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

FuYanjie 2021.10.27

Overview

Introduction

Challenges

1. Signal processing approaches rely on assumptions about the acoustic environment that may not hold well in real-world scenarios;
2. How to generalize sophisticated modeling of the complex environments;
3. The need of collecting a sufficient number of training data covering all variabilities in learning-based approaches.
   Making audio recordings and annotating audio recordings with the ground truth labels are particularly costly.

Introduction

Motivation

A popular way of obtaining training data for sound source localization is by acoustic simulation.

Domain adaptation, which uses both simulated and real data, may be applied to SSL.

Previous studies have investigated the unsupervised adaptation of neural networks for single-source sound source localization with entropy minimization

Introduction

Goals and Contributions

Goals: data collection at low cost and training models using domain adaptation

The contributions of this paper are:

- Propose a multi-source DOA estimation framework with domain adaptatoin so that the data collection workload can be significantly reduced.
- 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, which improves the performance of the weakly-supervised adaptation.

Related Work

A. NN-based Sound Source Localization

Different approaches differ in their input representation, output coding as well as their network structures.

More recent studies have shown that low-level signal representation without explicit feature extraction, whether in the time or time-frequency domains, can allow the networks to learn to extract the most informative high-level features for SSL.

Few of studies on DL based SSL clearly address issues related to the high cost of data collection, especially by applying domain adaptation to models trained with simulated data.

Related Work

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). 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.

DOA Estimation Model

A. Overview

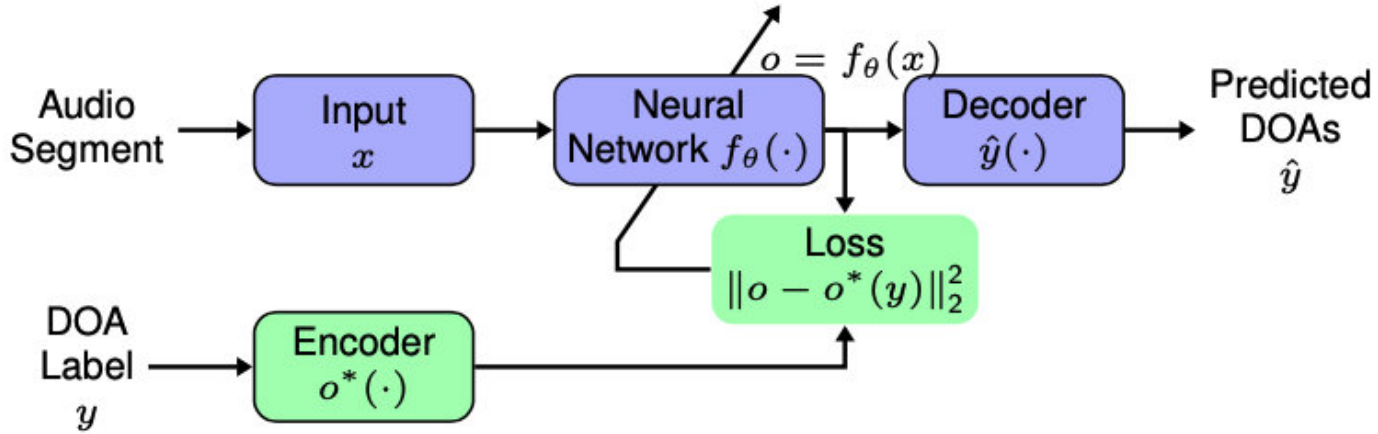


Fig. 2: Overview of our neural-network-based approach for multi-speaker direction of arrival estimation. The top part (blue) represents the prediction process, whereas the bottom part (green) indicates the supervised learning principle.

DOA Estimation Model

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 (通道间) cues (i.e. level/phase difference) and intra-channel (通道内) cues (i.e. spectral features).

Proposed Method

B. Network Input

We prepare the network input as follows:

First divide the 4-channel audio into 170 ms long segments (8192 samples in 48 kHz recordings).

Compute the STFT of the segments with a frame size of 43 ms (2048 samples) and 50% overlap. Thus, there are seven frames in each segment.

Only use the frequency bins between 100 and 8000 Hz, so that the number of frequency bins is reduced to 337.

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 $7 \times 337 \times 8$.

Proposed Method

C. Output Coding

The spatial spectrum coding:

A spatial spectrum is a function of the DOA and its value indicates how likely there is a sound source for a given DOA.

Thus, the localization problem becomes a spatial spectrum regression problem.

Fig. 2: Example of the output coding for multiple sources according to Eq. 3. It resembles a spatial spectrum: the peaks indicate the directions of the sources.

