DOA Estimation with Lightweight Network on LLM-Aided Simulated Acoustic Scenes

Haowen Li*, Zhengding Luo*, Dongyuan Shi†, Boxiang Wang*, Junwei Ji*, Ziyi Yang*and Woon-Seng Gan*
* Nanyang Technological University, Singapore
† Northwestern Polytechnical University, China
E-mail: haowen.li@ntu.edu.sg, luoz0021@e.ntu.edu.sg, dongyuan.shi@nwpu.edu.cn,
{boxiang001,junwei002, ziyi016}@e.ntu.edu.sg, ewsgan@ntu.edu.sg

***Abstract*—Direction-of-Arrival (DOA) estimation is critical in spatial audio and acoustic signal processing, with wide-ranging applications in real-world. Most existing DOA models are trained on synthetic data by convolving clean speech with room impulse responses (RIRs), which limits their generalizability due to constrained acoustic diversity. In this paper, we revisit DOA estimation using a recently introduced dataset constructed with the assistance of large language models (LLMs), which provides more diverse spatial audio scenes. We benchmark several representative neural network-based DOA methods on this dataset and propose LightDOA, a lightweight DOA estimation model based on depthwise separable convolutions and a gated recurrent unit, specifically designed for mutil-channel input in varying environments. Experimental results show that LightDOA achieves satisfactory accuracy and robustness across various acoustic scenes while maintaining low computational complexity. This study highlights the potential of LLM-assisted synthetic datasets for advancing DOA research, and demonstrates LightDOA as an efficient solution for resource-constrained applications.**

I. Introduction

Spatial sound perception and directional cues are crucial in real-world applications [1]–[3]. As a core technique in array signal processing, Direction-of-Arrival (DOA) estimation has recently shown great promise for enhancing learning-based ANC systems by providing spatial priors that improve noise suppression and robustness [4]–[6]. In recent years, neural network (NN)-based approaches have gained popularity due to their ability to model complex spatial cues from acoustic signals. Numerous architectures have been explored for this task, including multi-layer perceptrons (MLPs) [7]–[9], convolutional neural networks (CNNs) [10]–[12], and convolutional recurrent neural networks (CRNNs) [13], [14], all of which have shown promising results under controlled settings.

However, these methods face a key limitation: their success heavily relies on the diversity and realism of training data. Most existing datasets are constructed by convolving clean speech with simulated room impulse responses (RIRs), resulting in acoustically constrained and semantically limited scenarios [15], [16]. Such simplified data often fails to reflect the complexities of real-world environments, where background noise, reverberation, and diverse sound events are common. As a result, models trained on these datasets tend to struggle with generalization in unseen or acoustically diverse conditions.

To better bridge the gap between synthetic and real-world acoustic conditions, we leverage a recently proposed spatial audio dataset constructed with the assistance of large language models (LLMs)-*Both Ears Wide Open (BEWO)* [17]. Compared to traditional RIR-based corpora, this dataset introduces significantly greater diversity in both acoustic and semantic dimensions, including non-speech content, various environments, and dynamic contexts. This not only offers a more faithful proxy for real-world sound scenes, but also presents new challenges by exposing models to more complex conditions, raising the bar for network robustness and adaptability.

In this paper, we investigate single-source (SS) DOA estimation on the BEWO dataset, systematically evaluate several representative neural DOA models under this challenging setting, and propose *LightDOA*, a lightweight CRNN-based architecture that leverages depthwise separable convolutions to efficiently extract spatial cues from two-channel input. Inspired by the MobileNet family [18]–[21], *LightDOA* achieves a favorable balance between accuracy and computational cost. Experimental results show that our method achieves comparable accuracy to existing approaches while significantly reducing model complexity. Therefore, *LightDOA* holds promise for deployment in downstream tasks such as ANC system [22], [23], where both source localization and real-time noise suppression are required on edge devices.

Our contributions are threefold:

- We benchmark representative NN-based DOA models on an LLM-assisted dataset resembling real-world scenarios;
- We propose *LightDOA*, a lightweight CRNN-based architecture based on depthwise separable convolutions;
- We demonstrate that *LightDOA* maintains competitive accuracy with significantly lower complexity, benefiting from the diverse spatial data generated with LLM assistance. This highlights its potential for efficient DOA estimation and deployment in downstream applications.

