Comparison of Learning-Based DOA Estimation Between SH Domain Features

Yonggang Hu[1]
[1]*Audio and Acoustic Signal Processing Group*
Australian National University
Canberra, Australia

Sharon Gannot[2]
[2]*Faculty of Engineering*
Bar-Ilan University
Ramat-Gan, Israel

***Abstract*—Accurate direction-of-arrival (DOA) estimation in noisy and reverberant environments is a long-standing challenge in the field of acoustic signal processing. One of the promising research directions utilizes the decomposition of the multi-microphone measurements into the spherical harmonics (SH) domain. This paper presents an evaluation and comparison of learning-based single-source DOA estimation using two recently introduced SH domain features denoted *relative harmonic coefficients* (RHC) and *relative modal coherence* (RMC), respectively. Both features were shown to be independent of the time-varying source signal even in reverberant environments, thus facilitating training with synthesized, continuously-active, noise signal rather than with speech signal. The inspected features are fed into a convolutional neural network, trained as a DOA classifier. Extensive validations confirm that the RHC-based method outperforms the RMC-based method, especially under unfavorable scenarios with severe noise and reverberation.**

***Index Terms*—Learning-based direction-of-arrival estimation, relative harmonic coefficients, relative modal coherence.**

I. Introduction

Accurate knowledge of the DOA of an acoustic source is an important building block in many acoustic/audio signal processing techniques, including spatial beamforming, speech separation, speech recognition, and sound event detection [1].

Most unsupervised localization methods, such as time difference of arrival (TDOA) [2]–[4], steered response power (SRP) [5], [6], and subspace-based methods [7], are easy to implement and their localization accuracy is reasonable. However, their performance declines severely in complex acoustic environments, specifically as a result of strong acoustic reflections and low signal-to-noise ratios. In recent years, there is a growing interest in using deep learning-based approaches for obtaining improved localization performance in unfavorable environments. These methods are typically based on learning the patterns of the acoustic features and their relation to the source position using a training dataset measured in advance over a pre-defined source area of interest. Then, they utilize the learned patterns to estimate the position of an unknown sources in the test stage [8]. Deep learning localization methods may either classify the desired source DOA into one of candidate directions, or use regression to directly estimate the DOA. In this paper we adopt the former approach, namely to cast the localization as a classification task. Several neural networks approaches for data-driven source localization were introduced in a recent special issue [9].

Several neural network architectures were adopted to the task of acoustic source localization, e.g. deep neural networks (DNN) [10], [11], convolutional neural networks (CNNs) [12], [13], and convolutional and recurrent neural networks (CRNNs) [14], [15]. Different features were used, including but not limited to, binaural features [16], eigenvectors of the spatial covariance matrix [17], generalized cross-correlation (GCC) [10], and short-time Fourier transform (STFT) of the received signals [12]. A comprehensive list of recent approaches can be found in the survey paper [18].

Spherical microphone arrays are widely used in source localization tasks [15], [19], [20], as they are capable of recording multi-channel measurements over a large area, thus providing more relevant cues of the source(s) to be localized. The multi-channel measurements can be decomposed into the spherical harmonics (SH) domain using a set of orthogonal spatial functions [21]. Following the W-disjoint orthogonality assumption [22], Perotin *et al.* [20] adopted a CRNN architecture in the SH domain to estimate the DOAs of multiple sound sources. Then, Grumiaux *et al.* [23] proposed an improved scheme by changing the layout between convolutional and pooling layers of the CRNN in [20]. The input features of both algorithms in [20], [23] are pseudo-intensity vectors, which are denoted using the first-order Ambisonics in the SH domain. Initially, the pseudo-intensity vector was used by a closed-form DOA estimator in [24], [25]. Recently, Hu *et al.* in [26] proposed another closed-form DOA estimator using a new SH domain feature denoted *relative harmonic coefficients* (RHC) [27]. This estimator outperformed the intensity-based method under equivalent noisy and reverberant conditions. The RHCs are the SH domain counterparts of the relative transfer functions (RTFs) [28], [29], that are spatially more discriminative, as the spherical harmonic decomposition enhances the spatial resolution over space. Several localization schemes utilized the RHC such as [26], [27], [30]–[34].

