recurrent Neural Network (CRNN) combining CNNs for pattern capture and RNNs for temporal context. Using RHCs, they treated DOA estimation as a classification task, which limits the precision of the estimations.

In this work, we present RHC-ED (Relative Harmonics - Convolutional Encoder Decoder), which leverages stacked CNNs to compress and expand STFT (Short Time Fourier Transform) coefficients limited to the 1st order. This process aims to remove noise and reverberation from non-ideal RHCs. We demonstrate that RHC-ED effectively maps noisy and reverberant RHCs to their ideal counterparts. For training and testing, we created a synthetic dataset of RHCs derived from diverse acoustic environments to ensure generalization. Our results show that RHC-ED significantly enhances RHC-based localization and achieves lower localization errors compared to classification-based DOA methods.

II. Problem definition

Let us represent a point $\mathbf{x}$ in spherical coordinates, characterized by a radial distance r, an azimuthal angle $\phi \in [0, 2\pi)$, and an inclination angle $\theta \in [0, \pi]$ as depicted in Figure 1.

In the Spherical Harmonic (SH) representation [20], the sound field at an arbitrary point $\mathbf{x} = (r, \theta, \phi)$ can be expressed as a weighted sum of SH orthogonal basis functions

$$Y_{nm}(\theta, \phi) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi}, \quad (1)$$

where, n and m denote the order and degree of the SH function. The term $P_n^m(\cdot)$ is the real-valued associated Legendre function.

Let us consider a SMA with Q capsules, each positioned at a radial distance r from the origin. The sound pressure at the qth microphone, located at $\mathbf{x}_q = (r, \theta_q, \phi_q)$ for $q \in \{1, \ldots, Q\}$, is

$$p(\mathbf{x}_q, k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(k) b_n(kr) Y_{nm}(\theta_q, \phi_q), \quad (2)$$

where $k = \frac{2\pi f}{c}$ is the wave number, c is the speed of sound, and f is the frequency. The truncation order N is determined by $N = \lceil kr \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function.

Moreover, the function $b_n(kr)$ is defined as

$$b_n(kr) = \begin{cases} j_n(kr), & \text{for an open array,} \\ j_n(kr) - \frac{j_n'(kr)}{h_n'(kr)} h_n'(kr), & \text{for a rigid array.} \end{cases} \quad (3)$$

where, $j_n'(\cdot)$ and $h_n'(\cdot)$ are the partial derivatives of the spherical Bessel and Hankel functions, respectively.

According to [21], the SHCs $\alpha_{nm}(k)$ can be obtained by inverting (2), leading to

$$\alpha_{nm}(k) = \frac{1}{b_n(kr)} \sum_{q=1}^{Q} p(r, \theta_q, \phi_q) Y_{nm}^*(\theta_q, \phi_q) w_q. \quad (4)$$

Here, w_q represents a microphone weight that ensures consistency in the equation.

Assuming that a sound source satisfies the far-field condition [22], the sound field can be approximated as a planar wave. In this case, the SHCs [23] simplify into

$$\alpha_{nm}(k) = S(k) 4\pi i^n Y_{nm}^*(\theta_s, \phi_s), \quad (5)$$

where $S(k)$ is the source signal and θ_s, ϕ_s are the azimuth and inclination of the source with respect to the center of the SMA. In [16], authors introduce the RHC defined as the ratio between α_{nm} and α_{00}

$$\beta_{nm}(k) = \frac{\alpha_{nm}(k)}{\alpha_{00}(k)}. \quad (6)$$

Substituting the SHC expression from (5) into (6) we obtain

$$\beta_{nm}(k) = 2\sqrt{\pi} i^n Y_{nm}^*(\theta_s, \phi_s). \quad (7)$$

This formulation underscores the key theoretical properties of RHCs, including their independence from the driving signal $S(k)$, frequency, and spatial characteristics. Additionally, it highlights their exclusive dependence on the source DOA (θ_s, ϕ_s), making RHCs particularly valuable for localization algorithms.

