METHODOLOGY   **Open Access**

Dual input neural networks for positional sound source localization

Eric Grinstein[1*] , Vincent W. Neo[1] and Patrick A. Naylor[1]

Abstract

In many signal processing applications, metadata may be advantageously used in conjunction with a high dimensional signal to produce a desired output. In the case of classical Sound Source Localization (SSL) algorithms, information from a high dimensional, multichannel audio signals received by many distributed microphones is combined with information describing acoustic properties of the scene, such as the microphones' coordinates in space, to estimate the position of a sound source. We introduce Dual Input Neural Networks (DI-NNs) as a simple and effective way to model these two data types in a neural network. We train and evaluate our proposed DI-NN on scenarios of varying difficulty and realism and compare it against an alternative architecture, a classical Least-Squares (LS) method as well as a classical Convolutional Recurrent Neural Network (CRNN). Our results show that the DI-NN significantly outperforms the baselines, achieving a five times lower localization error than the LS method and two times lower than the CRNN in a test dataset of real recordings.

Keywords Sound source localization, Multichannel audio processing, Multimodal machine learning, Convolutional recurrent neural networks

1 Introduction

Most signals, such as audio and images, contain metadata. Metadata can be signal-based, which describes quantitative properties of the signal, such as its sampling rate, as well as semantic, which describes, for example, contextual properties. In speech processing, semantic metadata could consist of the speaker's language or gender. Whether signal-based or semantic, including metadata as a secondary input into neural network models may provide relevant information which would translate into an economy of training time, model parameters and flexibility. However, metadata typically has a different dimensionality than the input signals, making its incorporation into those models not trivial.

The main focus of this paper is to study the effectiveness of schemes to process signals and exploit metadata jointly using neural network models. We focus on the task of Sound Source Localization (SSL) [1] using distributed microphone arrays to demonstrate the effectiveness of our proposed approach. In the context of SSL, relevant metadata which is exploited by classical methods is the microphone positions, which can be acquired by manual measurement or using self-calibration [2] methods. Other relevant metadata is the room dimensions and its reverberation time.

SSL refers to the task of estimating the spatial location of a sound source, such as a human talker or a loudspeaker. In this scenario, metadata refers to properties of the acoustic scene such as the coordinates of microphones, dimensions of the room and, the reflection coefficient of the walls. SSL has many applications, including noise reduction and speech enhancement [3], camera steering [4] and acoustic Simultaneous Localization and Mapping (SLAM) [5]. In turn, distributed microphone arrays have become an active research topic in the

*Correspondence:
Eric Grinstein
e.grinstein@imperial.ac.uk
[1] Department of Electrical and Electronic Engineering, Imperial College London, London, UK
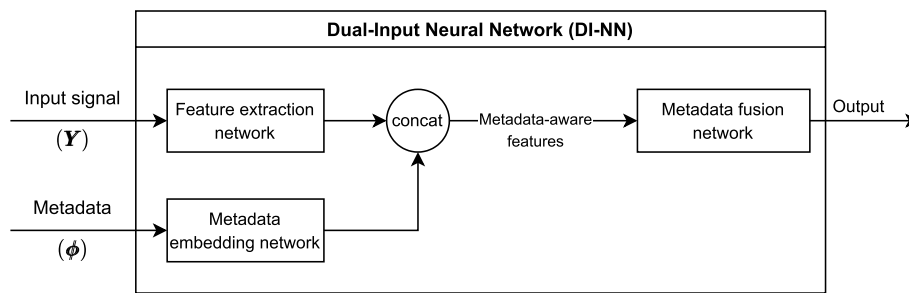
Fig. 1 Overview of the Dual-Input Neural Network (DI-NN) approach

signal processing community due to their versatility. Such arrays may be composed of multiple network-connected devices, including everyday devices such as cell phones, smart assistants, and laptops, for example. The array and the constituent devices may be configured as a Wireless Acoustic Sensor Network (WASN) [6].

SSL approaches may be divided into classical signal processing-based and data-driven neural network-based methods. By explicitly exploiting metadata describing microphone positions and room dimensions, classical approaches may be applied to different rooms and microphone configurations. Conversely, neural network approaches have recently achieved state of the art results for source localization [7–9], at the expense of requiring one network to be trained for every microphone topology. One reason current neural approaches do not incorporate the microphones' positional information is that the microphones' signal and positional data are very different from one another in nature and dimension.

Previous work which discusses the joint processing of signals and metadata is [10], where a single input neural network is used to process metadata in conjunction with a low-dimensional physical signal. However, unlike our work, the method of [10] is restricted to multilayer perceptron architectures and one-dimensional input and metadata, limiting its application in practical scenarios.

Another related field is multimodal fusion [11, 12], although this is usually concerned with learning representations using two types of signals, such as audio-visual data. Simultaneously processing signals and metadata have also been explored using non-neural models for sound source separation [13], where metadata consists of information about the type of sound (speech, music) and how the sources were mixed. However, none of the existing work discusses effective schemes for incorporating and evaluating signals and metadata of different dimensionality.

Our main contribution is the DI-NN neural network architecture, which is capable of processing high-dimensional signals, namely spectrograms, along with a relevant

metadata vector of lower dimensionality. An overview diagram of our approach is shown in Fig. 1, which will be discussed in Section 3.2. We compare our method against three baselines for the task of Positional Sound Source Localization (PSSL), namely, a metadata-unaware Convolutional Recurrent Neural Network (CRNN), a metadata-aware classical signal processing approach, as well as an alternate metadata-aware neural network. Our proposed method is able to outperform all baselines by a large margin in realistic scenarios. In contrast to previous approaches [9, 14], our network dispenses with the need for training a network for each scenario, broadening our method's applicability.

This work continues as follows. In Section 2, an overview of neural and non-neural SSL methods will be discussed. The approach for training our proposed DI-NN for SSL is described together with several baseline methods in Section 3. In Section 4, the experiments comparing our approach with the baselines using multiple datasets are described. Finally, results and conclusions are drawn in Section 5.

2 Prior art on sound source localization
2.1 Neural-based methods
In recent years, deep neural networks have been widely adopted for the task of sound source localization. The various approaches differ in the input features used, the network architectures and output strategies. Most studies focus on the task of Direction-of-Arrival (DOA) estimation, i.e., estimating the angle between the propagation direction of the acoustic wavefront due to the source and a reference axis of the array.