II. Dataset Description

A. Dataset Overview

We utilize the SS set of the BEWO dataset [17] [1], which is a large-scale synthetic spatial audio corpus created by combining captions generated by LLMs with physically based acoustic rendering. The process of constructing the SS-set from *AudioCaps* [24] is illustrated in Fig. 1, including caption

[1]Dataset: https://huggingface.co/datasets/spw2000/BEWO-1M

979-8-3315-7206-8/25/$31.00

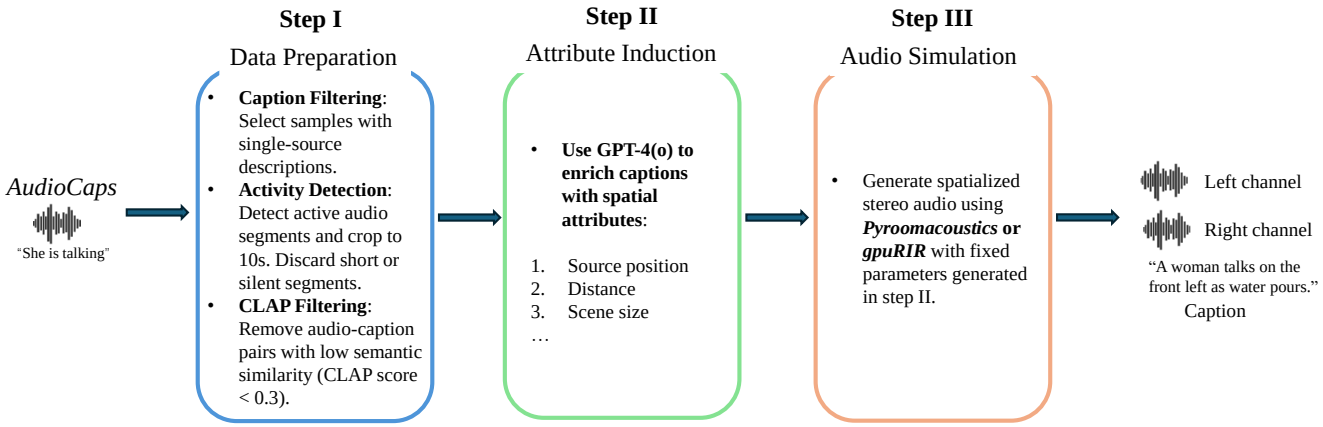


Fig. 1: Pipeline for Constructing the SS Subset from *AudioCaps* [17].

TABLE I: DOA dataset composition used in this study.

Subset	Samples	Description
Training	43,213	For model training
Validation	1,981	For early stopping, tuning
Test	4,301	For final evaluation
Source	AudioCaps [24] (caption-based retrieval)	
Spatialization	Simulated with room acoustics and IPD/ILD	
Content Type	Single dominant source per scene	
Labels	Azimuth DOA angles	

filtering, activity detection, CLAP [25]-based audio-text alignment filtering, spatial attribute enrichment via GPT-4(o) [26], and stereo audio simulation using Pyroomacoustics [27] or gpuRIR [28].

The DOA dataset used in our study is summarized in Table I, including train/val/test splits, audio source, and labels. All audio samples in the SS-subset contain only a single dominant source per scene, ensuring spatial labels are well-defined.

Limitation on Front–Back Resolution: Despite being dual-channel, the BEWO dataset does not incorporate head-related transfer functions (HRTFs) or any form of anatomical filtering. Consequently, directional cues are limited to interaural phase difference (IPD) and interaural level difference (ILD). As shown in Fig.2 (a), the raw DOA angles can result in front–back ambiguity. To mitigate this, a front–back mapping is applied to fold all angles into the $[0°, 180°]$ range. The resulting class distributions for the training, validation, and test subsets are shown in Fig.2 (b).

Imbalanced Class Distribution: Another limitation lies in the distribution of DOA angles across the dataset. Although the train, validation, and test splits maintain similar overall statistics, the number of samples per class (i.e., per direction) is not uniformly distributed. Certain angles are overrepresented, while others have relatively few examples. This internal class imbalance can affect model training and evaluation, especially in tasks requiring fine-grained angular resolution. This imbalance is illustrated in Figure 2.

