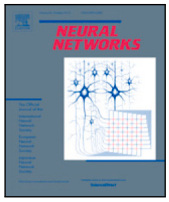
Neural Networks 181 (2025) 106679

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

Eliminating quantization errors in classification-based sound source localization

Linfeng Feng [a,b,c], Xiao-Lei Zhang [a,b,c,*], Xuelong Li [b]

[a] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[b] Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd, Beijing 100033, China
[c] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Guangdong 518063, China

ARTICLE INFO

Dataset link: https://github.com/linfeng-feng/ULD

Keywords:
Sound source localization
Quantization error
Label distribution
Decoding
Loss function

ABSTRACT

Sound Source Localization (SSL) involves estimating the Direction of Arrival (DOA) of sound sources. Since the DOA estimation output space is continuous, regression might be more suitable for DOA, offering higher precision. However, in practice, classification often outperforms regression, exhibiting greater robustness. Conversely, classification's drawback is inherent quantization error. Within the classification paradigm, the DOA output space is discretized into several intervals, each treated as a class. These classes show strong inter-class correlations, being inherently ordered, with higher similarity as intervals grow closer. Nevertheless, this characteristic has not been fully exploited. To address this, we propose Unbiased Label Distribution (ULD) to eliminate quantization error in training targets. Furthermore, we introduce Weighted Adjacent Decoding (WAD) to overcome quantization error during the decoding stage. Finally, we tailor two loss functions for the soft labels: Negative Log Absolute Error (NLAE) and Mean Squared Error without activation (MSE(wo)). Experimental results show our approach surpasses classification quantization limits, achieving state-of-the-art performance. Our code and supplementary material are available at https://github.com/linfeng-feng/ULD.

1. Introduction

Sound Source Localization (SSL) encompasses the task of determining the spatial coordinates of sound sources. Typically, this task is simplified to estimating the Direction of Arrival (DOA) of sound sources relative to microphones (Grumiaux, Kitić, Girin, & Guérin, 2022). The obtained DOA information can enhance the performance of various downstream applications. One common example is Sound Event Localization and Detection (SELD) (Bai et al., 2023; Shimada, Koyama, Takahashi, Takahashi, & Mitsufuji, 2021; Shimada et al., 2022). Accurate DOA estimates facilitate effective multichannel speaker separation (Wang & Wang, 2022) and can serve as a criterion for ordering labels of multiple speakers during training (Taherian, Tan, & Wang, 2022). It can aid speech recognition to reduce word error rates. Subramanian et al. (2021), Subramanian, Weng, Watanabe, Yu, and Yu (2022). In complex acoustic environments, speaker diarization systems leveraging DOA information have exhibited substantial improvements (Gburrek, Schmalenstroeer, & Haeb-Umbach, 2023; Taherian & Wang, 2023; Zheng et al., 2021).

1.1. Motivation and challenges

Over the past few decades, most researchers mainly focused on developing SSL algorithms based on traditional array signal processing techniques (DiBiase, 2000; Knapp & Carter, 1976; Schmidt, 1986). In recent years, SSL research has shifted towards deep learning methods, where Deep Neural Networks (DNNs) have demonstrated considerable promise and robustness in challenging acoustic environments, e.g. ambient noise, high reverberation, and multiple speakers (Grumiaux et al., 2022). Based on the training objectives of DNN, deep learning-based DOA estimation can be typically categorized into two categories: regression and classification.

One of the early regression-based methods (Vesperini, Vecchiotti, Principi, Squartini, & Piazza, 2016) designs an output layer with two neurons, which are used to estimate the x and y coordinates of a sound source in a Cartesian coordinate system. The works (Adavanne, Politis, Nikunen, & Virtanen, 2018; Vera-Diaz, Pizarro, & Macias-Guarasa, 2018) add a neuron to estimate the z coordinate. Vecchiotti, Pepe, Principi, and Squartini (2019) employs a similar output structure, but

* Corresponding author at: School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China.
E-mail addresses: fenglinfeng@mail.nwpu.edu.cn (L. Feng), xiaolei.zhang@nwpu.edu.cn (X.-L. Zhang), xuelong_li@ieee.org (X. Li).

https://doi.org/10.1016/j.neunet.2024.106679
Received 28 January 2024; Received in revised form 2 July 2024; Accepted 29 August 2024
Available online 31 August 2024

with an additional neuron for voice activity detection. Shimada et al. (2021, 2022) introduce the concept of Activity-coupled Cartesian DOA (ACCDOA), combining DOA with sound activity to form labels for the trajectory regression task in SELD. In addition, there are also regression methods that estimate the DOA of a source in the spherical coordinate system (Diaz-Guerra, Miguel, & Beltran, 2020, 2022). The soft-argmax regression-based method (Diaz-Guerra et al., 2022) can be considered a specific type of classification-based approach.

One of the early classification-based methods (Xiao et al., 2015) is implemented through a fully connected neural network. A common approach is to set the class corresponding to an active source to 1 and the rest to 0 (Grumiaux et al., 2022). Due to this property, classification-based methods have a natural advantage for multi-source localization, which can simply set the classes corresponding to active sources to 1. At the beginning, some work assumes that the number of sound sources are known. For example, Chakrabarty and Habets (2019) used phase spectra as the input of convolutional neural networks and selected the top classes with the highest probabilities from the network output as the predicted locations of multi sources. Subramanian et al. (2022) conducted experiments only considering scenarios with two speakers. It separates the problem of double-speaker localization into two independent single-speaker localization problems. Later on, some work shifts into the scenario where the source count is unknown. For example, He, Motlicek, and Odobez (2019) compares the predicted distributions produced by a DNN with a threshold, where classes exceeding the threshold are considered to have source activity. Nguyen, Gan, Ranjan, and Jones (2020) constructs a DNN with two output branches, one for outputting the locations of sound sources, and the other for outputting the number of sources. In Fu et al. (2022), an iterative SSL method was proposed, which extracts the DOA of each sound source iteratively from predicted distributions without using a threshold.

The experiments conducted by Tang, Kanu, Hogan, and Manocha (2019) demonstrate that the regression-based methods are inferior to the classification-based methods in the spherical coordinate system, while the regression-based methods outperform the classification-based methods in the Cartesian coordinate system. This finding is consistent with Perotin, Défossez, Vincent, Serizel, and Guérin (2019), which emphasizes the precision of regression-based approaches and the robustness of classification-based approaches in the Cartesian coordinate system. Feng, Gong, and Zhang (2023) pointed out that the localization error of classification-based methods can be decomposed into quantization error and learning error, where the quantization error refers to the localization error when the one-hot-encoding based classification reaches an accuracy of 100%. The reason why the classification-based models in Feng et al. (2023), Perotin et al. (2019), Tang et al. (2019) do not perform well is due to the large quantization errors.

The quantization errors in one-hot labels not only directly impact localization accuracy but also introduce non-smoothness in the labels. Here are some examples to illustrate. Given a classification resolution of 5 degrees. Two samples have ground-truth DOA values of 87.6 and 92.4, resulting in identical one-hot labels representing 90. Since their DOA difference is 4.8, this leads to low intra-class similarity. Conversely, if the ground-truth DOA values are 92.4 and 92.6, their respective labels represent 90 and 95, with a DOA difference of 0.2, indicating high inter-class similarity. To this end, Gaussian Label Coding (GLC) (He, Motlicek, & Odobez, 2018) and Soft Label Distribution (SLD) (Subramanian et al., 2022) were proposed as soft labels alternative to one-hot encoding. Both assign non-zero values to multiple classes near the ground-truth, smoothing labels. A GLC vector exhibits highly smooth intra- and inter-class transitions without the constraint of its elements sum to 1 for one speaker. Thus, this flexibility prohibits the use of softmax activation in the output layer. In contrast, a SLD has elements that sum to 1. Their advantage over one-hot encoding lies in the smoothness. Note that these methods do not completely resolve the issue of label quantization errors during the training phase. Furthermore, during the decoding phase, they still select the only peak class from the output vector, reintroducing quantization errors.