A. Source DOA estimator

Considering the characteristics of RHCs, various methods have been proposed for DOA estimation, as discussed in [24], [25]. However, leveraging the inherent properties of RHCs, it has been demonstrated in [25] that, in ideal conditions (anechoic and noiseless), by truncating the expansion to the first order, the DOA can be directly estimated from the unitary vector

$$\boldsymbol{\eta}(k) = \begin{bmatrix} \sqrt{1/6}\Im(\beta_{1,-1}(k) - \beta_{1,1}(k)) \\ -\sqrt{1/6}\Re(\beta_{1,-1}(k) + \beta_{1,1}(k)) \\ \sqrt{1/3}\Im(\beta_{1,0}(k)) \end{bmatrix} = \begin{bmatrix} \sin(\phi)\sin(\theta) \\ \cos(\phi)\sin(\theta) \\ \cos(\theta) \end{bmatrix}, \quad (8)$$

where $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and the imaginary parts of a number, respectively. However, in real-world scenarios, the reverberant component must also be considered and analyzed.

B. Signal model in diffuse environment

Let us consider a noiseless diffuse field scenario in which there is a point source located at $\mathbf{x}_s = (r_s, \theta_s, \phi_s)$. In this case, the sound field and the corresponding SHCs consist of both direct and reverberant contributions. The SHCs in $\mathbf{x}_s$ corresponding to the direct-path echo are

$$\alpha_{nm}^{dir}(k) = S(k) ikh_n(kr_s) Y_{nm}^*(\theta_s, \phi_s), \quad (9)$$

leading to the corresponding RHCs

$$\beta_{nm}^{dir}(k) = \frac{2\sqrt{\pi} h_n(kr_s) Y_{nm}^*(\theta_s, \phi_s)}{h_0(kr)}. \quad (10)$$

If we include the reverberant component [26], the SHCs become

$$\alpha_{nm}^{rev}(k) = \alpha_{nm}^{dir}(k) + \sum_{v=0}^{\tilde{N}} \sum_{u=-v}^{v} \hat{\alpha}_{nm}^{vu}(k) S(k) ikj_v(kr_s) Y_{vu}^*(\theta_s, \phi_s), \quad (11)$$

where $\hat{\alpha}_{nm}^{vu}$ corresponds to the SHC of order and degree vu generated by the incident reflection outgoing field of order and degree nm [27], and $\tilde{N}$ is its truncation order.

By substituting (11) into (6), the RHCs for the reverberant field become

$$\beta_{nm}^{rev}(k) = \frac{h_n(kr_s) Y_{nm}^*(\theta_s, \phi_s) + \sum_{v=0}^{N} \sum_{u=-v}^{v} \hat{\alpha}_{nm}^{vu}(k) j_v(kr_s) Y_{vu}^*(\theta_s, \phi_s)}{h_0(kr_s) Y_{00}^*(\theta_s, \phi_s) + \sum_{v=0}^{N} \sum_{u=-v}^{v} \hat{\alpha}_{00}^{vu}(k) j_v(kr_s) Y_{vu}^*(\theta_s, \phi_s)}. \quad (12)$$

Equation (12) reveals that in a reverberant field, RHCs are no longer solely dependent on the source DOA. Specifically, as shown in [27], $\hat{\alpha}_{nm}^{vu}$ is influenced by the Relative Transfer Function (RTF) between the source and receiver regions. This dependency causes the RHCs to deviate from their ideal behavior. This observation underscores the need for a system capable of mapping nonideal RHCs to their ideal anechoic counterparts.

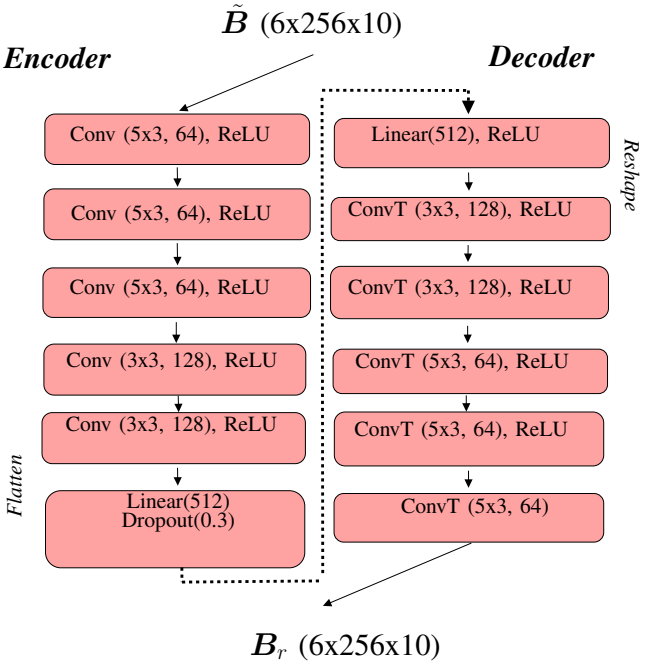

Fig. 2: RH-CED description, including kernels dimensions. The encoding and decoding architectures are mirrored, except that the latter uses transposed convolutions instead of standard convolutions to upsample the compressed features. $\boldsymbol{B}_r$ represents the anechoic reconstruction tensor in similarity to $\boldsymbol{B}$ in Sec.III.