Proposed Method

C. Output Coding

$$Eq.2. \qquad o^*(y) = \begin{cases} \max_{\phi' \in y} \left\{ e^{-d\left(\phi_i, \phi'\right)^2 / \sigma^2} \right\} & \text{if } |y| > 0 \\ 0 & \text{otherwise} \end{cases}$$

y : label, a set of locations $\quad |y|$: the number of sources

ϕ': one ground truth DOA

σ: beam width 波束宽度是主瓣两侧的两个最低值之间的间距（即主瓣的零点之间的宽度）[1]

$\sigma = \theta_{BW} = 2 \sin^{-1}\left(\frac{c}{Mdf}\right)$ M: num_mics, d: 麦克风间距, c: sound speed

[1] H.L. Van Trees, Detection, Estimation, and Modulation Theory, Optimum ArrayProcessing, John Wiley & Sons, New-York, USA, 2004.

Proposed Method

C. Output Coding

Decode when inferencing

When the number of sources z is unknown, the peaks above a given threshold ξ are taken as predictions:

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

When the number of sources z is known, the z highest peaks are taken as predictions:

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

$o = f_\theta(x)$ is the network output.

Proposed Method

D. Network Architecture

Fully-convolutional neural network structure.

Our network comprises two parts, which convolve along different axes.

In the first part, the network convolves along the time and frequency axes.

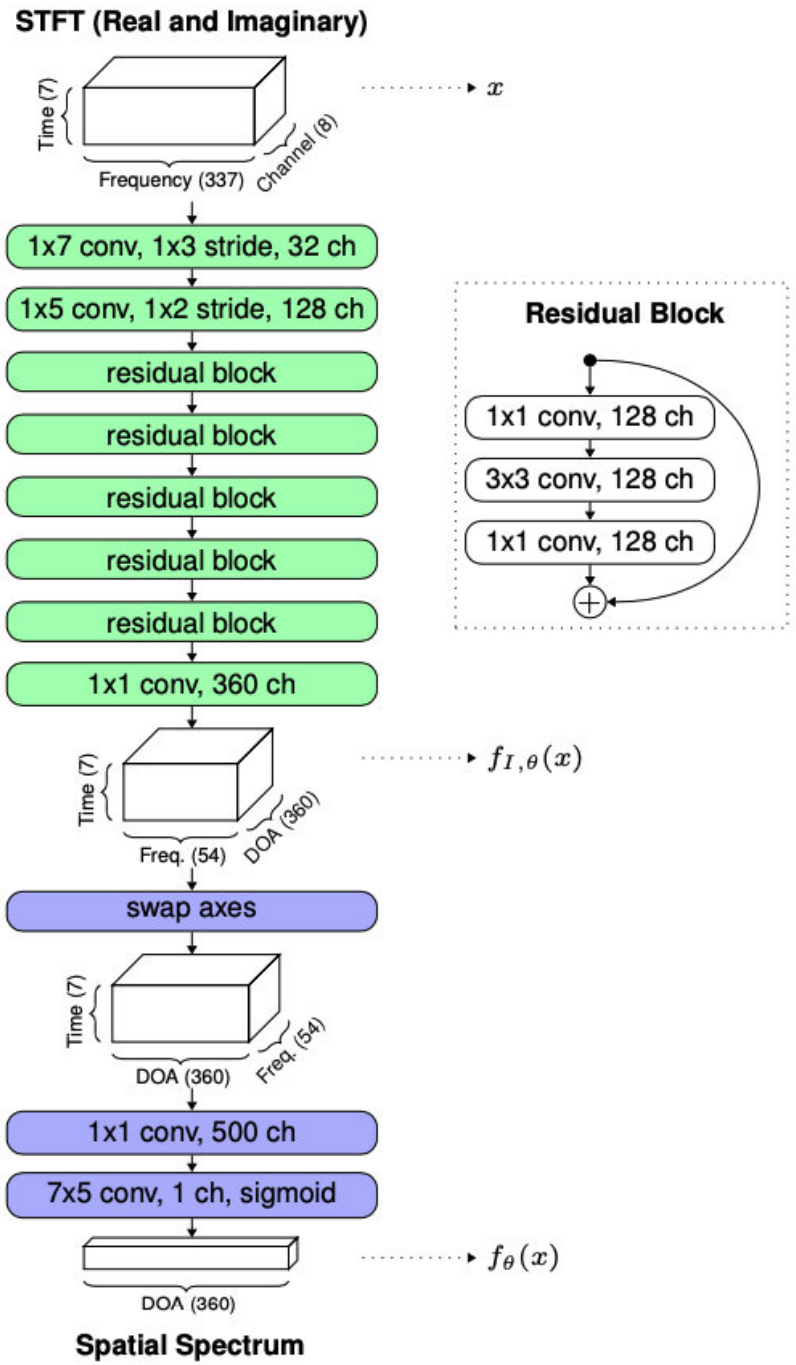
In the second part, the network convolves along the DOA axis.

