Neural Network Adaptation and Data Augmentation for Multi-Speaker Direction-of-Arrival Estimation

Weipeng He, *Student Member, IEEE,* Petr Motlicek, *Senior Member, IEEE,* and Jean-Marc Odobez, *Member, IEEE*

***Abstract*—Deep neural networks have been successfully applied to sound direction-of-arrival estimation under challenging conditions. However, such a learning-based approach requires a large amount of labeled training data, which is difficult to acquire. To address this problem, we propose a novel approach for multi-speaker direction-of-arrival estimation with data augmentation and weakly-supervised domain adaptation. We generate source domain data with simulation, and collect real data annotated with the number of sound sources as the weak labels. The real data are further augmented by mixing single-source segments. Then, weakly-supervised domain adaptation is applied to models pre-trained on the simulated data. We define a loss function for the adaptation process which exploits the weak labels and the mixture component information in the augmented data. Experiments with real robot audio data show that our proposed approach achieves similar performance as if the fully-labeled real data are used. This paper suggests an effective development procedure for DOA estimation models applied to new types of microphone arrays with minimal data collection efforts.**

***Index Terms*—DOA estimation, sound source localization, weakly-supervised learning, data augmentation**

I. Introduction

SOUND Source Localization (SSL) is the technology of estimating the locations or the Direction-of-Arrival (DOA) of sound sources from the signals captured by the audio sensors. It is an important research topic with many applications. For instance, it is an essential part of audio perception in Human-Robot Interaction (HRI), and it provides prerequisite information for speech enhancement methods, which are widely used in smart speakers, video conferencing systems, and hearing aids. This paper introduces a novel deep learning framework for SSL, which includes data collection and learning procedures that effectively reduce the burden of real data acquisition.

A. Motivation and Challenges

The studies of sound source localization have started with signal processing approaches [1–3], which frame the problem in mathematical forms based on the physical law of sound propagation, and are built around analytical solutions. These solutions rely on assumptions about the acoustic environments, which may include known transfer functions or steering vectors, free-field anechoic sound propagation, high Signal-to-Noise Ratio (SNR), spatial white noise, or a known number of sources. However, these assumptions may not hold well in real-world applications. For example, the transfer functions are not exactly known in practice due to the error in measuring the microphone placement and obstacles that can change the sound propagation (e.g. robot head). Furthermore, there are often multiple simultaneous sound sources in the environments. The discrepancy between assumptions and reality may lead to significant performance degradation. Sophisticated modeling of the complex environments may mitigate the problem, but it is not clear how to generalize it as exhaustive modeling of all environments is not possible.

As an alternative, researchers have recently proposed learning-based approaches that build machine learning models from training samples to avoid explicit sound propagation modeling. These approaches rely on probabilistic models [4, 5] or neural network models [6–9]. In particular, the neural network based approaches have been shown to handle strong background noise, reverberation and multiple speakers. Furthermore, neural networks can be extended to jointly solve other related tasks, such as speech/non-speech classification [10, 11] and sound event detection [12].

In the learning-based approaches, the difficulties have been shifted from modeling the complex environments in the signal processing approaches to the need of collecting sufficient number of training data covering all variabilities in the test environment. Such variabilities include various sound classes, samples per class, source locations, reverberation, noises, and solid objects in the scenes. In addition to making audio recordings, annotating them with the ground truth labels is also particularly costly. The complexity of annotation is due to the fact that audio data do not intrinsically contain direct information for researchers to annotate the sound source locations. The annotation requires complimentary sensor data to be recorded during data collection. These include sensor data from cameras [5, 9], motion capture systems [13], and robot motor sensor [14, 15]. Moreover, since multi-channel audio data are distinct among different types of microphone arrays, individual target-domain data collection is needed for each new type of microphone array.

One possible solution to avoid costly data collection is to develop device-independent SSL models, allowing real audio data to be reused for multiple devices. This is a difficult problem and little research has been conducted in this direction. Models using uniform input representation, such as the ambisonics intensity vectors [16, 17], could potentially be applied to multiple devices. However, this idea has not been verified by experiments and the conversion of multi-channel audio data to ambisonics intensity vectors is limited to non-coplanar microphone arrays.

W. He, and J.-M. Odobez are with Idiap Research Institute, Martigny, Switzerland and EPFL, Lausanne, Switzerland. P. Motlicek, is with Idiap Research Institute, Martigny, Switzerland. E-mail: {weipeng.he, petr.motlicek, odobez}@idiap.ch

A popular way of obtaining training data for sound source localization is by simulating artificial audio data. The most commonly used room acoustic simulation methods [18, 19], however, only handle over-simplified room settings [20, 21]. Recently, advanced simulation techniques considering sound reflection and scattering caused by solid objects (e.g. robot head or other objects in rooms) have been proposed [17, 22–24]. Nevertheless, in practice it is still difficult to measure and simulate the surfaces of the complex solid objects, thus simulation cannot perfectly reproduce the directivity and frequency response patterns of the real microphone arrays. Although we can obtain realistic impulse responses through measurement [25], their availability is almost restricted to binaural sound localization [26, 27].

As an alternative to synthesizing audio data, there has also been research using autonomous mobile robots for data collection in new and unconstrained environments. These robots are equipped with cameras as well as microphones. They autonomously record audio data and use the images captured by the cameras to annotate the sound source locations. The annotation from images either follows fixed data collection procedures [28] or is achieved by self-supervised learning [29]. However, these approaches are limited to some specific robots.

Finally, domain adaptation, which uses both simulated and real data, may be applied to SSL. It exploits the large variety of conditions from the simulated data and reduces the discrepancy between simulation and reality with the help of available real data. Although there are many studies on domain adaptation theory [30], there are only a few applications of it to sound source location. Previous studies have investigated the unsupervised adaptation of neural networks for single-source sound source localization with entropy minimization [31, 32]. However, such a principle can only be applied to classification problems, which is not suitable for multi-source localization. In our previous work [33], we applied domain adaptation to multi-source DOA estimation, and showed that using the numbers of active sound sources as weak labels for the real data we achieve much better performance than using simulated data alone. However, there is still a gap of performance between the weakly-supervised and supervised approaches, because when the initial model prediction is not reliable, the weak supervision may not generate positive change on the models.

B. Goals and Contributions

As we have mentioned in the previous section, one main challenge of learning-based approaches is acquiring suitable training data. In order to address this problem, this paper proposes a framework for multi-source DOA estimation with deep neural networks. The framework includes data collection at low cost and training models using domain adaptation (Fig. 1). First, we generate a large number of simulated data using sets of clean speech and background noise. These data serve as the source domain data and are used to pre-train a deep Convolutional Neural Network (CNN) model. Then, we collect a relatively small set of real audio data, of which only the number of sounds sources is manually labeled (as weak labels), thus the high cost of exact location annotation is avoided. These real audio data are also augmented by mixing single-source frames. Lastly, we adapt the pre-trained model to the real and augmented data in a weakly-supervised fashion.

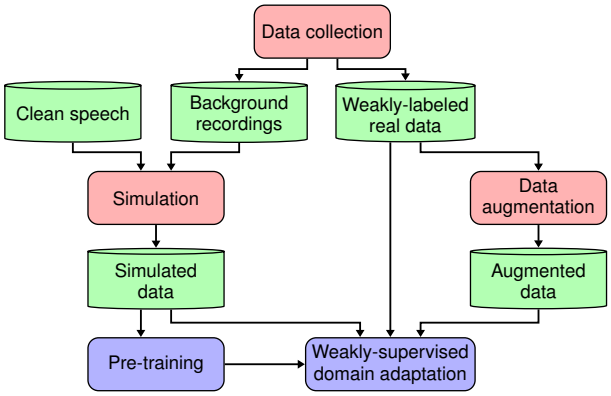


Fig. 1: Overview of our framework of neural-network-based multi-speaker DOA estimation with weakly-supervised domain adaptation. The arrows indicate which datasets (green) are required or generated by data preparation procedures (red), and which datasets are used for the training processes (blue).

We evaluated the effectiveness of the proposed framework on two versions of Pepper robots[1] in HRI settings. Our experiments show that our method achieves comparable performance to the supervised case (where the real data are fully-labeled). The contributions of this paper are:

- We propose a multi-source DOA estimation framework with domain adaptation so that the data collection workload can be significantly reduced.
- We propose a weakly-supervised adaptation scheme that minimizes the distance in the output coding space between the network output and all the predictions consistent with the weak labels.
- The weakly-supervised adaptation scheme is extended through data augmentation. This extension significantly improves the performance of the weakly-supervised adaptation.

This paper is an extension of our previous work [33]. It includes an improved version of the weakly-supervised adaptation scheme, a large and more comprehensive set of experiments using data of two robots, and quantitative analysis of the adaptation scheme.

The rest of the paper is structured as follows: Section II reviews the related work. Section III introduces the proposed neural network for multi-speaker DOA estimation including training in the supervised learning setting. Section IV introduces the proposed domain adaptation approaches. Experiments are described and discussed in Section V, and we conclude in Section VI.

II. Related Work

We review the state-of-the-art neural-network-based sound source localization approaches, and the domain adaptation methods which are closely related to our work.

[1] http://doc.aldebaran.com/2-5/home_pepper.html

A. Neural-network-based Sound Source Localization

These approaches build neural network models to approximate the mapping between the audio signal and the sound source locations. The model parameters are optimized with a set of training data, and the model is expected to make predictions on unseen data. Different approaches differ in their input representation, output coding as well as their network structures.

Previous methods have used various high-level input features, including inter-channel time (phase) difference [34, 35], inter-channel level difference [35], MUSIC eigenvectors [8, 36], GCC-PHAT coefficients [7, 9], or GCC-PHAT coefficients on filter bank [9]. With such high-level localization cues relatively simple neural networks can be used as the mapping functions. Nevertheless, more recent studies have shown that low-level signal representation without explicit feature extraction, whether in the time [37] or time-frequency [11, 12, 38, 39] domains, can allow the networks to learn to extract the most informative high-level features for SSL.

