Quantization-Error-Free Soft Label for 2D Sound Source Localization

Linfeng Feng[1,2,3], *Xiao-Lei Zhang*[1,2,3], *Xuelong Li*[1]

[1]Institute of Artificial Intelligence (TeleAI), China Telecom
[2]School of Marine Science and Technology, Northwestern Polytechnical University
[3]Research & Development Institute of Northwestern Polytechnical University in Shenzhen
fenglinfeng@mail.nwpu.edu.cn, xiaolei.zhang@nwpu.edu.cn, xuelong_li@ieee.org

2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 979-8-3315-1682-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISCSLP63861.2024.10800169

Abstract

One of the state-of-the-art direction of arrival (DOA) estimation techniques is formulated as a classification problem using deep learning. However, it inherently suffers from quantization errors during the classification formulation. This weakness is further amplified in two-dimensional (2D) sound source localization (SSL). To address this limitation in 2D SSL, this paper aims to develop a quantization-error-free training objective, named Unbiased Label Distribution (ULD), along with a corresponding decoding scheme for the predicted distribution. The key idea is to use multiple adjacent classes jointly to eliminate quantization error. Experimental results show that the proposed algorithm significantly breaks the quantization error limit when the classification model achieves high accuracy. It also demonstrates strong robustness in low signal-to-noise ratio, high reverberation, and far-field environments.

Index Terms: 2D sound source localization, soft label encoding, decoding, quantization-error-free

1. Introduction

Sound source localization (SSL) is a technique that uses multichannel signals captured by microphone arrays to infer the spatial coordinates of sound sources. This technology holds promise across diverse applications, serving as an auxiliary tool in contexts like human-robot interaction [1, 2] and speech separation [3, 4], and target speaker extraction [5].

In the past, the main focus was on traditional array signal processing methods [6–8]. In recent years, the infusion of deep neural networks (DNNs) into SSL has gained prominence owing to their nonlinear capabilities. Notably, DNNs employing classification as the output strategy exhibit enhanced resilience to interference signals [9].

Typically, compact microphone arrays concentrate solely on direction of arrival (DOA) estimation. Dissimilar from compact microphone arrays, distributed microphone arrays showcase the capacity to pinpoint Cartesian coordinates of sound sources. Facilitated by wireless interconnections, these arrays can significantly augment the processing efficiency for the sound emanating from remote sources [10]. Such configurations are commonly referred to as ad-hoc microphone arrays.

This paper mainly focuses on indoor localization. In a study conducted by [11], the experimental setup entails the alignment of room and microphone layouts during both training and testing phases. They introduced an end-to-end model and presented three distinct labeling strategies designed for two-dimensional (2D) SSL. Another approach is proposed by [12], advocating a stage-wise SSL methodology. In this framework, each node corresponds to a compact microphone array tasked with estimating the DOA. The intersections of these DOA estimates are subsequently clustered to ascertain the precise 2D positions of the sound sources. An alternative perspective is offered by [13], who introduce an end-to-end model distinct from that presented by [11]. Notably, in their methodology, random alterations are introduced to the room and microphone layout. Consequently, this mandates the provision of positional information for each microphone node to the deep neural network. Furthermore, [14] leverages the network presented in [13] as a foundational backbone for its research.

Both [11] and [13] share a common methodology of partitioning the room into multiple grids, as illustrated in Figure 1, while employing a classification model. The conventional and intuitive approach involves treating each grid as a distinct class and utilizing one-hot encoding for labels. In this paradigm, the grid housing the sound source is designated as 1, while others are set to 0. The decoding process entails identifying the center of the grid with the highest probability as the sound source location. However, this method conspicuously manifests significant quantization error. To reduce the quantization errors, [11] introduced a method called Refined Grid. This method employs a two-step localization process: the first step is to classify which grid the sound source is in, and the second step is to perform regression within the grid. Theoretically, this approach offers the potential of eliminating quantization errors entirely. However, its practical application requires equipping each grid with three output neurons, which represent the presence of a sound source and its x-axis and y-axis ratios. This approach significantly increases the complexity of the models. On the one hand, its implementation is complicated. On the other hand, in adverse environments, the convergence of training may be challenging.