Proposed Method

D. Network Architecture

The first part (green) applies convolution along the time and frequency axes.

The second part (blue) applies convolution along the DOA axis.
It aggregates features in the neighboring directions across all the time-frequency bins (global), and outputs a spatial spectrum.



Proposed Method

E. Two-stage Training

The goal of training is to make the network regress the ideal spatial spectrum with the MSE loss:

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

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.

Proposed Method

E. Two-stage Training

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 across time and frequency:

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

$f_{I,\theta}(x)[t,k]$ 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.

Domain Adaptation

A. Supervised Adaptation

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.
To apply supervised domain adaptation,
1.Use the simulated data to pre-train a model, which is the initialization of the subsequent optimization processes.
2.Then train a model that minimizes the loss on both the source domain and the target domain:

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

μ_t: weighting parameters for the loss on the target domain

μ_s: weighting parameters for the loss on the source domain

D_t: a set of fully-labeled real audio data

D_s: a set of fully-labeled simulated audio data

Domain Adaptation

B. Weakly-Supervised Adaptation

D_w: a set of weakly-labeled real audio data 弱标注是指仅给出声源数

D_s: a set of fully-labeled simulated audio data

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:

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

Domain Adaptation

B. Weakly-Supervised Adaptation

Define 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:

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

$r(z)$ is the set of all sound DOA labels that satisfy the weak label z, i.e. the number of sources in y is z:

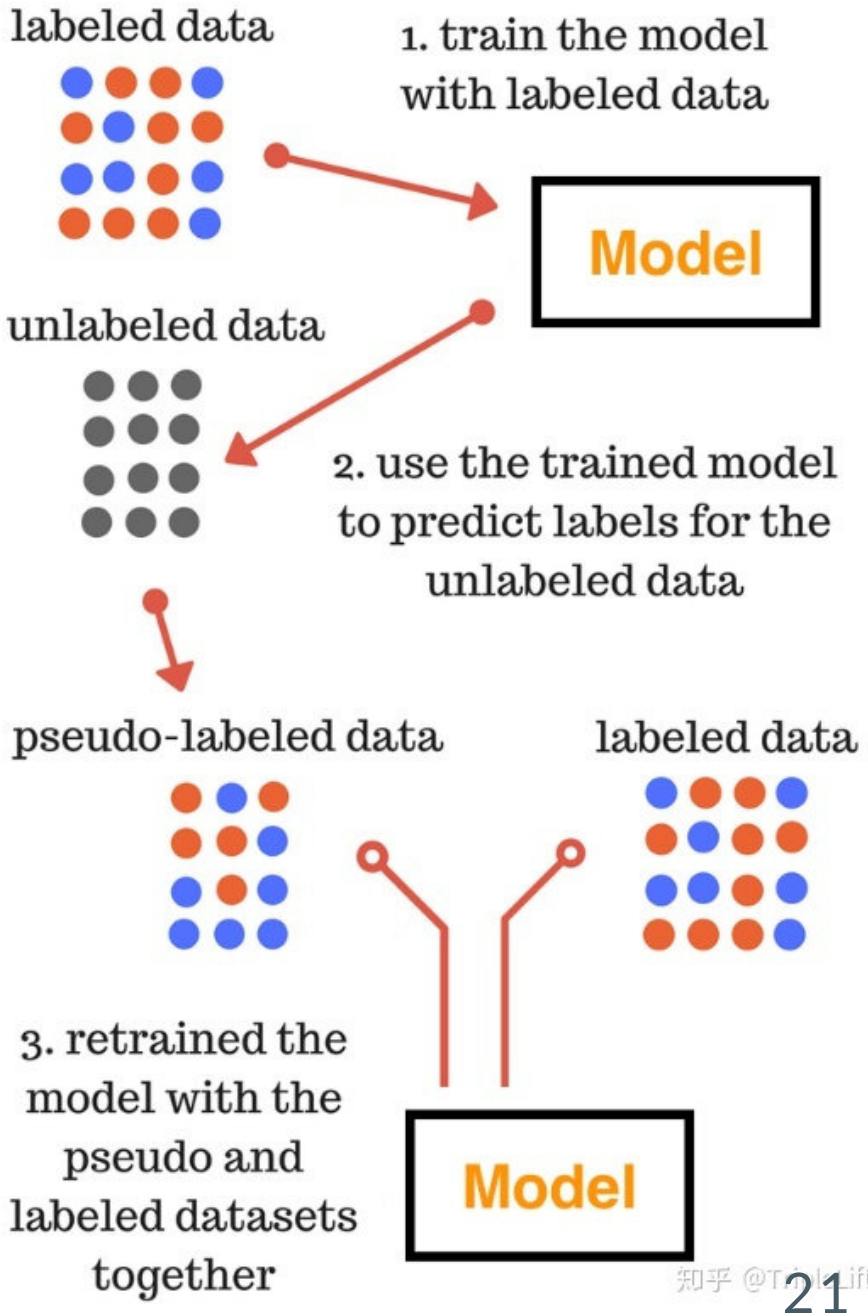
$$r(z) = \{y : |y| = z\}$$

Domain Adaptation

C. Pseudo-labeling with Data Augmentation

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.



Domain Adaptation

C. Pseudo-labeling with Data Augmentation

Pseudo-labeling:

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

First apply pseudo-labeling (Eq. 11) to its single-source components

Then, use the union of these pseudo-labels for the multi-source frame

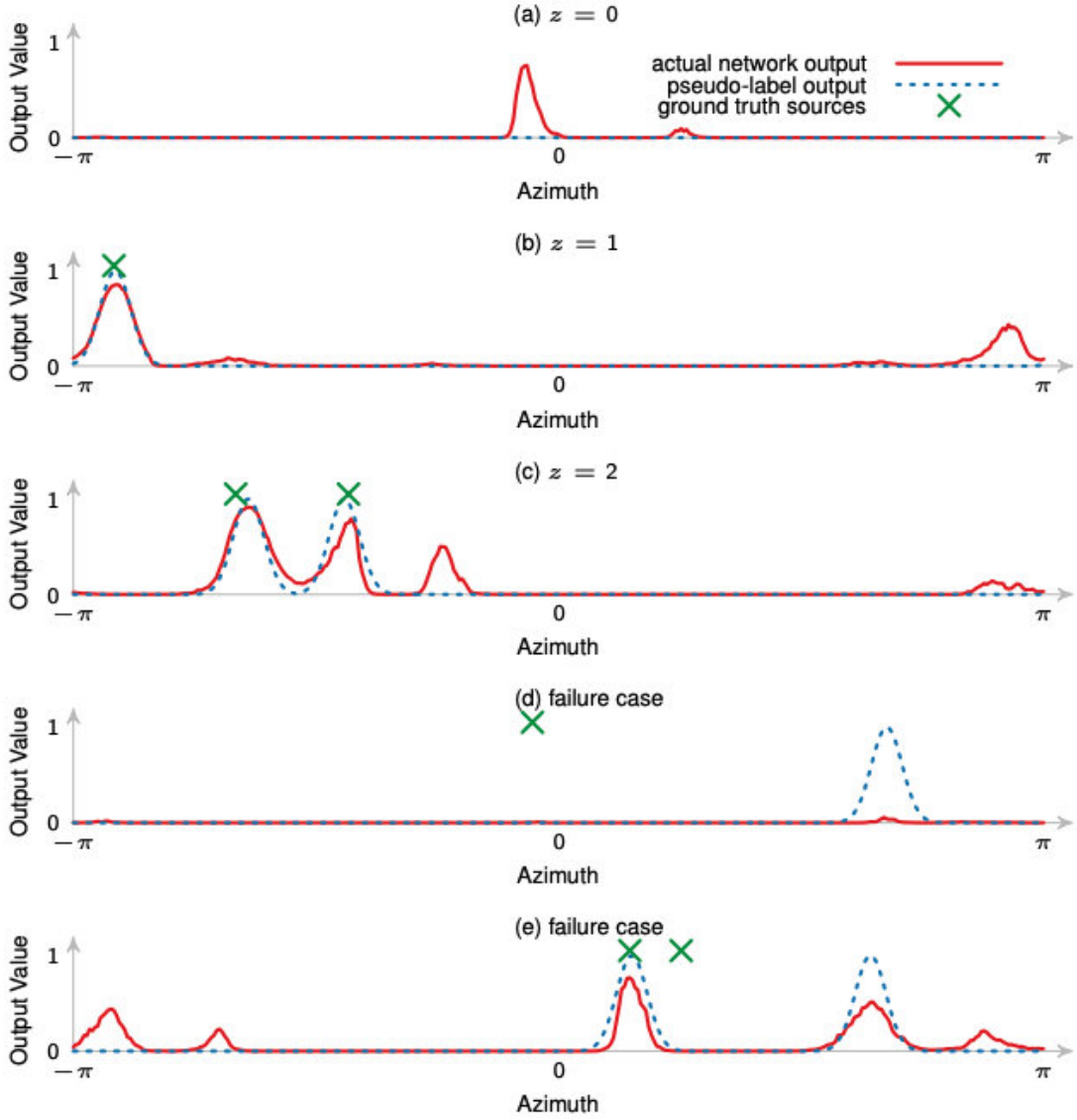


Fig. 5: Examples of weak supervision with a known number of sources on real audio segments. The ground truth locations are shown but are not used for weak supervision.