The output coding defines how labels are encoded into ideal network outputs, and how the network outputs are decoded into labels. It plays a significant role in a method. Single sound source localization is commonly considered as a classification problem, where the network output is interpreted as a posterior distribution of the classes that corresponds to sound locations or directions [7, 8, 37, 40], and additionally silence [8]. Such an approach can be extended with some difficulties to the localization of up to two sound sources by using the marginal posterior distributions of the sound sources [36], but to handle an arbitrary number of sound sources, the spatial spectrum coding we have introduced in [9] is a more suitable approach (see also [41]). It is explained in detail in Section III.

Although there have been an increasing number of studies on deep learning based sound source localization, few of them clearly address issues related to the high cost of data collection, especially by applying domain adaptation to models trained with simulated data.

B. Domain Adaptation

Domain adaptation explores how the knowledge from a dataset (source domain) can be exploited to help build machine learning models on another set (target domain) [30]. It is often applied to scenarios with abundant source data and limited target data. Domain adaptation approaches include re-weighting samples so that the loss function on the source samples are corrected to approximate that on the target domain [42]. Another way is to construct a common representation space, so that a model can be used for inference on both domains. This can be achieved through aligning correlation on the data from both domains [43], or using domain-adversarial training [44]. Moreover, pseudo-labeling — generating labels by applying the model to unlabeled data — is becoming a popular approach for domain adaptation [45, 46]. In addition, unlabeled data can also be used for entropy regularization of the model [47]. Nevertheless, few domain adaptation methods have been applied to SSL.

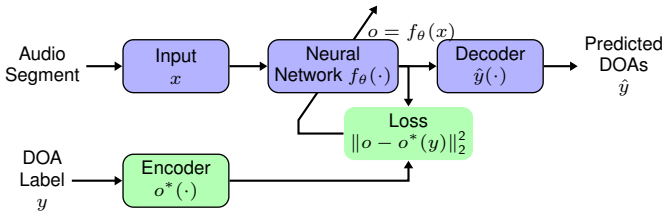


Fig. 2: Overview of our neural-network-based approach for multi-speaker direction of arrival estimation. The top part (blue) represents the prediction process, whereas the bottom part (green) indicates the supervised learning principle.

III. DOA Estimation Model

In this section, we introduce our neural network model for DOA estimation. Although we present it in the supervised learning setting, the same network model will be used in other settings throughout our experiments.

A. Overview

We aim to build a model parameterized by θ that approximates the mapping from an input audio signal in the input space X to location labels in the label space Y. Since we consider estimating the directions of an arbitrary number of sound sources, each label $y \in Y$ is a set of locations in Φ, including the empty set. That is:

$$Y = \{y \subset \Phi : |y| < \infty\},$$

where $|y|$ is the cardinality of y (i.e. the number of sources), and Φ is the set of target candidate locations or DOAs.

The model consists of three parts: feature representation, output coding, and network architecture (Fig. 2). The neural network maps the features to an output space O, from where the network output can be decoded into sound location labels. During the training process, we are given a set of labeled samples:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \subset X \times Y,$$

and the neural network parameters θ are trained to minimize a loss function $\mathcal{L}$ on these samples:

$$\theta^* = \arg\min_\theta \mathop{\mathbf{E}}_{(x,y)\in D} \mathcal{L}\left(f_\theta(x), y\right), \tag{1}$$

where $f_\theta(x)$ is the output of the network. We describe in the following the details of the model in term of input features, output coding and network structure.

B. Network Input

The network input comprises the real and imaginary parts of the time-frequency domain signal. In contrast to high-level features extraction, such a representation retains all the information of the signal and allows the network to implicitly extract informative features for localization, which potentially include both inter-channel cues (i.e. level/phase difference) and intra-channel cues (i.e. spectral features). In addition, as speech is known to be sparse in the time-frequency domain, with such a representation the network can learn to separate

overlapping sound sources in the mixed input signal. Using multiple input features (including both low-level and high-level features) may potentially improve the performance, as it has been shown in recent studies on sound event localization and detection [48, 49]. However, more feature extraction requires more computation cost in data pre-processing. As the domain adaptation scheme studied in this paper can be applied to any network design, other input representations are not studied.

Specifically, we prepare the network input as follows: We first divide the 4-channel audio into 170 ms long segments (8192 samples in 48 kHz recordings). This segment size provides a good balance between the amount of information and the time resolution. In addition, such a short input segment is suitable for real-time applications, as it takes 5 ms for our neural network to process an input segment of 170 ms on an NVIDIA GTX 1080 Ti GPU. We compute the Short-Time Fourier Transform (STFT) of the segments with a frame size of 43 ms (2048 samples) and 50% overlap. Thus, there are seven frames in each segment. We only use the frequency bins between 100 and 8 kHz, so that the number of frequency bins is reduced to 337. We take the real and imaginary part of the complex values instead of the phase and power, so that we avoid the discontinuity problem of the phase at π and $-\pi$. Eventually, the dimension of the input vector is $7 \times 337 \times 8$.

C. Output Coding

We use spatial spectrum coding to handle an arbitrary number of sound sources [9]. The spatial spectrum is a function of the DOA ($o : \Phi \to \mathbb{R}$), and its value indicates how likely there is a sound source for a given DOA. Unlike signal processing approaches, where the aim is to find the analytical solution for the spatial spectrum, our approach trains models to approximate an ideal spatial spectrum that we can arbitrarily define. Thus, the localization problem becomes a spatial spectrum regression problem.

In practice, the network outputs a vector $\mathbf{o} = \{o_l\}_{l=1}^{L}$ that indicates values of the spatial spectrum on the sampled directions $\{\phi_l\}_{l=1}^{L} \subset \Phi$, where l is the index of the DOA. In our experiments, $\{\phi_l\}$ are 360 evenly-spaced azimuth directions. We define the ideal spatial spectrum of a label y as the maximum of Gaussian curves centered at the sound source directions (Fig. 3).

$$o^*(y)_l = \begin{cases} \max\limits_{\phi' \in y} \left\{ e^{-d(\phi_l, \phi')^2 / \sigma^2} \right\} & \text{if } |y| > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where $d(\cdot, \cdot)$ is the angular distance, and σ is a constant that controls the width of the Gaussian curves. The ideal output values are close to zero at directions away from the sound sources. They peak at the ground truth directions with a value of one, and gradually decrease to zero as the distance to the sound source increases. Such a "soft assignment" design takes the uncertainty of the estimation into account.

During inference, the network output is decoded to the prediction in Y by finding the peaks in the predicted spatial spectrum. When the number of sources z is unknown, the peaks above a given threshold ξ are taken as predictions:

$$\hat{y}(o; \xi) = \left\{ \phi_l : o_l > \xi \quad \text{and} \quad o_l = \max_{d(\phi_i, \phi_l) < \sigma_n} o_i \right\}, \quad (3)$$

where $o = f_\theta(x)$ is the network output and σ_n is the neighborhood size for non-maxima suppression. When z is known, the z highest peaks are taken as predictions:

$$\hat{y}(o; z) = \left\{ \phi_l : \text{among the } z \text{ greatest } o_l = \max_{d(\phi_i, \phi_l) < \sigma_n} o_i \right\}. \quad (4)$$

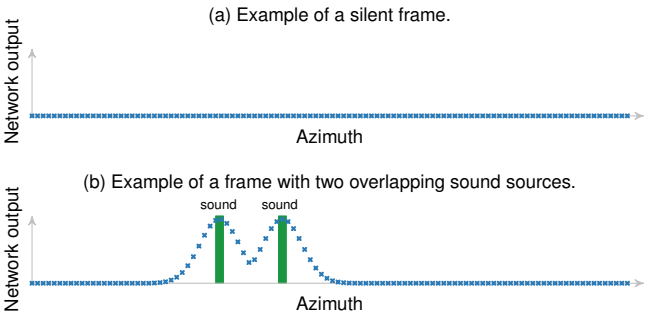


Fig. 3: Output coding for multiple sources.

D. Network Architecture

We design a fully-convolutional neural network structure for DOA estimation (Fig. 4). CNN facilitates weight sharing for deep neural network models, thus reducing the overall number of parameters as well as the risk of overfitting. CNN on the time-frequency domain signal has been proven effective for sound source localization [9, 17, 33, 39]. Recent studies have also suggested using recurrent structures in the network (e.g. recurrent CNN), so it can leverage the context information [16, 48, 49]. However, recurrent structure may introduce additional computational cost. Moreover, the CNN output as sound source detection results can be used as input for separate Recurrent Neural Networks (RNNs) or temporal filters to incorporate longer context information. Therefore, recurrent structures are not studied in this paper.

Our network comprises two parts, which convolve along different axes. In the first part, the network convolves along the time and frequency axes. Specifically, it includes two layers of strided convolution in the frequency axis for downsampling as well as feature extraction, five residual blocks for the extraction of higher level features, and a layer projecting the features to the DOA space. The residual connection allows the construction of very deep neural network models, and therefore increases their capabilities at extracting high-level features [50]. The output of the first part of the network is time-frequency local, meaning that each output value is derived from a local time-frequency region of the input.

In the second part, the network convolves along the DOA axis. It aggregates features in the neighboring directions across all time-frequency bins, and outputs the spatial spectrum.