1.2. Goals and contributions

Based on the aforementioned analysis, we propose a novel output architecture designed for classification. This architecture not only retains robustness of classification but also incorporates the high precision of regression. Across experiments in diverse environments, from ideal to challenging, we substantiate our proposed architecture's effectiveness. The contributions can be summarized as follows:

- **We introduce Unbiased Label Distribution (ULD) to eliminate quantization errors in the labels.** ULD's one-to-one encoding permits unbiased inverse mapping to ground truth position. Notably, ULD exhibits smooth transitions within and between classes. Furthermore, ULD retains advantages of one-hot encoding for classification.
- **We propose Weighted Adjacent Decoding (WAD) to address shortcomings of sole reliance on peak probability.** Selecting the class with peak probability as the estimated DOA, denoted as *Top-1 decoding*, suffers from quantization errors. We incorporate sidelobes of the peak class into decoding design, yielding WAD. This overcomes the quantization error limit of Top-1 decoding.
- **We utilize two loss functions for soft labels: Negative Log Absolute Error (NLAE) and Mean Squared Error without activation (MSE(wo)).** Cross Entropy (CE) loss may be suboptimal for soft labels because its optimization objective does not directly point to the soft label itself. We analyzed compatibility between cross-entropy-like loss functions and classification models, and advantages of MSE loss for soft labels. After analysis, our strategy is to combine these loss types for the soft label family.

The remainder of this paper is organized as follows. Section 2 outlines the classification paradigm to introduce the issues. In Sections 3 to 4, we provide a detailed description of our contributions. Sections 6 and 7 demonstrate the effectiveness of our proposed method through experimental results. Finally, Section 8 presents the conclusions of our study.

2. Supervised sound source localization

In this paper, we focus solely on azimuth DOA estimation. We begin with the single-source localization problem. The DOA is measured in degrees. Assuming the maximum output range of DOA is denoted as r, if microphones are collinear, then r is 180; if microphones are coplanar but not collinear, then r is 360. The classification model discretizes the output space of DOA into several cells, with the standard cell length denoted as l. We set $I = r/l$ with $I \in \mathbb{N}$, yielding the set of class values $\{0, 1, \dots, I-1, I\}$, so the output space of DOA is discretized into $\{0, l, \dots, (I-1)\cdot l, I \cdot l\}$. We use the $I+1$ classes to cover boundaries, and an explanation will be presented in Section 3.2.

Without loss of generality, we assume that u denotes an utterance. A DNN-based SSL model $f(\cdot)$ can be formulated as:

$$\begin{aligned} \boldsymbol{\kappa} &= f(u;\boldsymbol{\theta}) \\ \hat{\boldsymbol{y}} &= \sigma(\boldsymbol{\kappa}) \end{aligned} \tag{1}$$

where $\boldsymbol{\theta}$ denotes learnable parameters, and the operation $\sigma(\cdot)$ maps $\boldsymbol{\kappa} \in \mathbb{R}^{I+1}$ to a predicted distribution $\hat{\boldsymbol{y}} \in [0,1]^{I+1}$.

We assume that $\boldsymbol{y} \in [0,1]^{I+1}$ represents the label distribution of the sound source's ground-truth position $p \in [0, r]$, and the process of obtaining $\boldsymbol{y}$ can be formalized as follows:

$$\boldsymbol{y} = \text{Encoding}(p) \tag{2}$$

where $\text{Encoding}(\cdot)$ represents the mapping from p to $\boldsymbol{y}$. In general, if $\sum_{i=0}^{I} y_i = 1$, then softmax activation is suitable as a candidate of $\sigma(\cdot)$ in Eq. (1), otherwise sigmoid activation is more appropriate.

The training objective of a DNN is to find its optimal learnable parameters $\boldsymbol{\theta}$ that minimize the loss function $\mathcal{L}$, while producing an

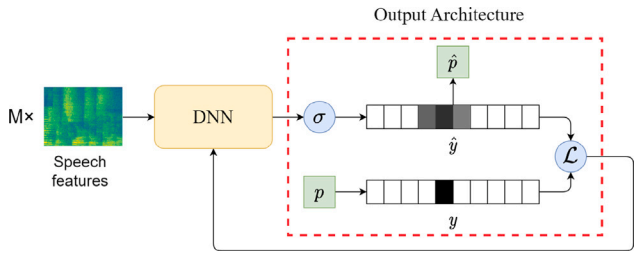


Fig. 1. Workflow of a supervised sound source localization, where M is the number of microphones. The output architecture is highlighted in the red box.

output $\hat{\boldsymbol{y}}$ that is as close as possible to the ground-truth label distribution $\boldsymbol{y}$. This can be achieved through supervised training, which can be formulated as:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}}; \boldsymbol{\theta}) \tag{3}$$

After obtaining $\hat{\boldsymbol{y}}$ in the test stage, it becomes feasible to map $\hat{\boldsymbol{y}}$ to a corresponding predicted position $\hat{p} \in [0, r]$, which can be referred to as a decoding process:

$$\hat{p} = \text{Decoding}(\hat{\boldsymbol{y}}) \tag{4}$$

where $\text{Decoding}(\cdot)$ represents the mapping from $\hat{\boldsymbol{y}}$ to $\hat{p}$.

The workflow for supervised single-source localization can be represented by Fig. 1, where the red box represents the output architecture that is our main focus. The ultimate goal of our design is to improve the predictive accuracy of $\hat{p}$.

Concerning the challenge of localizing multiple sound sources, we suggest utilizing the source splitting mechanism proposed in Subramanian et al. (2022) to decompose the problem into multiple single-source localization tasks.

3. Unbiased label distribution

This section primarily concentrates on the label encoding presented in Eq. (2), where we discuss the label distribution of a single source in a simple yet general manner.

3.1. Analysis

First, note that the true position of a sound source, p, is a real number, where $p \in [0, r]$. Given that $I = r/l$, we define a scaled variable $\gamma = p/l$, where $\gamma \in [0, I]$. Typically, a common classification method is to apply an operation, $\text{round}(\cdot)$, to assign γ to its nearest integer and then encode it as a one-hot label distribution. From the perspective of probability, it can be interpreted as that, the probability of the sound source located in the $\text{round}(\gamma)$-th class is 1, while the probabilities in other classes are all 0. Formally, the one-hot label distribution is $\boldsymbol{y}^{\text{1-hot}} = \{y_i^{\text{1-hot}}\}_{i=0}^{I}$, with the code for the ith class $y_i^{\text{1-hot}}$ defined as:

$$y_i^{\text{1-hot}} = \begin{cases} 1, & \text{if } i = \text{round}(\gamma) \\ 0, & \text{otherwise} \end{cases}, \quad \forall i = 0, \dots, I \tag{5}$$

From the above description, we can infer that the reason why the encoding from p to the one-hot distribution is not one-to-one mapping lies in the operation $\text{round}(\cdot)$. In other words, it is inevitable to have quantization errors when using a single integer, $\text{round}(\gamma) \in \mathbb{N}$, to represent a real number p.

Theorem 1. *The operation of* $\text{round}(\cdot)$ *results in an absolute quantization error whose mathematical expectation is* $l/4$.

Proof. See Appendix A for the proof. □

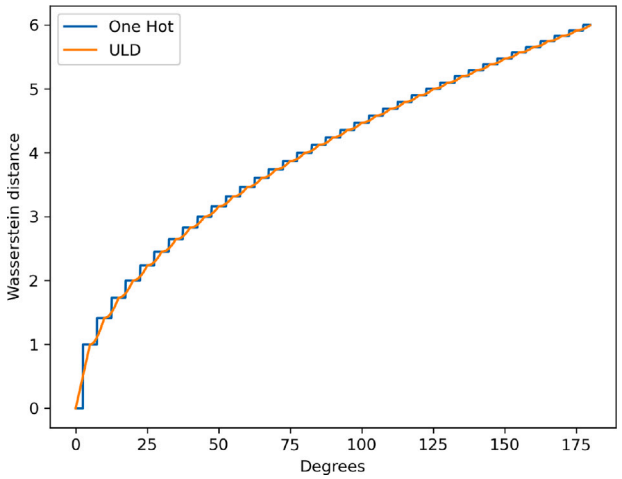


Fig. 2. Wasserstein distance between a distribution of any angle between $[0, 180]$ degrees and the distribution of 0 degree, where the label distribution is one-hot or ULD.

However, as $\sum_{i=0}^{I} y_i^{\text{1-hot}} = 1$, the one-hot distribution is consistent with probability theory interpretation, suitable for supervised models improving classification accuracy. Hence, we fine-tune this distribution to address strengths and weaknesses.