Domain Adaptation

C. Pseudo-labeling with Data Augmentation

Thus, the loss function of the modified adaptation is:

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

and the optimization target becomes:

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

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

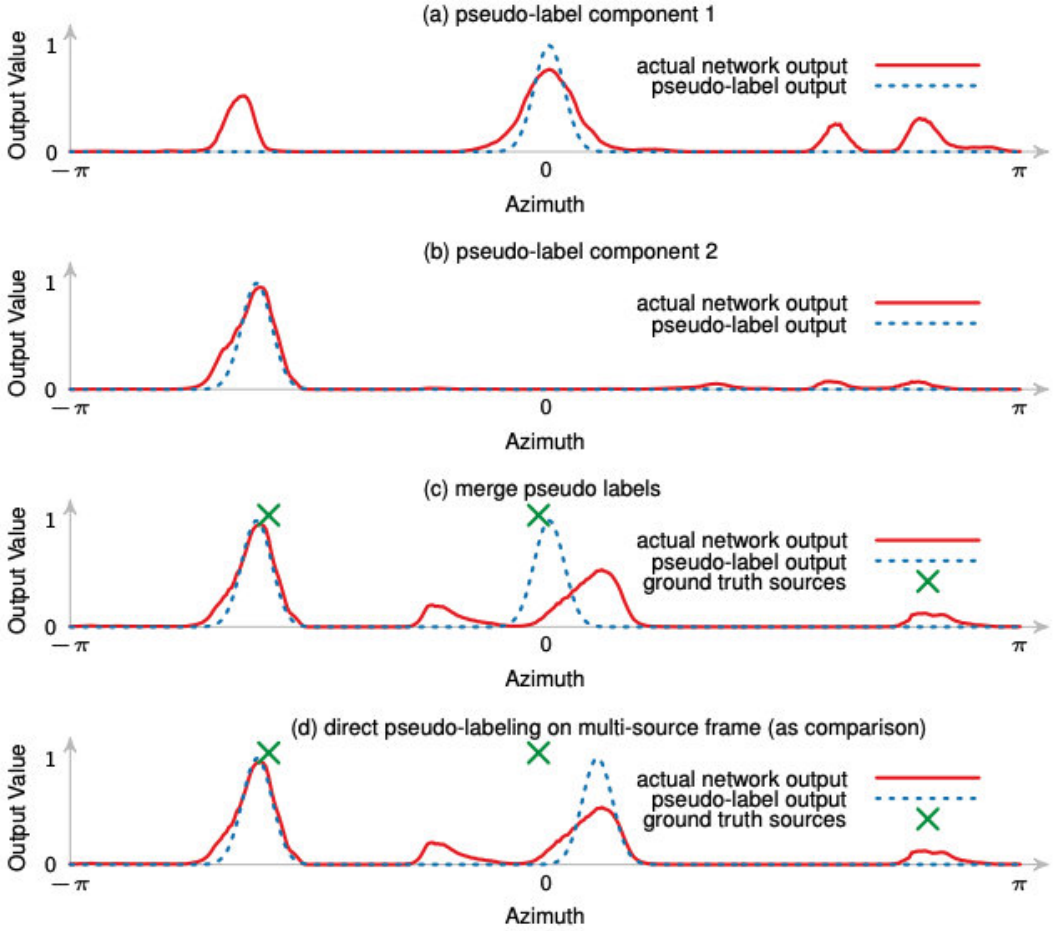


Fig. 6: Example of weak supervision from mixture components of an augmented multi-source frame. (a, b) The pseudo-labeling is applied first on the single-source components. (c) Then the pseudo-label of the two-source mixture is obtained by merging the pseudo-labels of its components. This approach is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

Experiment

A. Microphone Array and Data

Microphone array

2 versions of the robots: $P1$ and $P2$ differ in their microphone directivity patterns: directional and omni-directional

Source-domain data

Generated the source domain data by convolving clean speech audio with RIRs.
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.

Target-domain data (Real data: SSLR)

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.

