

REFERENCE-INVARIANT TRAINING FOR NEURAL ARRAY-GENERIC DOA ESTIMATION

Jackie Lin, Saurabhchand Bhati, Anton Ratnarajah, Mrudula Athi, Wontak Kim, and Minje Kim

University of Illinois Urbana-Champaign, Siebel School of Computing and Data Science, USA 61801
Amazon.com, Inc., Boston, MA, USA 02210

ABSTRACT

Deep learning methods for direction-of-arrival (DOA) estimation are often trained on a single microphone-array geometry and therefore cannot be transferred to a new array without costly array-specific data collection and retraining. This paper proposes an array-generic neural DOA estimator that uses inter-channel phase difference (IPD) features. Yet changing which channel is the reference rewrites the IPD representation even when the scene is unchanged, so a model trained with a fixed reference can latch onto reference-specific cues. To improve robustness to the reference-channel choice and to strengthen array-generic behavior, we investigate three training strategies: randomizing which microphone is used as the IPD reference, expressing microphone coordinates in a reference-relative frame consistent with that choice, and a consistency-enforcing loss that requires the same scene to yield the same DOA under two different references. We show that these strategies improve performance on an unseen array type on both a synthetic dataset and a real-world measured dataset.

Index Terms— Direction-of-arrival, inter-channel phase difference, array-generic, array-agnostic, reference-invariant

1. INTRODUCTION

Direction-of-arrival (DOA) estimation uses a microphone array to infer the spatial direction of one or more acoustic sources relative to the array. It is a fundamental task in spatial audio processing, with applications including teleconferencing, spatial audio capture, source-aware speech enhancement, and human-machine interaction. Classical signal processing approaches exploit inter-microphone spatial cues, such as time difference of arrival (TDoA), generalized cross-correlation (GCC) and the eigenstructure of the spatial covariance matrix (MUSIC) [1, 2, 3]. Given an appropriate steering model of array geometry, these methods are in principle adaptable to arbitrary arrays which have a variable number of microphones in varying locations, and have been shown to work remarkably well in simple acoustic scenarios. However, their performance may degrade substantially under reverberation, noise, unknown number of active sources, moving sources, etc. More recently, deep learning-based DOA estimation methods have demonstrated improved robustness in such challenging scenarios [4, 5, 6].

Despite substantial progress, practical deployment of neural DOA estimators remains challenging because the estimators fail to generalize when the array geometry changes. In addition, room characteristics such as reverberation time, sources’ geometric diversity, and acoustic reflections introduce substantial domain mismatch between training data and real-world recordings. Many neural DOA estimators are nonetheless trained using a fixed array configuration [7, 8], and thus fail to generalize to unseen array geometries and acoustic environments which results in costly re-collection of data

and re-training for each new array configuration. This is undesirable for applications in which the DOA model is deployed across multiple devices or product variants. We therefore consider the *array-generic* DOA problem, in which the estimator is conditioned on the known coordinates of a device whose geometry was not seen during training.

Several recent works have proposed DNNs that explicitly incorporate array geometry, aiming at their generalization to arbitrary, unseen microphone arrays. For example, FC-GA combines GCC-PHAT-based features with microphone coordinates and trains over randomized planar geometries [9]. Neural-SRP processes microphone pairs together with their coordinates using a permutation-invariant aggregation mechanism [10]. IPDnet instead learns direct-path inter-channel phase differences (IPDs), from which DOA can subsequently be recovered [11]. More recently, GI-DOAEnet incorporates microphone position encodings into channel-wise neural features [12], while PhaseCoder jointly processes multichannel audio and microphone coordinates to obtain geometry-agnostic spatial representations [13]. Approaches in sound event localization and detection (SELD) research circumvent geometry issues by relying on ambisonics representations [14, 15], though significant recent work has been underway to encode arbitrary arrays to ambisonics. Array-generic DOA has also been formulated using measured or simulated array transfer functions as explicit device metadata [16]. Overall, these works demonstrate that conditioning a neural estimator on array information can improve generalization to previously unseen arrays.