3.2. Definition

Theorem 2. *Let* γ *be a non-negative real number, with* $\text{int}(\gamma)$ *denoting its integer part, and* $\text{deci}(\gamma)$ *denoting its decimal part. Then, for any* γ *between two adjacent integers,* $\text{int}(\gamma)$ *and* $\text{int}(\gamma) + 1$*, an unbiased approximation is given by:*

$$\gamma = (1 - \text{deci}(\gamma)) \times \text{int}(\gamma) + \text{deci}(\gamma) \times (\text{int}(\gamma) + 1)$$

Proof. See Appendix B for the proof. □

Based on Theorem 2, we can easily derive our novel *unbiased label distribution* $\boldsymbol{y}^{\text{u}} = \{y_i^{\text{u}}\}_{i=0}^{I}$ as follows:

$$y_i^{\text{u}} = \begin{cases} 1 - \text{deci}(\gamma), & \text{if } i = \text{int}(\gamma) \\ \text{deci}(\gamma), & \text{if } i = \text{int}(\gamma) + 1 \\ 0, & \text{otherwise} \end{cases}, \quad \forall i = 0, \dots, I \tag{6}$$

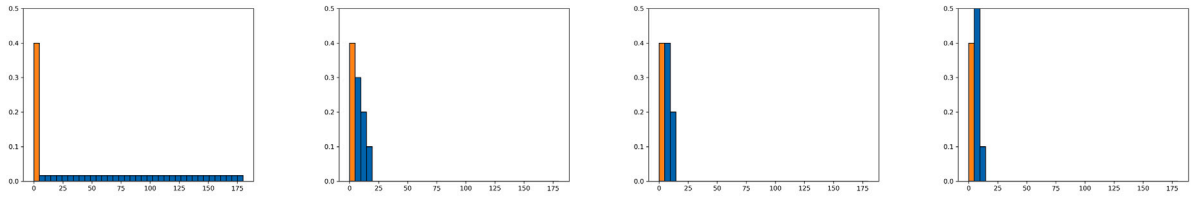
Theorem 2 justifies our use of $I + 1$ classes. Specifically, we always require two adjacent integers. Thus, we use the 0-th class and the Ith class, which represent the boundaries of the output space, to consider boundary cases. Clearly, the sum of the elements of a ULD vector, $\sum_{i=0}^{I} y_i^{\text{u}} = 1$, has a probabilistic interpretation, indicating the probability of a sound source appearing in the respective class. The main advantage of ULD lies that:

Theorem 3. *ULD is free of quantization errors.*

Proof. It is clear that the encoding mapping from p to $\boldsymbol{y}^{\text{u}}$ is a one-to-one mapping, such that when $p_1 \neq p_2$, we have $\boldsymbol{y}_1^{\text{u}} \neq \boldsymbol{y}_2^{\text{u}}$, enabling $\boldsymbol{y}^{\text{u}}$ to be accurately inverse-mapped to p. □

3.3. Connection between ULD and one-hot distribution

When $\text{deci}(\gamma) < 0.5$, $\text{int}(\gamma) = \text{round}(\gamma)$, otherwise $\text{int}(\gamma) + 1 = \text{round}(\gamma)$. Therefore, ULD can be regarded as a smoothed one-hot distribution, which can inherit the advantages of the one-hot distribution, while avoiding the problem of disproportionate distribution similarity to the DOA distance.



(a) Distribution A (b) Distribution B (c) Distribution C (d) Distribution D

Fig. 3. An example on the advantage of BCE over CE for SSL. Consider a sound source with r is 180, l is 5 and p is 0. In conventional classification problems, it is easy to have a predicted distribution like distribution A. However, for SSL, it is easy to have predicted distributions like the latter three. Distribution B has unwanted sidelobes, while distributions C and D even have pseudo peaks. The occurrence of pseudo peaks means that classification errors have already occurred. These four distributions have equal probability values in the ground-truth class, so the CE loss of these distributions is equal. However, the BCE losses of these four are 1.52, 1.60, 1.65 and 1.71 respectively, meaning that BCE gives greater penalty to these negative factors that are easily encountered in SSL. The theoretical analysis of the results above is provided in the supplementary material.

We use Wasserstein distance (WD) (Rubner, Tomasi, & Guibas, 2000) to analyze the connection between ULD and one-hot distribution. WD is a metric that can be used to measure the distance between two discrete probability distributions. Fig. 2 shows the WD between the any-angle label distribution and 0-angle distribution of either ULD or one-hot. We see that the WD curves of the ULD and one-hot are closely related. However, unlike the sudden changes in the WD curve of the one-hot distribution, the WD curve of ULD is smooth.

4. Weighted adjacent decoding

This section primarily concentrates on the decoding presented in Eq. (4). After training a DNN, we can obtain a predicted distribution $\hat{\boldsymbol{y}}$. This section describes our new approach for decoding $\hat{\boldsymbol{y}}$ into a DOA estimation $\hat{p}$.

4.1. Analysis

Naturally, we first extract the peak class $\hat{k}$ corresponding to the peak probability in $\hat{\boldsymbol{y}}$, which can be represented as:

$$\hat{k} = \arg\max_i \{\hat{y}_i\}_{i=0}^{I} \tag{7}$$

Then, the source location can be obtained from $\hat{k}$ as:

$$\hat{p} = \hat{k} \cdot l \tag{8}$$

which we refer to as the Top-1 Decoding. However, as discussed in Section 3.1, this formulation inevitably introduces quantization error. Even if a DNN achieves 100% classification accuracy, the mathematical expectation of the absolute quantization error using this decoding is $l/4$.

4.2. Definition

For the predicted distributions produced from DNN, the classes adjacent to the peak often have non-negligible probabilities due to strong correlation and ordering in the discretized DOA output space. Leveraging this, we propose Weighted Adjacent Decoding (WAD). The classes adjacent to $\hat{k}$ are $\hat{k}-1$ on the left and $\hat{k}+1$ on the right, respectively. Specifically, in cases where the index is out of bounds, such as $i < 0$ or $i > I$, we assign $\hat{y}_i = 0$. This can be interpreted as setting the probability of the source being outside the output space to 0.

Initially, we consider the scenario where only the peak class $\hat{k}$ and the adjacent class $\hat{k}_h$ with relatively high probability are used. Specifically, $\hat{k}_h$ is defined as $\arg\max_i \{\hat{y}_i\}_{i=\{\hat{k}-1,\hat{k}+1\}}$. Since two classes are involved, we refer to this as WAD-2. Hence, WAD-2 is formalized as:

$$\hat{p} = \frac{\sum_{i=\{\hat{k},\hat{k}_h\}} \hat{y}_i \times i \times l}{\sum_{i=\{\hat{k},\hat{k}_h\}} \hat{y}_i} \tag{9}$$

Similarly, WAD-3 can be formalized as follows:

$$\hat{p} = \frac{\sum_{i=\{\hat{k}-1,\hat{k},\hat{k}+1\}} \hat{y}_i \times i \times l}{\sum_{i=\{\hat{k}-1,\hat{k},\hat{k}+1\}} \hat{y}_i} \tag{10}$$

Substituting the ULD from Eq. (6) into Eq. (8), the quantization error remains the same as one-hot. However, substituting the ULD into Eq. (9) or Eq. (10) yields zero quantization error. Hence, this integrated output architecture is self-consistent.

Here is the generalization of WAD. We denote I_s as the number of classes selected for decoding. To simplify, we restrict I_s to be an odd number, which ensures that $\hat{k}$ is at the center of the I_s classes. Then, we can generalize to WAD-I_s with $i = \{\hat{k}-\frac{I_s-1}{2}, \dots, \hat{k}, \dots \hat{k}+\frac{I_s-1}{2}\}$. However, the actual classes are $\{0, 1, \dots, I-1, I\}$, so any out-of-bounds class will have a probability value of zero. We can thus compute the upper limit of I_s, where WAD can use all classes. Considering the extreme case where $\hat{k} = 0$, we have $\hat{k} + \frac{I_s-1}{2} = I$, thus the upper limit of I_s is $2I+1$.

4.3. Connection between WAD, Top-1, and Soft-argmax

From the above formulation, we see that Top-1 decoding, defined in Eq. (8), is a special case of WAD with $i = \{\hat{k}\}$. Essentially, WAD extends Top-1 decoding by using a weighted combination of multiple classes to reduce quantization error.

Soft-argmax is similar to WAD with all classes, eliminating the need for label encoding; instead, it directly uses the sound source location as the training target.

5. Loss functions

This section is to design a reasonable loss function $\mathcal{L}$ in Eq. (3) to train the DNN for yielding a predicted distribution closely matching the label distribution.

5.1. Analysis

We think that a reasonable $\mathcal{L}$ should have two key properties: (i) the direction of backpropagated gradient should always be correct, and (ii) the magnitude of backpropagated gradient should be proportional to the deviation from the training target to the DNN's output $\boldsymbol{\kappa}$. However, none of the common loss functions, including CE, BCE, and MSE, satisfy both properties for soft labels, as will be analyzed in Section 5.1. This analysis motivates deriving the NLAE loss and MSE (wo) loss in Section 5.2, satisfying both properties simultaneously.