[15] introduced the *Unbiased Label Distribution* (ULD), a quantization-error-free soft label tailored for azimuth DOA estimation with compact microphone arrays. Inspired by this work, we extend ULD for estimating 2D sound source coordinates with ad-hoc microphone arrays. The proposed soft labels serve as a plug-and-play alternative to conventional one-hot encoding. It maintains the simplicity of the one-hot codes where one output neuron of the classification-based deep model corresponds to a grid. Empirical results showcase that the proposed algorithm markedly transcends the limitations of quantization errors, especially in scenarios where the classification model achieves a high level of accuracy. Furthermore, the algorithm demonstrates robustness in challenging environments.

Xiao-Lei Zhang is the corresponding author.

This work was supported in part by the National Science Foundation of China (NSFC) under Grant 62176211, and in part by the Project of the Science, Technology, and Innovation Commission of Shenzhen Municipality, China under Grant GJHZ20240218114401004 and JSGG20210802152546026.

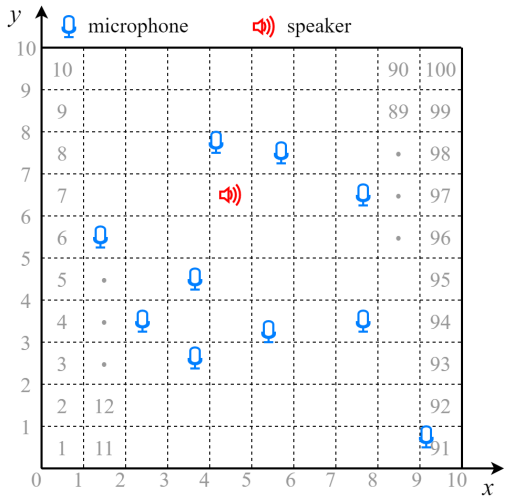

Figure 1: *The 2D plane of a room is divided.*

2. Method

2.1. Label encoding

Taking the x-axis as an example, the y-axis follows the same principles. Assuming the room has a length of L along the x-axis with a resolution of I, each segmented unit length is $l = L/I$. To account for the boundaries, the output space along the x-axis is discretized into $\{0, l, \ldots, (I-1)\cdot l, I\cdot l\}$.

Assuming the ground-truth position of a sound source is denoted as p^x, it can be classified into the χ-th class where $\chi = p^x/l$. It is emphasized that χ is a real number and not necessarily an integer. It can be encoded using a 1D-ULD vector denoted as $\boldsymbol{x} = \{x_i\}_{i=0}^{I}$. This vector can be formulated as follows:

$$x_i = \begin{cases} 1-\mathrm{deci}(\chi), & \text{if } i = \mathrm{int}(\chi) \\ \mathrm{deci}(\chi), & \text{if } i = \mathrm{int}(\chi)+1 \\ 0, & \text{otherwise} \end{cases}, \quad \forall i = 0, \ldots, I \tag{1}$$

where $\mathrm{deci}(\cdot)$ represents extracting the decimal portion, while $\mathrm{int}(\cdot)$ represents extracting the integer portion. Eq. (1) can be intuitively understood as that two adjacent integers are employed to approximate a real number that lies between them. This is in contrast to the One-hot that involves rounding a real number to the nearest integer. As illustrated in Figure 1, the position of the speaker along the x-axis can be represented through the approximation by the 4th and 5th classes.

The process described earlier can be interpreted as transforming SSL into a probability prediction problem, where the model's output signifies the probability distribution of the speaker's presence along different axes. It is crucial to underscore a fundamental aspect: the x-axis and y-axis are orthogonal. Consequently, from a probabilistic standpoint, the ULDs along these two axes represent independent 1D edge probability distributions. This independence allows us to easily derive the 2D-ULD we seek. Consequently, their multiplication results in the ground-truth 2D joint probability distribution. The subsequent content delineates the algorithmic flow.

Reshaping $\boldsymbol{y}$ as an $(I+1)\times 1$ column vector and $\boldsymbol{x}$ as a $1\times(I+1)$ row vector, we obtain:

$$\boldsymbol{y} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_I \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} x_0 & x_1 & \cdots & x_I \end{bmatrix} \tag{2}$$

The matrix multiplication of $\boldsymbol{y}$ and $\boldsymbol{x}$ results in a 2D-ULD matrix:

$$\boldsymbol{\rho} = \begin{bmatrix} y_0x_0 & y_0x_1 & \cdots & y_0x_I \\ y_1x_0 & y_1x_1 & \cdots & y_1x_I \\ \vdots & \vdots & \ddots & \vdots \\ y_Ix_0 & y_Ix_1 & \cdots & y_Ix_I \end{bmatrix} \tag{3}$$

Each element in $\boldsymbol{\rho}$ represents the probability of the corresponding grid having a sound source.