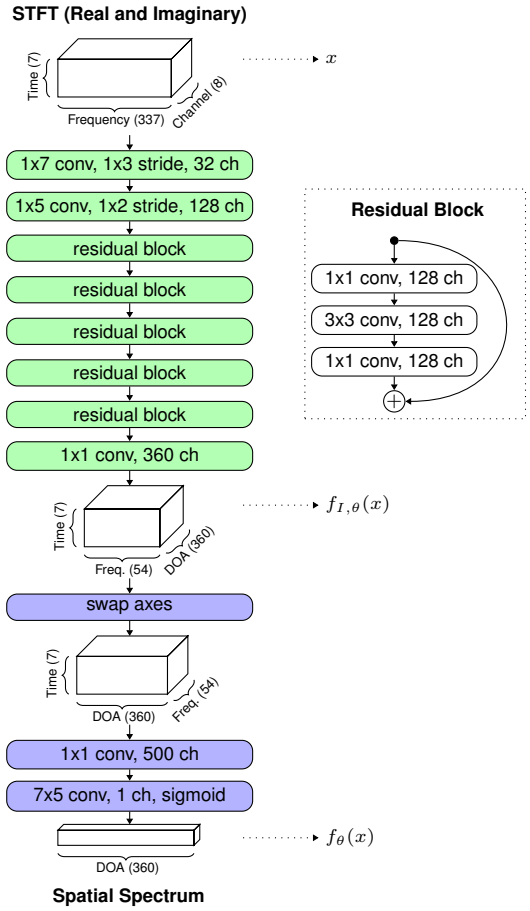


Fig. 4: Neural network architecture for multi-speaker DOA estimation. It uses STFT of the audio signals as input and predicts the spatial spectrum of the sound sources. It consists of two parts: the first part (green) applies convolution along the time and frequency axes, and the second part (blue) applies convolution along the DOA axis.

E. Two-stage Training

The goal of training is to make the network regress the ideal spatial spectrum with the Mean Squared Error (MSE) loss:

$$\mathcal{L}(f_\theta(x), y) = \|f_\theta(x) - o^*(y)\|_2^2 . \tag{5}$$

This is achieved in two stages. In the first stage, we train the first part of the network, by considering its output as the short-term narrow-band predictions of the spatial spectrum. The loss function for the first stage is replicating the ultimate loss function (Eq. 5) across time and frequency:

$$\mathcal{L}_I\left(f_{I,\theta}(x), y\right) = \sum_{t,k} \mathcal{L}\left(f_{I,\theta}(x)[t,k], y\right), \tag{6}$$

where $f_{I,\theta}(x)[t,k]$ is the output of the first part of the network at time t and frequency k. The pre-trained parameters are then used to initialize the network for the second stage where the whole network is trained with the loss function Eq. 5.

Previous experiments have shown that the two-stage training is necessary, as the network is deep and directly training it from scratch is prone to local optima [11].

IV. Domain Adaptation

In this section, we discuss the domain adaptation methods for the proposed neural network architecture for DOA estimation. The idea of domain adaptation is to train a model using both simulated (source domain) and real (target domain) data so that the model has the best performance in real test scenarios. As it is costly to collect real data while generating simulated data at a large scale is cheap, we implicitly assume that the latter cover more variabilities than the real dataset.

A. Supervised Adaptation

We first consider the supervised domain adaptation in which we are given a set of labeled simulated data $D_s \subset X \times Y$, together with a set of labeled real audio data $D_t \subset X \times Y$. To apply supervised domain adaptation, we first use the simulated data to pre-train a model, which is the initialization of the subsequent optimization processes. Then, we train a model that minimizes the loss on both the source domain and the target domain:

$$\theta^* = \arg\min_\theta \mu_t \mathop{\mathbf{E}}_{(x,y)\in D_t} \mathcal{L}\left(f_\theta(x), y\right) + \mu_s \mathop{\mathbf{E}}_{(x,y)\in D_s} \mathcal{L}\left(f_\theta(x), y\right), \tag{7}$$

where μ_t and μ_s are the weighting parameters for the loss on the two domains. In practice, the weighting is implemented by changing the proportion of source and target domain samples in each mini-batch. The added loss term relying on source domain data can reduce the bias caused by data insufficiency in the target domain.

B. Weakly-Supervised Adaptation

Although we can reduce the number of real samples with supervised domain adaptation, annotation of the real samples still requires a heavy workload. Therefore, we propose a weakly-supervised adaptation scheme to further reduce the effort for data collection. In the weakly-supervised adaptation setting, instead of fully-labeled data D_t, we are given a set of weakly-labeled real data:

$$D_w = \{(x_i, z_i)\}_{i=1}^{N_w} \subset X \times Z.$$

accompanied by a set of fully-labeled simulated (source domain) data D_s. Each value z_i from the weak label domain Z indicates the number of sources in the audio frame x_i.

We apply the adaptation by minimizing a weak supervision loss $\mathcal{L}_w$ on the target domain as well as the supervised loss (Eq. 5) on the source domain:

$$\theta^* = \arg\min_\theta \mu_w \mathop{\mathbf{E}}_{(x,z)\in D_w} \mathcal{L}_w\left(f_\theta(x), z\right) + \mu_s \mathop{\mathbf{E}}_{(x,y)\in D_s} \mathcal{L}\left(f_\theta(x), y\right), \tag{8}$$

where μ_w and μ_s are weighting parameters. We propose the *minimum distance adaptation* scheme by defining the weak supervision loss as the minimum distance in the output space between the network output and all possible labels that satisfy the weak label:

$$\mathcal{L}_w(f_\theta(x), z) = \min_{y\in r(z)} \|f_\theta(x) - o^*(y)\|_2^2 , \tag{9}$$

where $o*(\cdot)$ is the output encoding defined by Eq. 2, and $r(z)$ is the set of all sound DOA labels that satisfy the weak label z, i.e. the number of sources in y is z:

$$r(z) = \{y \in Y : |y| = z\}.$$

The heuristic is that among all the models that predict well the DOAs on the source domain, we favor those that can make correct detection of the number of sources in the target domain.

The weak supervision can also be viewed as a *pseudo-labeling* approach. The loss function L_w can be rewritten as:

$$\begin{aligned}\mathcal{L}_w\left(f_\theta(x), z\right) &= \left\| f_\theta(x) - o^*(\arg\min_{y \in r(z)} \|f_\theta(x) - o^*(y)\|_2^2) \right\|_2^2 \\ &= \mathcal{L}\left(f_\theta(x), \arg\min_{y \in r(z)} \|f_\theta(x) - o^*(y)\|_2^2\right) \\ &= \mathcal{L}\left(f_\theta(x), p_\theta(x, z)\right),\end{aligned} \tag{10}$$

with

$$p_\theta(x, z) = \arg\min_{y \in r(z)} \|f_\theta(x) - o^*(y)\|_2^2 \tag{11}$$

as the pseudo-labeling function. We can see that the weak supervision loss function is equivalent to the supervised loss if $p_\theta(x, z)$ is used as the label.

Furthermore, we can visualize the pseudo-labels in the output space to see how the weak supervision works (Fig. 5). When the number of sources is zero, the network is supervised to output zero, thus reducing the false positives caused by unseen noise (Fig. 5a). When the number of sources is one or more, the network is supervised to give more certain prediction on the most prominent peaks, thus increasing the recall (Fig. 5c). At the same time, the other peaks that are caused by unseen conditions are suppressed (Fig.5b, c). However, the effectiveness of the weakly-supervised adaptation depends on the initial performance of the network model. If the network initial output is too far away from the ground truth, the weak supervision will lead to incorrect pseudo-labels (Fig.5d, e).

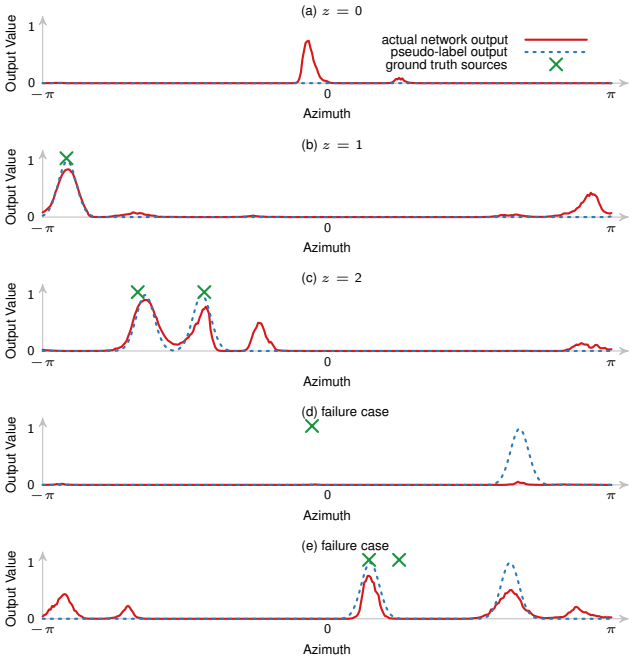


Fig. 5: Examples of weak supervision with a known number of sources on real audio segments. The ground truth locations are shown but are not used for weak supervision.

C. Pseudo-labeling with Data Augmentation

In practice, we observe that the network trained on simulated data initially performs worse on the multi-source audio segments as illustrated in Fig. 5e. Thus, in order to increase the correctness of the pseudo-labeling on multi-source audio frames, we augment real data by generating mixture data with known single-source components, and extend the weak supervision method using a modified pseudo-labeling approach. The idea is that we apply the pseudo-labeling to the easier single-source components rather than to the multi-source mixtures, so that we can obtain more effective weak supervision.