One of the main attributes of the RHC as a feature for source localization is its independence of the time-varying source signal. This is also true for the *relative modal coherence* (RMC), another SH domain feature recently defined by Fahim *et al.* [35], that uses the covariance of the decomposed spherical harmonic coefficients. The RMC in [35] was used as a feature for a CNN-based multi-source DOA estimator in reverberant soundfield, while not considering noisy environments.

The aim of the current contribution is the evaluation and comparison of learning-based, single source, DOA estimation procedures with either the RHC or the RMC as input features, in noisy and highly reverberant environments. Compared with past research examining these two features [26], [27], [30]–[33], [35], the contributions of this paper are: (i) presenting a performance evaluation and comparison between both features in the source localization task; (ii) casting the RHC-based localization scheme as a classification task rather than regression [30] and implementing the classifier using CNN; and (iii) considering more challenging acoustic environments, that is, source features are more distorted by severe noise and reverberation.

II. Acoustic model

Assume a spherical microphone array with M microphones capturing a soundfield in a reverberant and noisy environment. The polar coordinates of the microphones are given by $\boldsymbol{x}_j = (r, \theta_j, \phi_j)$, $j = 1, \ldots, M$ with respect to its local origin O. Assume a single far-field sound source propagating from an unknown DOA, e.g., $\Phi = (\vartheta_s, \varphi_s)$ where $0 < \vartheta_s < \pi$, $0 < \varphi_s < 2\pi$ with respect to the origin of the microphone array. Hence, the measured sound pressure in the STFT domain, as measured by the j-th microphone, is the sum of direct-path signal, the sound reflections and the noise signal,

$$P_{\boldsymbol{x}_j}(t,k) = S(t,k)\left[G_d(k)e^{i\boldsymbol{k}^\top \boldsymbol{x}_j} + \int_{\hat{\boldsymbol{y}}} G_r^{\hat{\boldsymbol{y}}}(k) e^{i\boldsymbol{k}_{\hat{\boldsymbol{y}}}^\top \boldsymbol{x}_j} d\hat{\boldsymbol{y}}\right] + V_{\boldsymbol{x}_j}(t,k) \tag{1}$$

where $t \in \{1, \ldots, T\}$ and $k \in \{1, \ldots, K\}$ denote the time and frequency index in the STFT domain, $k = 2\pi f/c$, f is the frequency and c is the speed of sound, $S(t,k)$ denotes the source signal, $G_r^{\hat{\boldsymbol{y}}}$ denotes the reflection gain along an arbitrary direction of $\hat{\boldsymbol{y}}$, $G_d(k)$ denotes the source's direct gain, the wavenumber vector is represented by $\boldsymbol{k} = (k\cos\phi\sin\theta, k\sin\phi\sin\theta, k\cos\theta)^\top$, and $V_{\boldsymbol{x}_j}(t,k)$ denotes the additive noise signal. Note that the additive noise in (1) is assumed to be non-directional, otherwise it could be regarded as an additional source to be localized.

III. Spatial decomposition of the soundfield

This section introduces the spatial decomposition of a measured soundfield. The sound pressure measured by the spherical microphone array can be decomposed into the SH domain using a set of orthogonal spatial functions [21],

$$P(\boldsymbol{x}_j, k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(k)\, b_n(kr)\, Y_{nm}(\theta_j, \phi_j) \tag{2}$$

where the time index t is omitted for brevity, $N = \lceil kr \rceil$ is the truncated order of the soundfield [36], and

$$Y_{nm}(\theta,\phi) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} P_{nm}(\cos\theta) e^{im\phi} \tag{3}$$

is the spherical harmonic function with order n and mode m, $P_{nm}(\cdot)$ denotes the real-valued associated Legendre function, $b_n(\cdot)$ is a function based on the array configuration,

$$b_n(kr) = \begin{cases} j_n(kr), & \text{for an open array} \\ j_n(kr) - \frac{j_n'(kR)}{h_n'(kR)} h_n(kr), & \text{for a rigid array} \end{cases} \tag{4}$$