Experiment

B. Training Parameters

Pretrain: one epoch in the first stage (Eq. 6) and four epochs on the second stage (Eq. 5).

Then the pretrained 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 μ_w = 0.9, μ_a = 0.1, and μ_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.

lr: 0.001 and reduced it by half once the training loss no longer decreased

Adam optimizer MSE loss mini-batch size: 100

Experiment

C. Analysis of Pseudo-Labeling

Computed the loss gain between the MSE loss (Eq. 5) of the model prediction and that of the pseudo-label:

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

A positive loss gain indicates the pseudo-labeling is beneficial for the model.

Experiment

C. Analysis of Pseudo-Labeling

The green bars indicate positive gain (correct weak supervision), while the red bars indicate negative gain (incorrect weak supervision).

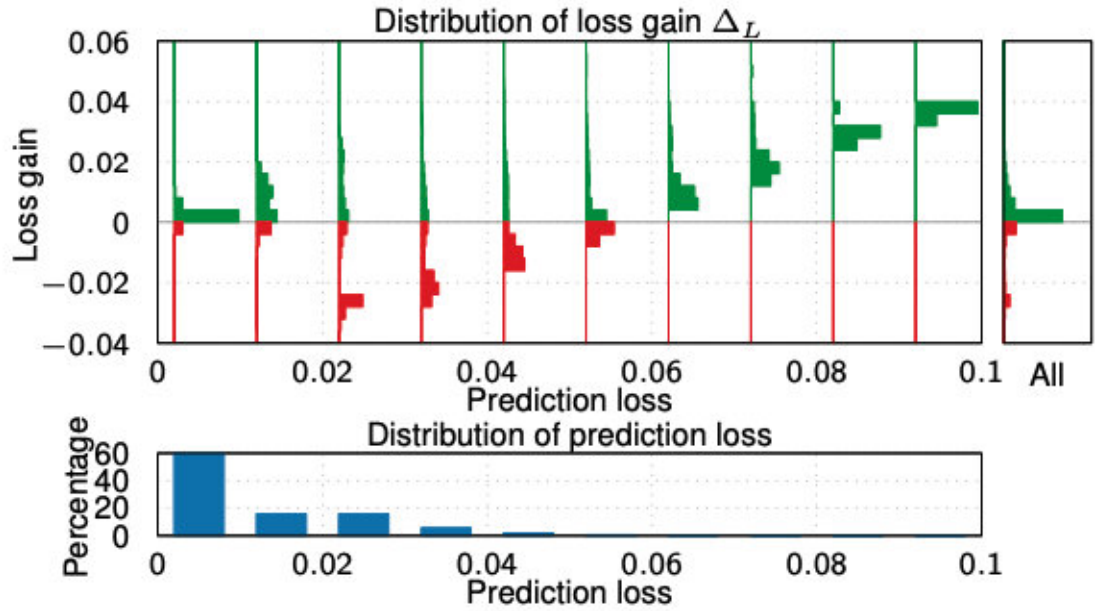


Fig. 9: Analysis of the minimum distance adaptation of the single-source samples from the P1 training data.