Data augmentation. The augmented mixture dataset D_a consists of a set of mixture x_i and their single-source components $\mathbf{u}_i = \{u_{ij}\}_{j=1}^{z_i}$:

$$D_a = \{(x_i, \mathbf{u}_i)\}_{i=1}^{N_a} \subset X \times 2^X.$$

Here, the mixtures are generated by linear combination:

$$x_i = \sum_{j=1}^{z_i} \alpha_{ij} u_{ij}, \tag{12}$$

where $\{u_{ij}\}$ are single-source segments randomly sampled from the weakly-labeled dataset D_w, z_i is the number of components (sources), and $\{\alpha_{ij}\}$ are random scaling factors. Since all the real recordings include background noise, mixing real audio segments may change the characteristics of the background noise. To minimize the discrepancy between the generated mixtures and real recordings, we choose the scaling factors in a way that the power of the background noise in the generated mixtures is equal to that of the real recordings, that is:

$$\sum_{j=1}^{z_i} \alpha_{ij}^2 = 1. \tag{13}$$

Here, we assume the power of the background noise in the real recordings is constant and the segments are mutual independent.

A benefit of such data augmentation is that it increases the number of realistic multi-source segments, which is difficult to obtain by recording. In addition, as the combinations of sound directions increases exponentially with the number of sources, we need a large number of multi-source training samples to cover such variabilities.

Pseudo-labeling on components. The other benefit is that the knowledge of the single-source components allows us to apply reliable pseudo-labeling on this dataset: we first apply pseudo-labeling (Eq. 11) to its single-source components, that are $p_\theta(u_{ij}, 1), j = 1 \ldots z_i$ (Fig. 6a,b). Then, we use the union of these pseudo-labels for the multi-source frame (Fig. 6c). Thus, the loss function of the modified adaptation is:

$$\mathcal{L}_a(f_\theta(x_i), \mathbf{u}_i) = \mathcal{L}\left(f_\theta(x_i), \cup_{j=1}^{z_i} p_\theta(u_{ij}, 1)\right), \tag{14}$$

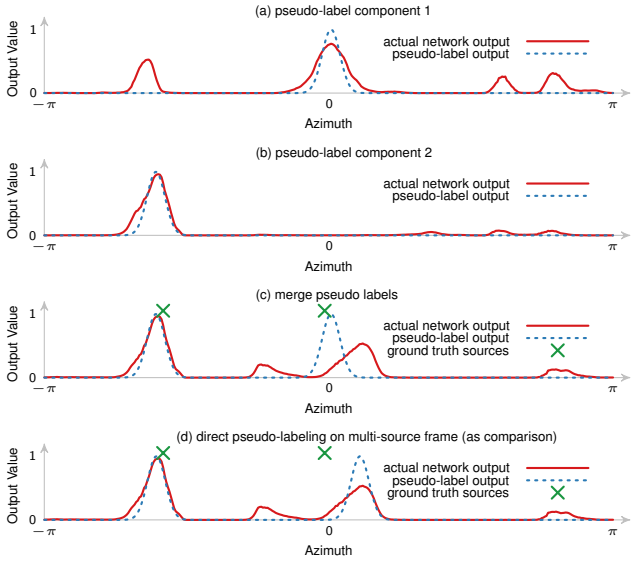


Fig. 6: Example of weak supervision from mixture components of an augmented multi-source frame. (a, b) The pseudo-labeling is applied first on the single-source components. (c) Then the pseudo-label of the two-source mixture is obtained by merging the pseudo-labels of its components. This approach is more reliable than directly applying the pseudo-labeling to the mixture as shown in (d).

and the optimization target becomes:

$$\theta^* = \arg\min_{\theta} \mu_a \mathop{\mathbf{E}}_{(x,\mathbf{u})\in D_a} \mathcal{L}_a\left(f_\theta(x), \mathbf{u}\right) + \mu_w \mathop{\mathbf{E}}_{(x,z)\in D_w} \mathcal{L}_w\left(f_\theta(x), z\right) + \mu_s \mathop{\mathbf{E}}_{(x,y)\in D_s} \mathcal{L}\left(f_\theta(x), y\right), \tag{15}$$

where μ_a controls the weight of the modified weak-supervision loss on the augmented dataset.

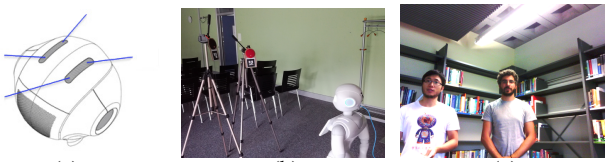
V. Experiments

We applied the proposed approach with both simulated data and real data from two robots, and we verified its effectiveness in two ways — by analyzing the correctness of pseudo-labeling and evaluating the performance of DOA estimation.

A. Microphone Array and Data

Microphone array. We used the Pepper robot in our experiments. There are four microphones placed on the robot head in a rectangular shape with a dimension of 5.80 cm by 6.86 cm (Fig. 7a). Our experiments involved two versions of the robots, which differ in the microphone directivity pattern: directional and omni-directional. We use *P1* and *P2* to denote the two versions respectively.

Source-domain data. We generated the source domain data by convolving clean speech audio with simulated room impulse responses (Table I). The room impulse responses are simulated with the RIR-Generator [21]. The clean audio speech data were the close talking recordings randomly selected from the AMI corpus [51]. We first generated spatialized audio data of single speech in cuboid rooms of random sizes. Both the microphone array and the sound source were randomly placed in the room. The distances between the microphone array, the sound source and the walls were at least 0.5 m. The microphone array geometry was set according to that on the robot. We tried to simulate both omni-directional and cardioid directivity patterns of the microphones, and found out that the models trained with omni-directional simulation have in general better performance, even for the robot P1, whose actual microphones are directional. We hypothesize that this is because the simulation cannot replicate exactly the directivity patterns of the real microphones. Therefore, we used the omni-directional simulation for both robots throughout our experiments. Then the single-source simulated audio frames are mixed randomly at runtime with other frames as well as the real robot background recordings. During mixing, there was no constraint on the distances between the sound sources. We generated one million mixture frames (47 hours), the number of sources of which varies from zero to four. This includes a significant number of source locations and audio content for training. We experimented with both anechoic and reverberant room conditions. For the reverberant simulation, the reverberation time (RT60) is randomly selected between 200 and 800 ms. The absorption coefficients of all six walls in a room are the same, and are set according to the selected reverberation time as well as the size of the room.



(a) (b) (c)

Fig. 7: (a) Microphone placement in the robot head. (b) The loudspeaker recording scenes. (c) Camera view from the robot during human talker recordings.

TABLE I: Specifications of the simulated data.

	Simulation P1 & P2
Total duration	47 hours
# of frames	1 million
- subset($z=0$)	200k
- subset($z=1$)	400k
- $subset(z=2)$	300k
- subset($z=3$)	50k
- subset($z=4$)	50k
# of male speakers	105
# of female speakers	43
SNR (dB)†	10
Azimuth (°)	[-180, 180.0]
Elevation (°)	[-74, 75]
Distance (m)	[0.5, 10.8]
Room length (m)	[8.0, 12.0]
Room width (m)	[6.0, 9.0]
Room height (m)	[2.0, 5.0]
RT60 (ms)*	[200, 800]

†Average SNR of single-source frames.
*RT60 values apply only to the reverberant simulation.

Target-domain data. We collected real data with the robots. For P1, we used the public SSLR dataset[2] from [9], which was collected in a semi-automatic fashion (Table II). It includes recordings of loudspeakers for both training and evaluation (denoted by *P1-LSP*), as well as an additional evaluation set with human talkers (denoted by *P1-HUM.E*). For the loudspeaker data recordings, we played clean speech audio on loudspeakers from various locations (Fig. 7b). These recordings were prepared using the same set of selected utterances from the AMI corpus [51] as the simulated data. During each piece of recording the sound source locations are fixed, therefore the coverage in terms of source locations in the real recordings is considerably less than that of the simulated data. For the human talker data, we recorded participants speaking in front of the robot. This test set is an extended version of the human talker data used in [9] and [33]. In this extension, we added recordings of human voices with overlapping sounds from a loudspeaker, and collected audio data with up to four overlapping sound sources. The sounds played from loudspeaker include non-speech sounds (from Audio Set [52]) in addition to speech (from the AMI courpus). Both P1-LSP and P1-HUM.E were collected in various ordinary office rooms. The reverberant time (RT60) of these rooms ranges from 400 to 800 ms.

The locations of the sound sources were automatically labeled for both the training and evaluation data using the video data captured by the camera on the robot. Specifically, we marked the loudspeakers with QR code, so that their locations can be detected. For human talkers, we applied the human pose detection with convolutional pose machines [53] to detect the nose locations, which indicates the azimuth of the speaker directions. For P2, we conducted the data collection in the same way and obtained a set of loudspeaker data (denoted by *P2-LSP*, Table III).

The weak labels were derived from the sound source location labels, and used for the weakly-supervised approaches. The fully-labeled training data allowed us to analyze quantitatively the effectiveness of the weak supervision, and compare the weakly-supervised approaches to supervised ones.

TABLE II: Specifications of the target-domain data for P1.

	P1-LSP		P1-HUM.E
	Training	Evaluation	Evaluation
Total duration	16 hours	8 hours	12 minutes
# of frames	507k	262k	7410
- subset($z=0$)	106k	54k	1538
- subset($z=1$)	350k	179k	2740
- subset($z=2$)	51k	29k	2212
- subset($z=3$)	—	—	636
- subset($z=4$)	—	—	284
# of male speakers	105	8	19
# of female speakers	43	8	2
SNR (dB)†	5	7	5
Azimuth (°)	[-180, 180]	[-180, 180]	[-82, 133]
Elevation (°)	[-39, 56]	[-29, 45]	[-14, 14]
Angular separation (°)*	[13, 149]	[18, 90]	[9, 147]
Distance (m)	[0.5, 1.8]	[0.5, 1.9]	[0.8, 2.8]
RT60 (ms)	[400, 800]	[400, 800]	[400, 800]

†Average SNR of single-source frames, estimated by assuming constant background noise power.
*Angular separation in azimuth between overlapping sound sources.