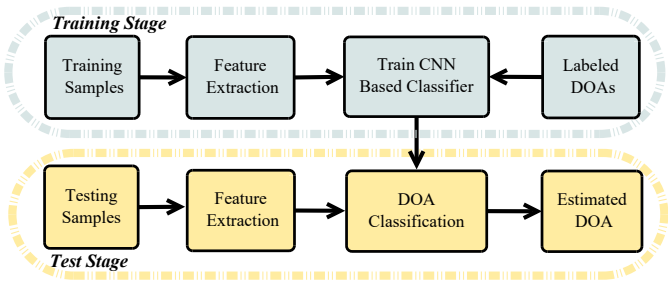


Fig. 1: Block diagram of the proposed localization approach, which comprises of a training and test stage, respectively.

where R denotes the radius of the spherical microphone array, $j_n'(\cdot)$ and $h_n'(\cdot)$ denote the partial derivative of the spherical Bessel and Hankel functions, respectively, and $\alpha_{nm}(k)$ denotes the spherical harmonic coefficients that characterize the measured soundfield in the SH domain. Assume far-field scenarios, namely that the aperture of the recording area is much smaller compared to its distance to the source [37]. In this case, the spherical harmonic coefficients in (1) are given by,

$$\alpha_{nm}^{\text{rev}}(t,k) = \alpha_{nm}^{\text{dir}}(t,k) + \underbrace{\sum_{v=0}^{N}\sum_{u=-v}^{v} \alpha_{nm}^{\text{dir}}(k)\hat{\alpha}_{nm}^{vu}(k)}_{\text{Reverberant-path}} + \gamma_{nm}(t,k) \tag{5}$$

where $\gamma_{nm}(t,k)$ denotes the decomposed noise signal, and

$$\alpha_{nm}^{\text{dir}}(t,k) = S(t,k)G_d(k)4\pi i^n Y_{nm}^*(\vartheta_s, \varphi_s) \tag{6}$$

denotes the direct-path signals, $(\cdot)^*$ denotes the complex conjugate operator and $\hat{\alpha}_{nm}^{vu}(k)$ denote the coupling coefficients that are independent of the sound source, and remain fixed in a static reverberant acoustic environment, where the settings of the environment and the microphone array are not changing over time [38].

IV. Proposed learning-based Localization

A. Framework of the Algorithm

This paper presents a data-driven solution to localize an unknown sound source in adverse acoustic conditions, including noise and reverberation. Figure 1 depicts the compact block diagram of the algorithm, consisting of two disjoint stages, as described below.

Training stage: (i) Select $\mathcal{N}_L$ labeled training samples within a defined area of interest (AoI); (ii) Measure the recordings due to each training source in the AoI using a spherical microphone array, decompose the soundfield into the SH domain, and then extract the corresponding source features; (iii) Use the training feature set and labeled DOAs to train the deep learning-based classifier (i.e., CNN) that will be used in the test stage.

Test stage: (i) Measure the recordings from an unknown test source position within the AoI; (ii) Decompose the soundfield into the SH domain, and then extract the test source features for localization; (iii) Estimate the source's unknown DOA using the classifier obtained in the training stage.

The learning-based DOA classifier adopts two spatial features defined using the spherical harmonic coefficients in (5), as introduced below.