Experiment

C. Analysis of Pseudo-Labeling

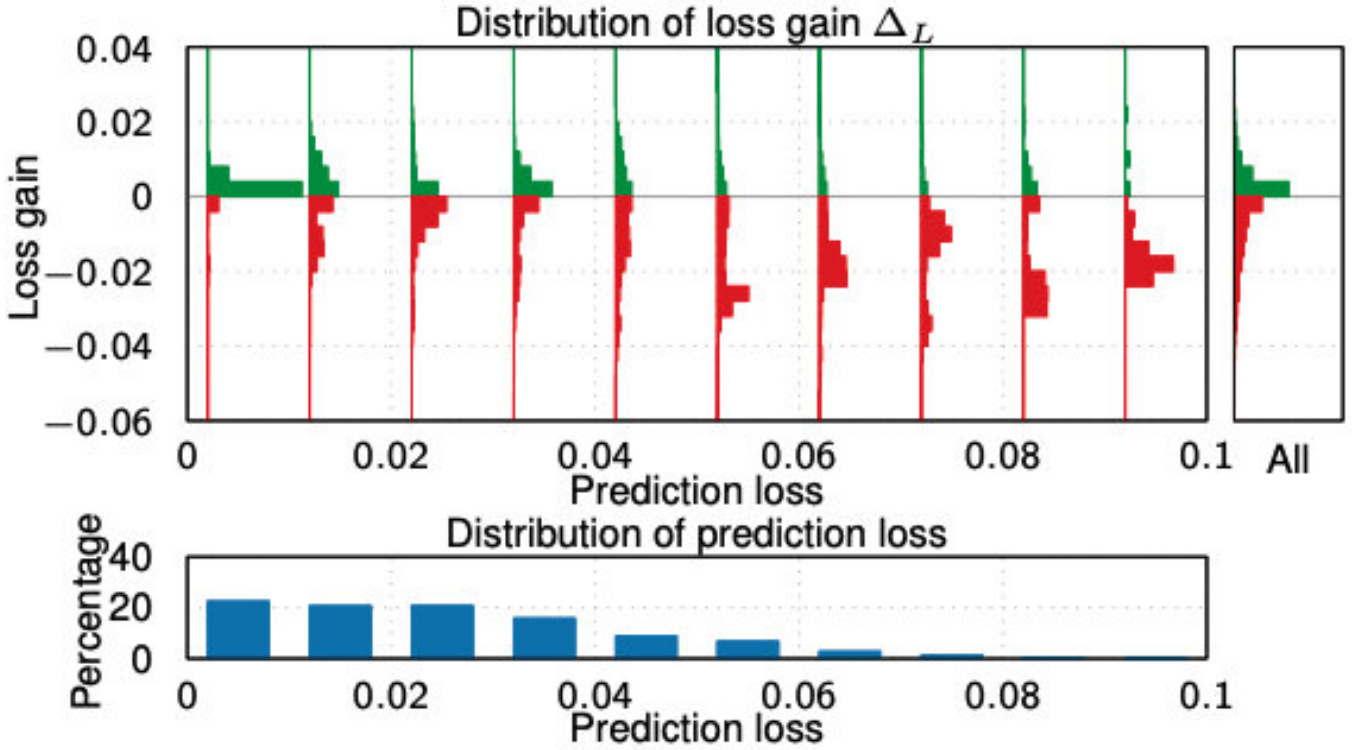


Fig. 10: Analysis of the minimum distance adaptation of the multi-source samples from the P1 training data.

Experiment

D. DOA Estimation Evaluation Protocol

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.

Two evaluation settings: (a) the number of sound sources is known, or (b) unknown

(a) MAE(°) and ACC(%)

$$\mathrm{MAE} = \frac{\sum_i \sum_{j=1}^{z_i} d\left(\hat{\phi}_{ij}, \phi_{ij}\right)}{\sum_i z_i}$$

Experiment

D. DOA Estimation Evaluation Protocol

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

SUPREAL fully-supervised approach using only fully-labeled real data for two-stage training.

SUPSIM trained with only the simulated data. (This is also the pre-trained model for the domain adaptation approaches.)

Experiment

D. DOA Estimation Evaluation Protocol

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 pre-trained with the simulated data and then adapted using the weakly-labeled real data and augmented data with the adaptation scheme (Eq. 15).

z = 2, the performance of ADPROP is significantly better compared to ADWEAK.