5.1.1. CE loss function

The CE loss function is commonly used in conventional classification problems:

$$\mathcal{L}^{\mathrm{CE}} = -\sum_{i=0}^{I} y_i \log \hat{y}_i \tag{11}$$

where CE does not directly impose penalties on the classes with zero values in ground-truth label $\boldsymbol{y}$.

Table 1
Specifications of the simulated data.

Dataset		C1	C2	A1	L1	L2
Shape of array (m)		Circular, radius = 0.05			Linear, aperture = 0.08	
Self-rotation angle of array (degree)		0	0	0	[0, 180]	[0, 180]
Distance from speaker to array (m)		1.5	1.5	1.5	[0, 14.1]	[0, 14.1]
Minimum distance from speaker to wall (m)		0.5	0.5	0.5	0.0	0.0
Number of sound sources		1	2	1	1	2
Reverberation (s)	train	[0.2, 0.7]	[0.2, 0.7]	anechoic	[0.2, 1.2]	[0.2, 1.2]
	validation	[0.2, 0.7]	[0.2, 0.7]	anechoic	[0.2, 1.2]	[0.2, 1.2]
	test	[0.2, 0.8]	[0.2, 0.8]	anechoic	[0.2, 1.2]	[0.2, 1.2]
Segments	train	36 000	36 000	18 000	36 000	72 000
	validation	3600	3600	1800	3600	7200
	test	4320	4320	1800	3600	7200

5.1.2. BCE loss function

In contrast to CE, the BCE loss function imposes a penalty for each class:

$$\mathcal{L}^{\mathrm{BCE}} = -\sum_{i=0}^{I} y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i) \tag{12}$$

Generally, BCE is paired with sigmoid activation for multi-label classification, which treats each class as an independent binary classification problem, such as in object detection (Wu, Xu, Yang, & Li, 2024) and image segmentation (Jing et al., 2024). However, in this paper, BCE is not restricted to pairing with sigmoid. For one-hot and ULD vectors, where the elements sum to 1, the output layer's activation is softmax.

Fig. 3 provides an example to visually convey our motivation for calculating a loss for each class. Unlike the conventional classification, the SSL problem is more likely to occur in distributions that significantly deviate from Fig. 3(a). Therefore, we recommend using a loss function with a global view, such as BCE, to suppress pseudo peaks.

When the optimization target is soft labels, focusing on the non-zero value classes, we find that BCE simultaneously propagates gradients in two opposing directions. The first part directs $\hat{y}_i$ towards 1, while the second part directs $\hat{y}_i$ towards 0, rather than towards y_i.

5.1.3. MSE loss function

The MSE loss function, which also possesses a global view, is defined as follows:

$$\mathcal{L}^{\mathrm{MSE}} = \sum_{i=0}^{I} (y_i - \hat{y}_i)^2 \tag{13}$$

It is self-evident that the gradient direction passed back through the MSE function ensures that $\hat{y}_i$ consistently points towards y_i. In Eq. (1), the function $\sigma(\cdot)$ is a highly nonlinear transformation when κ_i is either very large or very small. As a result, the MSE function often suffers from the problem of gradient disappearance when $\hat{y}_i$ approaches 0 or 1, and is therefore not frequently used for optimizing classification models.

5.2. Definition

5.2.1. NLAE loss function

Given the aforementioned analyses, we have devised a loss function called Negative Log Absolute Error (NLAE). It can be defined as follows:

$$\mathcal{L}^{\mathrm{NLAE}} = -\sum_{i=0}^{I} \log(1 - |y_i - \hat{y}_i|) \tag{14}$$

It is evident that the optimization direction of NLAE is to make $1 - |y_i - \hat{y}_i|$ approach 1, which is equivalent to make $y_i = \hat{y}_i$. In particular, when y_i is 0 or 1, $\mathcal{L}^{\mathrm{NLAE}}$ and $\mathcal{L}^{\mathrm{BCE}}$ are equivalent, so the magnitude of the gradient passed back through NLAE is reasonable. Hence, the NLAE loss function is theoretically more suitable for the family of soft labels.

5.2.2. MSE(wo) loss function

Alternatively, we can use a trick to address the problem. Since the nonlinearity is caused by $\sigma(\cdot)$, we can consider discarding it. The MSE loss function can be modified to directly operate on $\boldsymbol{\kappa}$, formulated as:

$$\mathcal{L}^{\mathrm{MSE(wo)}} = \sum_{i=0}^{I} (y_i - \kappa_i)^2 \tag{15}$$

where $\mathcal{L}^{\mathrm{MSE(wo)}}$ represents the MSE loss function without $\sigma(\cdot)$. However, if we take this approach, we need an additional operation to ensure that the predicted distribution does not exceed the boundary values, which is $\hat{\boldsymbol{y}} = \{\min(\max(0, \kappa_i), 1)\}_{i=0}^{I}$.

The above discussion is suit for loss functions for single label. If necessary, calculating the loss for multiple labels separately and then adds them with weights to jointly optimize the DNN, may improve performance.

6. Experimental setup

6.1. Datasets

In this section, we conducted experiments on both simulated and real-world data. All source speech came from the LibriSpeech corpus (Panayotov, Chen, Povey, & Khudanpur, 2015). The train-clean-360, dev-clean and test-clean subsets were used to generate corresponding subsets of simulated datasets. We utilized the Pyroomacoustics (Scheibler, Bezzam, & Dokmanić, 2018) module to generate room impulse responses. For each utterance, we randomly set a room size and selected a 2-second segment. Each multi-source speech signal was mixed from different speakers. Additionally, we introduced additive noise to the reverberant speech. The additive noise was randomly selected from a large-scale noise set (Tan & Zhang, 2021) containing 126 h of various types of noises. Without specific instructions, the signal-to-noise ratio (SNR) of each utterance was randomly selected from a range of $[10, 20]$ dB. The training, validation, and testing sets had non-overlapping subsets of additive noise.

As shown in Table 1, we created five sets of simulated datasets, denoted as C1, C2, A1, L1, and L2 respectively, with different acoustic conditions to test the effectiveness and reliability of the proposed method. All datasets use microphone arrays consisting of 4 microphones. The complexity of the acoustic environment is mainly reflected in the reverberation and far-field, covering a wide range from anechoic to highly reverberant. The impact of number of sound sources is also considered in our datasets.

The length and width of a room were randomly chosen within $[4, 10]$ m, the height of the room was fixed at 3.2 m, and the height of the sound sources and microphones was fixed at 1.3 m. The reverberation time T_{60} of the room was randomly selected within a given range, or the room was set to be anechoic.

For the C1, C2, and A1 datasets, each random room produces a single utterance. For the L1 and L2 datasets, a random room plays an audio once, but with 10 microphone arrays in the room to capture signals. Consequently, each room in L1 and L2 generates 10 distinct

DOA segments. The placement of both microphones and sound sources is randomized. As a result, L1 and L2 are two datasets without any constraints on the distance between sound sources and microphone arrays or between sound sources and walls.

We recorded a real-world dataset (Liu, Gong, & Zhang, 2022) in two scenarios: an office and a conference room respectively. The office room is approximately $10.3 \times 9.8 \times 4.2$ m with a T_{60} of approximately 1.39 s. The conference room is approximately $4.26 \times 5.16 \times 3.16$ m with a T_{60} of approximately 1.06 s. The ambient noise in both rooms can be ignored. We used the test-clean subset of LibriSpeech as the source sound to play back in the room, with different speakers corresponding to sound sources played at different locations. The equipment used to record the dataset was one speaker and 10 linear arrays, each with the same shape as those used in L1 and L2. After being divided into 2-second segments, each room had a total of 97,480 samples. We randomly selected segments to generate a subset with two speakers. Therefore, each room contains a total of 7200 multi-speaker samples. We used the real-world data only for testing, while the simulated datasets of L1 and L2 were used for training and selecting models.

6.2. Comparison among different methods

The comparison methods include both traditional methods and DNN-based methods. The traditional methods include MUSIC (Schmidt, 1986) and SRP-PHAT (DiBiase, 2000). For DNN-based methods, we compared with different training targets and loss functions.