B. SH Domain Features

1) Relative harmonic coefficients (RHC): In [27], [30], [31] the RHC were formally defined in a noise-free environment as the ratios between spherical harmonic coefficient at the individual modes (n, m) and mode $(0, 0)$. Assume a noise-free environment, the ratio between $\alpha_{nm}^{\text{rev}}(t,k)$ and $\alpha_{00}^{\text{rev}}(t,k)$ in (5), defines the feature expression in reverberant environment,

$$\beta_{nm}^{\text{rev}}(k) = 2\sqrt{\pi} i^n Y_{nm}^*(\vartheta_s, \varphi_s)\lambda(k) \tag{7}$$

where

$$\lambda(k) = \frac{1 + \sum_{v=0}^{N}\sum_{u=-v}^{v} \hat{\alpha}_{nm}^{vu}(k)}{1 + \sum_{v=0}^{N}\sum_{u=-v}^{v} \hat{\alpha}_{00}^{vu}(k)} \tag{8}$$

depends on the fixed coupling coefficients in the assumed static environment. The expression in (7) confirms that the RHCs only depend on the source DOA, even in a reverberant environment. In practice, we adopt the biased RHC estimator in [30] to extract the RHC in noisy environments, using the ratio between the CPSD (cross power spectral density) and the PSD of the spherical harmonic coefficients,

$$\bar{\beta}_{nm}^{\text{rev}}(k) = \frac{\mathbb{E}\left\{\alpha_{nm}(k)\alpha_{00}^*(k)\right\}}{\mathbb{E}\left\{\alpha_{00}(k)\alpha_{00}^*(k)\right\}} \tag{9}$$

where $\mathbb{E}\left\{\cdot\right\}$ denotes the statistical average over the time-varying signal. Note that, in non-stationary noise environments, the RHC contains a time-dependence due to the noise, as the biased estimator in (9) cannot fully remove the noise influence. Finally, for the N-th order array, define the $(N+1)^2 \times 1$ feature vector:

$$\boldsymbol{\beta}^{\text{rev}}(k) = \left[1, \frac{\mathbb{E}\left\{\alpha_{1,-1}(k)\alpha_{00}^*(k)\right\}}{\mathbb{E}\left\{\alpha_{00}(k)\alpha_{00}^*(k)\right\}}, \cdots, \frac{\mathbb{E}\left\{\alpha_{NN}(k)\alpha_{00}^*(k)\right\}}{\mathbb{E}\left\{\alpha_{00}(k)\alpha_{00}^*(k)\right\}}\right]^\top. \tag{10}$$

2) Relative modal coherence (RMC): The covariance of the time-varying spherical harmonic coefficients in (5) is denoted:

$$\mathbb{E}\left\{\alpha_{nm}(k)\alpha_{n'm'}^*(k)\right\}. \tag{11}$$

Recent work [35] defined the relative modal coherence (RMC) as the normalized covariance of the spherical harmonic coefficients in (11), i.e.,

$$\mathbb{E}\left\{\alpha_{nm}(k)\alpha_{n'm'}^*(k)\right\} / \mathbb{E}\left\{\alpha_{00}(k)\alpha_{00}^*(k)\right\} \tag{12}$$

where the ratio between the numerator and denominator removes the influence of the source signal. For the N-th order microphone array, it has a two-dimensional feature matrix, i.e.,

$$\begin{bmatrix} 1 & \cdots & \frac{\mathbb{E}\left\{\alpha_{NN}(k)\alpha_{00}^*(k)\right\}}{\mathbb{E}\left\{\alpha_{00}(k)\alpha_{00}^*(k)\right\}} \\ \cdots & \cdots & \cdots \\ \frac{\mathbb{E}\left\{\alpha_{00}(k)\alpha_{NN}^*(k)\right\}}{\mathbb{E}\left\{\alpha_{00}(k)\alpha_{00}^*(k)\right\}} & \cdots & \frac{\mathbb{E}\left\{\alpha_{NN}(k)\alpha_{NN}^*(k)\right\}}{\mathbb{E}\left\{\alpha_{00}(k)\alpha_{00}^*(k)\right\}} \end{bmatrix} \tag{13}$$

which actually denotes the covariance of RHC in (9) (refer to [35] for the expression of the RMC).