Practicioners have experimented with many types of neural input features, such as the raw audio samples of the microphone signals [9], their frequency-domain representation through the Short Time Fourier Transform (STFT) [15], their cross-spectra [16] or cross-correlation [8]. Multiple architectures have been also tested, including the Multi-layer Perceptron (MLP) [8], Convolutional Neural Networks (CNNs) [17] and residual networks

[18]. In this work, we focus on the Convolutional Recurrent Neural Network (CRNN) architecture, which has received widespread adoption in the field [7, 19, 20]. Finally, approaches differ in terms of the network's output strategy. While regression-based approaches directly estimate the source's coordinates, classification based-approaches discretize the source locations to a grid of available positions. We refer to [21] for a discussion on the merits of both approaches. We also refer the reader to a substantial survey of neural SSL papers [22].

In this paper, we focus on the task of estimating the absolute Cartesian coordinates of the source, which we shall refer to as Positional Sound Source Localization (PSSL), and has applications in robot navigation [5] and noise reduction [23]. The PSSL task has been much less studied using neural methods. To the best of our knowledge, only [14] and [9] focus on PSSL. However, both these approaches only work for the same room with fixed relative microphone positions. We believe this shortage of studies to be at least in part due to the lack of an architecture capable of incorporating the scene's metadata, which is addressed by our proposed DI-NN. We also refer to the recent L3DAS22 challenge [24], where practitioners were invited to develop 3D PSSL algorithms for a realistic office environment containing a pair of microphone arrays.

2.2 Classical signal processing methods

Classical approaches to SSL have been widely studied within the signal processing community. In PSSL approaches, the source's coordinates are estimated using a model involving signal processing, physics and geometry. By measuring differences in the microphone signals' amplitudes and phases, distance metrics between the microphones and source can be estimated. These estimates can in turn be combined to estimate the source's coordinates [1]. Besides the microphones' signals, the positions of microphones are usually needed for the position of the source to be estimated. Available approaches for SSL may be classified as delay-based [1, 25], energy-based [26, 27], subspace-based [28] and beamforming-based [29, 30] approaches. We shall focus on delay-based approaches and will provide background for our baseline method.

Delay-based SSL methods usually rely on computing the Time-Difference-of-Arrival (TDOA) between each microphone pair within the system, which corresponds to the difference in time taken for the source signal to propagate to different microphones. The locus of candidate source positions with the same TDOA with respect to a microphone pair is, when considering planar coordinates, a hyperbola [1, 25]. The source is located at the intersection of the hyperbolae defined by all microphone pairs. The multiple TDOAs can be combined using a Least-Squares (LS) framework [31], or using a Maximum Likelihood (ML) approach if some noise properties of the system are known [1]. In general, TDOAs are estimated using cross-correlation based methods such as Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [32], which are shown to be somewhat robust to reflections produced in the room due to, for example, the walls, ceiling and furniture, i.e. reverberation [33].

3 Method

3.1 Signal model and scope of this work

Our scope is restricted to the localization of a static source at the planar coordinates $\boldsymbol{p}_s = [p_s^x, p_s^y]^T$. The source emits an intermittent signal $s(t)$ at time t. In our experiments, $s(t)$ may consist of White Gaussian Noise (WGN) as well as of speech utterances. Also, M static microphones with known positions are present in the room, each placed at coordinates $\boldsymbol{m}_i = [m_i^x, m_i^y]^T$. Both source and microphones are enclosed in a room of planar dimensions $\boldsymbol{d} = [d^x, d^y]^T$. The amount of reverberation in the room is modeled by its reverberation time r, a measure of the amount of time it takes for a sound to decay by 60 dB from its original level. The signal y_i received at microphone i is

$$y_i(t) = a_i s(t - \tau_i) + \epsilon_i(t) . \tag{1}$$

In (1), a_i is a scaling factor representing the attenuation suffered by the wave propagating from $\boldsymbol{p}_s$ to $\boldsymbol{m}_i$. We assume that the gains between the microphones are approximately calibrated, although we show in Section 4.3 that our method is robust to uncalibrated microphones of the same kind. τ_i is the time taken for a sound wave to propagate from the source to microphone i, and $\epsilon_i(t)$ models the noise. We assume τ_i to be equal to $\|\boldsymbol{m}_i - \boldsymbol{p}_s\|_2/c$, where $\|\boldsymbol{m}_i - \boldsymbol{p}_s\|_2$ is the Euclidean distance between the source and the microphone located at $\boldsymbol{m}_i$, c is the speed of sound and $\|\cdot\|_2$ represents the L_2-norm.

We also define $\boldsymbol{y}(t) = [y_1(t), \ldots, y_M(t)]^T$ as the vector containing all microphone signals at discrete time index t. The Short Time Fourier Transform (STFT) of $y_i(t)$ is $Y_i(\ell, f)$, for frequency f and time frame ℓ, and $\boldsymbol{Y}(\ell, f) = [Y_1(\ell, f), \ldots, Y_M(\ell, f)]^T$. The STFT [34] represents the frequency content of a signal over time, and is a widely used feature for source localization using neural networks [15, 19]. Figure 2 shows the magnitude representation of $\boldsymbol{Y}$ at the input.

Finally, the *metadata vector* $\boldsymbol{\phi} \in \mathbb{R}^{N_\phi}$ is the concatenation of the coordinates of the microphones, the room dimensions and reverberation time, as shown in Fig. 2. We chose the three aforementioned types of metadata as the room dimensions and microphone coordinates are explicitly exploited in classical localization methods
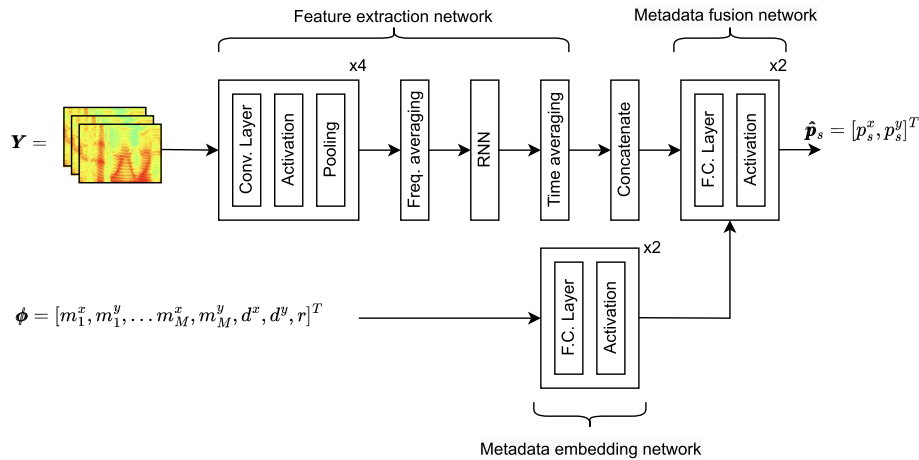
Fig. 2 Detailed DI-NN architecture for the task of PSSL

such as the LS. Furthermore, we included the reverberation time as an additional metadata to verify whether its knowledge can reduce the detrimental effect of reverberation in localization methods. However, other metadata could have been exploited such as the energy ratio between the microphone signals, or the absoption coefficients of the walls.