When implementing the aforementioned 2D-ULD, if the DNN output is a 1D vector (for instance, a fully connected output layer), the 2D-ULD can be reshaped into a 1D vector for supervising the training of the DNN.

2.2. Decoding

Upon completing the training of the DNN model using ground-truth label supervision, the decoding process for the DNN output $\hat{\boldsymbol{\rho}}$ also necessitates consideration of quantization errors. Firstly, the DNN output needs to be reshaped into a $(I+1)\times(I+1)$ matrix. This matrix serves as the predicted 2D joint distribution, denoted as $\hat{\boldsymbol{\rho}} \in \mathbb{R}^{(I+1)\times(I+1)}$. According to probability theory, it is straightforward to obtain the 1D marginal distribution in the remaining direction by summing over individual directions of the joint distribution. This process can be formulated as follows:

$$\hat{\boldsymbol{x}} = \sum_{i=0}^{I} \hat{\boldsymbol{\rho}}_{i:}, \quad \hat{\boldsymbol{y}} = \sum_{j=0}^{I} \hat{\boldsymbol{\rho}}_{:j} \tag{4}$$

where $\hat{\boldsymbol{\rho}}_{i:}$ represents the i-th row of $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\rho}}_{:j}$ represents the j-th column of $\hat{\boldsymbol{\rho}}$. After obtaining $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$, we can apply the processing steps from the following Eq. (5) to derive refined predicted source coordinates $(\hat{p}^x, \hat{p}^y)$.

If only the peak class is chosen for decoding the DNN outputs $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$, it is inevitable that quantization errors will still arise. Taking the x-axis as an example, the y-axis follows the same principles. Aligning with the aforementioned complementary concept, the proposed approach involves taking into account multiple adjacent classes and executing a weighted approximation to refine $\hat{p}^x$. Here, the peak class is denoted as $\hat{k} = \arg\max_i \{\hat{\boldsymbol{x}}\}_{i=0}^{I}$. Subsequently, the weighted adjacent decoding (WAD) can be formulated as follows:

$$\hat{p}^x = \frac{\sum_{i=\{\hat{k}-1,\hat{k},\hat{k}+1\}} \hat{x}_i \times i \times l}{\sum_{i=\{\hat{k}-1,\hat{k},\hat{k}+1\}} \hat{x}_i} \tag{5}$$

Specifically, if there is an out-of-bounds situation, i.e., when $i<0$ or $i>I$, we set $\hat{x}_i = 0$. By substituting the ground-truth distribution $\boldsymbol{x}$ from Eq. (1) as a predicted distribution $\hat{\boldsymbol{x}}$ into Eq. (5) for decoding, the process achieves a lossless reconstruction of p^x. Therefore, this algorithmic procedure of the output architecture is self-consistent.

3. Experiments

3.1. Datasets

We evaluate our method on both simulated and real-world datasets. We employed the Pyroomacoustics module [16] to

generate three sets of simulated datasets, progressively increasing in localization difficulty, denoted as L-1, Ad-1, and Ad-2, respectively. The source speech is sourced from the LibriSpeech corpus [17], and the source noise is extracted from an extensive noise set [18]. Specifically, the source speech for the training, validation, and test sets is derived from train-clean-360, dev-clean, and test-clean, respectively. The dimensions of the simulated rooms maintain a fixed height of 4.2m, the microphones are situated at a constant height of 0.9m, and the sources are positioned at a fixed height of 0.95m.

The configurations for the L-1 dataset followed the setting in [11], which is a simple single-source dataset. Specifically, all speech segments share the same room with no echo or noise. Two 4-channel linear arrays were placed along the left wall and the bottom wall of the room. The speaker was randomly positioned throughout the room. Each segment is 160ms long. The training, validation, and test sets of the dataset consist of 100,000, 5,000, and 5,000 segments, respectively.