Here, additional explanations about the features are provided: (i) Both spatial features are separately used as inputs to the deep learning architecture, hence, a fair comparison of the localization performance can be presented (see Section V). For a fair comparison with RHC, the $(N+1)^2 \times (N+1)^2$ RMC matrix in (13) is also flattened to a $(N+1)^4 \times 1$ vector. (ii) Since only a single-source case is evaluated, the average operator, in (10) and (13), utilizes all frames to extract the features. Hence, in the test stage, the time-domain recordings correspond to two features for localization, one is the $K \times (N+1)^2$ RHC and the other is the $K \times (N+1)^4$ RMC, respectively (K denotes the number of frequency bins in (1)). (iii) Both spatial features are complex-valued. We therefore concatenate the real and imaginary parts of the features, to finally obtain $K \times (N+1)^2 \times 2$ RHC and $K \times (N+1)^4 \times 2$ RMC tensors, that are used as inputs of the networks.

C. Convolutional Neural Networks

The localization algorithm adopts a CNN architecture to infer the underlying relations between the spatial features and the source DOA. Typical CNN architectures comprise multiple layers of convolution as well as pooling operations for significantly reducing the number of parameters. For simplicity, the CNN used by this paper uses two convolution layers (64 local filters), directly followed by one fully-connected layers (1000 nodes) to produce a 360×1 output vector, without using any pooling operations. The supporting reason is that the spatial features are directly related to the source position, thus only imposing mild requirements on the network complexity. The CNN requires all the SH modes, as each of them contains unique characterization of the soundfield. Hence, the 64 local filters are set with the size of 3×1.

V. Simulation Study

In this section we simulate signals in challenging environments to validate the effectiveness of the learning-based algorithm, and present a full comparison between the localization accuracy using the two features.

A. CNN Training with Synthesized White Noise Signals

As analyzed above, both features are independent of the time-varying signals, allowing to generate the training samples using a synthesized white noise signal, rather than speech recordings with non-negligible silent periods. The acoustic paths between the sound source and an open-sphere spherical microphone array with 32 channels and 4.2 cm radius are simulated by the room impulse response (RIR) generator toolbox [39]. We use a convolution between the simulated RIR and the synthesized noise to generate the measured recordings. Then, the time-domain recordings are transformed into the STFT domain using a 0.5 s window, 90% overlap, 4096-point discrete Fourier transform (DFT), and 8 KHz sampling frequency. Then, the sound pressure is decomposed into the SH domain to obtain the spherical harmonic coefficients. Finally, using (10) and (13) the RHC and RMC, respectively, are calculated. Thirty frequency bins ranging from 1600 Hz to 2500 Hz, measuring the soundfield up to second-order ($N = 2$), are exploited.

TABLE I: Training parameters

Signal	Synthesized noise signals
Room size	$(6 \times 4 \times 3)$ m
Array position	$(2 \times 2 \times 2)$ m
Source-array distance	1 m
T_{60}	800 ms
Labeled DOAs	$\{0°, 1°, 2°, \cdots, 358°, 359°\}$

For simplicity, we assume the source elevation to be fixed. Hence only the azimuth angle is sampled. We use one-degree resolution, i.e., $\mathcal{N}_L = 360$. See Table I for more parameters of the setup. Finally, we construct two training feature sets, with dimensions $360 \times 30 \times 9 \times 2$ for RHC and $360 \times 30 \times 81 \times 2$ for RMC, respectively. We set the number of epochs to 100 when implementing the CNN using tensorflow package.

B. Baseline Approach and Evaluation Metric

For a comprehensive evaluation, in addition to the CNN based approaches, we adopt a simple baseline approach. Specifically, the DOA is estimated by calculating the distances between the test feature and all candidate features associated with the labeled DOAs. We localize the source with the minimum distance calculated below,

$$\mathcal{T}(\boldsymbol{B}^{\star}, \boldsymbol{B}_{n_L}) = \frac{||\boldsymbol{B}^{\star} - \boldsymbol{B}_{n_L}||_2}{||\boldsymbol{B}^{\star}||_2 ||\boldsymbol{B}_{n_L}||_2}, \quad 1 \leq n_L \leq \mathcal{N}_L \qquad (14)$$

where $\boldsymbol{B}^{\star}$ and $\boldsymbol{B}_{n_L}$ denote the test and training features respectively, and $||\cdot||_2$ denotes an ℓ_2 norm of the inputs. In total, we compare four localization approaches which we denote 'RHC-CNN', 'RMC-CNN', 'RHC-distance' and 'RMC-distance', respectively. In the test stage, we apply all approaches to $\mathcal{M}_{\text{tot}} > 1$ test samples. Each test uses a speech source located at randomly selected DOA. To evaluate the algorithms, we use the success-ratio (SR) quality measure:

$$\text{SR} = \frac{\mathcal{M}_{\text{suc}}}{\mathcal{M}_{\text{tot}}} \times 100\% \qquad (15)$$