3.2 Proposed method: dual input neural network
Our proposed DI-NN architecture is comprised of two neural networks, a **feature extraction network** and a **metadata fusion network** as can be seen in Fig. 1. An additional third network, called the **metadata embedding network** is also used in the alternative DI-NN-Embedding network, which will be presented in Section 3.3 .

The input of the network consists of the STFTof the microphone signals as defined in Section 3.1. Instead of using the complex representation generated by the STFT, we split the real and imaginary parts of the STFT Y use them as separate channels as in [19], giving rise to $2 * M$ input channels. The role of the feature extraction network is to transform this high dimensional tensor into a one dimensional feature vector which compactly represents relevant information for the task in hand. In our experiments, we adopt a CRNN [35] as our feature extraction network, due to its wide adoption for SSL [7, 20, 36].

This metadata-unaware vector is then concatenated to the available metadata, thus creating a metadata-aware feature vector. For our application, the metadata is a one-dimensional vector consisting of the positions of the microphones, the dimensions of the room, and its reverberation time. This metadata-aware feature vector is then fed to a metadata fusion network, whose role is to merge the metadata and feature vector to produce the result. In

our experiments, we adopt a two-layer Fully Connected Neural Network (FC-NN) which maps the metadata-aware features to a two dimensional vector corresponding to the estimated coordinates of the source.

Our feature extractor CRNN is divided into two sequential sub-networks: a CNN block, responsible for extracting local patterns from the input data and a Recurrent Neural Network (RNN), responsible for combining these pattens into global, time-independent features. A diagram representing the components of the DI-NN network is shown in Fig. 2.

The convolutional block receives a tensor of shape (M, L, F) representing a multi-channel complex STFT, where M represents the number of audio channels, L represents the number of time frames generated by the STFT, and F is the number of frequency bins used. The role of this block is two-fold: firstly, to combine local information across all microphone channels, and secondly to reduce the dimensionality of the data to make it more tractable for the RNN layer.

The convolutional block consists of four sequential layers, where each performs three sequential operations. Firstly, a set of K convolutional filters is applied to the input signal, resulting in K output channels. Secondly, a non-linear activation function is applied to the result. Finally, an average pooling operation is applied to the width and height of the activations, generating an output of reduced size. After passing the input through the four convolutional layers, we perform a global average pooling operation across all frequencies, generating a two-dimensional output matrix.

After the convolutional block, the resulting matrix serves as input to a bidirectional, gated recurrent unit neural network (GRU-RNN) [37]. As sound may not be present throughout the whole duration of the audio
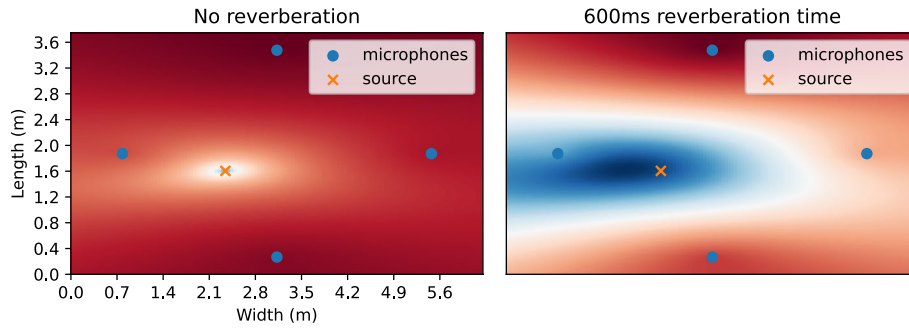
Fig. 3 Error grid produced by the LS algorithm for an anechoic and a reverberant room of the same dimensions and microphone coordinates

signal, such as during speech pauses, the RNN is important for propagating location information to silent time-steps. After this network, we reduce the dimensions of the features once again by performing average pooling on the time dimension, resulting in a vector of time-independent features.

The output of the feature extraction network are then concatenated to the available metadata and serve as input to the metadata fusion network. This network consists of a set of two fully connected layers which map the metadata-aware features to a two-dimensional vector corresponding to the estimated cartesian coordinates of the active source. We jointly train both networks using the same loss function, defined as the L_1-norm or the sum of the absolute error between the network's estimate of the source coordinates $\hat{\boldsymbol{p}}_s$ and the target $\boldsymbol{p}_s$, given by

$$\mathcal{L}(\boldsymbol{p}_s, \hat{\boldsymbol{p}}_s) = |\boldsymbol{p}_s - \hat{\boldsymbol{p}}_s| . \tag{2}$$

We also considered using the more common squared error loss. Although both losses yielded similar results in our experiments, we chose the absolute error for its easier interpretability, since it corresponds to the distance in metres between target and estimated coordinates.

3.3 DI-NN-Embedding

To test whether it is advantageous to process the metadata before combining it with the microphone features, we also propose a variant of the DI-NN model, where the metadata $\boldsymbol{\phi}$ is processed by a *metadata embedding network* to produce an embedding, which is then concatenated to the microphone features. This network is represented by the *metadata embedding network* block in Fig. 1.

3.4 Baseline: least-squares based source localization

Our final comparative baseline is the Least-Squares (LS) algorithm [1] which uses the signal model defined in Section 3.1. We provide an overview of the algorithm below.