6.2.1. Training targets

We compare ULD with four training targets, which are:

- **One-hot**.
- **Gaussian Label Coding (GLC)** (He et al., 2018): We followed (He et al., 2018) and set the standard deviation to 8.
- **Soft Label Distribution (SLD)** (Subramanian et al., 2022).
- **Soft-argmax (SA)** (Diaz-Guerra et al., 2022).

6.2.2. Loss functions

We compare NLAE and MSE (wo) with four loss functions, which are:

- **Cross Entropy (CE)**.
- **Binary Cross Entropy (BCE)**.
- **Mean Squared Error (MSE)**.
- **Wasserstein Distance (WD)** (Subramanian et al., 2022).

Both CE and WD only apply to label with a sum of 1, so they are not suitable for GLC. During training, soft-argmax is equivalent to regression, thus only MSE can be used.

6.3. Neural networks

Four neural networks served as backbone networks, and the specific architectures are detailed in the supplementary material. The first network is the Phase Neural Network (PNN) (Chakrabarty & Habets, 2019), comprising three convolutional layers and three dense layers. The second network is PNN-Split (Subramanian et al., 2022), a modified version of PNN. Notably, PNN-Split routes the output of the first dense layer through a recurrent layer for implicit speech separation. Finally, the separated features are passed through another dense layer. The third network is SNet (He, Motlicek, & Odobez, 2021), while the fourth network is a hybrid model combining PNN-Split and SNet, known as SNet-Split. SNet-Split adopts all feature extraction modules of SNet, flattens the embedding features from the last residual block, and follows the subsequent operations consistent with PNN-Split.

Table 2
Experimental results on the dataset of A1, where the loss function is NLAE, and the backbone network is PNN. QE is short for quantization error, which is calculated by directly decoding the ground-truth one-hot labels.

	ACC	MAE		
QE limit	100.00	1.223		
Regression	–	0.642		
SA	–	0.310		
		Top-1	WAD-2	WAD-3
One-hot	98.56	1.225	0.920	0.924
GLC	97.50	1.231	1.036	0.697
SLD	97.44	1.237	1.437	1.144
ULD	**98.67**	**1.224**	**0.065**	**0.061**

6.4. Inputs of networks

We used 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 to extract Short-Term Fourier Transform (STFT) features. For PNN, the input is a single frame of the phase spectrum. For SNet, the input is 7 consecutive frames of STFT, with the real and imaginary parts of the STFT concatenated along the microphone channel dimension.

6.5. Training and evaluation details

For all experiments, we employed the AdamW (Loshchilov & Hutter, 2019) optimizer with a batch size of 32 and a maximum of 30 training epochs. The learning rate was initialized at 0.001 and reduced to 0.0001 if the validation loss did not improve over 3 consecutive epochs. The PNN and SNet were trained and tested on single-source datasets, while the Split networks was used for multi-source datasets. Since the DOA space is inherently ordered, it is easy to train multi-source models using location-based training (Taherian et al., 2022). By default, for C1 dataset, l was set to 3; for C2, l was set to 8; for A1, L1, and R1, l was set to 5; and for L2 and R2, l was set to 7.5.

6.6. Evaluation metrics

Suppose a dataset has N test speakers. As we primarily discusses classification models, a natural evaluation metric is classification accuracy (ACC), which can be formalized as follows:

$$\mathrm{ACC}(\%) = \frac{N^{\mathrm{acc}}}{N} \times 100 \tag{16}$$

where N^{acc} is the number of speakers for which the peak class of the predicted distribution equals to the ground truth class.

The most intuitive evaluation metric for SSL should be the mean absolute error (MAE) between the predicted source position and ground-truth source position, which can be described as follows:

$$\mathrm{MAE}(\circ) = \frac{1}{N} \sum_{n=1}^{N} \min(|\hat{p}_n - p_n|, 360 - |\hat{p}_n - p_n|) \tag{17}$$

7. Experimental results

7.1. Empirical study on WAD

7.1.1. Results on anechoic environment

As the aforementioned, the data for A1 is anechoic. The distance between the source and the microphone array is fixed at 1.5 m. Overall, this is a very simple dataset that we primarily use to investigate the issue of quantization error. As shown in Table 2, the quantization error limit is 1.223, which confirms our theoretical analysis that the quantization error is approximately $l/4$ (in this case, 5/4=1.250). If we use classical Top-1 decoding, then the MAE is bounded by the quantization error limit. However, when we use ULD in conjunction with WAD, the quantization error limit has been significantly broken

Table 3
Results on the dataset of C1, the loss function is NLAE, the encoding method is ULD, and the backbone network is PNN.

l		Regression	SA	360	180	90	45	20	10	5	3	2	1
ACC		–	–	99.21	97.18	97.64	96.30	96.39	94.68	92.27	87.59	80.51	54.77
QE limit		–	–	89.056	45.740	22.519	11.407	5.059	2.481	1.244	0.752	0.504	0.249
MAE	Top-1	21.725	14.996	89.056	46.012	22.666	11.522	5.102	2.538	1.307	0.853	0.676	**0.625**
	WAD-2			17.562	7.336	3.226	1.847	1.028	0.793	0.649	0.557	0.547	**0.541**
	WAD-3			17.562	10.055	3.756	1.865	1.005	0.719	0.540	**0.476**	0.486	0.561

Table 4
Results on the lightly-reverberant dataset C1, where the backbone network is PNN.

Method	Loss	ACC	MAE		
			Top-1	WAD-2	WAD-3
SA (Diaz-Guerra et al., 2022)	MSE	–	14.996		
One-hot	CE	86.57	0.869	0.572	0.517
	MSE	85.93	0.881	0.558	0.522
	WD	53.61	1.681	1.716	1.701
	NLAE	86.39	0.868	0.565	**0.514**
	MSE (wo)	84.63	0.905	0.591	0.558
GLC (He et al., 2018)	BCE	81.34	0.954	0.895	0.894
	MSE (He et al., 2018)	80.23	0.961	0.889	0.898
	NLAE	80.42	0.950	0.882	0.880
	MSE (wo)	82.59	0.921	0.881	**0.846**
SLD (Subramanian et al., 2022)	CE	80.00	0.947	0.800	0.765
	BCE	81.62	0.929	0.794	0.744
	MSE	83.84	0.881	0.791	0.713
	WD (Subramanian et al., 2022)	55.51	1.536	1.665	1.636
	NLAE	82.22	0.918	0.775	0.716
	MSE (wo)	84.95	0.881	0.781	**0.688**
ULD	CE	86.71	0.852	0.574	0.490
	BCE	87.18	0.852	0.572	0.489
	MSE	87.41	0.853	0.561	0.487
	WD	55.69	1.567	1.629	1.606
	NLAE	87.59	0.853	0.557	**0.476**
	MSE (wo)	87.69	0.854	0.600	0.503

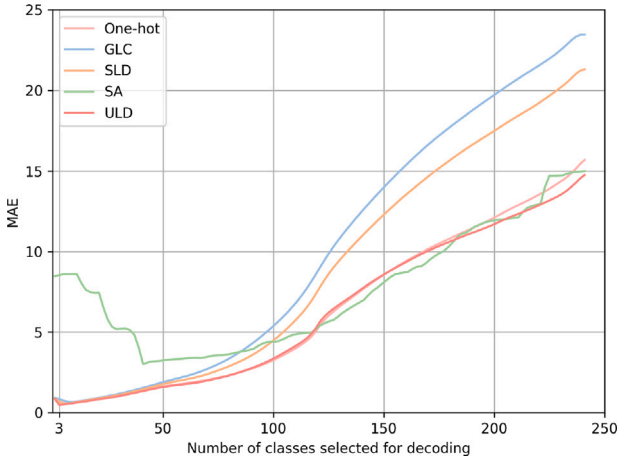


Fig. 4. The impact of selecting different numbers of classes on the performance of weighted adjacent decoding. The dataset is C1. Each training target is paired with the best loss function as indicated in Table 4. Given the inter-class interval l of 3 degrees, we have $I = \frac{360}{l} = 120$. Consequently, the upper limit for selected classes I_s is $2I + 1 = 241$.

through. Interestingly, applying WAD to the one-hot encoding actually reduces the MAE and breaks the quantization error limit as well, due to the presence of sidelobes (the loss is not 0).

7.1.2. Results on reverberant environment

In the previous section, we have observed that WAD could break the quantization error limit in the anechoic environment, here we study WAD in a reverberant environment. To study the effect of WAD integratively, we tune the parameter l in a wide range. Specifically, the parameter l determines the azimuth range of a cell (i.e. a class), so as to the number of classes. When l decreases from 360 to 1, the number of classes naturally increases, which results in an increased model complexity accordingly.