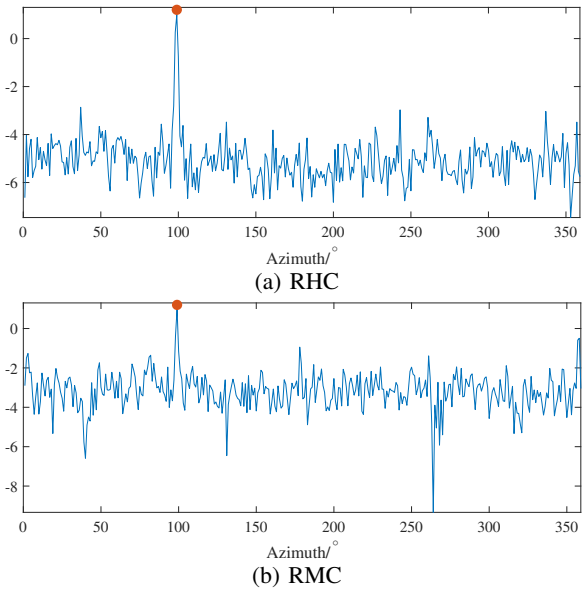
where $\mathcal{M}_{\text{suc}}$ denotes the number of cases in which the sound source was successfully localized. Since a dense DOA sampling resolution is used, we declare successful localization when the absolute error between the estimated and original azimuth angles is less than three degrees, i.e., $|\varphi_{\text{ori}} - \varphi_{\text{est}}| \leq 3°$.

C. Performance in Noise and Reverberation Conditions

The algorithms are evaluated and compared in adverse noise and reverberation conditions. Speech signal randomly selected from the TIMIT database (down-sampled to the frequency of 8 KHz) is used as the input signal. Figure 2 demonstrates examples of the predicted output as a 360×1 vector obtained from the CNN, and the peak point denotes the estimated DOA. We observe that the RHC-based method exhibits a more

TABLE II: Performance in various noisy environments.

Methods	SNR = -6 dB	SNR = -3 dB	SNR = 0 dB
RHC-distance	38%	50%	78%
RMC-distance	4%	16%	30%
RMC-CNN	18%	34%	60%
RHC-CNN	**60**%	**84**%	**92**%



(a) RHC

(b) RMC

Fig. 2: Example of the CNN's output vectors (0 dB noise), where the red circle denotes the source's true azimuth.

significant peak as compared to the RMC-based method, indicating a better capability to classify the test candidate over the training feature set. Table II depicts the localization accuracy of all approaches at low SNR levels, namely $\{-6, -3, 0\}$ dB, contaminated by a Gaussian white noise. For both features, we see that the CNN-based approaches outperform the distance-based methods. This should be attributed to the sensitivity of distance based estimator in (14) to the acoustic reflections and distortions in complex environments, as compared with the CNN based methods that infer the underlying mapping between the source feature(s) and position(s). We observe that the 'RHC-CNN' achieves the best performance under all scenarios, accurately localizing 84% of all the test sources, even with SNR $= -3$ dB noise and $T_{60} = 800$ ms. The 'RMC-distance' degradation due to noise is significantly more pronounced than the 'RHC-distance' degradation, confirming that the RMC is less effective as a feature for localization. This may be attributed to the RMC, being equivalent to the covariance matrix of RHC, suppresses/complicates the direct and simple relation between the reverberant feature and source location, as originally preserved within RHC.