We investigate a complementary source of device dependence that has been underexplored: *the choice of reference microphone* and its implications. Relative spatial features, such as IPD, are inherently defined with respect to a reference microphone:

$$\text{IPD}_{i,r}(f, t) = \angle X_i(f, t) - \angle X_r(f, t), \quad (1)$$

where r denotes the selected reference microphone, which governs the pairwise IPD of other channels i , even if the physical acoustic scene remains unchanged. Consequently, a model trained with a fixed reference channel, such as IPDnet, may inadvertently learn reference-dependent patterns rather than reference-independent spatial relationships. Instead, we seek reference-channel flexibility to ensure our models can adapt to different channel ordering or the lack of a designated reference microphone. Related issues have been studied in multichannel speech enhancement, where improved array generalization was achieved by varying the reference channel and microphone ordering during training [17], and where VarArray avoids a fixed reference by defining spatial features relative to the array-average spectrum [18]. Yet, this idea remains largely unexplored in neural DOA estimation, where both acoustic and geometric representations must remain consistent under reference changes.

In this work, we investigate the effect of reference channel con-

sistency on neural DOA estimation in the form of structured train-time randomization. This is done via three strategies: 1) randomly selecting the reference r and computing the IPD features relative to that microphone 2) applying a compatible transformation of the microphone coordinates to a reference-relative frame that, expresses microphone locations relative to the chosen IPD reference microphone, and 3) enforcing consistency of DOA estimates between different choices of reference channel during training. These interventions expose the model to physically equivalent views and encourage it to learn the spatial relationship between the source and the array rather than reference-specific correlations.

Our main contributions can be summarized as follows:

- We identify reference-channel dependence as a distinct, complementary generalization problem in neural DOA estimation.
- We propose *reference-randomized training* that jointly randomizes IPD and microphone-coordinate representations, providing multiple physically equivalent views of each training example.
- We propose a permutation-invariant neural DOA estimator that combines shared per-microphone audio and geometry encoders, cross-microphone self-attention, and temporal modeling, and learns from an additional consistency objective across different reference choices.
- Experiments show that our reference channel strategies improve generalization on unseen, arbitrary arrays.

2. METHOD

2.1. Reference Consistency

The proposed model takes multichannel reverberant speech from an arbitrary microphone array in an arbitrary room, and predicts the DOA of a single static sound source. We use the short-time Fourier transform (STFT) representation of the multichannel audio. To each complex STFT coefficient, a phase-preserving magnitude compression is applied where $f(c) = \frac{\log(1+|c|)}{|c|+\epsilon}c$. The input to the model consists of each microphone channel’s compressed STFT real and imaginary coefficients and the IPD calculated channel r as the reference concatenated along the frequency axis, yielding the shape $[M, 3F, T]$, where M , F , and T denote the numbers of channels, frequency bins, and time frames. Channel r ’s IPD is against itself.

We consider three strategies toward improving consistent DOA prediction regardless of the reference channel choice: IPD reference randomization, reference-relative coordinates, and a consistency-enforcing training scheme and loss function.

IPD Reference Randomization: During training, for each input sample, we randomly choose a microphone channel r to be the reference channel and compute the IPD of every channel relative to it (1). Hence, the model is encouraged to learn spatial relationships invariant to the choice of reference channel, while also increasing the effective diversity of array layouts in the training dataset.

Reference-Relative Coordinates: To maintain physical consistency with the randomized IPD representation, the microphone coordinates are transformed relative to the same randomly selected reference, going from the “canonical” frame to the “reference-relative” frame, as illustrated in Fig 1. An array’s canonical frame is predefined as follows: the origin O is placed at the array centroid p_c (the mean position of all microphones in the array), the upward direction $+z$ is parallel to the world z -axis, and a frontal direction $+x$ is assigned, i.e., the $(1, 0, 0)$ vector of the canonical frame. All microphone locations $p_i \in \mathbb{R}^3$ are defined in this canonical frame.