We define the *theoretical* TDOA between microphones i and j with respect to the source coordinates $\boldsymbol{p}_s$ as

$$\tau_{ij}(\boldsymbol{p}_s) \triangleq \frac{\|\boldsymbol{m}_i - \boldsymbol{p}_s\|_2 - \|\boldsymbol{m}_j - \boldsymbol{p}_s\|_2}{c} , \tag{3}$$

where c is the speed of sound. Next, the *measured* TDOA between microphones $\hat{\tau}_{ij}$ is estimated from the cross-correlation peak between the received signals according to

$$\hat{\tau}_{ij} \triangleq \arg\max_t \left(\mathcal{C}(t; y_i, y_j) \right) , \tag{4}$$

where $\mathcal{C}$ denotes the cross-correlation operator, usually computed in the frequency domain using the GCC-PHAT algorithm [32]. We then aggregate the total error for all microphone pairs using

$$E(\boldsymbol{p}_s) \triangleq \sum_{i=1}^{m} \sum_{j \neq i} E_{ij}(\boldsymbol{p}_s) , \tag{5}$$

where $E_{ij}(\boldsymbol{p}_s) \triangleq |\tau_{ij}(\boldsymbol{p}_s) - \hat{\tau}_{ij}|^2$ is the squared difference between the theoretical and measured TDOA of each microphone pair in (3) and (4), respectively. To estimate the source's location, we compute the values of E for a set of candidate locations $\boldsymbol{p}_s$ within the room. In the absence of noise and reverberation, the location with the minimum error corresponds to the true position of the source [1]. Figure 3 shows the heatmaps or error grids generated using the LS algorithm in an anechoic and a reverberant room. The position of the source is estimated by selecting the positions that minimize the total error,

$$\hat{\boldsymbol{p}}_s = \arg\min_{\boldsymbol{p}_s} E(\boldsymbol{p}_s) . \tag{6}$$

Figure 3 illustrates the limitations of the LS algorithm when the reverberation time is large. The two figures show the results of our algorithm for two simulations, where one source and four microphones are placed in a room with the same dimensions. When the

room is simulated to be anechoic, i.e., all the reflections are absorbed, the algorithm produces a sharp blue peak in the heatmap. Conversely, when the simulated room is reverberant, the peak becomes much more dispersed. An explanation for this is that the model used by the LS method assumes anechoic propagation between the source and microphones, i.e., no reflections are assumed. Conversely, we will show that the DI-NN model is able to localize sources in reverberant environments, as it is trained using a reverberant dataset. A study conducted in [38] shows that speech inteligibility is maximized in rooms with a reverberation time between 0.4 and 0.5 ms, therefore limiting the practical application of the LS method on those environments.

4 Experimentation

This section describes our experiments with DI-NNs with three SSL datasets representing scenarios of varying difficulties. For each dataset, our approach is compared to two other methods. The first method is a CRNN with the same architecture but without using the available metadata, i.e., without the "Concatenate" block in Fig. 2. By comparing this network's performance to the DI-NN, we can see the performance gains of our proposed method. The second comparative method is the classical LS source localization method described in Section 3.4. The experiments will be described below.

All of our experiments consisted of randomly placing one source and four microphones within a room. The height of the microphones, source and room were fixed for all experiments. For each experiment, the goal of the proposed method and baselines was to estimate the planar coordinates of the source within the room using a one-second multichannel audio signal as well as the positions of the microphones. We emphasize that the training and testing samples do not overlap, and hence demonstrate our method's effectiveness for handling unseen scenes and metadata. We refer the reader to Appendix A for a discussion on the independence of our datasets.

To simulate sound propagation in a reverberant room, we used the image source method [39] implemented by the Pyroomacoustics Python library (MIT license) [40]. We trained our neural networks using PyTorch (BSD license) [41] along with the PyTorch Lightning (Apache 2.0 license) library [42]. The models were trained using a single NVIDIA P100 GPU with 16 GB of RAM memory. The configuration of our experiments is managed using the Hydra (MIT license) library [43]. We release the code used for generating the data and training the networks on GitHub[1],

Table 1 Hyperparameters

Parameter	Value
Num. parameters (DI-NN)	3.5M
Num. conv. kernels	64, 128, 256, 512
Conv. kernel size	2x2
Conv. layer pooling size	2x2
GRU output size	256
Metadata fusion net. layer out. sizes	$512 + N_\phi, 2$
Metadata embedding layer out. sizes	$2N_\phi, N_\phi$
Activation func. last layer	None
Activation func. other layers	Rectified Linear Unit (ReLU)
Num. Discrete Fourier Transform (DFT) bins (for STFT)	1024
DFT hop length (for STFT)	512
Input duration	0.5 secs.
Sampling rate	16kHz
Grid resolution of LS method	2 cm
Learning rate	0.0005
Batch size	32
Num. epochs	40
Batch normalization [44]	Only after conv. layers
Optimizer	Adam [45]

as well as a Kaggle notebook [2] to allow reproduction of the experiments without the need for any local software installation. The hyperparameters used for training the proposed method and baselines are shown in Table 1.

4.1 Simulated anechoic rooms

The goal of this experiment is to evaluate the performance of the DI-NN and baselines in multiple rooms and microphone positions in the absence of reverberation. Our dataset generation procedure is shown in Fig. 4a. For each dataset sample, we randomly select two numbers from a uniform distribution in the interval [3, 6] m representing the room's width and length. The height of the rooms is fixed at 3 m. Next, we randomly place one microphone along a line segment 0.5 m away and parallel to each room's walls. We chose to place the microphones close to the wall as a simplified localization scenario, as our main goal is to test the effectiveness of our metadata fusion procedure. Nonetheless, this scenario is realistic in the context of smart rooms, where the microphones are usually placed in or near the room's walls.

Finally, the source is randomly placed in the room, following a uniform distribution while respecting a minimum margin of 0.5 m from the walls. In this

---

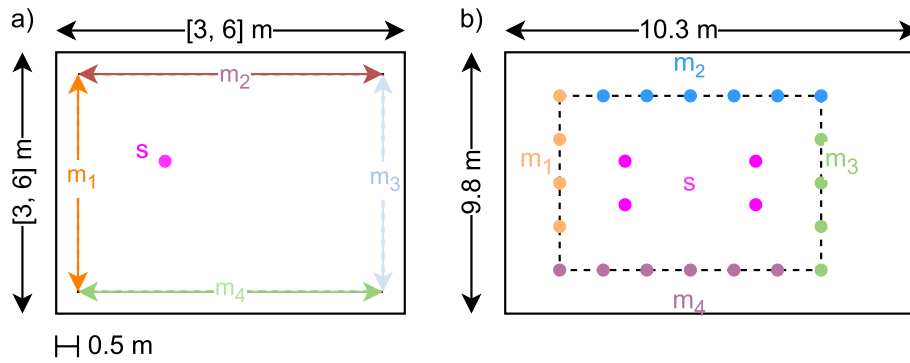[1] Code: https://github.com/egrinstein/di_nn

Fig. 4 Experimental setup. **a** For the anechoic and reverberant simulations, each of the four microphones m_i is placed on a random point along the the coloured arrows, while the source s is randomly placed on a point within the rectangle defined by them. **b** The sampling procedure for Section 4.3, where positions of the microphones and source are randomly drawn from each differently coloured set of points

experiment, the source signal is WGN, and 30 dB Signal-to-Noise Ratio (SNR) sensor noise, simulated using WGN, is also added to each microphone. A dataset of 15,000 samples is generated, from which 10,000 samples are used for training, 2,500 for validation, and 2,500 for testing.