VI. Conclusion

This paper presents an evaluation and comparison of learning-based source localization algorithms in adverse noisy and reverberant environments using two spherical harmonics domain features. We also simplify the operational complexity of the data-driven algorithms by measuring the training samples using synthesized noise signal in the training stage. Extensive evaluations, under various unfavorable acoustic conditions, confirms that the relative harmonic coefficients is more discriminative, thus provides better localization accuracy than the relative modal coherence based method. The current study only examined source localization under single-source scenarios, while further efforts are required in the future for generalizing the algorithm to multi-source scenarios.

REFERENCES

[1] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Tran. on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 1620–1643, 2020.

[2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Tran. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, 1976.

[3] M. Azaria and D. Hertz, "Time delay estimation by generalized cross correlation methods," *IEEE Tran. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 280–285, 1984.

[4] T. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[5] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays*. Springer, 2001, pp. 157–180.

[6] K. Yao, J. C. Chen, and R. E. Hudson, "Maximum-likelihood acoustic source localization: experimental results," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2002, pp. 2949–2952.

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

[8] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C. Deledalle, "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, pp. 3590–3628, 2019.

[9] S. Gannot, M. Haardt, W. Kellermann, and P. Willett, "Introduction to the issue on acoustic source localization and tracking in dynamic real-life scenes," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 3–7, 2019.

[10] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2814–2818.

[11] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1335–1345, 2019.

[12] S. Chakrabarty and E. A. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 136–140.

[13] L. Comanducci, F. Borra, P. Bestagini, F. Antonacci, S. Tubaro, and A. Sarti, "Source localization using distributed microphones in reverberant environments based on deep learning and ray space transform," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 28, pp. 2238–2251, 2020.

[14] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1462–1466.

[15] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[16] X. Wu, D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Spatial feature learning for robust binaural sound source localization using a composite feature vector," in *IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings (ICASSP)*, 2016, pp. 6320–6324.

[17] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 405–409.

[18] P. Grumiaux, S. Kitic, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *arXiv:2109.03465*, 2021.

[19] N. Poschadel, R. Hupke, S. Preihs, and J. Peissig, "Direction of arrival estimation of noisy speech using convolutional recurrent neural networks with higher-order ambisonics signals," in *29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 211–215.

[20] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "CRNN-based multiple DOA estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[21] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic Press, 1999.

[22] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Tran. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[23] P. A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "Improved feature extraction for CRNN-based multiple sound source localization," in *29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 231–235.

[24] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *2010 18th European Signal Processing Conference*, pp. 442–446.

[25] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *IEEE Tran. on Signal Processing*, vol. 42, no. 9, pp. 2481–2491, 1994.

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

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

[28] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Tran. on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[29] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised source localization on multiple manifolds with distributed microphones," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1477–1491, 2017.

[30] 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 Tran. on Audio, Speech, and Language Processing*, vol. 28, pp. 3108–3123, 2020.

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

[32] Y. Hu, T. D. Abhayapala, P. N. Samarasinghe, and S. Gannot, "Decoupled direction-of-arrival estimations using relative harmonic coefficients," in *IEEE 28th European Signal Processing Conference (EUSIPCO)*, 2020, pp. 246–250.

[33] Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, "Evaluation and comparison of three source direction-of-arrival estimators using relative harmonic coefficients," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 815–819.

[34] Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Acoustic signal enhancement using relative harmonic coefficients: spherical harmonics domain approach," in *INTERSPEECH*, 2020, pp. 5076–5080.

[35] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "Multi-source DOA estimation through pattern recognition of the modal coherence of a reverberant soundfield," *IEEE/ACM Tran. on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 605–618, 2020.

[36] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 9, no. 6, pp. 697–707, 2001.

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

[38] P. N. Samarasinghe, T. Abhayapala, M. Poletti, and T. Betlehem, "An efficient parameterization of the room transfer function," *IEEE/ACM Tran. on Audio Speech and Language Processing*, vol. 23, no. 12, pp. 2217–2227, 2015.

[39] E. A. Habets, "Room impulse response (RIR) generator," 2006. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.