Training with reference-relative coordinates requires re-defining a reference-relative frame around the chosen reference microphone

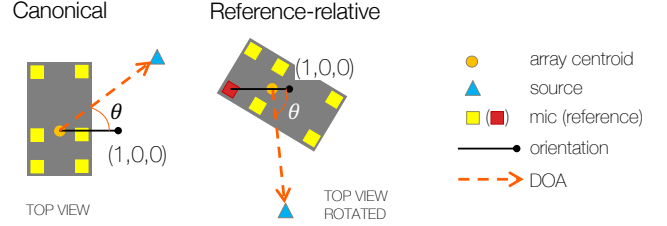


Fig. 1. Device-native coordinate systems. (Left) Canonical coordinates: the origin is the array centroid and $+x$ is the predefined front of the device. (Right) Reference-relative coordinates: the origin is the reference microphone and the front points from the reference microphone to the array centroid. The DOA points from the array centroid to the source. Coordinates are in meters.

r ’s position and expressing all microphone coordinates in this new frame. Our proposed method sets the reference-relative origin to the reference-microphone location $O := p_r$, and the reference-relative “front” to the vector pointing from the reference to the array centroid, i.e., $+x := p_r - p_c$. The coordinates of microphone i are then the difference between its canonical coordinates and those of the reference, rotated by the angle between the canonical and reference-relative front, i.e., $p_i = Q_r(p_i - p_r)$, where p_i is the position of microphone i in the reference-relative frame and Q_r is the rotation that maps $+x \rightarrow +x$. The reference-relative $+z$ is left unchanged, pointing upward in world coordinates, which is appropriate for our application.

To express the DOA in the reference-relative coordinate frame, the canonical DOA $d \in \mathbb{R}^3$ (defined as the unit vector pointing from p_c to the source position) is rotated into the new frame, i.e., $d_r = Q_r d$. At inference time, the predicted reference-relative target $\hat{d}_r$ is rotated back to the canonical frame, $\hat{d} = Q_r^T \hat{d}_r$, where $Q_r^{-1} = Q_r^T$ for rotation matrices. Note that the target DOA vector in the reference-relative frame d_r does not point from the reference microphone to the source. Rather, d_r represents the same centroid-to-source vector shifted to the new origin O .

Thus, under this strategy, both the acoustic and geometric representations undergo a consistent change of reference, encouraging the model to learn spatial relationships that are invariant to the reference microphone selection.

Consistency Loss To further encourage the model to predict the same DOA no matter which microphone is chosen as the reference, we introduce a two-view consistency training scheme and objective. For each training example, two distinct reference indices r and s are sampled uniformly from the M microphones (with $s \neq r$). The network is evaluated twice under these references, yielding unit-vector predictions $\hat{d}_r$ and $\hat{d}_s$. The consistency training loss is

$$\mathcal{L}_{\text{consistency}} = \mathcal{L}_{\text{doa}}(\hat{d}_r, d) + \mathcal{L}_{\text{doa}}(\hat{d}_s, d) + \lambda (1 - \hat{d}_r \cdot \hat{d}_s), \quad (2)$$

$$\mathcal{L}_{\text{DOA}}(x, y) = 1 - x \cdot y \quad (3)$$

where $\mathcal{L}_{\text{DOA}}$ is the cosine loss and λ controls the strength of the consistency penalty (we use $\lambda = 0.1$). The first two terms supervise each view against the label; the third term penalizes disagreement between the two predictions. In effect, the model is trained to recover the same canonical-frame DOA regardless of which microphone is used as the IPD reference.