TABLE III: Specifications of the target-domain data for P2.

	P2-LSP	
	Training	Evaluation
Total duration	3.8 hours	2.1 hours
# of frames	122k	67k
- subset($z=0$)	26k	14k
- subset($z=1$)	90k	46k
- subset($z=2$)	6k	7k
# of male speakers	101	8
# of female speakers	41	8
SNR (dB)†	6	5
Azimuth (°)	[-180, 180]	[-178, 180]
Elevation (°)	[-39, 56]	[-29, 48]
Angular separation (°)*	[29, 126]	[28, 128]
Distance (m)	[0.5, 1.8]	[0.9, 2.0]
RT60 (ms)	[400, 800]	800

†Average SNR of single-source frames, estimated by assuming constant background noise power.
*Angular separation in azimuth between overlapping sound sources.

B. Training Parameters

According to the proposed framework (Fig. 1), we first pre-trained a model on the simulated data using the two-stage training. The model was trained for one epoch in the first stage (Eq. 6) and four epochs on the second stage (Eq. 5). Then the pre-trained model was used as the initial model for the weakly-supervised domain adaptation. We controlled the weights of the components in the optimization target Eq. 15 to be $\mu_w = 0.9$, $\mu_a = 0.1$, and $\mu_s = 1.0$. This is equivalent as composing mini-batches using 45%, 5% and 50% of the samples from the weakly-labeled dataset, augmented dataset, and the simulated dataset, respectively. We used a learning rate of 0.001 and reduced it by half once the training loss no longer decreased. We continued the adaptation until the plateau was reached four times. During all the training processes, the models were optimized with the Adam optimizer [54] and a mini-batch size of 100.

[2]https://www.idiap.ch/dataset/sslr

C. Analysis of Pseudo-Labeling

To better understand the minimum distance adaptation on weakly-labeled data, we analyzed how the effectiveness of the pseudo-labeling depends on the initial model performance and the number of overlapping sources. Our expectation is that good pseudo-labels will have a positive impact on the learned model if the pseudo-labels are on average closer to the ground truth than the actual model predictions are. Therefore, we computed the loss gain between the MSE loss (Eq. 5) of the model prediction and that of the pseudo-label:

$$\Delta_L = \mathcal{L}\left(f_\theta(x), y\right) - \mathcal{L}\left(o^*(p_\theta(x,z)), y\right), \qquad (16)$$

where y and z are, respectively, the location label and the weak label corresponding to the audio segment x. A positive loss gain indicates the pseudo-labeling is beneficial for the model.

We applied the minimum distance adaptation (Eq. 9) to the pre-trained model on the target-domain data, and computed the

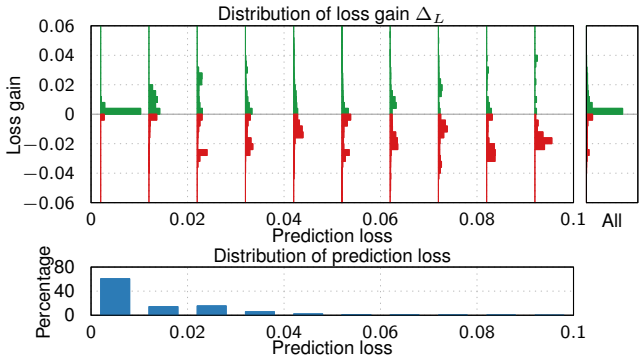


Fig. 8: Analysis of the minimum distance adaptation of all samples in the P1 training data. Top figure: Each histogram (plotted vertically) shows a distribution of the loss gain (Eq. 16) on the samples with the indicated prediction loss and on all samples (right-most histogram). The green bars indicate positive gain (correct weak supervision), while the red bars indicate negative gain (incorrect weak supervision). Bottom figure: The distribution of the initial prediction loss. The network is pre-trained with the anechoic simulation data.

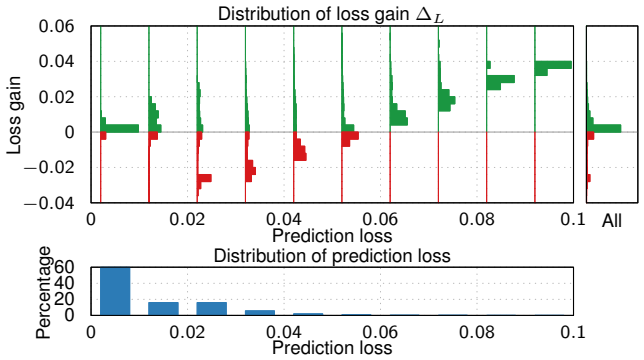


Fig. 9: Analysis of the minimum distance adaptation of the single-source samples from the P1 training data.

distributions of the loss gain (Eq. 16) on samples with different prediction loss. Result (Fig. 8) shows that weak supervision is mostly correct when the prediction loss is small (below 0.02), and becomes unreliable as the prediction loss increases.

By comparing the loss gain distributions on single-source (Fig. 9) and multi-source samples (Fig. 10), we can verify the assumption that weak supervision is more reliable on single-source frames. The pre-trained model initially performs better on the single-source frames. Moreover, even on the single-source frames with large prediction loss the weak supervision is more likely to generate correct pseudo-labels.

We also compared the minimum distance adaptation (Eq. 9) to its modified version relying on mixture components (Eq. 14) on the mutli-source augmented data (Fig. 11 and 12). Since the modified adaptation relies on pseudo-labels of the single-source components, it generates more reliable results than the direct application of pseudo-labeling on the multi-source frames. Even when the initial prediction loss is larger than 0.02, the pseudo-labels are more likely to have positive gain.

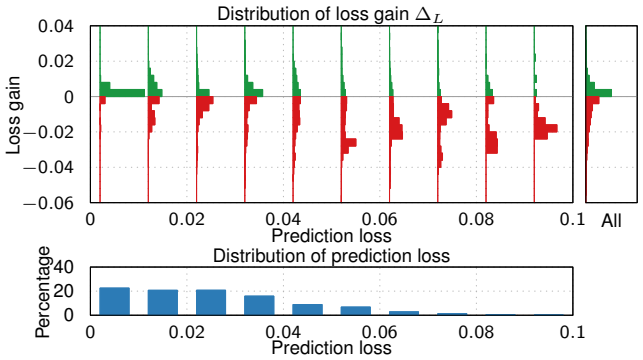


Fig. 10: Analysis of the minimum distance adaptation of the multi-source samples from the P1 training data.

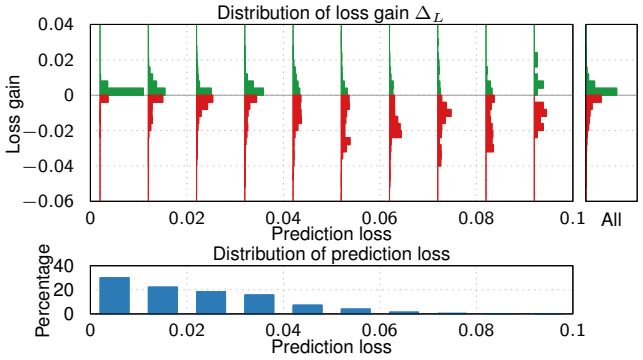


Fig. 11: Analysis of the minimum distance adaptation on the P1 augmented data.

D. DOA Estimation Evaluation Protocol

We evaluated the DOA estimation performance of the proposed approaches and compared it with other baseline methods as listed in Section V-E. The neural network models were trained on fully-labeled simulated data, weakly-labeled (for weakly-supervised approaches) or fully-labeled (for supervised approaches) real data, and augmented data if applicable. Their performance is evaluated on the test set of the real data.

We considered two evaluation settings: (a) when the number of sound sources is known, or (b) when it is not. When the number of sources is known, we evaluate how close the predicted DOAs are from the ground truth. In this case, the predictions $\hat{y}_i = \{\hat{\phi}_{ij} : j = 1, \ldots, z_i\}$ are the DOAs of the z_i (number of sources) highest peaks in the output spatial spectrum (according to Eq. 4). The indices js are selected such that the predicted DOA $\hat{\phi}_{ij}$ is nearest to the ground truths DOA ϕ_{ij} in label $y_i = \{\phi_{ij} : j = 1, \ldots, z_i\}$. As performance measure, we compute the Mean Absolute Error (MAE) in terms of angular distance between the predictions and the ground truth:

$$\text{MAE} = \frac{\sum_i \sum_{j=1}^{z_i} d(\hat{\phi}_{ij}, \phi_{ij})}{\sum_i z_i}. \tag{17}$$

We also compute the Accuracy (ACC), that is the percentage of the predictions of which the error is less than a given

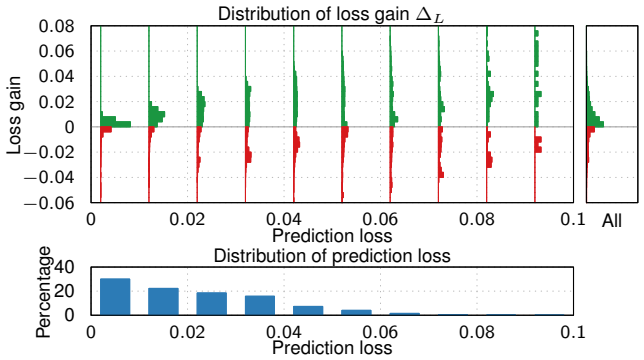


Fig. 12: Analysis of the modified adaptation on the P1 augmented data.

admissible error E_a:

$$\text{ACC} = \frac{\sum_i \sum_{j=1}^{z_i} \mathbf{1}_{d(\hat{\phi}_{ij},\phi_{ij})<E_a}}{\sum_i z_i}, \tag{18}$$

where $\mathbf{1}$ is the indicator function.