B. Room and Microphone Configuration

Each scene is simulated within a cuboid room $[R_0, R_1, R_2]$ based on a scene type-specific base size r, perturbed with uniform noise:

$$[R_0,\ R_1,\ R_2] = [r + \xi_{r0},\ r + \xi_{r1},\quad r + \xi_{r2}],$$
$$\xi_{ri} \sim \mathcal{U}(-0.1r,\ 0.1r) \tag{1}$$

Room size r and reverberation time (RT60) are detailed in Table II. We use $\mathcal{U}(a, b)$ to denote a uniform distribution over the range $[a, b]$.

a) Microphone Placement.: A stereo microphone array is positioned near the room center:

$$[M_0,\ M_1,\ M_2] = \left[\frac{R_0}{2} + \xi_{m0},\ \frac{R_1}{2} + \xi_{m1},\quad \frac{R_2}{2} + \xi_{m2}\right],$$
$$\xi_{mi} \sim \mathcal{U}(-0.1r,\ 0.1r) \tag{2}$$

with microphones placed along axis M_1 using a fixed inter-microphone spacing sampled from the range $[0.16, 0.18]$m, $\mathcal{U}(a, b)$ to represent a uniform distribution over $[a, b]$.

b) Source Placement.: The source azimuth θ is sampled around class-centered angles, and converted to Cartesian coordinates based on d and microphone center:

$$\mu_{\text{begin}} = [M_0 + d\sin\theta,\ M_1 + d\cos\theta,\ M_2] \tag{3}$$

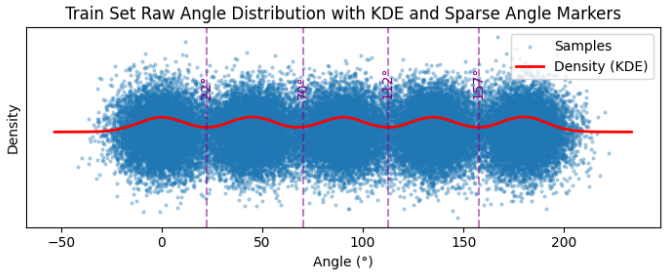
Class-based distributions for θ are listed in Table II.

C. Source Distance Modeling

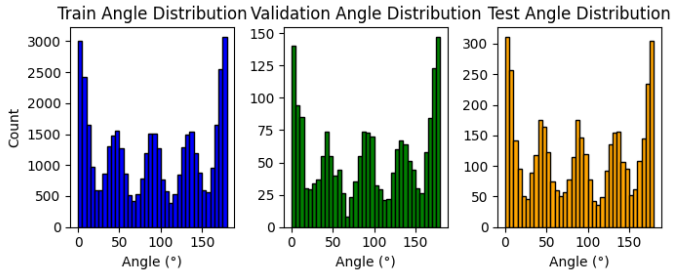
The source–array distance d is sampled by applying a relative ratio α_d to the shortest axis-aligned distance from the microphones to the room boundaries:

$$d = \alpha_d \cdot \min(R_0 - M_0,\ R_1 - M_1,\ M_0,\ M_1) \tag{4}$$

Here, α_d is selected to represent near-, moderate-, and far-field distances. See Table II for detailed sampling settings.



(a) KDE of training angles with sparse regions highlighted



(b) Histogram of angle classes across train/val/test splits

Fig. 2: Angle distribution of the BEWO dataset. (a) shows the kernel density estimation (KDE) of raw DOA angles in the training set, where sparse regions (e.g., around 22°, 67°, 112°, and 157°) are visually identifiable. (b) shows the histogram of DOA angles after applying the front–back mapping across training, validation, and test sets, which are globally consistent but imbalanced within each set.

TABLE II: Scene simulation parameters in the BEWO dataset.