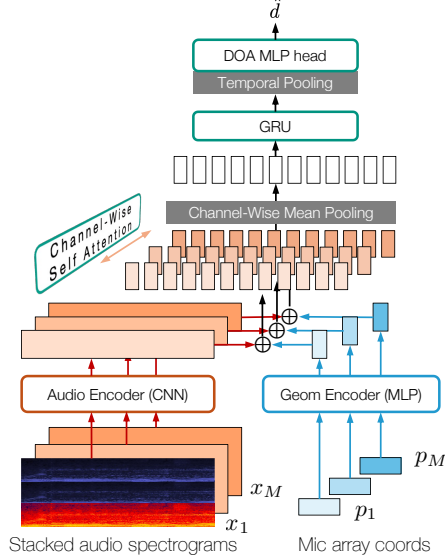


Fig. 2. Neural DOA Estimator Architecture. Inputs are stacked spectrograms and coordinates. The model processes each microphone channel independently, then applies self-attention across channels.

2.2. Neural DOA Estimator Architecture

We propose a base deep neural network (DNN) to estimate the DOA regressively. The inputs are processed by a shared lightweight convolutional audio encoder that acts on each microphone independently. It consists of two Conv1D layers, each followed by a GeLU activation, with a kernel size of 3 and strides of 1 and 2, respectively, and each with 128 output channels. It encodes the input audio into shape $[M, C, N]$ where $C = 128$ is the number of hidden channels and N is the number of time frames. Separately, to incorporate explicit geometry information, an MLP-based microphone geometry encoder maps the microphone coordinates to geometry embeddings $[M, 3] \mapsto [M, C]$; again, the weights are shared across microphones. Then the microphone geometry embeddings are repeated across the time dimension and added to the encoded audio.

Following the general idea in most DOA approaches, we garner useful information across microphones using a 4-head self-attention layer to attend to the fused audio-geometry features over the microphone channel dimension [19]. The output is mean-pooled across microphones into a single time-sequence channel representing the entire array of shape $[1, C, N]$. We purposefully design the model to be permutation-invariant to the input channel order, via independently applying the audio and geometry encoder to each mic channel, the self-attention, and the pooling across microphones. Finally, the sequence is processed in the time dimension by a 2-layer bi-directional GRU, followed by a mean-pooling across time, and a linear layer mapping the GRU hidden dimension to the target DOA dimension. The training loss function is the cosine loss (3) unless otherwise noted. The model has a total of 2.3M parameters.

2.3. Dataset

The training and evaluation datasets are simulated using the LibriSpeech dataset [20] and Pyroomacoustics [21] synthesized room impulse responses (RIR). To simulate the sound captured by the microphone array, RIRs are generated for each array geometry and con-

volved with the clean speech signal. For each training sample, the array is randomly chosen from six common array types (cellphone, laptop, circular, hexagonal, random planar, and L-shaped) with randomly sampled dimensions. All types have six channels, and each type is assigned a reasonable frontal direction. Because many of these common devices lie flat on table tops, $+z$ in the canonical frame always points upward. We generate 200k multi-channel training samples over 10k unique shoebox rooms with sizes uniformly sampled from $2.4 \times 3 \times 3$ to $4.5 \times 12 \times 12$ m and reverberation times from 0.15 to 1.0 s. The array’s location, its rotation about the z -axis, and the source location are randomized within the room. To simulate room noise, we add spatially isotropic [22] brown noise to each channel’s reverberant speech so that the SNR ranges from -60 to -40 dBFS. Each audio input is 4 s long at 16 kHz.

We create a synthetic test set with two splits using the aforementioned pipeline: Split I is an in-distribution set of 3.2k samples over 200 unseen shoebox rooms. Split II contains 800 samples from linear arrays of length 12–50 cm, an array type unseen during training. We also evaluate on RSL2019 [23], a real-world DOA dataset whose Dev split uses a stand-mounted, symmetric 4-channel array in a relatively easy acoustic condition, and whose Test split uses tabletop recordings in a conference room in which the direct path is sometimes occluded. Non-speech segments in RSL Test were removed with the WebRTC¹ voice activity detector (VAD) before DOA estimation.

