ADAPTATION OF MULTIPLE SOUND SOURCE LOCALIZATION NEURAL NETWORKS WITH WEAK SUPERVISION AND DOMAIN-ADVERSARIAL TRAINING

Weipeng He*†*, Petr Motlicek** *and Jean-Marc Odobez**†

*Idiap Research Institute, Switzerland. {weipeng.he, petr.motlicek, odobez}@idiap.ch
†École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

ABSTRACT

Despite the recent success of deep neural network-based approaches in sound source localization, these approaches suffer the limitations that the required annotation process is costly, and the mismatch between the training and test conditions undermines the performance. This paper addresses the question of how models trained with simulation can be exploited for multiple sound source localization in real scenarios by domain adaptation. In particular, two domain adaptation methods are investigated: weak supervision and domain-adversarial training. Our experiments show that the weak supervision with the knowledge of the number of sources can significantly improve the performance of an unadapted model. However, the domain-adversarial training does not yield significant improvement for this particular problem.

Index Terms— Sound source localization, DOA estimation, domain adaptation, weakly-supervised learning.

1. INTRODUCTION

Recent studies have shown that deep neural networks (DNNs) have become the state-of-the-art for sound source localization (SSL) and directions of arrival (DOA) estimation [1–8]. Although these DNN-based approaches outperform the classical signal processing-based techniques [9–11] under certain conditions, they suffer two major drawbacks.

First, these approaches require a large amount of device-specific training data and obtaining real data with annotation is arduous. Since signals captured by microphone arrays with different geometries are radically different, individual data collection is required for each specific type of microphone array. In addition to recording data, the annotation of the ground truth sound source locations in real data is also particularly difficult. The costly data recording and annotation process hinders the application of DNN-based SSL systems.

Another drawback of the DNN-based approaches is their sensitivity to the mismatch between the training and test conditions. The acoustic environments vary considerably in terms of background noise, reverberation, signal-to-noise ratio (SNR) as well as distribution of source locations. In case of using simulated training data, there is even larger mismatch between the virtual and real environments, such as difference in sensor properties, device physical bodies, etc.

This paper seeks to solve the aforementioned problems by adapting the models developed on simulated data for real scenarios. By using simulation, we can easily produce sufficient training data for any device. At the same time, we can acquire a large amount of unlabelled or weakly labelled real data, which can be exploited for adaptation. By applying unsupervised or weakly supervised adaptation, we can minimize the efforts for data collection.

This paper focuses on two domain adaptation techniques. First, we propose weak supervision by output regularization. Specifically, we examine the weak supervision with known number of sources. The number of sources contains crucial information for SSL, and is much easier to annotate compared to the exact location of each source. Based on the available weak labels, we can significantly reduce the dimension of the desired output space, and the output regularization aims to bring the network output closer to the reduced space. Second, we study the use of domain-adversarial training [12]. This unsupervised adaptation method seeks to train the network to extract domain-invariant features. This is achieved by a domain classifier that distinguishes the domains of the features, whereas the first part of the network attempts to extract features that are indistinguishable by the domain classifier.

2. RELATION TO PRIOR WORK

Previous studies have investigated the unsupervised adaptation of neural networks for SSL with entropy minimization [13, 14]. These methods attempt to modify part of the network parameters so that the entropy of the network output, namely the predicted posterior probability, is minimized.

The limitation of the previous studies is that, entropy minimization can only be applied to classification problems, where the network output is interpreted as a probability distribution. Two types of output coding have been proposed for multiple sound source localization: joint posterior prob-

*This research was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 688147 (MultiModal Mall Entertainment Robot, MuMMER, mummer-project.eu).

ability [15] and likelihood-based coding (or spatial spectrum) [7, 8]. The latter type is more advantageous as it is not limited to a predefined maximum number of sources. However, this type of output cannot be considered a probability distribution. Therefore, entropy minimization is not applicable to these multiple sound source localization approaches.

In this paper, we focus on the adaptation of multiple sound source localization neural networks. Specifically, our approach differs from previous studies in: (1) the neural network for adaptation does not predict posterior probability; (2) in addition to unsupervised adaptation, we also investigate weakly supervised adaptation to produce more stable results; (3) we examine the regularization not only on the output, but also on the features.

3. PROPOSED DOMAIN ADAPTATION APPROACH

We consider the problem of multiple sound source localization as learning the mapping from the audio segments X to the sound locations Y, with source domain data S and target domain data T. In the experiments, X are 170ms long audio segments (8192 samples with 48kHz sampling rate). We compute the short-time Fourier transforms (STFT) of these segments with frame size of 43ms (2048 samples) and 50% overlap, and use both their real and imaginary parts as the network input. The sound location space Y is the set of all finite subsets of the horizontal directions $\Phi = [-\pi, \pi)$. The source domain data are samples of audio segments with location labels: $S = \{(x_i, y_i)\}_{i=1}^{n} \subset X \times Y$, and the target domain data are samples with weak labels: $T = \{(x_i, z_i)\}_{i=n+1}^{n+n'} \subset X \times Z$. The weak labels Z provide inexact but related information about the source locations.