4.2 Simulated reverberant rooms
The data for the simulated reverberant rooms experiment is generated similarly to the anechoic experiment. However, instead of simulating sound propagation in an anechoic environment, each dataset sample is randomly assigned a reverberation time value for its corresponding room from a uniform distribution within the range of $[0.3 - 0.6]$ s. This value is used to simulate reverberation using the image source method [39]. For the source signal, we use speech recordings from the VCTK corpus [46]. The number of training, testing and validation samples is same as in the above section.

4.3 Real recordings
For this experiment, instead of simulations, we use measurements from the LibriAdhoc40 dataset [47] (GPL3 license). The signals were recorded in a highly reverberant room containing a grid of forty microphones and a single loudspeaker, which was placed in one of four available locations. The microphones recorded speech sentences taken from the Librispeech [48] corpus, which were played back through the loudspeaker. The reverberation time measured by the dataset authors was of approximately 900 ms.

To generate each dataset sample, we subselect four of the forty available microphones. We restrict our microphone selection to the outermost microphones of the grid, where one microphone per side is selected. A visual explanation of our microphone selection procedure is

provided in Fig. 4b. There are four available positions for the microphones near each of the west and east walls and seven positions near each of the north and south walls. Furthermore, there are four available source positions. There are, therefore, $4 \times 4 \times 7 \times 7 \times 4 = 3{,}136$ source/ microphone combinations available for selection. Finally, we randomly select four speech utterances for each combination, resulting in a dataset of 12,544 samples. We use 50% of those combinations for training, 25% for validation and 25% for testing. To create the training dataset for this experiment, we augment the aforementioned training split with the training data of the reverberant dataset described in Section 4.2, resulting in a dataset consisting of $10{,}000 + 6{,}272 = 16{,}272$ signals.

4.4 Metadata sensitivity study
In practical scenarios, the metadata, e.g., microphone coordinates and room reverberation time in PSSL, are uncertain because they are typically estimated or measured. To investigate the robustness of our approach to such uncertainties, we conducted a sensitivity study using the test dataset in Section 4.2. We modify the dataset by introducing different levels of perturbations to the input metadata, followed by a computation of the mean localization error for each level using the model trained on Section 4.2.

Our first three studies consist of perturbing the microphone coordinates of the testing dataset with increasing levels of random Gaussian noise. The reported precision of microphone coordinates measured optically is under a millimeter [49]. Conversely, when these are estimated using self-localization algorithms, the reported errors are under 7 cm [50, 51]. We therefore choose the standard deviation levels of the introduced noise to 1, 10 and 50 cm. In our fourth study, we introduced random Gaussian noise to the reverberation time with a standard
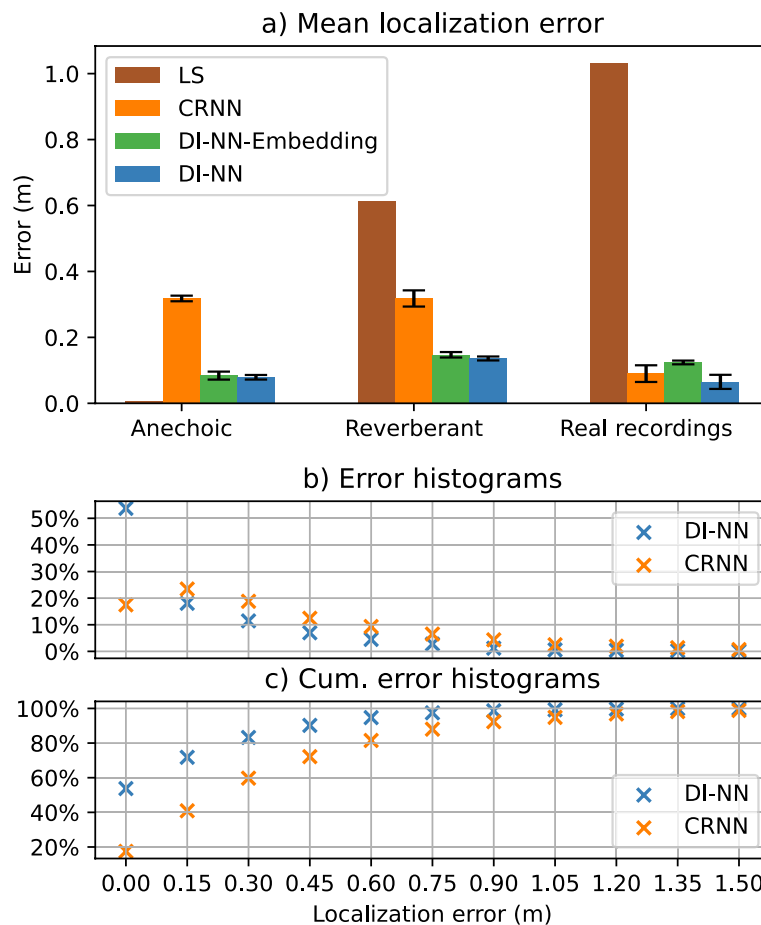
Fig. 5 **a** Mean localization error for DI-NNs and baselines on different datasets. **b** Normalized histogram comparison between the DI-NN and the CRNN baseline on the recorded dataset. **c** Cumulative version of (**b**)

deviation of 200 ms, based on reported errors obtained on reverberation estimation procedures [52, 53].

4.5 Metadata relevance study

To quantify the contribution of each metadata category to the improvement in localization performance, we conducted a metadata relevance study where we trained the DI-NN network using six different combinations of the microphone positions, room dimensions and reverberation time. The results are summarized in Table 3.