The configurations for the Ad-1 and Ad-2 datasets were slightly modified from [13], representing the single-source and two-source scenarios, respectively. For each utterance in the dataset, we generated a room with random dimensions between $[4, 10]$m in length and width. The reverberation time T_{60} was randomly chosen in the range $[0.2, 1.2]$s. In the training set, the signal-to-noise ratio (SNR) was randomly chosen in the range $[0, 50]$dB, while in the validation and test sets, the SNR was chosen in the range $[10, 20]$dB. For each utterance, its room has 30 microphone nodes, with only one microphone per node. The room was divided into a 16×16 grid, where some grids were randomly selected with each grid used to place either a microphone or a speaker. Each utterance is 2s long. The training, validation, and test sets of the dataset consist of 18,000, 1,800, and 1,800 utterances, respectively.

We formulated two real-world datasets from Libri-adhoc40 [19], where most configurations are consistent with the Ad-1 dataset. These two datasets are referred to as Ad-r1 and Ad-r2, representing single-source and two-source datasets, respectively. The geometry of the real-world room has approximate dimensions of $9.8 \times 10.3 \times 4.2$m. The room's T_{60} is approximately 0.9s, with negligible presence of additive noise. For each utterance, 30 out of the 40 microphone nodes are randomly selected. The training, validation, and test sets contain 18,000, 1,800, and 2,468 utterances, respectively.

3.2. Comparison among different labels

We compared ULD with three different labels. It should be noted that the first three baseline labels use a number of classes $I \times I$, while ULD uses $(I+1) \times (I+1)$ to cover the boundaries.

- **One-hot** [13]: In conjunction with the utilization of Cross Entropy loss, the peak class is selected for decoding.
- **Heat map (HM)** [11]: In adherence to the loss function and decoding specified in [11], the standard deviation of the Gaussian distribution is likewise set to 0.1.
- **Refined grid (RG)** [11]: The replication of this follows the loss function and decoding set in [11].
- **Unbiased label distribution (ULD)**: In conjunction with the utilization of Cross Entropy loss, the WAD is selected for decoding.

3.3. Experimental settings

In our experiments, three neural networks served as the backbone networks. The model proposed by [11] was directly employed in experiments conducted on L-1. Meanwhile, the model proposed by [13] was applied directly to experiments on Ad-1. Notably, this network incorporates three fully-connected layers at the end. For experiments on Ad-2, the second layer was substituted with two parallel Bi-directional Long Short-Term Memory (BiLSTM) layers, featuring the same number of output neurons. This modification aims at learning masks for separating mixed speech features, aligning with the implicit source separation concept introduced in [20]. Given that the model entails multiple outputs, training this network necessitates the application of permutation invariant training (PIT) [21]. For experiments on real-world datasets, the models can be pre-trained on the simulated datasets and then fine-tuned on the real-world training sets.

A sampling rate of 16 kHz, a window length of 512 samples, a hop length of 256 samples, a Hanning window, and 512 FFT points were used for extracting Short-Term Fourier Transform (STFT) features. Notably, the real and imaginary parts of the STFT were concatenated and then input into the DNN.

For the Ad-hoc dataset, the frame-level replication of the one-hot coding of node positions from [13] was employed and fed into the DNN. What sets this study apart is that when using different label codings, the input and output maintained the same type of position coding.

Drawing inspiration from [22, 23], during training on the Ad-hoc dataset, only 15 nodes were randomly selected for each utterance in each epoch. This entails using fewer microphone nodes during training compared to testing, contributing to an enhanced generalization capability of the model.

The AdamW optimizer [24] was employed with a maximum of 50 training epochs. When conducting training on L-1, the batch size was set to 128, while for other experiments it was set to 32. For all experiments on simulated datasets, the learning rate was initialized at 10^{-3} and reduced to 10^{-4} if the validation loss did not decrease over 3 epochs. During the fine-tuning stage, the learning rate was initialized at 10^{-4}. Training was terminated early if the model's loss on the validation set did not reduce for 10 epochs. The model with the minimum localization error on the validation set was selected for test.

3.4. Evaluation metrics

Since all experiments are based on classification models, the most common and intuitive metric is classification accuracy (ACC), which can be described as follows:

$$\mathrm{ACC}(\%) = \frac{N^{\mathrm{acc}}}{N} \times 100 \tag{6}$$

where N is the number of test speakers, and N^{acc} represents the number of speakers whose predicted position are correctly classified into their corresponding ground-truth grid.