We propose a deep convolutional network, the main structure of which is adopted from [16], using only a single-task output (Fig. 1). Instead of directly predicting labels in Y, the network outputs the likelihood values of each sampled direction. The output coding, which defines the mapping between the labels Y and the output space O, is explained in details in Section 3.1. The neural network consists of a feature extractor $G_f(\cdot;\theta_f)$, a DOA estimator $G_y(\cdot;\theta_y)$, and a domain classifier $G_d(\cdot;\theta_d)$.

During adaptation, depending on the part of the network, the following objective function is either minimized or maximized:

$$\begin{aligned} E(\theta_f,\theta_y,\theta_d) = &\mathop{\mathbf{E}}_{x,y\in S} \mathcal{L}_y\left(F_y(x),y\right) + \mu \mathop{\mathbf{E}}_{x,z\in T} \mathcal{L}_z\left(F_y(x),z\right) \\ &- \lambda\left(\mathop{\mathbf{E}}_{x\in S} \mathcal{L}_d\left(F_d(x),0\right) + \mathop{\mathbf{E}}_{x\in T} \mathcal{L}_d\left(F_d(x),1\right)\right), \end{aligned} \tag{1}$$

where $F_y(x) = G_y(G_f(x;\theta_f);\theta_y)$ is the output of the DOA estimator, and $F_d(x) = G_d(G_f(x;\theta_f);\theta_d)$ is the output of the domain classifier. The loss terms (namely the prediction loss $\mathcal{L}_y$, the weak supervision loss $\mathcal{L}_z$, the domain loss $\mathcal{L}_d$), and their optimization targets are introduced in the following sections.

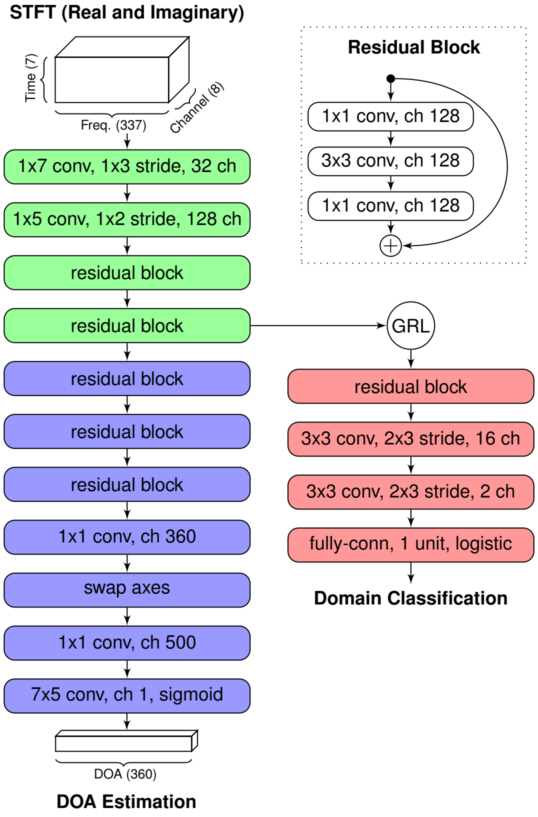

Fig. 1: The proposed network architecture consists of the feature extractor G_f (green), the DOA estimator G_y (blue), and the domain classifier G_d (red). GRL is the gradient reversal layer [12]. Batch normalization layers and rectified linear unit (ReLU) after each hidden layer are omitted in this graph.

3.1. Prediction Loss and Output Coding

We choose prediction loss to be the mean squared error (MSE) loss between the network output and the likelihood-based coding $o(y)$ [7]:

$$\mathcal{L}_y(F_y(x), y) = \|F_y(x) - o(y)\|_2^2 \,. \tag{2}$$

The output is encoded as the likelihood of a source existence on the 360 sampled directions $\{\phi_i\}_{i=1}^{360} \subset \Phi$ (Fig. 2):

$$o(y)_i = \begin{cases} \max\limits_{\phi' \in y} \left\{ e^{-d(\phi_i,\phi')^2/\sigma^2} \right\} & \text{if } |y| > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where $|y|$ is the number of sources, $d(\cdot,\cdot)$ is the angular distance, and σ is parameter for the beam width. We map the network output to the prediction in Y by finding the peaks

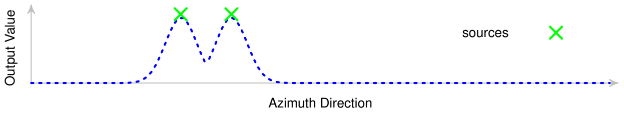

Fig. 2: Example of the output coding for multiple sources according to Eq. 3. It resembles a spatial spectrum: the peaks indicate the directions of the sources.

above a given threshold ξ:

$$\hat{y}(o) = \left\{ \phi_i : o_i > \xi \quad \text{and} \quad o_i = \max_{d(\phi_j, \phi_i) < \sigma_n} o_j \right\}, \tag{4}$$

with σ_n being the neighborhood distance.