5 Results and discussions

5.1 Results

Figure 5a compares the average error of our proposed DI-NN and DI-NN-Embedding methods to the CRNN and LS baselines. To obtain statistically significant results, we train the DI-NN, DI-NN-Embedding and CRNN models four times independently for each experiment using random initial network parameters. The results shown in Fig. 5 are averaged across the four times,

with error bars showing the standard deviation across the runs. Conversely, as the LS method is deterministic, it does not require multiple runs.

A first remark is that although the LS approach is very effective in the anechoic scenario, its performance is degraded on the other datasets, indicating its sensitivity to reverberation. The CRNN outperforms the LS method in reverberant scenarios without knowledge of the microphone's coordinates. Interestingly, the CRNN baseline is also obtains good localization performance on the recorded dataset, indicating that the network is able to infer the metadata to an extent when trained on a single room.

However, by exploiting the microphone coordinates, the DI-NN is shown to significantly improve the performance compared to the CRNN. The most significant difference is observed in the anechoic case, where an improvement close to three times is obtained. In this case, the microphone coordinates are more useful as this information cannot be derived from the signals.

Table 2 Metadata sensitivity analysis

Coord. std. (m)	Reverb. std. (ms)	Err. increase (%)
0.01	0	0.05
0.1	0	1.02
0.5	0	32.9
0	200	0.4

Table 3 Metadata relevance analysis

Mic. coords.	Room dims.	RT60	% performance
✓	✓	✓	100
✓	✓	✗	102
✓	✗	✓	100
✗	✓	✓	61
✓	✗	✗	104
✗	✓	✗	60
✗	✗	✓	47

In a reverberant room, however, the network might be able to use reflections to its advantage, as discussed in [54], to infer the microphone coordinates and making the metadata less useful. Figure 5a also shows the errors obtained using the alternative DI-NN-Embedding architecture were similar to the DI-NN in all scenarios, indicating no advantage in the proposed embedding, although it still allows the network to exploit the metadata.

In turn, Fig. 5b compares the normalized error histograms between our approach and the CRNN baseline on the real recordings test dataset. The mode of the DI-NN's error is centred on the 0-15 cm bin compared to the 15-30 cm bin for CRNN's error. In other words, only the DI-NN is median-unbiased. The cumulative distribution for the same data is shown in Fig. 5c. While the DI-NN is shown to locate over 50% and 80% of the dataset samples with less than 15 and 45 cm error, the CRNN achieves the same errors for less than 20% and 60% of the data, respectively.

The results of the sensitivity study conducted in Section 4.4 are displayed in Table 2. The last column refers to the relative error increase between the perturbed case and the noiseless experiment conducted in Section 4.2. The results show that our approach is robust to the uncertainty inherent in practical measurements of the microphone coordinates and reverberation time estimates. The case where the microphone coordinates are disturbed by an extreme error of 0.5 m (more than five times above typical errors) has been included to demonstrate the impact of including microphone

coordinates for PSSL, reiterating the importance and improved performance of metadata in our proposed fusion approach.

Finally, the results of the metadata relevance analysis study described in Section 4.5 are displayed in Table 3. Each line represents a version of the DI-NN model trained on the reverberant dataset. The first three columns describe which metadata types are used in the model, and the last column shows the model performance relative to the model using all metadata, represented in the first line. The results show that the microphone coordinates are the most relevant for the model. In fact, using the microphone coordinates alone provides the best results. The results also indicate that the room dimensions are more relevant than the reverberation time in the absence of the microphone coordinates.

5.2 Limitations and extensions
Our approach exploits the metadata, such as the microphone coordinates and reverberation time and therefore this data must be known a priori or somehow measured. We have, however, shown that using this additional information is justified by a significant improvement in performance. While we have also assumed that the gains of the microphones are calibrated in our experiments, which may not be verifiable in practical scenarios, we have shown in Section 4.3 that our model can perform well even when using uncalibrated microphones of the same kind. If calibration cannot be ensured, extracting gain invariant features from the signal pairs such as the cross spectra [16] may be used as a preprocessing step.

We have also limited our scope to the localization of one static sound source using static microphones to focus on metadata fusion. However, extensions to moving sources and microphones could be possible by using smaller processing frames, for example. Another extension would be to estimate the three dimensional coordinates of the source. Finally, a possible extension for multiple source localization is expanding the output of DI-NN to a vector of size $2N$, where N is the number of maximum sources, and performing Permutation Invariant Training (PIT) [55].

6 Conclusion
In this work, we proposed DI-NN, a simple yet effective way of jointly processing signals and relevant metadata using neural networks. Our results for the task of SSL on multiple simulated and recorded scenarios indicate that the DI-NN is able to exploit successfully the metadata, as its inclusion reduced the mean localization error by a factor of at least two compared to the CRNN baseline, as well as significantly improving localization results in comparison with the classical LS algorithm in
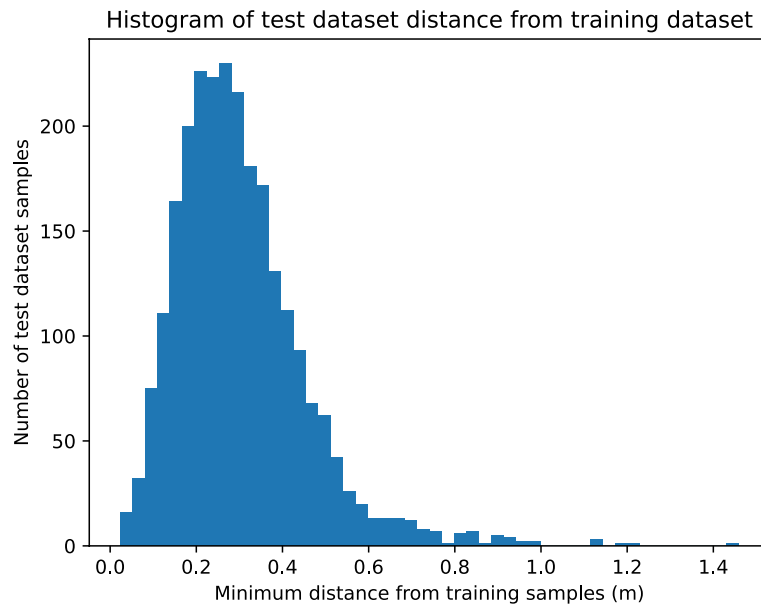
Fig. 6 Distance between test dataset's microphone coordinates and training dataset

reverberant environments. Additional relevance and sensitivity studies revealed that the microphone coordinates the most important metadata, and that the DI-NN is robust to realistic noise in the metadata.