The most intuitive metric for SSL is the mean absolute error (MAE) of the straight-line distance between the predicted position and the ground-truth position, which can be described as follows:

$$\mathrm{MAE}(m) = \frac{1}{N} \sum_{n=1}^{N} \sqrt{(p_n^x - \hat{p}_n^x)^2 + (p_n^y - \hat{p}_n^y)^2} \tag{7}$$

where (p_n^x, p_n^y) represents the ground-truth coordinates of the n-th speaker position.

Table 1: *The experimental results on the L-1 dataset, with a uniform network output class count of* 6×6.

	QE	One-hot	HM	RG	ULD
ACC	100.00	96.50	96.74	92.00	96.72
MAE	0.381	0.384	0.453	0.104	**0.031**

Table 2: *The experimental results on the Ad-1 dataset, where "/" indicates cases where the model training did not converge or was not feasible.*

	I	1	2	4	8	16
QE	MAE	2.705	1.340	0.687	0.340	0.174
One-hot	ACC	/	90.77	81.66	63.20	38.08
	MAE	/	1.395	0.765	0.460	**0.361**
HM	ACC	/	84.32	74.21	41.75	/
	MAE	/	1.468	0.691	**0.652**	/
RG	ACC	/	97.28	91.72	78.60	39.69
	MAE	/	0.253	0.221	**0.206**	0.322
ULD	ACC	96.11	90.38	83.66	71.15	56.20
	MAE	0.292	0.270	0.217	0.202	**0.193**

In scenarios where a single utterance involves multiple speakers, there exist various permutations between the predicted positions and the ground-truth positions. We select the permutation that minimizes the total sum of MAE.

3.5. Main results

For all the results tables, QE refers to *Quantization Error*, representing the MAE achieved on the test set when the one-hot models attain 100% ACC.

In Table 1, the discernible trend is the exemplary performance of the classification model in pristine environments, yielding remarkably high ACC. Notably, in this context, the constraints on one-hot encoding primarily stem from quantization errors. However, both RG and ULD notably transcend these limitations. ULD emerges as the frontrunner, exhibiting the most impressive performance by diminishing MAE by a remarkable 91.93% in comparison to one-hot.

Table 2 illuminates a compelling relationship between resolution I, quantization error, and ACC. As resolution increases, QE diminishes, albeit accompanied by a rapid surge in the number of classes, precipitating a notable decline in ACC. Notably, as I attains a value of 32, models fail to converge. This impediment curtails the efficacy of one-hot, resulting in a MAE plateauing at a modest 0.361m. RG and ULD, functioning as representations of microphone spatial features, surpass one-hot in precision, achieving superior ACC for identical I values. RG, endowed with I^2 classes, slightly outpaces ULD at lower I values. However, the imposition of $3I^2$ output neurons for RG, when used as a label, yields a marked decrease in ACC as I escalates from 8 to 16. Both methodologies, nonetheless, notably surmount quantization error constraints, with ULD showcasing supremacy by reducing MAE by 46.54% relative to one-hot.

Table 3 elucidates the impact of implicit separation on model performance. One-hot closely approximates single-source scenarios, while the intricate implementation and augmented complexity of the loss function combined with PIT render RG significantly less accurate than one-hot. Despite RG's conceptual regression within correctly classified grids, its heavy reliance on high ACC results in suboptimal performance. Conversely, ULD, akin to a refined iteration of one-hot with minor modifications, inherits the ease of training advantages, leading to comparable performance in both single and multi-source scenarios. Across diverse resolutions, ULD consistently achieves smaller MAEs relative to one-hot, maintaining a substantial advantage in mitigating quantization errors and reducing MAE by 33.56% compared to one-hot.

Table 3: *The experimental results on the Ad-2 dataset, where "/" indicates cases where the model training did not converge or was not feasible.*

	I	1	2	4	8	16
QE	MAE	2.695	1.346	0.664	0.339	0.170
One-hot	ACC	/	91.14	78.81	61.39	31.98
	MAE	/	1.401	0.775	0.507	**0.435**
HM	ACC	/	78.61	64.94	40.25	/
	MAE	/	1.713	**1.097**	1.222	/
RG	ACC	/	70.36	55.64	42.47	17.05
	MAE	/	**1.499**	1.725	1.862	1.948
ULD	ACC	88.75	77.44	66.75	57.39	41.83
	MAE	1.018	0.650	0.462	0.331	**0.289**

Table 4: *The experimental results on the real-world datasets, where "/" indicates cases where the model training did not converge or was not feasible.*