III. Proposed model

Let us define $\boldsymbol{\beta}(k) = [\beta_{00}(k), \ldots, \beta_{nm}(k)]^T$ and $\boldsymbol{B} = [\boldsymbol{\beta}(k_1), \ldots, \boldsymbol{\beta}(k_\ell), \ldots, \boldsymbol{\beta}(k_L)]$ where ℓ is the frequency bin index and L is the number of frequency bins. We can formulate the mapping problem as

$$\begin{aligned} \exists\, f_{\boldsymbol{\Omega}} :\ \boldsymbol{B}_r &= f_{\boldsymbol{\Omega}}\,(\widetilde{\boldsymbol{B}}), \\ \boldsymbol{\Omega} &= \underset{\boldsymbol{\Omega}}{\operatorname{argmin}} \mathcal{F}(\widehat{\boldsymbol{B}}, f_{\boldsymbol{\Omega}}(\widetilde{\boldsymbol{B}})), \end{aligned} \tag{13}$$

where $f_{\boldsymbol{\Omega}}$ is a parameterized function that maps the noisy and reverberant coefficients $\widetilde{\boldsymbol{B}}$ to their anechoic counterparts $\widehat{\boldsymbol{B}}$, as the same $\boldsymbol{B}_r$ represents the reconstructed RHCs, and $\boldsymbol{\Omega}$ represents the set of parameters that define this mapping. The function $\mathcal{F}(\widehat{\boldsymbol{B}}, \cdot)$ denotes the mean square error loss function used to optimize $\boldsymbol{\Omega}$.

Thus, we developed the Relative Harmonic Coefficients Encoder Decoder (RHC-ED) to optimize $\boldsymbol{\Omega}$ as depicted in Figure 2. Indeed, after a training phase, applying $f_{\boldsymbol{\Omega}}$ to the measured RHCs $\tilde{\boldsymbol{\beta}}(k)$, we aim for the RHC-ED to produce anechoic reconstructions, denoted as $\boldsymbol{\beta}_r(k)$ in

$$\boldsymbol{\beta}_r(k) = f_{\boldsymbol{\Omega}}\,(\tilde{\boldsymbol{\beta}}(k)). \tag{14}$$

As input we use 10 consecutive time frames of the multichannel STFT with 256 frequency bins of the RHCs up to the 1st order expansion without the channel 0, which is 1 by definition.

A. RHC-ED description

The input features consist of time-frequency representations of RHC limited to order $N = 1$, excluding β_{00}. Each selected coefficient is decomposed into its real and imaginary components. The input features are structured into a tensor

$$\overline{\overline{B}} = \begin{bmatrix} \Re(\boldsymbol{\beta}_{1,-1}) \\ \Im(\boldsymbol{\beta}_{1,-1}) \\ \Re(\boldsymbol{\beta}_{1,0}) \\ \Im(\boldsymbol{\beta}_{1,0}) \\ \Re(\boldsymbol{\beta}_{1,1}) \\ \Im(\boldsymbol{\beta}_{1,1}) \end{bmatrix} \vdots\vdots \in \mathbb{R}^{6\times 128\times 10} \tag{15}$$

Given that the representation spans 256 frequency bins, the final input tensor has dimensions $6 \times 256 \times 10$.

The first processing stage is data compression, which takes place in the encoder. Here, the input tensor is passed through a sequence of five consecutive convolutional layers, progressively reducing its dimensionality while preserving essential features (see Fig. 2). The decoding process mirrors the encoding structure, employing transposed convolutions instead of standard convolutions. These upsampling operations restore the compressed latent representation to its original dimensions, aiming at reconstructing the anechoic coefficients. The detailed architecture of the RHC-ED, as well as the number of learnable parameters, is depicted in Figure 2.