Attribute	Options List	Sampling Value
Room size r	Outdoors	100 m
	Large	$\mathcal{U}(40, 90)$
	Moderate	$\mathcal{U}(20, 40)$
	Small	$\mathcal{U}(5, 20)$
Direction θ	Left	$\mathcal{N}(180°,\ 11°)$
	Front-left	$\mathcal{N}(135°,\ 11°)$
	Front	$\mathcal{N}(90°,\ 11°)$
	Front-right	$\mathcal{N}(45°,\ 11°)$
	Right	$\mathcal{N}(0°,\ 11°)$
Distance ratio α_d	Far	$\mathcal{U}(0.6,\ 0.9)$
	Moderate	$\mathcal{U}(0.3,\ 0.6)$
	Near	$\mathcal{U}(0.1,\ 0.3)$
RT60	—	$\mathcal{U}(0.3,\ 0.6)$ (or 0 for outdoors)
Mic spacing	—	0.16–0.18 m

Note: $\mathcal{U}(a,\ b)$ indicates a uniform distribution over $[a,\ b]$, and $\mathcal{N}(\mu,\ \sigma)$ denotes a normal distribution with mean μ and standard deviation σ.

III. Proposed Method

A. Feature Extraction

We formulate the dual-channel DOA estimation task as a classification problem over discrete azimuth angles. Given a dual-channel audio waveform $(x_1(t), x_2(t))$ captured by a two-microphone array, the objective is to predict the azimuth angle $\theta \in [0°, 180°]$ of the dominant sound source.

We first transform the time-domain waveforms into the time-frequency domain using the Short-Time Fourier Transform (STFT):

$$X_1(f,t) = \text{STFT}[x_1(t)], \tag{5}$$
$$X_2(f,t) = \text{STFT}[x_2(t)], \tag{6}$$

where $X_1(f,t)$ and $X_2(f,t)$ are the complex-valued spectrograms of the two input channels.

To capture inter-channel spatial cues, we compute the IPD, defined as the difference in phase between the two channels at each time-frequency (TF) bin:

$$\text{IPD}(f,t) = \angle X_1(f,t) - \angle X_2(f,t), \tag{7}$$

where $\angle X$ denotes the phase of a complex value. The resulting feature $\text{IPD}(f,t)$ is a real-valued matrix with the same shape as the input spectrogram, and serves as the input to our neural network.

B. Network Architecture

To efficiently model the directional information embedded in IPD, we design a lightweight neural network architecture named *LightDOA*, illustrated in Figure 3. The model consists of a convolutional frontend followed by a recurrent backend, tailored for time-frequency DOA classification using two-channel input.

The input IPD feature map, with shape $(B, 1, F, T)$, is passed through a stack of depthwise separable convolutional blocks. These layers progressively extract spatial representations while maintaining low computational cost. After three convolutional stages, we apply a 2×2 adaptive average pooling to compress the feature map, yielding a compact representation of shape $(B, 32, 2, 2)$.

The pooled features are then reshaped and permuted to match the input format of a uni-directional gated recurrent unit (GRU), treating the time axis as the sequence dimension. Specifically, the GRU operates on features of shape $(B, T = 32, F = 4)$, producing a sequence embedding that captures temporal dependencies across frames.

The GRU output is flattened and passed through two fully connected (FC) layers to produce a 37-dimensional classification output, corresponding to azimuth classes in $[0°, 180°]$ with $5°$ resolution. The final prediction is computed as the expected angle over the softmax probabilities, enabling smooth regression during inference.

This architecture balances representational capacity and efficiency, making it well-suited for real-time or resource-constrained spatial audio applications.

C. Loss Function

As discussed in Section II, the BEWO dataset lacks HRTFs, resulting in front-back ambiguity. To address this issue, we apply a symmetric angle mapping during preprocessing. Specif-