3. EXPERIMENTS

The effect of each reference consistency strategy, IPD reference randomization, reference-relative coordinates, and consistency loss, is examined in a series of ablation studies. First, we train the base model `Base` without any reference consistency strategy, where the IPD is always calculated with fixed reference channel $r = 0$ and mic coordinate inputs are in the canonical frame. A second model, `Coor`, is trained with fixed $r = 0$ but with reference-relative coordinates so that the array geometry is always expressed in a relative manner in congruence with the IPD reference mic.

Next, we employ IPD randomization in models `IPD` and `IPD-Coor`, and ablate reference-relative coordinates. Finally, we add the two-view consistency loss under randomized IPD reference, and ablate reference-relative coordinates (`IPD-Loss` vs `IPD-Coor-Loss`). All models are trained for 40 epochs on batch size $b = 256$ using the AdamW optimizer with learning rate $lr = 1e-4$ and weight decay $\delta = 1e-4$.

Three classical, data-agnostic baselines are implemented: `SRP-PHAT` which aggregates DOA predictions per frequency band based on direct path arrival from a steered beamformer; `GCC-PHAT` which selects the time sample with the largest peak in the whitened cross-correlation function between microphone channels; and `MUSIC` which uses the eigenstructure of the microphone array’s spatial covariance matrix to form a high-resolution spatial pseudo-spectrum.

4. RESULTS

Table 1 shows the performance of the proposed method, ablations, and classical baselines on the seen and unseen array type splits of the synthetic test set. It reports mean absolute error (MAE), median absolute error, 90th percentile absolute error (P90), MAE and median error of the azimuthal and the elevation angle in degrees, along with accuracy within 10° (Acc_{10}).

¹<https://github.com/wiseman/py-webrtcvad>.

Model Configuration				Test I: Seen Array Types								Test II: Unseen Array Type		
ID	r	Ref. Rel. Coords.	Consist. Loss	MAE ($^\circ \downarrow$)			Median ($^\circ \downarrow$)			P90 ($^\circ \downarrow$)	Acc ₁₀ (%) $\uparrow$	Azimuthal Error		
				3D	az	el	3D	az	el			MAE	Median	Acc ₁₀
Base	ch 0	$\times$	—	7.8	4.9	4.9	6.4	3.2	3.6	15.0	74.9	18.3	13.5	40.3
Coor	ch 0	$\checkmark$	—	6.2	3.3	4.5	5.0	2.2	3.4	12.1	83.8	13.2	9.0	53.6
IPD	rand	$\times$	$\times$	5.9	3.1	4.3	4.8	2.1	3.3	11.3	86.3	16.3	11.9	44.2
IPD-Coor	rand	$\checkmark$	$\times$	5.0	2.3	3.9	4.2	1.5	3.0	9.6	91.0	10.5	6.3	64.1
IPD-Loss	rand	$\times$	$\checkmark$	5.7	2.9	4.2	4.8	2.0	3.2	11.0	86.9	16.9	12.7	42.6
IPD-Coor-Loss	rand	$\checkmark$	$\checkmark$	4.9	2.4	3.8	4.0	1.6	2.8	9.4	91.5	10.8	6.5	63.4
SRP-PHAT	—	—	—	37.7	3.2	37.2	34.1	1.7	33.7	58.0	10.9	7.5	4.3	75.0
GCC-PHAT	—	—	—	35.5	5.0	34.6	30.1	1.9	29.6	64.3	14.5	9.9	5.2	69.0
MUSIC	—	—	—	58.6	13.4	56.2	58.3	7.0	56.7	82.5	5.1	16.1	10.0	49.2

Table 1. DOA performance on synthetic test splits I. Seen Array Types and II. Unseen Array Type (azimuthal error only)—linear array. 3D = $\angle(\hat{d}, d)$ absolute angular error, az = azimuthal error, el = elevation error. P90 and Acc₁₀ are computed on 3D angle errors. Top rows are the proposed DNN ablations; reported DNN errors average over all choices of reference channel r 0–6. Bottom rows are classical methods.