When the number of sources is unknown, we evaluate the DOA estimation in terms of sound source detection. The predictions $\hat{y}_i = \{\hat{\phi}_{ik} : k = 1, \ldots, \hat{z}_i\}$ decoded from the network output by Eq. 3, are matched with the ground truth DOAs. We use $m(\hat{\phi}_{ik}, \phi_{ij})$ to denote a match. The number of predicted sound sources $\hat{z}_i$ may not be equal to the number of ground truth sources z_i, and each ground truth source is matched with at most one prediction (could be none), which is the nearest prediction with an error less than E_a:

$$m(\hat{\phi}_{ik}, \phi_{ij}) = \begin{cases} 1 & \text{if } d(\hat{\phi}_{ik}, \phi_{ij}) < E_a \quad \text{and} \\ & k = \arg\min_{k=1}^{\hat{z}_i} d(\hat{\phi}_{ik}, \phi_{ij}), \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

We vary the prediction threshold ξ (see Eq. 3) and plot the precision vs. recall curve. The precision is the percentage of the correct predictions among all predictions:

$$\text{Precision} = \frac{\sum_i \sum_{j=1}^{z_i} \sum_{k=1}^{\hat{z}_i} m(\hat{\phi}_{ik}, \phi_{ij})}{\sum_i \hat{z}_i}. \tag{20}$$

The recall is the percentage of the correct detections out of all ground truth sources:

$$\text{Recall} = \frac{\sum_i \sum_{j=1}^{z_i} \sum_{k=1}^{\hat{z}_i} m(\hat{\phi}_{ik}, \phi_{ij})}{\sum_i z_i}. \tag{21}$$

We chose $E_a = 5°$ for the evaluation.

E. Compared Methods

The following approaches were included for comparison:

- **SRP-PHAT** steered response power with phase transform [3].
- **SUPREAL** fully-supervised approach described in Section III using only fully-labeled real data for training (two-stage training with loss functions Eq. 5 and 6).
- **SUPSIM** The model trained with only the simulated data. This is also the pre-trained model for the domain adaptation approaches.

TABLE IV: MAE(°) and ACC(%) on the P1-LSP dataset. Performance is evaluated on different subsets: all frames, single-source frames and two-source frames. The source-domain data are simulated with two different room conditions (anechoic and reverberant).

Dataset	P1-LSP					
Subset	All		$z = 1$		$z = 2$	
	MAE	ACC	MAE	ACC	MAE	ACC
SRP-PHAT	20.9	79.4	17.6	83.9	41.0	51.5
SUPREAL	3.0	93.9	2.6	95.5	5.6	83.9
Anechoic Sim.						
SUPSIM	13.1	80.2	11.6	82.4	22.6	66.4
ADSUP	3.3	93.8	2.7	95.0	7.1	86.2
ADWEAK	7.9	86.1	3.8	91.4	33.2	53.8
ADPROP	4.5	93.0	3.3	94.7	12.2	82.8
Reverb. Sim.						
SUPSIM	11.7	85.5	10.0	88.2	22.7	69.2
ADSUP	3.8	93.8	3.1	95.4	7.9	84.4
ADWEAK	8.6	85.0	4.5	90.3	33.4	52.8
ADPROP	5.2	92.1	3.8	94.3	14.2	78.8

- **ADSUP** The supervised adapted model, i.e. pre-trained with the simulated data and then adapted using the fully-labeled real data in a supervised fashion (Eq. 7).
- **ADWEAK** The weakly-supervised adapted model without using augmented data, i.e. pre-trained with the simulated data and then adapted using the weakly-labeled real data with the minimum distance adaptation scheme (Eq. 8).
- **ADPROP** the proposed weakly-supervised adaptation approach, i.e. first pre-trained with simulated data and then adapted using the weakly-labeled real data and augmented data with the adaptation scheme (Eq. 15).

We experimented with both anechoic and reverberant simulation for the methods that use simulation.

F. DOA Estimation Results

We applied these approaches to both the robots P1 and P2, and evaluated them on their respective test sets. We report their performance on single-source and two-source frames, as well as the overall performance on all test frames.

Learning-based vs SRP-PHAT. From the performance of the approaches on the P1-LSP data (Table IV and Fig. 13), we see that all learning-based approaches outperform the traditional SRP-PHAT. Because there is strong background noise in the robot audio data, the SRP-PHAT method, which assumes the target signal is dominant across all frequencies, is more affected. The learning-based approaches, on the other hand, learn from the training samples to implicitly suppress the noise.

Simulation vs Real Data. Comparing the models trained with simulated data (SUPSIM) to those trained with real data (SUPREAL), we see the expected performance degradation caused by the discrepancy between the acoustic simulation and real recordings.

Supervised Adaptation. The model first pre-trained with simulated data and then adapted with fully-labeled real data (ADSUP) achieves similar performance as that directly trained

TABLE V: MAE(°) and ACC(%) on the P1-HUM.E dataset. The source-domain data are simulated with anechoic condition.

Dataset	P1-HUM.E									
Subset	All		$z=1$		$z=2$		$z=3$		$z=4$	
	MAE	ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE	ACC
SRP-PHAT	15.4	69.0	3.9	88.4	19.8	57.9	34.9	42.0	48.0	30.3
SUPREAL	10.1	82.9	3.9	93.0	10.5	79.7	19.2	69.6	47.3	39.8
SUPSIM	12.3	76.5	6.4	87.3	14.4	72.6	21.8	61.1	32.2	36.3
ADSUP	11.0	84.9	4.3	93.4	12.0	83.5	19.7	74.1	48.9	37.4
ADWEAK	20.4	73.0	5.3	91.5	27.4	63.2	39.2	46.9	69.2	29.8
ADPROP	12.5	83.2	4.6	91.8	13.7	81.9	23.3	71.7	54.4	34.9

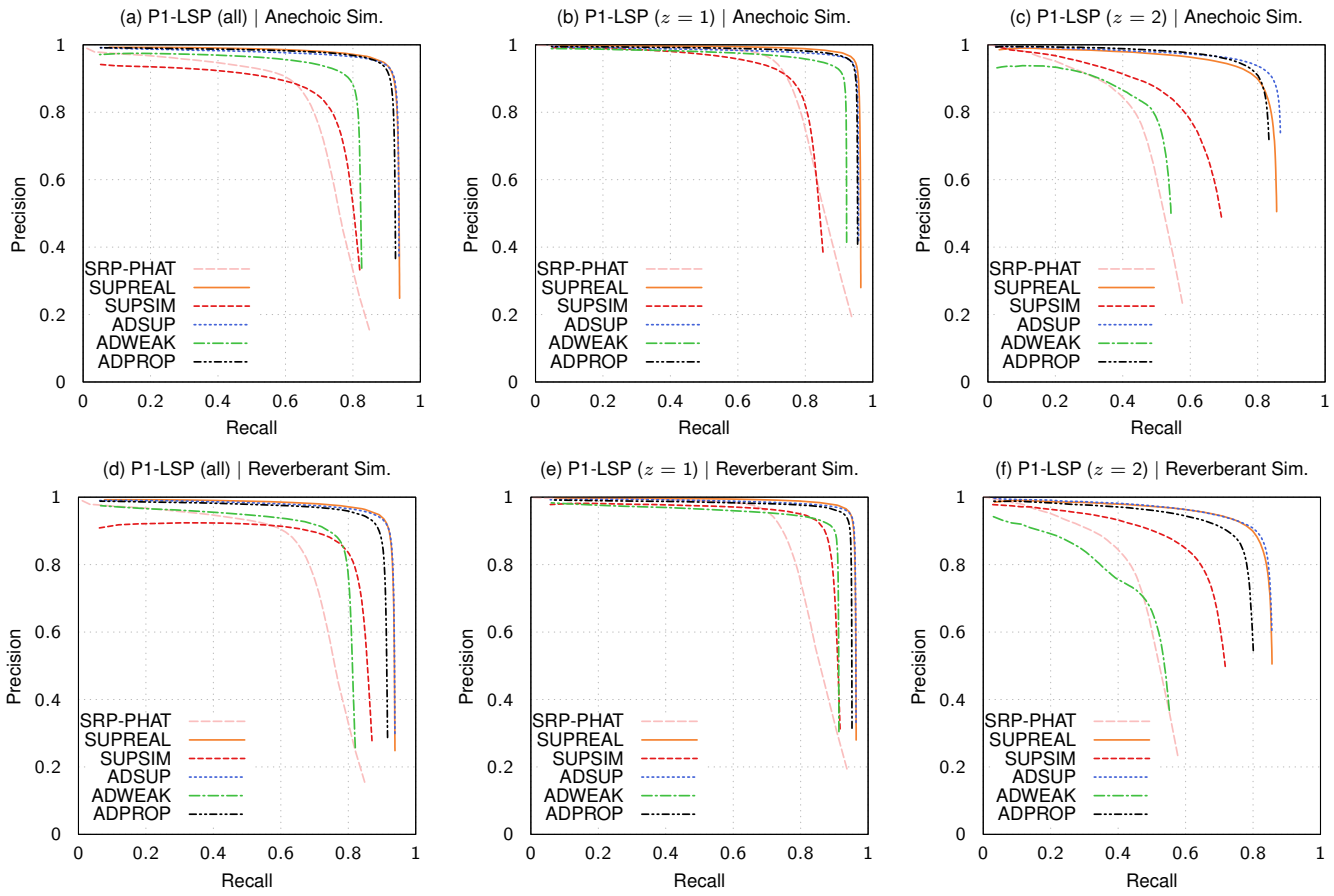


Fig. 13: Precision-recall curves as a sound source detection problem on the P1-LSP dataset. The curves are generated by varying the prediction threshold ξ in Eq. 3. DOA estimation with less than 5° error is considered correct. The room conditions (anechoic or reverberant) for simulation are indicated in the figure titles.

on real data (SUPREAL) in the P1-LSP test set. Nevertheless, a noticeable difference (see Fig. 13c) is that the adapted model has better precision and recall in the two-source frames. This is because the simulated data provide a broader coverage of sound source directions, especially in the multi-source case, than the real data.