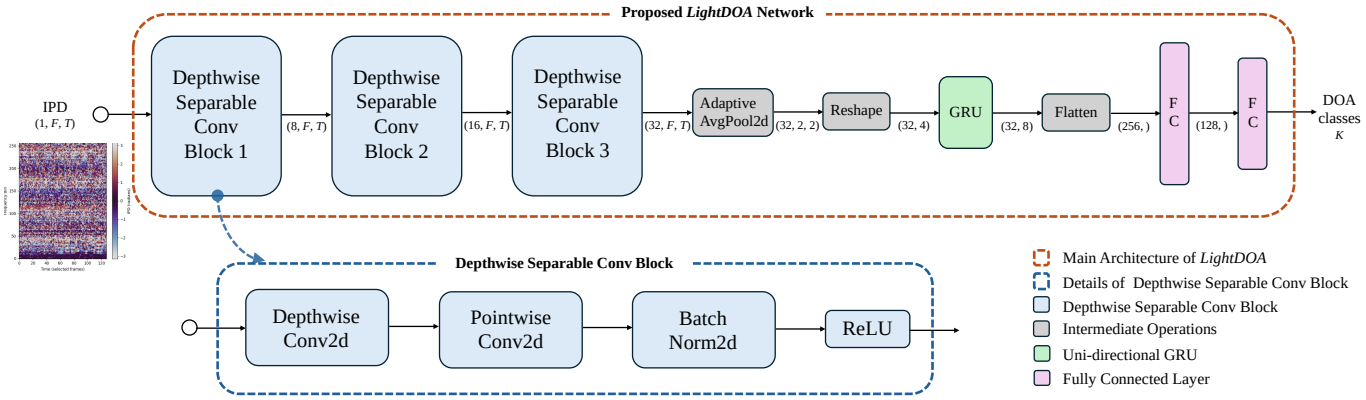


Fig. 3: Overview of the proposed *LightDOA* architecture. The input IPD feature is processed by a stack of depthwise separable convolutional blocks, followed by temporal modeling and classification. The region enclosed in the orange dashed line indicates the overall architecture of the proposed *LightDOA* network. The blue dashed box illustrates the internal structure of a single depthwise separable convolutional block.

ically, we first normalize all angles into the range $[0°, 360°)$ using:

$$\theta_{\text{norm}} = \theta - 360° \cdot \left\lfloor \frac{\theta}{360°} \right\rfloor, \tag{8}$$

where $\lfloor \cdot \rfloor$ denotes the floor operation.

Then, we fold the full-circle range into the front-facing half-circle $[0°, 180°]$ as:

$$\theta_{\text{mapped}} = \begin{cases} \theta_{\text{norm}}, & \text{if } \theta_{\text{norm}} \leq 180° \\ 360° - \theta_{\text{norm}}, & \text{if } \theta_{\text{norm}} > 180°. \end{cases} \tag{9}$$

This mapping ensures that front–back symmetric directions are assigned to the same target class, for example, $60°$ and $300°$, or $150°$ and $210°$. We formulate DOA estimation as a multi-class classification task, where the half-plane is discretized into K uniform angular bins. The number of classes K depends on the chosen angular resolution (e.g., $K = 37$ for $5°$ spacing, $K = 19$ for $10°$, etc.). The ground-truth label $y \in \{0, 1, \ldots, K-1\}$ is assigned based on the discretized azimuth.

The model is trained using the standard cross-entropy loss [29]:

$$\mathcal{L}_{\text{CE}} = -\sum_{k=0}^{K-1} \delta_{k,y} \log p_k, \tag{10}$$

where p_k is the predicted probability for class k, and y is the ground-truth class label. Note that the symmetric mapping is applied prior to training and evaluation, and is not explicitly incorporated into the loss function.

IV. EXPERIMENTS AND ANALYSES

A. Experimental Setup

We conduct all experiments using the BEWO dataset, which simulates diverse spatial audio environments with dual-channel input. Details of the dataset generation process and spatial coverage are described in Section II. Each sample contains a single dominant source with a ground-truth azimuth angle, mapped to the $[0°, 180°]$ range via symmetric folding.

Our experiments are conducted under four angular resolution settings, using $K \in \{9, 13, 19, 37\}$ classes corresponding to $20°$, $15°$, $10°$, and $5°$ spacing, respectively. The model is trained to minimize cross-entropy loss, and evaluated using classification accuracy in degrees.

We use the Adam optimizer [30] with a learning rate of 5×10^{-3}, a batch size of 256, and train for up to 150 epochs. Early stopping is applied based on the validation accuracy: training is terminated if no improvement is observed over 10 consecutive epochs. No data augmentation is used.

Model selection is based on the checkpoint that achieves the highest validation accuracy. All reported results are evaluated on the test set using this best-performing checkpoint. Each experiment is repeated with three random seeds to ensure result stability.