Appendix
Validation of metadata independence between training and testing datasets
The datasets used in Sections 4.1 and 4.2 are created entirely synthetically by generating random training, validation and testing samples. The attributes generated for each sample are the room's width and length, the coordinates of the four microphones, and the source coordinates. Additionally, in Section 4.2, the room reverberation time is also randomly sampled. These values are then used to simulate the microphone recordings using the image source method. The only difference in the procedure for generating the training and testing sets is the random seed used for sampling values. Although highly unlikely, generating a test sample with the exact room dimensions, reverberation time, microphone and source coordinates as a sample in the training set could be possible and would violate the machine learning principle of having independent training and testing sets.

To assure the reader that this has not occurred in our experiments, we compute a distance metric D between each testing sample and the entire training dataset. We focus on comparing the microphone coordinates between the training and testing sets and show that our approach has been validated against unseen metadata. Each sample comprises four microphone coordinates, each placed near one of the room's walls. We define the distance $d(i, j)$ between the i-th testing sample and j-th training sample as the sum of the distances of the microphone coordinates between the samples given by

$$d(i,j) = \sum_{k=1}^{4} \|\boldsymbol{m}_i^k - \boldsymbol{m}_j^k\|_2 \,, \tag{7}$$

where $\boldsymbol{m}_i^1$, $\boldsymbol{m}_i^2$, $\boldsymbol{m}_i^3$ and $\boldsymbol{m}_i^4$ refer to the coordinates of the microphones located near the north, south, east and western walls of the room from the i-th sample and $\|\cdot\|$ denotes the L_2-norm.

To measure the distance between the i-th testing sample and the entire training dataset, we compute (7) for every training sample j. We define the smallest distance $D(i)$ between the i-th testing sample and the entire training set as the minimum distance between i and all training samples j, expressed as

$$D(i) = \min_j \{d(i,j)\} \,. \tag{8}$$

This measure quantifies the worst-case similarity between the i-th testing sample and the most similar sample in the entire training set. By plotting a histogram of $D(i)$ for every i-th sample in the testing set, we observe in Fig. 6 that no training microphone configuration appeared in the testing set. Moreover, the average

minimum distance between the testing and training sets is around 30 cm. Besides having different microphone coordinates, we like to emphasize that the room dimensions and reverberation time also vary from sample to sample, increasing the differences between training and testing sets even further.

Declarations

Competing interests
The authors declare that they have no competing interests.

References
1. H.C. So, *Source Localization: Algorithms and Analysis* (Wiley, Hoboken, 2011)
2. D.B. Haddad, M.V.S. Lima, W.A. Martins, L.W.P. Biscainho, L.O. Nunes, B. Lee, Acoustic Sensor Self-Localization: Models and Recent Results. Wirel. Commun. Mob. Comput. 2017, e7972146 (2017). https://doi.org/10.1155/2017/7972146
3. M. Brandstein, D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications* (Springer Science & Business Media, Berlin, 2001)
4. H. Wang, P. Chu, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. Voice Source Localization for Automatic Camera Pointing System in Videoconferencing (IEEE, USA, 1997), pp. 187–190
5. C. Evers, P.A. Naylor, *Acoustic SLAM. IEEE Trans. Audio Speech Lang*. Process, vol. 26 (IEEE, USA, 2018) p. 1484–1498
6. A. Bertrand, in *IEEE Symp. on Commun. and Veh. Technol. in the Benelux (SCVT)*. Applications and Trends in Wireless Acoustic Sensor Networks: A Signal Processing Perspective (IEEE, USA, 2011), pp. 1–6
7. S. Adavanne, A. Politis, T. Virtanen, in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*. Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network (IEEE, USA, 2018), pp. 1462–1466
8. W. He, P. Motlicek, J.M. Odobez, in *Proc. Int. Conf. Robotics and Automation*. Deep Neural Networks for Multiple Speaker Detection and Localization (IEEE, USA, 2018), pp. 74–79
9. J.M. Vera-Diaz, D. Pizarro, J. Macias-Guarasa, Towards End-to-End Acoustic Localization using Deep Learning: From Audio Signal to Source Position Coordinates. Sensors. 18(10), 3418 (2018)
10. P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, D. Whiteson, Parameterized Neural Networks for High-Energy Physics. Eur. Phys. J. C. 76(5), 235 (2016)
11. P.K. Atrey, M.A. Hossain, A. El Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: A survey. Multimed. Syst. 16(6), 345–379 (2010)
12. T. Baltrušaitis, C. Ahuja, L.P. Morency, Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41(2), 423–443 (2019)
13. A. Ozerov, E. Vincent, F. Bimbot, A General Flexible Framework for the Handling of Prior Information in Audio Source Separation. IEEE Trans. Audio, Speech, Language Process. 20(4), 1118–1133 (2012)
14. H. Sundar, W. Wang, M. Sun, C. Wang, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources (IEEE, USA, 2020), pp. 4642–4646
15. A. Anonymous, in *Under Review*. Deep Complex-Valued Convolutional-Recurrent Networks for Single Source DOA Estimation (2022)
16. W. Xue, Y. Tong, C. Zhang, G. Ding, X. He, B. Zhou, in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*. Sound Event Localization and Detection Based on Multiple DOA Beamforming and Multi-Task Learning (ISCA, France, 2020), pp. 5091–5095
17. S. Chakrabarty, E.A.P. Habets, in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*. Broadband DoA Estimation Using Convolutional Neural Networks Trained with Noise Signals (2017), pp. 136–140
18. N. Yalta, K. Nakadai, T. Ogata, Sound Source Localization Using Deep Learning Models. J. Robot. Mechatron. 29(1), 37–48 (2017)
19. D. Krause, A. Politis, K. Kowalczyk, in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*. Comparison of Convolution Types in CNN-based Feature Extraction for Sound Source Localization (IEEE, USA, 2021), pp. 820–824
20. L. Perotin, R. Serizel, E. Vincent, A. Guérin, CRNN-Based Multiple DoA Estimation Using Acoustic Intensity Features for Ambisonics Recordings. IEEE J. Sel. Topics Signal Process. 13(1), 22–33 (2019)
21. L. Perotin, A. Défossez, E. Vincent, R. Serizel, A. Guérin, in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*. Regression Versus Classification for Neural Network Based Audio Source Localization (IEEE, USA, 2019), pp. 343–347
22. P.A. Grumiaux, S. Kitić, L. Girin, A. Guérin, A Survey of Sound Source Localization with Deep Learning Methods. J. Acoust. Soc. Am. 152, 107–151 (2022)
23. M. Taseska, E.A.P. Habets, in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*. Spotforming using distributed microphone arrays (IEEE, USA, 2013)
24. E. Guizzo, C. Marinoni, M. Pennese, X. Ren, X. Zheng, C. Zhang, B. Masiero, A. Uncini, D. Comminiello, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. L3DAS22 Challenge: Learning 3D Audio Sources in a Real Office Environment (2022). Comment: Accepted to 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022). arXiv admin note: substantial text overlap with arXiv:2104.05499
25. F. Gustafsson, F. Gunnarsson, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. Positioning Using Time-Difference of Arrival Measurements (IEEE, USA, 2003)
26. D. Li, Y.H. Hu, Energy Based Collaborative Source Localization Using Acoustic Micro-Sensor Array. EURASIP J Appl. Signal Process. 985029 (2003)
27. Z. Liu, Z. Zhang, L.W. He, P. Chou, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. Energy-Based Sound Source Localization and Gain Normalization for Ad Hoc Microphone Arrays (IEEE, USA, 2007)
28. R.O. Schmidt, Multiple Emitter Location and Signal Parameter Estimation. IEEE Trans. Antennas Propag. 34(3), 276–280 (1986)
29. J.P. Dmochowski, J. Benesty, S. Affes, A Generalized Steered Response Power Method for Computationally Viable Source Localization. IEEE Trans. Audio Speech Lang. Process. 15(8), 2510–2526 (2007)
30. R. Lebarbenchon, E. Camberlein, D. di Carlo, C. Gaultier, A. Deleforge, N. Bertin, *in Proc. of the LOCATA Challenge Workshop.* Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge. (2018).