Weakly-supervised Adaptation. Using the weakly-labeled real data, both the weakly-supervised domain adaptation approaches (ADWEAK and ADPROP) significantly outperforms the pre-trained model (SUPSIM). The discrepancy between the simulation and real data is mitigated. Between both the approaches, the performance of our proposed approach (ADPROP) is significantly better, especially on the two-source frames, with an accuracy of 82.8 vs 53.8 for instance. In fact, directly applying the minimum distance adaptation (ADWEAK) on the multi-source frames is not reliable and generates wrong pseudo-labels. Therefore, its performance on the two-source frames is worse than the pre-trained model. Applying the adaptation on the single-source components of the augmented data prevents unreliable pseudo-labeling and improves the adaptation result. As a result, our approach achieves comparable results, in terms of accuracy as well as precision and recall (Fig. 13(a,d)), as those using fully-labeled real data. This shows that we can substitute exact labels in the real data with weak labels, thus the workload of annotation can be significantly reduced.

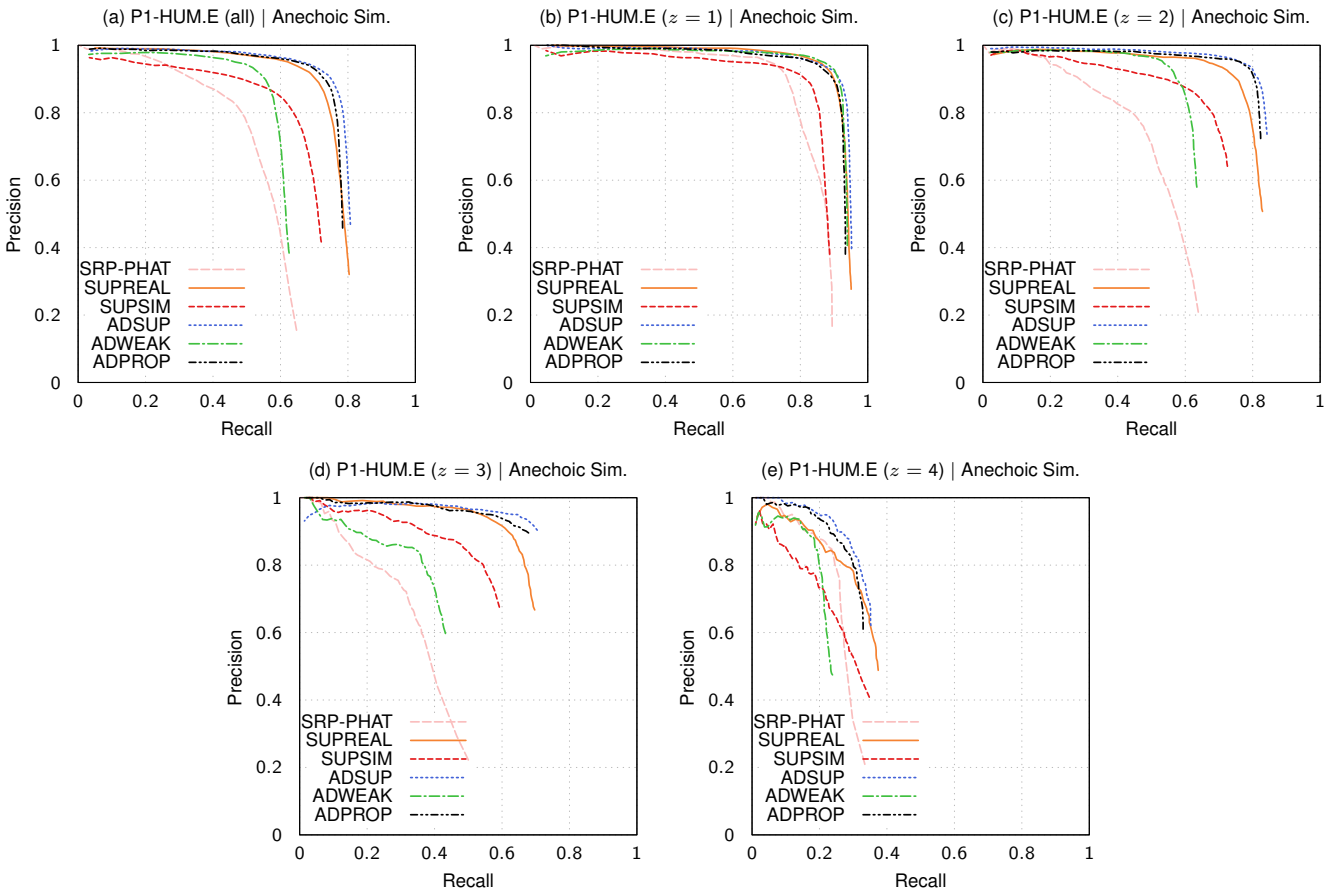


Fig. 14: Precision-recall curves as a sound source detection problem on the P1-HUM.E dataset.

TABLE VI: MAE(°) and ACC(%) on the P2-LSP dataset. The source-domain data are simulated with anechoic condition.

Dataset	P2-LSP					
Subset	All		$z = 1$		$z = 2$	
	MAE	ACC	MAE	ACC	MAE	ACC
SRP-PHAT	13.5	72.0	11.1	75.3	29.3	50.6
SUPREAL	5.5	86.6	4.5	89.2	12.0	69.5
SUPSIM	7.2	76.5	6.5	77.5	12.3	70.2
ADSUP	3.5	92.4	3.2	93.5	5.6	85.0
ADWEAK	5.1	80.6	4.4	82.1	9.7	71.1
ADPROP	4.7	82.2	4.4	83.0	7.1	76.9

Anechoic vs Reverberant Simulation. Comparing the different simulation conditions, we find that the pre-trained models with reverberant simulation in general outperform those with anechoic simulation, as they matches the evaluation data better, which are collected in reverberant environments. However, after domain adaptation, the models with the anechoic simulation achieves better performance in most conditions. This is probably because the models with simpler source-domain conditions (anechoic simulation) are more capable to adapt, while the models with reverberant simulation might overfit to the difficult conditions in the simulated training data.

P1-HUM.E Data. There are more condition mismatch between this test set and the target-domain training data, as it includes real human voices instead of sounds from loudspeakers, non-speech sounds, as well as more overlapping sound sources. Under such a condition, the incorporation of simulated data increases the coverage of various training conditions, thus can help the model with better generalization. This is seen from the results (Table V and Fig. 14), as ADSUP outperform SUPREAL under all conditions using most of the criteria (ACC, precision and recall). Although the proposed approach (ADPROP) does not use any exact labels, it achieves better overall performance than the supervised approach SUPREAL. This shows that the proposed approach can generalize well under mismatched conditions with the help of data augmentation as well as data simulation.

The results on P1-HUM.E also demonstrate the performance of all approaches when there are more than two overlapping sources. The overall performance degrades as the number of overlapping sound sources increases. This is expected as the average Signal-to-Interference-plus-Noise Ratio (SINR) is lower when there are more number of sources, and the input window duration of 170 ms is too short to have enough information for localizing more sound sources. In addition, most of the our training samples include less than two simultaneous sound sources (because this is more common under real HRI conditions). Nevertheless, SUPREAL, which is trained with maximum two overlapping sound sources using the real loudspeaker recordings, achieves reasonable performance (90% precision at 60% recall) when tested with three overlapping

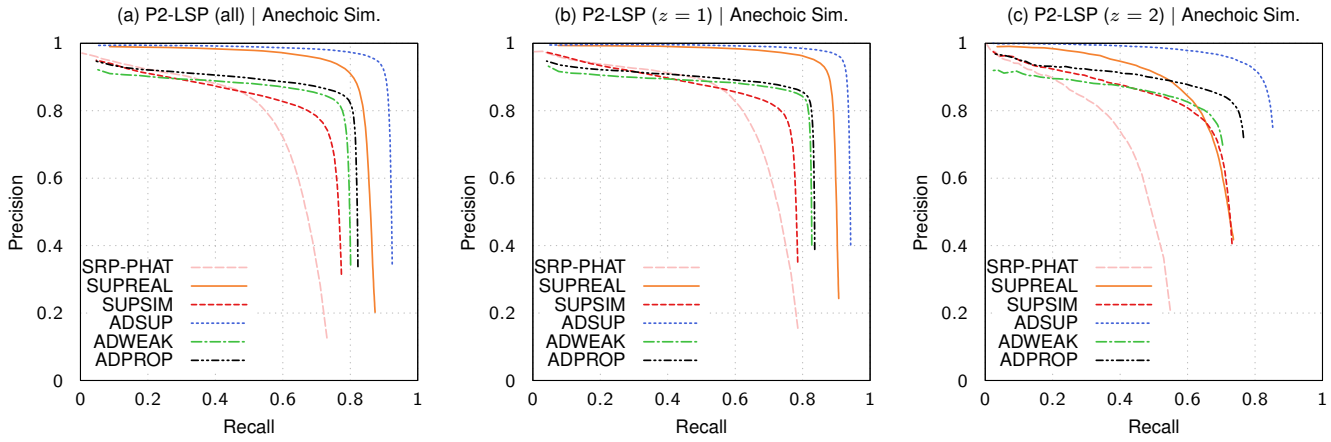


Fig. 15: Precision-recall curves as a sound source detection problem on the P2-LSP dataset.

sound sources. This shows the network can generalize to test conditions where there are more overlapping sound sources than it has seen during training. Several techniques can be employed to improve the localization performance of more than two overlapping sound sources. For example, using longer input segment can increase the chance that each sources is dominant in at least one frame. This has been shown in [55]. Moreover, integrating temporal filtering, such as tracking or recurrent network structure, can also be used to increase the amount of information for localizing more sound sources.