Table 3 lists the performance of decoding methods along with the parameter l in the reverberant data C1. We see that WAD breaks the quantization error limit. Specifically, when $l \geq 3$, WAD yields smaller MAE than the quantization error limit, while the Top-1 decoding always yields larger MAE than the limit. The best performance of WAD appears at $l = 3$, and when $l = 1$, the MAE of WAD becomes larger than the quantization error limit. This phenomenon is caused by the decrease in ACC with the increase in the number of classes. Specifically, reducing l from 3 to 2 requires increasing the number of classes from 121 to 181, and reducing l from 2 to 1 even requires increasing the number of classes from 181 to 361. As ACC decreases, classification error gradually replaces quantization error as the dominant factor. Another interesting result is that, even when l is as large as 360 (i.e., it is a binary classification), the MAE of the proposed WAD is still smaller than the regression model.

7.1.3. Effect of selecting different numbers of classes for WAD

Fig. 4 illustrates the effect of selecting different numbers of classes for WAD on performance. The model trained with ULD achieves the lowest error when the number of classes I_s is 3 (i.e. WAD-3). As I_s increases, the MAE gradually rises because the classes far from the peak class $\hat{k}$ become less reliable. When I_s reaches 241, incorporating all classes in the decoding process is akin to using soft-argmax decoding. At this point, the model trained with SA shows an MAE of approximately 15, consistent with the results in Table 4.

Table 5
Results on the heavily-reverberant dataset of L1, where the backbone network is PNN.

Method	Loss	ACC	MAE		
			Top-1	WAD-2	WAD-3
SA (Diaz-Guerra et al., 2022)	MSE	–	17.430		
One-hot	CE	68.17	3.947	3.696	3.649
	MSE	68.53	3.985	3.751	3.693
	WD	62.94	4.408	4.184	4.154
	NLAE	65.50	3.814	3.592	**3.539**
	MSE (wo)	65.78	4.040	3.874	3.787
GLC (He et al., 2018)	BCE	62.08	3.938	3.921	3.789
	MSE (He et al., 2018)	60.94	3.924	3.883	3.761
	NLAE	62.75	3.936	3.932	3.760
	MSE (wo)	59.56	3.693	3.652	**3.552**
SLD (Subramanian et al., 2022)	**CE**	62.67	3.670	3.694	**3.537**
	BCE	61.61	3.804	3.791	3.669
	MSE	62.19	4.011	3.973	3.866
	WD (Subramanian et al., 2022)	57.78	4.337	4.313	4.185
	NLAE	60.67	4.763	4.725	4.630
	MSE (wo)	63.78	4.005	4.006	3.871
ULD	CE	67.03	3.988	3.713	3.649
	BCE	69.08	3.689	3.459	3.363
	MSE	68.53	4.012	3.735	3.663
	WD	65.75	4.427	4.233	4.137
	NLAE	65.44	4.632	4.385	4.295
	MSE (wo)	65.39	3.481	3.179	**3.148**

Table 6
Main results on the real-world data, where the backbone neural network is PNN.

Test data	Method	ACC	MAE		
			Top-1	WAD-2	WAD-3
Office	MUSIC	–	50.866		
	SRP-PHAT	–	44.890		
	SA	–	20.505		
	One-hot + CE	62.60	3.168	3.181	3.077
	One-hot + NLAE	63.17	3.071	3.087	2.978
	GLC + MSE(wo)	57.93	3.518	3.926	3.500
	SLD + MSE(wo)	60.43	3.184	3.582	3.163
	ULD + MSE(wo)	64.12	3.008	3.116	**2.925**
Conference	MUSIC	–	48.724		
	SRP-PHAT	–	43.709		
	SA	–	23.176		
	One-hot + CE	54.32	6.888	6.905	6.805
	One-hot + NLAE	53.84	5.192	5.268	**5.138**
	GLC + MSE(wo)	49.90	5.816	6.078	5.814
	SLD + MSE(wo)	53.71	6.052	6.498	6.052
	ULD + MSE(wo)	54.18	5.431	5.635	5.420

7.2. Single-source localization

7.2.1. Results on simulated data

Table 4 lists the performance of various combinations of encoding methods, loss functions, and decoding methods on the lightly-reverberant dataset C1. From the table, it can be seen that ULD is the best label encoding method in this environment; NLAE is the best loss function; and WAD-3 is the best decoding method in almost all cases except with the WD loss function. The reason for the poor performance of the "ULD+WD loss" scheme, we think that WD can essentially be viewed as a special form of global regression, therefore inheriting the vulnerability of global regression when used for SSL. Compared to the most common paradigm of one-hot encoding with CE and Top-1 decoding, the combination of ULD with NLAE and WAD-3 reduces the MAE by 45.22%.

Table 5 further lists the performance of comparison methods on the heavily reverberant L1 dataset, where the distances between sound sources and microphone arrays are unconstrained. The table shows the proposed strategy of ULD, MSE(wo), and WAD-3 achieves top performance, 20.24% above conventional one-hot paradigm. We find NLAE performs poorly in this adverse condition.

7.2.2. Results on real-world data

First, we note that, due to space constraint of the paper, we only report main experimental results in the following sections, leaving full results in the supplementary material.

Table 6 lists the performance of comparison models on two real-world datasets, trained on simulated data L1. From the table, we see that: (i) the proposed ULD, MSE(wo), and WAD-3 strategy performs best on the office room dataset; (ii) the proposed NLAE and WAD-3, combined with one-hot encoding, performs best on the conference room dataset, closely followed by the proposed ULD, MSE(wo), and WAD-3 strategy; (iii) as analyzed earlier, the CE loss function underperforms compared to the proposed NLAE and MSE(wo), highlighting the importance of using a loss function with a global view.

7.2.3. Effect of SNR on performance

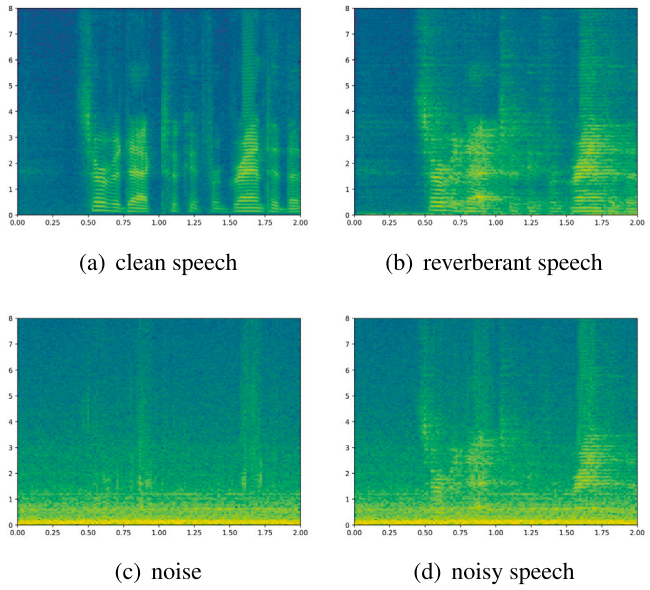
We studied the effectiveness of various methods in different noisy environments. We added noise at different SNRs to the test set of C1 dataset, creating three test subsets with SNRs of $\{-20, -10, 0\}$ dB. For each utterance, the same noise segment was used three times, only changing the SNR.[1]

Table 7 shows the performance of the comparison methods at various SNR levels. When the SNR level increases, the ACC of all classification-based models increases, leading to smaller MAEs. ULD remains the optimal one. For example, it produces an MAE reduction of 14.56%, 5.04%, and 6.26% relatively over the method of one-hot with WAD-3 in −20dB, −10dB, and 0dB. Additionally, it can be observed that the MAE for all classification models remains low, even at very low SNR levels. We think this is due to that (i) the speakers in the C1 dataset are only 1.5 meters away from the microphones, and (ii) the individual noise segments in the noisy dataset do not cover all time–frequency bins. As shown in Fig. 5, even at an SNR of −20 dB, the speech is not completely masked by noise, allowing the DNNs to still be able to locate the speaker in the noisy audio.

[1] Some testing audio samples are available at https://github.com/linfeng-feng/ULD.

Table 7
Results on different SNR, where the original test set is C1, and the backbone models that perform the best in Table 4 are selected.