Experiment

F. DOA Estimation Results

(b) Precison and Recall

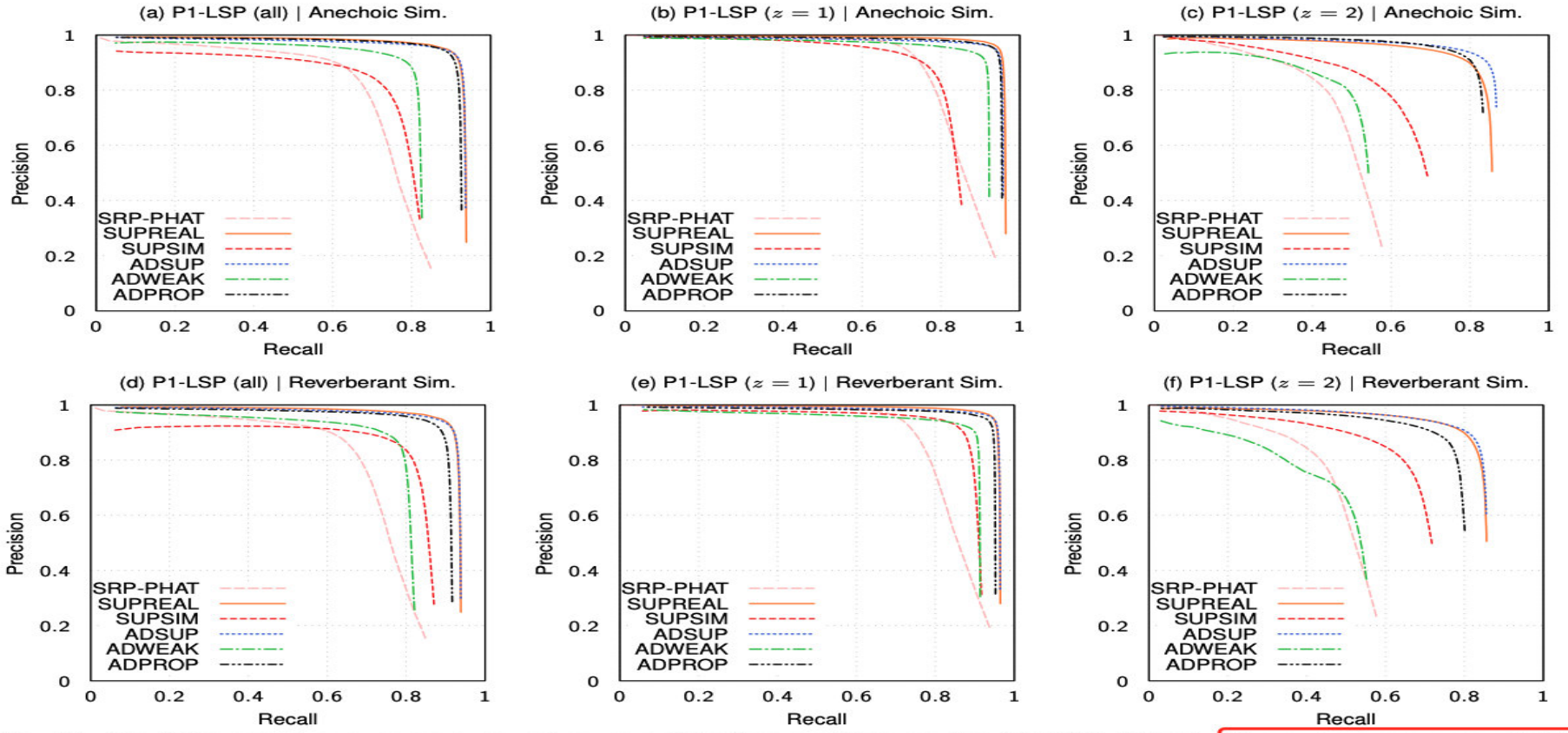


Fig. 13: Precision-recall curves as a sound source detection problem on the P1-LSP dataset. The curves are generated by varying the prediction threshold ξ in Eq. 3. DOA estimation with less than 5° error is considered correct. The room conditions

Experiment

G. Scalability with Data Size

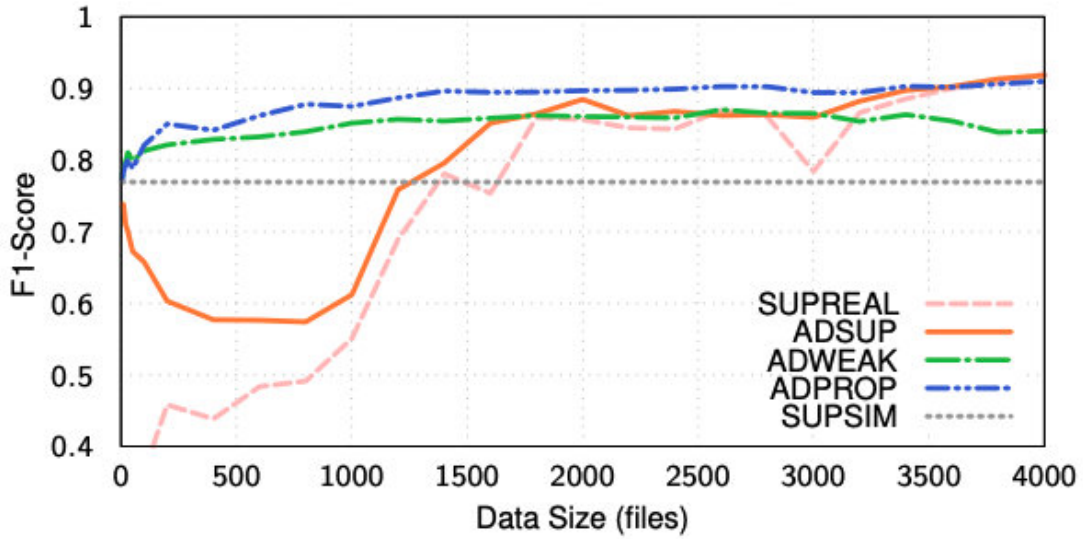


Fig. 16: Sound source detection F1-score on the P1-LSP evaluation set versus the training data size (the number of files ranging from 5 to 4000). Source-domain data are simulated under the anechoic condition. The pre-trained model (SUPSIM), which does not use any real data, is presented as a reference.

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 similar per- formance as 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.

Thanks for your time!