B. Baselines

To evaluate the effectiveness of our proposed *LightDOA* model, we compare it against three representative baseline methods, each reflecting a different design philosophy in neural-based DOA estimation:

- **CRNN** [13]: A popular end-to-end neural network for DOA estimation, originally designed to jointly estimate azimuth and elevation angles using acoustic intensity vectors as input [13]. Since our dataset provides only dual-channel recordings, we adapt the model to predict azimuth only and replace the original input with the input consists of the real and imaginary parts of the complex STFT, concatenated along the channel dimension, which called STFT (Re+Im) in table III. This baseline serves as a strong full-capacity reference using spectral features.
- **MTL-DOA** [16]: A multi-task learning framework that uses separate branches to process IPD and log-magnitude

TABLE III: Classification accuracy (%) and model parameters under different angular resolutions. Parameters vary across resolutions due to different numbers of output classes.

Model	**Feature**	**5°**		**10°**		**15°**		**20°**	
		Acc (%)	Param	Acc (%)	Param	Acc (%)	Param	Acc (%)	Param
CRNN [13]	STFT (Re+Im)	57.29	311k	71.08	309k	76.61	308k	84.26	308k
MTL-DOA [16]	IPD+logMag	56.68	285m	70.24	285m	78.15	285m	85.00	285m
CP-Mobile [21]	IPD	57.48	64.0k	71.03	62.1k	78.14	61.5k	84.40	61.1k
CP-Mobile [21]	STFT (Re+Im)	54.48	64.2k	69.05	62.3k	76.68	61.7k	85.00	61.4k
Proposed *LightDOA*	**IPD**	**57.96**	**39.0k**	**71.66**	**36.7k**	**77.98**	**35.9k**	**85.45**	**35.5k**

Note: The baseline models were re-implemented and slightly adjusted to fit the current DOA estimation task. In particular, their output layers were adapted to match the number of DOA classes under each angular resolution setting. Other parts of the architecture remained unchanged.

spectrograms (logMag), followed by a joint fusion module to estimate azimuth and elevation [16]. In our setting, we evaluate only the azimuth estimation performance, adapting the model to our dual-channel setup. This baseline reflects the benefit of multi-view spatial representation learning.

- **CP-Mobile** [21]: To assess performance–efficiency trade-offs, we adapt the compact CP-MobileNet [21], widely used in acoustic scene classification, for DOA classification. The model is tested with both IPD and STFT (Re+Im) inputs, and its output layer is modified to predict azimuth angle classes. This baseline provides insight into the potential of lightweight architectures for DOA tasks.

These baselines cover a diverse spectrum of design choices, from full-capacity CRNNs to efficient mobile models and multi-branch feature encoders. All models are trained and evaluated under the same experimental conditions as *LightDOA* to ensure fair comparison.

C. Results and Discussion

Table III presents the classification accuracy and corresponding parameter counts of various DOA estimation models under different angular resolutions (5°, 10°, 15°, and 20°). The performance is evaluated on the test set using the best validation checkpoint, as described in Section 4.1. Since the number of output classes varies with the angular resolution (e.g., 37 classes for 5°, 9 classes for 20°), the output layers of the models are adjusted accordingly. However, these changes have minimal impact on the overall parameter count, and the complexity remains relatively stable across different configurations, ensuring fair comparison among models.

Among all methods, our proposed *LightDOA* model consistently achieves the highest accuracy across all resolution settings, with performance gains especially notable at coarser resolutions (e.g., 85.45% at 20°). Despite its lightweight design with only (or less than) 39k parameters, it outperforms larger models such as CRNN (around 311k) [13] and the MTL-DOA model (around 285m) [16], indicating superior efficiency–accuracy trade-off.

In comparison, CRNN-based model[13] perform competitively but exhibit lower accuracy at finer resolutions, and their relatively large parameter size may limit deployment in real-time or embedded applications. CP-Mobile variants [21], adapted from acoustic scene classification tasks, also demonstrate reasonable accuracy with moderate complexity, but underperform compared to our proposed model, particularly when using STFT(Re+Im) input.

Notably, the performance drop observed in baseline methods relative to their original results is likely due to the greater diversity and inherent class imbalance in BEWO dataset as shown in Section II, which poses a more challenging evaluation setting.

These results demonstrate that our *LightDOA* architecture not only achieves state-of-the-art accuracy, but also sets a new standard for low-complexity DOA estimation. The results validate the effectiveness of IPD-based input and depthwise separable convolution for compact spatial audio modeling.