ID	RSL Dev			RSL Test		
	MAE	Median	Acc ₁₀	MAE	Median	Acc ₁₀
Base	5.90	5.08	74.5	28.0	9.1	52.1
Coor	4.41	3.92	87.4	23.6	<u>5.8</u>	61.3
IPD	4.61	4.14	88.0	22.2	6.1	63.2
IPD-Coor	4.87	4.53	86.9	22.7	5.9	62.0
IPD-Loss	<u>4.40</u>	4.13	<u>91.5</u>	22.4	6.4	62.2
IPD-Coor-Loss	4.97	4.83	91.2	22.6	<u>5.8</u>	63.3
SRP-PHAT	4.36	4.50	100.0	21.5	3.0	67.2
GCC-PHAT	6.47	5.50	79.4	30.3	7.0	57.4
MUSIC	4.84	4.50	97.8	23.1	9.0	57.6

Table 2. Azimuthal DOA performance on RSL2019 [23]. RSL Dev contains obstruction free recordings every 5° at 1 & 1.5m from the array; RSL Test samples contain occluding objects and near-field surfaces. **Bold** = best among all rows. Underline = best among DNNs. MAE ($^\circ \downarrow$), Median ($^\circ \downarrow$), Acc₁₀ (% $\downarrow$).

First, the Base model without any train-time strategies performs reasonably well on DOA estimation, likely in part to the diversity of the array geometries seen during training and input permutation-invariant structure. Next, the effect of reference-relative coordinates on its own is significant, on average improving the MAE and median error by 1.25° and 1.0° respectively in both the static reference channel case (Base $\rightarrow$ Coor) and the randomized reference channel case (IPD $\rightarrow$ IPD-Coor). The effect of the proposed coordinate system is even larger on the unseen linear arrays, improving MAE and median error by 5.5° and 5.1° respectively on average. We see that randomizing the reference channel has a positive effect on performance, on average improving MAE and median error by 1.6° and 1.2° respectively, as seen by comparing (Base $\rightarrow$ IPD) and (Coor $\rightarrow$ IPD-Coor). Again, the gains were larger on the unseen array split improving MAE and median error by 2.4° and 2.2° respectively on average.

The consistency loss marginally improved the random reference case when reference relative coordinates were not used, as seen in (IPD $\rightarrow$ IPD-Loss). When reference relative coordinates were on (IPD-Coor $\rightarrow$ IPD-Coor-Loss), the consistency loss training scheme reduced outlier predictions (better P90) while slightly worsening azimuthal MAE and median error. This suggests that performance gains from the consistency loss are greater when reference-relative coordinates are not used than when they are, where ultimately reference-relative coordinates cannibalize most of the gains from consistency loss. As a result, our complete proposed method IPD-Coor and IPD-Coor-Loss perform the best.

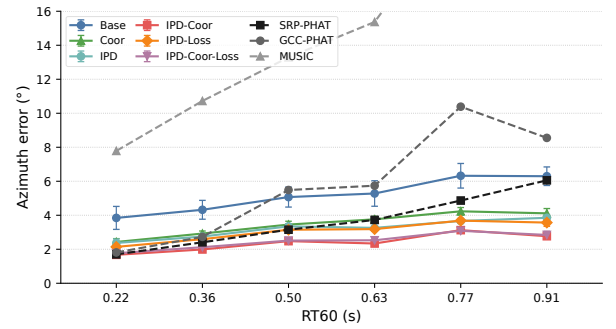


Fig. 3. Reverb time vs azimuthal MAE on synthetic test split I.

Compared to the classical baseline in Tab. 1, IPD-Coor and IPD-Coor-Loss do perform better than or equal to SRP-PHAT respectively in azimuthal prediction; furthermore, all of the ablations outperform GCC-PHAT and MUSIC. One thing of note, the classical methods' elevation performance is very poor because the test arrays are flat and have a very small aperture in the z dimension, leading to high-variance estimates. Lastly we show in Fig 3 that the neural methods are more robustness to strong room reverberation, where the classical methods degrade faster than the proposed method.