		I	1	2	4	8	16
Dataset	QE	MAE	4.312	1.796	0.720	0.651	0.289
Ad-r1	One-hot	ACC	/	90.97	89.99	80.80	75.44
		MAE	/	1.908	0.755	0.687	**0.351**
	ULD	ACC	99.92	86.33	58.27	81.13	74.63
		MAE	0.445	0.305	0.234	**0.180**	0.207
Ad-r2	One-hot	ACC	/	88.55	71.89	62.99	47.00
		MAE	/	1.951	1.078	0.946	**0.729**
	ULD	ACC	91.37	72.51	18.29	59.54	57.05
		MAE	1.354	1.199	0.881	0.540	**0.385**

Table 4 presents the results on real-world datasets. As can be seen, the experimental results are consistent with those on the simulated datasets. ULD significantly overcomes the quantization error limitations at high ACC. In single-source and two-source scenarios, it reduces the relative error by 48.72% and 47.19%, respectively, compared to one-hot paradigm.

4. Conclusion

In this paper, we introduce a novel labeling algorithm tailored for indoor 2D sound source localization. This algorithm not only surpasses the limitations of one-hot encoding but also retains its inherent ease of implementation. Starting with a one-dimensional unbiased label distribution, we systematically expand it into a two-dimensional joint distribution to serve as the ground-truth label for supervising DNN training. This approach involves decomposing the two-dimensional predicted distribution into two distinct one-dimensional probability distributions. This strategy enables probabilistic weighted adjacent decoding, thereby achieving a remarkably precise determination of the sound source position. Empirical findings from our experiments substantiate the efficacy of this algorithm, demonstrating a notable improvement in performance without increasing model complexity.

5. References

[1] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system'hark'—open source software for listening to three simultaneous speakers," *Advanced Robotics*, vol. 24, no. 5-6, pp. 739–761, 2010.

[2] B. Alenljung, J. Lindblom, R. Andreasson, and T. Ziemke, "User experience in social human-robot interaction," in *Rapid automation: Concepts, methodologies, tools, and applications*. IGI Global, 2019, pp. 1468–1490.

[3] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 27, no. 2, pp. 457–468, 2018.

[4] ——, "Localization based sequential grouping for continuous speech separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 281–285.

[5] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "L-spex: Localized target speaker extraction," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7287–7291.

[6] 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, 1976.

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

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[9] S. Chakrabarty and E. A. Habets, "Multi-speaker doa estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[10] X.-L. Zhang, "Deep ad-hoc beamforming," *Computer Speech & Language*, vol. 68, p. 101201, 2021.

[11] G. Le Moing, P. Vinayavekhin, T. Inoue, J. Vongkulbhisal, A. Munawar, R. Tachibana, and D. J. Agravante, "Learning multiple sound source 2d localization," in *21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–6.

[12] S. Liu, Y. Gong, and X.-L. Zhang, "Deep learning based two-dimensional speaker localization with large ad-hoc microphone arrays," *arXiv preprint arXiv:2210.10265*, 2022.

[13] Y. Gong, S. Liu, and X.-L. Zhang, "End-to-end two-dimensional sound source localization with ad-hoc microphone arrays," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1944–1949.

[14] L. Feng, Y. Gong, and X.-L. Zhang, "Soft label coding for end-to-end sound source localization with ad-hoc microphone arrays," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] L. Feng, X.-L. Zhang, and X. Li, "Eliminating quantization errors in classification-based sound source localization," *Available at SSRN 4715294*, 2024.

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

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[18] X. Tan and X.-L. Zhang, "Speech enhancement aided end-to-end multi-task learning for voice activity detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6823–6827.

[19] S. Guan, S. Liu, J. Chen, W. Zhu, S. Li, X. Tan, Z. Yang, M. Xu, Y. Chen, C. Liang *et al.*, "Libri-adhoc40: A dataset collected from synchronized ad-hoc microphone arrays," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 1116–1120.

[20] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Computer Speech & Language*, vol. 75, p. 101360, 2022.

[21] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] C. Liang, Y. Chen, J. Yao, and X.-L. Zhang, "Multi-Channel Far-Field Speaker Verification with Large-Scale Ad-hoc Microphone Arrays," in *Proc. Interspeech*, 2022, pp. 3679–3683.

[24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.