V. Conclusions

In this work, we revisited single-source DOA estimation using a recently introduced dual-channel dataset BEWO, where LLMs were used to assist spatial audio generation. To address the limitations of existing models in terms of generalizability and complexity, we proposed *LightDOA*, a lightweight neural network architecture based on depthwise separable convolutions and IPD features. Extensive experiments demonstrated that *LightDOA* achieves competitive or superior accuracy compared to existing CRNN-based model, multi-branch architectures and CP-Mobile variants, while maintaining significantly lower model complexity. The lightweight nature of our model makes it a promising candidate for real-time deployment on edge devices. Future work may explore extensions to elevation estimation, multi-source localization, or adaptation to real-world recordings beyond synthetic datasets.

References

[1] J. Cheer, V. Patel, and S. Fontana, "The application of a multi-reference control strategy to noise cancelling headphones," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. 3095–3103, 2019.

[2] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, IEEE, vol. 2, 2003, pp. 1228–1233.

[3] A. Xenaki, J. Bünsow Boldt, and M. Græsbøll Christensen, "Sound source localization and speech enhancement with sparse bayesian learning beamforming," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3912–3921, 2018.

[4] Z. Luo, D. Shi, and W.-S. Gan, "A hybrid sfanc-fxnlms algorithm for active noise control based on deep learning," *IEEE Signal Processing Letters*, vol. 29, pp. 1102–1106, 2022.

[5] Z. Luo, D. Shi, W.-S. Gan, and Q. Huang, "Delayless generative fixed-filter active noise control based on deep learning and bayesian filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1048–1060, 2024.

[6] D. Shi, W.-S. Gan, B. Lam, Z. Luo, and X. Shen, "Transferable latent of cnn-based selective fixed-filter active noise control," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2910–2921, 2023.

[7] Y. Kim and H. Ling, "Direction of arrival estimation of humans with a small sensor array using an artificial neural network," *Progress In Electromagnetics Research B*, vol. 27, pp. 127–149, 2011.

[8] H. Tsuzuki, M. Kugler, S. Kuroyanagi, and A. Iwata, "An approach for sound source localization by complex-valued neural network," *IEICE TRANSACTIONS on Information and Systems*, vol. 96, no. 10, pp. 2257–2265, 2013.

[9] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2013, pp. 2927–2932.

[10] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*, Audio Engineering Society, 2015.

[11] Y. Hao, A. Küçük, A. Ganguly, and I. M. Panahi, "Spectral flux-based convolutional neural network architecture for speech source localization and its real-time implementation," *IEEE Access*, vol. 8, pp. 197 047–197 058, 2020.

[12] G. Bologni, R. Heusdens, and J. Martinez, "Acoustic reflectors localization from stereo recordings using neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 1–5.

[13] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2018, pp. 241–245.

[14] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, "Event-independent network for polyphonic sound event localization and detection," *arXiv preprint arXiv:2010.00140*, 2020.

[15] H. Li, W. Zhang, and L. Zhang, "Doa estimation of room reflections using nn-based music algorithm," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2023, pp. 1960–1965.

[16] Y. Yang, J. Xi, W. Zhang, and L. Zhang, "Full-sphere binaural sound source localization using multi-task neural network," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2020, pp. 432–436.

[17] P. Sun, S. Cheng, X. Li, *et al.*, "Both ears wide open: Towards language-driven spatial audio generation," *arXiv preprint arXiv:2410.10676*, 2024.

[18] A. G. Howard, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[20] A. Howard, M. Sandler, G. Chu, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[21] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "Distilling the knowledge of transformers and cnns with cp-mobile," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 161–165.

[22] Z. Luo, H. Ma, D. Shi, and W.-S. Gan, "Gfanc-rl: Reinforcement learning-based generative fixed-filter active noise control," *Neural Networks*, p. 106 687, 2024.

[23] Z. Luo, D. Shi, J. Ji, X. Shen, and W.-S. Gan, "Real-time implementation and explainable ai analysis of delayless cnn-based selective fixed-filter active noise control," *Mechanical Systems and Signal Processing*, vol. 214, p. 111 364, 2024.

[24] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[25] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[26] J. Achiam, S. Adler, S. Agarwal, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[27] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 351–355.

[28] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[29] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[30] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.