31. Y. Huang, J. Benesty, G. Elko, R. Mersereati, Real-Time Passive Source Localization: A Practical Linear-Correction Least-Squares Approach. IEEE Trans. Audio Speech Lang. Process. 9(8), 943–956 (2001)
32. C. Knapp, G. Carter, The Generalized Correlation Method for Estimation of Time Delay. IEEE Trans. Acoust. Speech Signal Process. 24(4), 320–327 (1976)
33. C. Zhang, D. Florencio, Z. Zhang, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. Why Does PHAT Work Well in Lownoise, Reverberative Environments? (IEEE, USA, 2008), pp. 2565–2568
34. J.B. Allen, Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform. IEEE Trans. Acoust. Speech Signal Process. 25(3), 235–238 (1977)
35. K. Choi, G. Fazekas, M. Sandler, K. Cho, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. Convolutional recurrent neural networks for music classification (IEEE, USA, 2017), pp. 2392–2396
36. Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, M.D. Plumbley, in *Proc. Detect. and Classific. of Acoust. Scenes and Events (DCASE)*. Polyphonic Sound Event Detection and Localization using a Two-Stage Strategy (2019), pp. 30–34
37. J. Chung, C. Gulcehre, K. Cho, Y. Bengio, in *Proc. Neural Inform. Process. Conf.* Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling (NeurIPS Foundation, USA, 2014)
38. S. Bistafa, J. Bradley, Reverberation time and maximum background-noise level for classrooms from a comparative study of speech intelligibility metrics. J. Acoust. Soc. Am. 107, 861–75 (2000). https://doi.org/10.1121/1.428268
39. J.B. Allen, D.A. Berkley, Image Method for Efficiently Simulating Small-Room Acoustics. J. Acoust. Soc. Am. 65(4), 943–950 (1979)
40. R. Scheibler, E. Bezzam, I. Dokmanić, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms (IEEE, USA, 2018), pp. 351–355
41. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, et al., in *Proc. Neural Inform. Process. Conf.* Pytorch: An imperative style, high-performance deep learning library (2019)
42. W. Falcon, The PyTorch Lightning Team. PyTorch Lightning. (2019). https://www.pytorchlightning.ai. Accessed 28 Aug 2023
43. O. Yadan. Hydra - A Framework for Elegantly Configuring Complex Applications. (2019). https://www.hydra.cc. Accessed 28 Aug 2023
44. S. Ioffe, C. Szegedy, in *Proc. Int. Conf. Machine Learning (ICML)*. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift (PMLR, USA, 2015), pp. 448–456
45. D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. (2017). arXiv:1412.6980
46. J. Yamagishi, C. Veaux, K. MacDonald. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (2019). University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2645.
47. S. Guan, S. Liu, J. Chen, W. Zhu, S. Li, et al., in *Asia-Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf. (APSIPA)*. Libri-Adhoc40: A Dataset Collected from Synchronized Ad-Hoc Microphone Arrays (IEEE, USA, 2021)
48. V. Panayotov, G. Chen, D. Povey, S. Khudanpur, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. Librispeech: An ASR Corpus Based on Public Domain Audio Books (IEEE, USA, 2015), pp. 5206–5210
49. A.M. Aurand, J.S. Dufour, W.S. Marras, Accuracy map of an optical motion capture system with 42 or 21 cameras in a large measurement volume. J. Biomech. 58, 237–240 (2017)
50. N.D. Gaubitch, W.B. Kleijn, R. Heusdens, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (IEEE, USA, ICASSP)*. Auto-localization in ad-hoc microphone arrays (2013), pp. 106–110
51. P. Pertilä, M. Mieskolainen, M.S. Hämäläinen, in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*. Passive self-localization of microphones using ambient sounds (IEEE, USA, 2012), pp. 1314–1318
52. H. Gamper, I.J. Tashev, in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*. Blind Reverberation Time Estimation Using a Convolutional Neural Network (IEEE, USA, 2018), pp. 136–140
53. P.S. López, P. Callens, M. Cernak, in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*. A Universal Deep Room Acoustics Estimator (IEEE, USA, 2021), pp. 356–360
54. F. Ribeiro, D. Ba, C. Zhang, D. Florêncio, in *IEEE International Conference on Multimedia and Expo*. Turning enemies into friends: Using reflections to improve sound source localization (IEEE, USA, 2010), pp. 731–736
55. D. Yu, M. Kolbæk, Z.H. Tan, J. Jensen, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Permutation invariant training of deep models for speaker-independent multi-talker speech separation (IEEE, USA, 2017), pp. 241–245. https://doi.org/10.1109/ICASSP.2017.7952154

Publisher's Note