B. Dataset

We generate the dataset by applying spherical decomposition to the convolution of speech signal sources and Room Impulse Responses (RIRs) simulating various acoustic environments. Source signals are randomly selected from a Librispeech [28] subset for Task 1 [29] of the L3DAS23 dataset, sampled at 16 kHz with up to 12 s of clean speech, including 53 % male and 47 % female voices.

RIRs are synthesized using the SMIR generator [30], simulating interactions between randomly placed sources and a Spherical Microphone Array (SMA), configured as an Eigenmike with 32 microphones. The SMA is positioned at the center of the xy-plane at a height of 1.3 m. Sources are placed taking random inclinations θ_s from 60° to 130°, azimuths ϕ_s from 0° to 360°, and distances r_s from 1.5 m to 3.5 m from the SMA's center, generating 500 spatial samples per room. Table I summarizes the parameter ranges for the synthetic dataset. Given the SHCs of each sample, we computed RHCs using (6), then applied the STFT with a 512-sample Hamming window and a 320-sample hop size, yielding 427 time frames. Input data was created using a sliding

Azimuth (ϕ_s)	$[0°, 360°]$
Inclination (θ_s)	$[60°, 130°]$
Distance from microphone (r_s)	$[1.5, 3.5]$m
Room size (w, h, d)	$[4, 8]$m $\times$ $[5, 10]$m $\times$ $[3, 5]$m
T_{60}	$[0.25, 1.0]$s
SNR	$[5, 60]$dB

TABLE I: Summary of Parameters used for dataset generation.

window of 10 frames, resulting in audio representations of size $6 \times 256 \times 418$, discarding the last samples as they fall short of 10 frames. For training the RHC-ED model, we selected four rooms with randomly assigned reverberation times T60 in the pool $[0.25, 0.50, 0.75, 1.00]$s. Each room contained 517 recordings, resulting in a total training dataset of 2,068 samples. To facilitate model evaluation, $20\,\%$ of the training dataset was set aside as validation set. The test set is constructed by randomly selecting 50 audio samples for each of the 10 different T60 and SNR values, as detailed in the following section.

IV. Validation

To evaluate the performance of RHC-ED, we propose two distinct experiments: one utilizing the synthetic test set and another based on real measurements from [31]. Moreover, a voice activity detector [32] was implemented to discard the time frames in which the source is not active. Thus, we compare the DOA estimates obtained using (8) under three different conditions: RHCs processed using the proposed model, unprocessed RHCs, and a data-driven approach proposed in [19].

A. Metrics

To estimate the accuracy, we define the Angular Error (AE)

$$\mathrm{AE}(\boldsymbol{\eta}) = |\arccos\Big(\sum_{\ell=1}^{L}\frac{\boldsymbol{\eta}^T(k_\ell)}{L}\boldsymbol{\eta}_{\mathrm{GT}}\Big)|^\circ, \tag{16}$$

where $\boldsymbol{\eta}_{\mathrm{GT}}$ is the ground truth DOA derived directly from geometric data. Since RHC-ED provides us with an estimate of DOA for each time frame τ, we obtain a set of $T = 418$ estimates of AE for each audio sample. Therefore, we also measure the standard deviation

$$\sigma = \sqrt{\frac{1}{T}\sum_{\tau=1}^{T}(\mathrm{AE}(\boldsymbol{\eta}(\tau)) - \widehat{\mathrm{AE}})^2}, \tag{17}$$

where $\widehat{\mathrm{AE}} = \sum_{\tau=1}^{T}\mathrm{AE}(\boldsymbol{\eta}(\tau))/T$.

B. Performance Analysis on Synthetic Dataset

This evaluation was carried out using 50 samples (i.e. a total of 20800 estimates) from the synthetic dataset described in Section III-B. Specifically, we selected 10 different rooms with reverberation times T60= $[0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9, 1]$s that were not part of the training set. The results of the evaluation of the synthetic dataset for different T60 values are presented in Fig. 3. The findings indicate that RHC-ED is consistently more accurate that the other methods. Notably, the method in [19] demonstrates reduced accuracy in our tested scenario, primarily due to two key factors: the expansion order, which is constrained to the first order in this study, whereas prior work extends it up to the fourth order, and the inherent limitations of the method itself. More specifically, this approach is classification-based, relying on discrete class intervals spaced every 5°. Thus, any misclassification results in a minimum error of 5°.