Table 2 reports the performance on the measured dataset RSL2019. On RSL Dev, all of the DNN methods do reasonably well; Coor achieves the best median angular error and IPD-Loss performs well on MAE and Acc₁₀ at 91.5%. The substantial increase in Acc₁₀ when consistency loss is used further support its outlier-suppression effect. Base falls the shortest of the ablated versions by a clear margin, indicating that some combination of the proposed strategies transfers gains from simulation to real-world settings. Though classical method SRP-PHAT still outperforms the proposed DNNs on most metrics, the results demonstrate that the trained DNNs are adaptable and reasonably robust to an unseen number of channels and a new array type.

5. CONCLUSION

This paper proposed an IPD-based deep neural network for DOA estimation. Three strategies, IPD reference channel randomization, reference-relative coordinates, and a two-view consistency loss were demonstrated to improve the performance of the model on unseen synthetic linear arrays and on a real-world unseen array.

6. REFERENCES

- [1] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar. 1986.
- [2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.
- [3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., Digital Signal Processing, pp. 157–180. Springer, Berlin, Heidelberg, 2001.
- [4] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, Sept. 2018, pp. 1462–1466.
- [5] 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, July 2022.
- [6] B. Yang et al., "RealMAN: A Real-Recorded and Annotated Microphone Array Dataset for Dynamic Speech Enhancement and Localization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 105997–106019, Dec. 2024.
- [7] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting Temporal Context in CNN Based Multisource DOA Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1594–1608, 2021.
- [8] B. Yang, H. Liu, and X. Li, "SRP-DNN: Learning Direct-Path Phase Difference for Multiple Moving Sound Source Localization," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 721–725, ISSN: 2379-190X.
- [9] U. Kowalk, S. Doclo, and J. Bitzer, "Geometry-aware DOA estimation using a deep neural network with mixed-data input features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.
- [10] E. Grinstein, T. van Waterschoot, M. Brookes, and P. A. Naylor, "The neural-SRP method for positional sound source localization," in *Proc. 57th Asilomar Conf. Signals, Systems, and Computers*, 2023, pp. 1318–1323.
- [11] Y. Wang, B. Yang, and X. Li, "IPDnet: A universal direct-path IPD estimation network for sound source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 5051–5064, 2024.
- [12] M.-S. Baek, J.-H. Chang, and I. Cohen, "DNN-based geometry-invariant DOA estimation with microphone positional encoding and complexity gradual training," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 2360–2376, 2025.
- [13] A. Dementyev et al., "PhaseCoder: Microphone geometry-agnostic spatial audio understanding for multimodal LLMs," in *Proc. 43rd Int. Conf. Machine Learning (ICML)*, 2026.
- [14] 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, Mar. 2019.
- [15] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-Coupled Cartesian Direction of Arrival Representation for Sound Event Localization and Detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 915–919, ISSN: 2379-190X.
- [16] M. Heikkinen, A. Politis, K. Drossos, and T. Virtanen, "Neural array-generic direction-of-arrival estimation exploiting array transfer functions," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2026.
- [17] S. Zhang and X. Li, "Microphone array generalization for multichannel narrowband deep speech enhancement," in *Proceedings of Interspeech 2021*, Brno, Czechia, Aug. 2021, pp. 666–670.
- [18] T. Yoshioka et al., "VarArray: Array-geometry-agnostic continuous speech separation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, pp. 6027–6031.
- [19] M.-S. Baek, J.-H. Chang, and I. Cohen, "DNN-Based Geometry-Invariant DOA Estimation With Microphone Positional Encoding and Complexity Gradual Training," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2360–2376, 2025.
- [20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [21] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, p. 351–355.
- [22] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating non-stationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.
- [23] R. Sheelvant et al., "RSL2019: A Realistic Speech Localization Corpus," in *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Oct. 2019, pp. 1–6, ISSN: 2472-7695.