SNR	−20 dB	−10 dB	0 dB	−20 dB	−10 dB	0 dB
		ACC			MAE	
SA		—		39.044	27.913	20.338
One-hot	76.44	80.22	82.31	0.886	0.655	0.591
GLC	76.28	79.06	80.22	1.159	0.964	0.928
SLD	79.08	81.86	83.11	1.072	0.909	0.885
ULD	77.47	81.50	83.64	**0.757**	**0.622**	**0.554**



(a) clean speech (b) reverberant speech

(c) noise (d) noisy speech

Fig. 5. A set of spectrograms is shown, with time (seconds) and frequency (kHz). (a) is clean speech; (b) is the reverberant speech generated by playing (a) as the source in a room; (c) is a noise segment; (d) is the combination of (b) and (c) with the SNR of −20 dB.

Fig. 6 shows the impact of I, the number of quantization levels, on the localization performance with different SNR levels. The three curves exhibit a similar trend. For $I < 120$, the MAE gradually decreases as I increases. At $I = 120$ (i.e. $l = 3$), the MAE reaches its minimum. However, as I continues to increase, the MAE rises instead of falling.

7.2.4. Effect of backbone networks on performance

In previous experiments, all backbone networks were PNN. In this subsection, we study how backbone networks affect performance. Table 8 lists comparison method performance using SNet as the backbone network on both simulated data L1 and real-world data. Compared to Table 6, we see the proposed ULD, MSE(wo), and WAD-3 strategy performs best on L1 and the office room, consistent with results using the PNN backbone network. Although the best performance in the conference room appears with proposed MSE(wo) and WAD-3 combined with GLC, proposed ULD, MSE(wo), and WAD-3 performance follows closely, consistent with the PNN backbone results.

7.3. Multi-source localization

Table 9 lists the performance of the comparison methods on the multi-source data C2. From the table, we see that WAD-3 remains the best decoding method in almost all cases; MSE(wo) still fits the soft labels best, including GLC, SLD, and the proposed ULD. We also observe that GLC is slightly better than the proposed ULD. This phenomenon may be caused by the extreme smoothness of GLC, making it more conducive to model training in challenging scenarios.

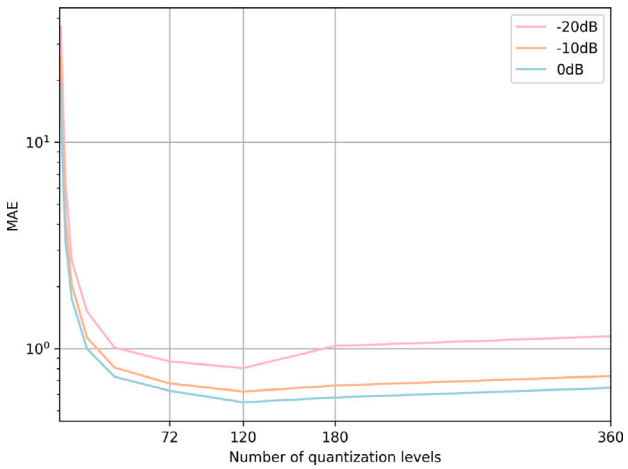


Fig. 6. The impact of I, the number of quantization levels, on localization performance under different SNR levels.

Table 8
Main results on the single-source data, where the backbone network is Snet, and the loss function for soft labels is MSE (wo).

Test data	Method	ACC	MAE		
			Top-1	WAD-2	WAD-3
Simulated data L1	SA	–		8.074	
	One-hot	75.79	2.612	2.265	2.236
	GLC	71.44	2.292	2.214	2.055
	SLD	69.93	2.543	2.544	2.360
	ULD	76.88	2.176	1.782	**1.696**
Office	SA	–		16.247	
	One-hot	67.48	2.285	2.197	2.169
	GLC	61.11	2.474	2.766	2.413
	SLD	59.88	2.555	2.894	2.521
	ULD	68.96	2.177	2.208	**2.145**
Conference	SA	–		17.439	
	One-hot	56.57	4.616	4.488	4.483
	GLC	53.70	4.337	4.462	**4.257**
	SLD	52.97	4.519	4.696	4.465
	ULD	57.18	4.583	4.476	4.456

Table 9
Main results on the simulated multi-source dataset C2, where the backbone network is PNN-Split.

Method	Loss	ACC	MAE		
			Top-1	WAD-2	WAD-3
SA	MSE	–		23.163	
One-hot	WD	68.22	7.189	6.573	6.576
GLC	**MSE (wo)**	77.09	5.855	5.391	**5.043**
SLD	MSE (wo)	71.88	5.588	5.386	5.203
ULD	MSE (wo)	79.33	6.089	5.291	5.114

Note that GLC and ULD exhibit distinct advantages—while one emphasizes greater smoothness, the other prioritizes higher precision, suggesting their combined use. We designed a joint training method using weighted loss with training objectives for both ULD and GLC: "αULD+$(1-\alpha)$GLC", where $\alpha \in [0, 1]$ is a tunable parameter. Table 10 lists the performance of the combined encoding method with respect to α. From the table, we see the combined encoding method significantly outperforms its components, i.e. GLC and ULD. The best performance of the combined method appears at $\alpha = 0.2$.

Finally, Table 11 lists the results on a set of highly challenging datasets, L2, and real-world datasets featuring two speakers with significant reverberation and considerable distance from microphones. Given the adverse conditions, ACC is notably low. Classification error predominates over quantization error, limiting performance gains. As shown

Table 10
Results of the combined encoding method with respect to the parameter α on C2, where the backbone network is PNN-Split, and the loss function is MSE(wo).

Method	α	ACC	MAE		
			Top-1	WAD-2	WAD-3
αULD+$(1-\alpha)$GLC	0.0	77.09	5.855	5.391	5.043
	0.2	78.40	5.334	4.815	**4.465**
	0.4	78.65	5.595	5.073	4.710
	0.6	77.29	5.939	5.347	5.049
	0.8	79.22	6.082	5.423	5.167
	1.0	79.33	6.089	5.291	5.114

Table 11
Results on the highly-reverberant multi-source data, where the backbone network is the SNet-Split. When using one-hot encoding, the training loss function is WD, while for all others it is MSE(wo). The parameter α in "αULD+$(1-\alpha)$GLC" is set to 0.2.

Subset	Method	ACC	MAE		
			Top-1	WAD-2	WAD-3
Simulated data L2	SA	–		15.647	
	One-hot	62.56	6.145	5.892	5.896
	GLC	70.74	4.760	4.386	4.093
	SLD	70.64	4.798	4.765	4.495
	ULD	73.53	5.296	4.554	4.524
	αULD+$(1-\alpha)$GLC	73.01	4.446	4.008	**3.739**
Office	SA	–		27.577	
	One-hot	60.47	6.422	6.309	6.291
	GLC	64.23	6.130	6.756	6.180
	SLD	63.62	6.204	7.005	6.244
	ULD	62.79	6.505	6.500	6.434
	αULD+$(1-\alpha)$GLC	65.22	6.107	6.620	**6.081**
Conference	SA	–		29.279	
	One-hot	44.96	12.723	12.552	12.543
	GLC	53.86	11.349	11.442	11.211
	SLD	53.63	11.714	11.941	11.640
	ULD	54.58	12.776	12.655	12.591
	αULD+$(1-\alpha)$GLC	53.68	10.519	10.459	**10.229**

in Table 11, results are overall consistent with previous experiments, showcasing sustained superior performance of our method.

8. Conclusions

In this paper, we propose a novel output architecture for SSL, incorporating three components: (i) ULD as the encoding method, (ii) WAD as the decoding method, and (iii) NLAE and MSE(wo) as the training loss functions. Specifically, unlike one-hot encoding, ULD is a one-to-one encoding method, resulting in the unbiased inverse mapping between label distribution and sound source position. Unlike Top-1 decoding, WAD considers not only peak class but also sidelobes during decoding. NLAE and MSE(wo) integrate benefits of cross-entropy-like and MSE-like functions. They can be viewed as regression-based loss functions applied per class. Experimental results on both simulated and real-world data show WAD significantly outperforms quantization error limits, especially with ULD encoding. The proposed NLAE is best for soft labels in simple environments, while MSE(wo) performs best for soft labels in challenging environments. The overall output architecture performs best in most cases.

CRediT authorship contribution statement

Linfeng Feng: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **Xiao-Lei Zhang:** Writing – review & editing, Supervision, Conceptualization. **Xuelong Li:** Writing – review & editing, Supervision.

Declaration of competing interest

We declare that there is no conflict of interests.

Data availability

Our code and supplementary materials are available at https://github.com/linfeng-feng/ULD.

Acknowledgements

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 JCYJ20210324143006016 and JSGG20210802152546026.