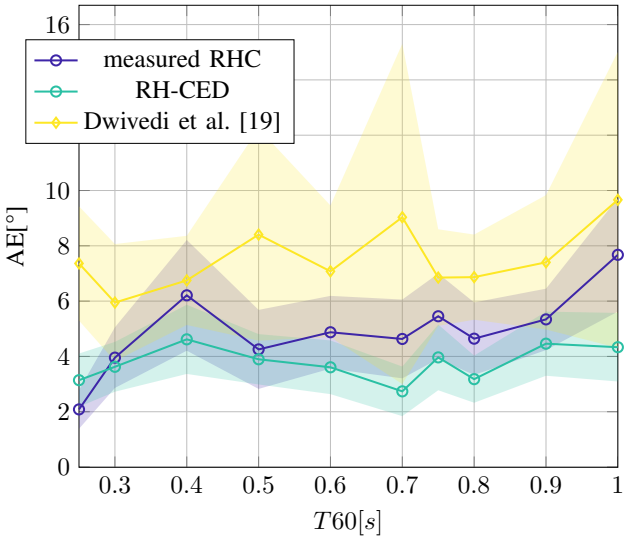

Fig. 3: AE distribution derived from 50 audio samples from the test set for each $T60$ with SNR in the whole considered range. Solid lines indicate the mean value of $\widehat{\mathrm{AE}}$ across all 50 samples, while the shaded regions represent the standard deviation σ of AE.

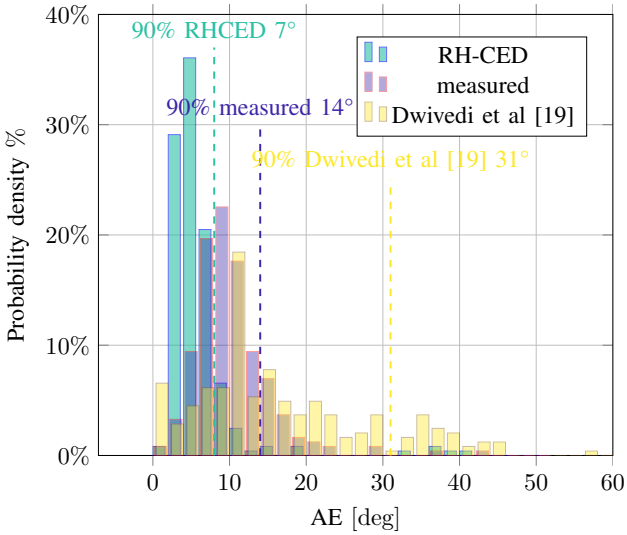

Fig. 4: Probability density function of the AE distribution obtained from the [31] dataset. Bars indicates the probability of AE for each 2[deg] sector, dashed lines indicate the 90% confindence interval of AE values.

C. Performance Analysis on Real-World Measurements

Figure 4 illustrates the histogram of AE obtained from the experiment using the dataset [31]. The dispersion is assessed by analyzing the AE of each time frame across the three methods. It is evident that RHC-ED exhibits a higher probability of maintaining lower errors. Specifically, the $90\,\%$ confidence interval highlights the effectiveness of RHC-ED processing, with an AE confidence interval of 7°,

compared to 14° for measured RHCs and 31° for the method in [19].

V. Conclusion

We implemented a deep learning-based denoising and dereverberating system to mitigate the effects of non-idealities on RHCs, theoretically depending only on source DOA. In this work, we proposed an encoder-decoder capable of mapping RHCs resulting from noisy and reverberated sound fields to their ideal counterparts. We demonstrated that localization using RHC-ED processed features outperformed a recent state-of-the-art technique for DOA classification, assuming same input conditions (1st order SH expansion).

References

[1] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, no. 1, p. 3956282, 2017.

[2] W. Shi, J. Huang, and Y. Hou, "Fast doa estimation algorithm for mimo sonar based on ant colony optimization," *Journal of Systems Engineering and Electronics*, vol. 23, no. 2, pp. 173–178, 2012.

[3] Y. Wu, C. Li, Y. T. Hou, and W. Lou, "Real-time doa estimation for automotive radar," in *2021 18th European Radar Conference (EuRAD)*, 2022, pp. 437–440.