P2-LSP Data. We notice that this dataset is in general more challenging than the P1-LSP data, as indicated by the accuracy as well as precision and recall of SRP-PHAT, SUPREAL and SUPSIM in the results (Table VI and Fig. 15). The proposed approach relies on initial performance of the pre-trained model, therefore it does not perform as well as that in the P1 data. In spite of this, the proposed approach (ADPROP) shows a significant improvement over the pre-trained model. We also find that the model trained with both simulated and real data (ADSUP) outperforms significantly the models using only real data. This is because there are less real training data for P2 (compared to P1), and adding the simulated data may help especially when the real training data are not sufficient.

G. Scalability with Data Size

We analyzed the scalability of the different approaches. Specifically, we examined on P1-LSP (Fig. 16) and P1-Human (Fig. 17) how their F1-scores evolve with the size of the target-domain training data. The data size is represented by the number of files, which we experimented from 5 to 4000. The number of files indicates the variabilities of sound source positions in the dataset, as sound source locations are fixed in each file. The F1-scores are computed with the precision and recall values that generate the best F1-scores.

Both figures show that the performance of all approaches generally increases as more real data are used. One exception is the model adapted using fully-labeled data (ADSUP) when tested on P1-LSP. It has better performance using less than 100 files than using between 200 and 1000 files. This is because that in the former case the model does not variate much from

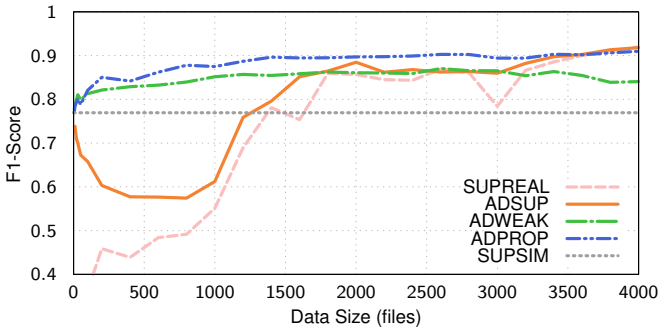


Fig. 16: Sound source detection F1-score versus the training data size (the number of files ranging from 5 to 4000) on the P1-LSP evaluation set. Source-domain data are simulated under the anechoic condition. The pre-trained model (SUPSIM), which does not use any real data, is presented as a reference.

the pre-trained model, while in the latter case the model starts to overfit the presented real data, which does not cover enough variabilities in conditions. The results indicate that the supervised approaches (SUPREAL and ADSUP) require more than 1500 files to outperform the model trained with simulated data (SUPSIM), whereas the weakly-supervised domain adaptation approaches (ADWEAK and ADPROP) achieves significantly better performance than SUPSIM using as few as 100 files. This suggests that in the case of very few real audio samples, the weakly-supervised approaches may also be used to prevent the overfitting problem of the supervised approach.

VI. Conclusion

We have proposed a framework to train deep neural networks for multi-source DOA estimation. The framework uses simulated data together with weakly labeled data under a domain adaptation setting. We have also proposed a data augmentation scheme combining our weakly-supervised adaptation approach with reliable pseudo-labeling of mixture components in the augmented data. This approach prevents incorrect adaptation caused by difficult multi-source samples. The proposed weakly-supervised method achieves almost equal

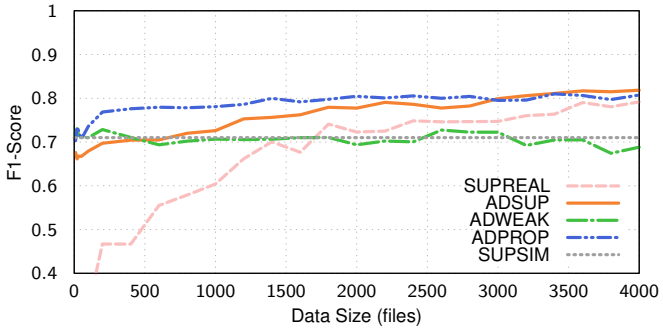


Fig. 17: Sound source detection F1-score versus the training data size (the number of files ranging from 5 to 4000) on the P1-HUM.E evaluation set.

performance to the fully-labeled case under certain conditions. Overall, the proposed framework can be used for deploying learning-based sound source localization approaches to new microphone arrays with a minimal effort for data collection.

ACKNOWLEDGMENT

This work was supported by the European Union under the EU Horizon 2020 Research and Innovation Action MuMMER (MultiModal Mall Entertainment Robot), grant agreement no. 688147. The authors gratefully thank the EU for their financial support, and all project partners for a fruitful collaboration. More information about MuMMER is available at http://mummer-project.eu.

REFERENCES

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[2] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[3] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Apr. 1997, pp. 375–378.

[4] T. May, S. v. d. Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[5] A. Deleforge, R. Horaud, Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 718–731, Apr. 2015.

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

[7] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "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 (ICASSP)*, Apr. 2015, pp. 2814–2818.

[8] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 405–409.

[9] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE Int. Conf. on Robotics and Automation (ICRA)*, Brisbane, Australia, 2018.

[10] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Proc. Audio Eng. Soc. Conv. 138*, Warsaw, Poland, May 2015.

[11] W. He, P. Motlicek, and J.-M. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," in *Proc. Interspeech 2018*, Hyderabad, India, 2018.

[12] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, Mar. 2019.

[13] H. W. Lllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Jul. 2018, pp. 410–414.

[14] A. Deleforge and R. Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2012, pp. 431–438.

[15] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International Journal of Neural Systems*, vol. 25, no. 01, 2015.

[16] L. Perotin, R. Serizel, E. Vincent, and A. Gurin, "CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector," in *IWAENC 2018 - 16th International Workshop on Acoustic Signal Enhancement*, Sep. 2018.

[17] Z. Tang, J. D. Kanu, K. Hogan, and D. Manocha, "Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks," in *Interspeech 2019*, Sep. 2019, pp. 654–658.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[19] A. Kulowski, "Algorithmic representation of the ray tracing technique," *Applied Acoustics*, vol. 18, no. 6, pp. 449–469, Jan. 1985.

[20] D. Campbell, K. Palomaki, and G. Brown, "A matlab simulation of "shoebox" room acoustics for use in research and teaching," *Computing and Information Systems*, vol. 9, no. 3, p. 48, 2005.

[21] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[22] K. Nakadai, D. Matsuura, H. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: robust sound source localization and extraction," in *2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA, 2003, pp. 1147–1152.

[23] D. P. Jarrett, E. a. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.

[24] C. Schissler and D. Manocha, "Interactive sound propagation and rendering for large multi-source scenes," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, Sep. 2016.

[25] S. Mller, "Measuring transfer-functions and impulse responses," in *Handbook of Signal Processing in Acoustics*, D. Havelock, S. Kuwano, and M. Vorlnder, Eds. New York, NY: Springer, 2008, pp. 65–85.

[26] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, Jun. 1995.

[27] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2001, pp. 99–102.

[28] J. Le Roux, E. Vincent, J. R. Hershey, and D. P. W. Ellis, "Micbots: Collecting large realistic datasets for speech and audio research using mobile robots," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5635–5639.

[29] H. Liu, Z. Zhang, Y. Zhu, and S.-C. Zhu, "Self-supervised incremental learning for sound source localization in complex indoor environment," in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 2599–2605.

[30] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, 2010.

[31] R. Takeda and K. Komatani, "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," in *2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[32] R. Takeda, Y. Kudo, K. Takashima, Y. Kitamura, and K. Komatani, "Unsupervised adaptation of neural networks for discriminative sound

source localization with eliminative constraint," in *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[33] W. He, P. Motlicek, and J. Odobez, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[34] M. S. Datum, F. Palmieri, and A. Moiseff, "An artificial neural network for sound localization using binaural cues," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 372–383, Jul. 1996.

[35] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[36] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2016, pp. 603–609.

[37] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019, pp. 451–455.

[38] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, Feb. 2017.

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

[40] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," *Proceedings of Interspeech 2015*, pp. 3302–3306, 2015.

[41] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proceedings of 2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, Sep. 2018.

[42] J. Huang, A. Gretton, K. M. Borgwardt, B. Schlkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," *Advances in neural information processing systems.*, 2007.

[43] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI Conf. on Artificial Intelligence*, 2016.

[44] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Int. Conf. on Machine Learning*, 2015.

[45] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *Workshop on challenges in representation learning, ICML*, 2013.

[46] J. Choi, M. Jeong, T. Kim, and C. Kim, "Pseudo-labeling curriculum for unsupervised domain adaptation," *arXiv:1908.00262 [cs]*, 2019.

[47] G. Yves and B. Yoshua, "Entropy regularization," in *Semi-Supervised Learning*, O. Chapelle, B. Scholkopf, and A. Zien, Eds. The MIT Press, 2006.

[48] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Wang, T. Chen, J. Pan, J. Du, and C.-H. Lee, "The USTC-IFLYTEK system for sound event localization and detection of DCASE2020 challenge," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020.

[49] T. N. T. Nguyen, D. L. Jones, and W. S. Gan, "DCASE 2020 task 3: Ensemble of sequence matching networks for dynamic sound event localization, detection, and tracking," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385 [cs]*, 2015.

[51] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, and others, "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[52] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[53] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.

[54] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Int. Conf. on Learning Representations (ICLR)*, San Diego, 2015.

[55] T. T. N. Nguyen, W.-S. Gan, R. Ranjan, and D. L. Jones, "Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2020.