Appendix A. Proof of Theorem 1

Proof. Let I be a positive integer. We have $\gamma \sim \mathcal{U}(0, I)$, and $n = \lfloor \gamma \rfloor$. Further, we have $\text{round}(\gamma) = \arg\min_i \{|\gamma - i|\}_{i=\{n,n+1\}}$. Then, the expected value of $|\gamma - \text{round}(\gamma)|$ is:

$$\mathbb{E}(|\gamma - \text{round}(\gamma)|) = \int_n^{n+1} |x - \text{round}(x)| dx$$

We can split the interval $[n, n+1]$ into two parts: $[n, n+0.5)$ and $[n+0.5, n+1]$. In the interval $[n, n+0.5)$, $\text{round}(x) = n$; in the interval $[n+0.5, n+1]$, $\text{round}(x) = n+1$. Therefore:

$$\begin{aligned}
&\mathbb{E}(|\gamma - \text{round}(\gamma)|) \\
&= \int_n^{n+0.5} |x - n| dx + \int_{n+0.5}^{n+1} |x - (n+1)| dx \\
&= \int_n^{n+0.5} (x - n) dx + \int_{n+0.5}^{n+1} ((n+1) - x) dx \\
&= \left[\frac{(x-n)^2}{2} \right]_{x=n}^{x=n+0.5} + \left[(n+1)x - \frac{x^2}{2} \right]_{x=n+0.5}^{x=n+1} \\
&= \frac{1}{8} + \frac{1}{8} \\
&= \frac{1}{4}
\end{aligned}$$

Through the above formula, we can further calculate the mathematical expectation of the quantization error as follows:

$$\begin{aligned}
\mathbb{E}(qe) &= \mathbb{E}(|p - \sum_{i=0}^{I} y_i^{\text{1-hot}} \times i \times l|) \\
&= \mathbb{E}(|\gamma \times l - 1 \times \text{round}(\gamma) \times l|) \\
&= \mathbb{E}(|\gamma - \text{round}(\gamma)|) \times l \\
&= \frac{l}{4} \quad \square
\end{aligned}$$

Appendix B. Proof of Theorem 2

Proof. Let γ be a non-negative real number. We have $\text{int}(\gamma) = \lfloor \gamma \rfloor$, and $\text{deci}(\gamma) = \gamma - \lfloor \gamma \rfloor$, which derives:

$$\begin{aligned}
&(1 - \text{deci}(\gamma)) \times \text{int}(\gamma) + \text{deci}(\gamma) \times (\text{int}(\gamma) + 1) \\
&= \text{int}(\gamma) - \text{deci}(\gamma) \times \text{int}(\gamma) + \text{deci}(\gamma) \times \text{int}(\gamma) + \text{deci}(\gamma) \\
&= \text{int}(\gamma) + \text{deci}(\gamma) \\
&= \gamma \quad \square
\end{aligned}$$

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2024.106679.

References

Adavanne, S., Politis, A., Nikunen, J., & Virtanen, T. (2018). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, *13*(1), 34–48.

Bai, J., Huang, S., Yin, H., Jia, Y., Wang, M., & Chen, J. (2023). 3D audio signal processing systems for speech enhancement and sound localization and detection. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–2). IEEE.

Chakrabarty, S., & Habets, E. A. (2019). Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, *13*(1), 8–21.

Diaz-Guerra, D., Miguel, A., & Beltran, J. R. (2020). Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 300–311.

Diaz-Guerra, D., Miguel, A., & Beltran, J. R. (2022). Direction of arrival estimation of sound sources using icosahedral CNNs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 313–321.

DiBiase, J. H. (2000). *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University.

Feng, L., Gong, Y., & Zhang, X.-L. (2023). Soft label coding for end-to-end sound source localization with ad-hoc microphone arrays. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–5). IEEE.

Fu, Y., Ge, M., Yin, H., Qian, X., Wang, L., Zhang, G., et al. (2022). Iterative sound source localization for unknown number of sources. In *Proc. interspeech 2022* (pp. 896–900).

Gburrek, T., Schmalenstroeer, J., & Haeb-Umbach, R. (2023). Spatial diarization for meeting transcription with ad-hoc acoustic sensor networks. arXiv preprint arXiv:2311.15597.

Grumiaux, P.-A., Kitić, S., Girin, L., & Guérin, A. (2022). A survey of sound source localization with deep learning methods. *Journal of the Acoustical Society of America*, *152*(1), 107–151.

He, W., Motlicek, P., & Odobez, J.-M. (2018). Deep neural networks for multiple speaker detection and localization. In *2018 IEEE international conference on robotics and automation* (pp. 74–79). IEEE.

He, W., Motlicek, P., & Odobez, J.-M. (2019). Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 770–774). IEEE.

He, W., Motlicek, P., & Odobez, J.-M. (2021). Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1303–1317.

Jing, L., Ding, Y., Gao, Y., Wang, Z., Yan, X., Wang, D., et al. (2024). HPL-ESS: Hybrid pseudo-labeling for unsupervised event-based semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23128–23137).

Knapp, C., & Carter, G. (1976). The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*(4), 320–327.

Liu, S., Gong, Y., & Zhang, X.-L. (2022). Deep learning based two-dimensional speaker localization with large ad-hoc microphone arrays. arXiv preprint arXiv:2210.10265.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International conference on learning representations*.

Nguyen, T. N. T., Gan, W.-S., Ranjan, R., & Jones, D. L. (2020). Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2626–2637.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing* (pp. 5206–5210). IEEE.

Perotin, L., Défossez, A., Vincent, E., Serizel, R., & Guérin, A. (2019). Regression versus classification for neural network based audio source localization. In *2019 IEEE workshop on applications of signal processing to audio and acoustics* (pp. 343–347). IEEE.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99.

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

Schmidt, R. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, *34*(3), 276–280.

Shimada, K., Koyama, Y., Takahashi, N., Takahashi, S., & Mitsufuji, Y. (2021). 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* (pp. 915–919). IEEE.

Shimada, K., Koyama, Y., Takahashi, S., Takahashi, N., Tsunoo, E., & Mitsufuji, Y. (2022). Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 316–320). IEEE.

Subramanian, A. S., Weng, C., Watanabe, S., Yu, M., Xu, Y., Zhang, S.-X., et al. (2021). Directional ASR: A new paradigm for E2E multi-speaker speech recognition with source localization. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing* (pp. 8433–8437). IEEE.

Subramanian, A. S., Weng, C., Watanabe, S., Yu, M., & Yu, D. (2022). Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition. *Computer Speech and Language*, *75*, Article 101360.

Taherian, H., Tan, K., & Wang, D. (2022). Multi-channel talker-independent speaker separation through location-based training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2791–2800.

Taherian, H., & Wang, D. (2023). Multi-channel conversational speaker separation via neural diarization. arXiv preprint arXiv:2311.08630.

Tan, X., & Zhang, X.-L. (2021). Speech enhancement aided end-to-end multi-task learning for voice activity detection. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing* (pp. 6823–6827). IEEE.

Tang, Z., Kanu, J. D., Hogan, K., & Manocha, D. (2019). Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks. In *Proc. interspeech 2019* (pp. 654–658).

Vecchiotti, P., Pepe, G., Principi, E., & Squartini, S. (2019). Detection of activity and position of speakers by using deep neural networks and acoustic data augmentation. *Expert Systems with Applications*, *134*, 53–65.

Vera-Diaz, J. M., Pizarro, D., & Macias-Guarasa, J. (2018). Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates. *Sensors*, *18*(10), 3418.

Vesperini, F., Vecchiotti, P., Principi, E., Squartini, S., & Piazza, F. (2016). A neural network based algorithm for speaker localization in a multi-room environment. In *2016 IEEE 26th international workshop on machine learning for signal processing* (pp. 1–6). IEEE.

Wang, Z.-Q., & Wang, D. (2022). Localization based sequential grouping for continuous speech separation. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 281–285). IEEE.

Wu, Z., Xu, Y., Yang, J., & Li, X. (2024). Misclassification in weakly supervised object detection. *IEEE Transactions on Image Processing*, *33*, 3413–3427.

Xiao, X., Zhao, S., Zhong, X., Jones, D. L., Chng, E. S., & Li, H. (2015). A learning-based approach to direction of arrival estimation in noisy and reverberant environments. In *2015 IEEE international conference on acoustics, speech and signal processing* (pp. 2814–2818). IEEE.

Zheng, S., Huang, W., Wang, X., Suo, H., Feng, J., & Yan, Z. (2021). A real-time speaker diarization system based on spatial spectrum. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing* (pp. 7208–7212). IEEE.