[4] D. Albertini, G. Greco, A. Bernardini, and A. Sarti, "Diffusion-based sound source localization using networks of planar microphone arrays," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[5] M. Olivieri, A. Bastine, M. Pezzoli, F. Antonacci, T. Abhayapala, and A. Sarti, "Acoustic imaging with circular microphone array: A new approach for sound field analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1750–1761, 2024.

[6] A. Alexandridis, D. Pavlidi, N. Stefanakis, and A. Mouchtaris, *Parametric Spatial Audio Techniques in Teleconferencing and Remote Presence*. John Wiley Sons, Ltd, 2017, ch. 15, pp. 363–386. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119252634.ch15

[7] M. Sekikawa and N. Hamada, "Doa estimation of multiple sources using arbitrary microphone array configuration in the presence of spatial aliasing," in *2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2014, pp. 080–083.

[8] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," pp. 157–180, 2001. [Online]. Available: https://doi.org/10.1007/978-3-662-04619-7_8

[9] P. Stoica and A. Nehorai, "Music, maximum likelihood, and cramer-rao bound," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 5, pp. 720–741, 1989.

[10] 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.

[11] J. Thomas, L. Scharf, and D. Tufts, "The probability of a subspace swap in the svd," *IEEE Transactions on Signal Processing*, vol. 43, no. 3, pp. 730–736, 1995.

[12] M. Pezzoli, J. Carabias-Orti, P. Vera-Candeas, F. Antonacci, and A. Sarti, "Spherical-harmonics-based sound field decomposition and multichannel nmf for sound source separation," *Applied Acoustics*, vol. 218, p. 109888, 2024.

[13] M. Cobos, M. Pezzoli, F. Antonacci, and A. Sarti, "Acoustic source localization in the spherical harmonics domain exploiting low-rank approximations," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] M. Pezzoli, M. Cobos, F. Antonacci, and A. Sarti, "Sparsity-based sound field separation in the spherical harmonics domain," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1051–1055.

[15] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 221–224, 2009.

[16] Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Sound source localization using relative harmonic coefficients in modal domain," *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 348–352, 2019.

[17] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 221–224, 2009.

[18] Y. Hu, T. D. Abhayapala, and P. N. Samarasinghe, "Multiple source direction of arrival estimations using relative sound pressure based music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 253–264, 2021.

[19] P. Dwivedi, S. B. Hazare, G. Routray, and R. M. Hegde, "Long-term temporal audio source localization using sh-crnn," *2023 National Conference on Communications (NCC)*, pp. 1–6, 2023.

[20] E. G. Williams and J. A. Mann, "Fourier acoustics: Sound radiation and nearfield acoustical holography," 1999. [Online]. Available: https://api.semanticscholar.org/CorpusID:121699111

[21] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II–1949–II–1952, 2002.

[22] D. Marković, F. Antonacci, L. Bianchi, S. Tubaro, and A. Sarti, "Extraction of acoustic sources through the processing of sound field maps in the ray space," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2481–2494, 2016.

[23] Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and G. Dickins, "Modeling characteristics of real loudspeakers using various acoustic models: Modal-domain approaches," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 561–565, 2019.

[24] Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and S. Gannot, "Unsupervised multiple source localization using relative harmonic coefficients," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 571–575, 2020.

[25] Y. Hu and S. Gannot, "Closed-form single source direction-of-arrival estimator using first-order relative harmonic coefficients," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 726–730.

[26] Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, "Semi-supervised multiple source localization using relative harmonic coefficients under noisy and reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3108–3123, 2020.

[27] P. N. Samarasinghe, T. D. Abhayapala, M. A. Poletti, and T. Betlehem, "An efficient parameterization of the room transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 2217–2227, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:15236337

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

[29] E. Guizzo, C. Marinoni, M. Pennese, X. Ren, X. Zheng, C. Zhang, B. Masiero, A. Uncini, and D. Comminiello, "L3das22 challenge: Learning 3d audio sources in a real office environment," May 2022.

[30] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, 09 2012.

[31] O. Olgun and H. Hacihabiboglu, "Metu sparg eigenmike em32 acoustic impulse response dataset v0.1.0," Apr. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2635758

[32] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," in *17th European Signal Processing Conference, EUSIPCO 2009, Glasgow, Scotland, UK, August 24-28, 2009*. IEEE, 2009, pp. 2549–2553. [Online]. Available: https://ieeexplore.ieee.org/document/7077834/