3.2. Weak Supervision by Output Regularization

We apply weak supervision with the target domain data by fine-tuning the network output to be coherent with the available information. We know that, the encoded outputs lie in a space $o(Y) = \{o(y) : y \in Y\}$, which is much more restricted than the network output space $O = [0, 1]^{360}$. Moreover, the weak labels can further reduce the dimension of the space by filtering out incoherent predictions. Therefore, we design the weak supervision as constraining the network output to be closer to the reduced output space:

$$\mathcal{L}_z(F_y(x), z) = \min_{y \in r(z)} \| F_y(x) - o(y) \|_2^2 , \tag{5}$$

where $r(z)$ is the set of predictions coherent with the weak label z. We formulate the candidate selection as:

$$r(z) = \{ y \in Y : |y| = z \} . \tag{6}$$

The weak supervision with known number of sources helps with the adaptation in several ways. When the number of sources is zero, the network is supervised to output all zero, thus reducing the false positives caused by unseen noise (Fig. 3a). 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. 3c). At the same time, the other peaks that are caused by unseen conditions are suppressed (Fig 3b,c). However, the weak supervision does not always yield correct results, mostly due to the inaccurate initial output (Fig 3d).

3.3. Domain-Adversarial Training

In addition to weak supervision, we apply regularization on the feature space by domain-adversarial training [12]. We introduce the domain loss as the binomial cross entropy loss:

$$\mathcal{L}_d(F_d(x), d) = -d \log F_d(x) - (1 - d) \log(1 - F_d(x)). \tag{7}$$

Domain-adversarial training tries to make the feature extractor and the DOA estimator minimize the objective function (Eq. 1), while the domain classifier maximize the that function. Specifically, it seeks to find the saddle point such that:

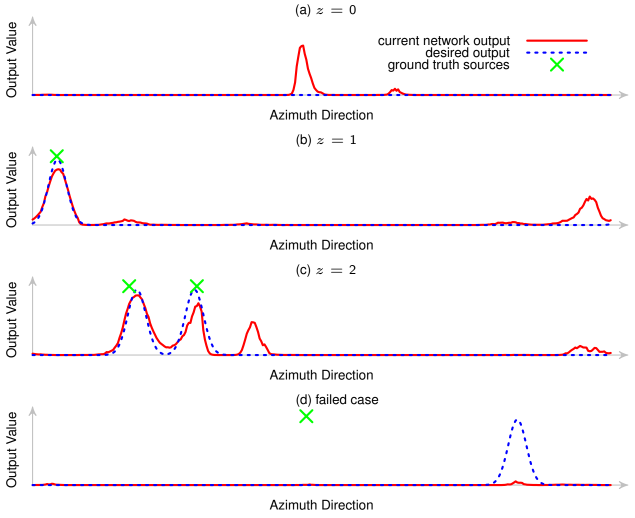

Fig. 3: Examples of output regularization with known number of sources. The desired output is the closest output in the reduced space to the actual network output. The ground truth locations are presented but not used for weak supervision.

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg\min_{\theta_f, \theta_y} E(\theta_f, \theta_y, \hat{\theta}_d), \tag{8}$$

$$\hat{\theta}_d = \arg\max_{\theta_d} E(\hat{\theta}_f, \hat{\theta}_y, \theta_d). \tag{9}$$

On such saddle point, the network yields low prediction and weak supervision loss. At the same time, the feature extractor attempts to fool the domain classifier. In this way, the network extracts domain-invariant features.

4. EXPERIMENTS

We evaluate the proposed method on the adaptation of simulation-based SSL neural networks to real robot data.

4.1. Data

The target domain data are the publicly available real robot recordings from [7]. They are recorded by the robot Pepper, which has four microphone on its head. The data include overlapping speech sources (maximum two sources) corrupted by the robot ego noise, and the frame-level source location ground truth obtained with the robot camera and markers. We use the loudspeaker training data (506k samples, 16 hours) for adaptation, and the weak label (number of sources) are derived from the location ground truth. For evaluation, we use both the loudspeaker and human talker test data.

We generate the source domain data by simulation. The microphone positions are set according to the real microphone

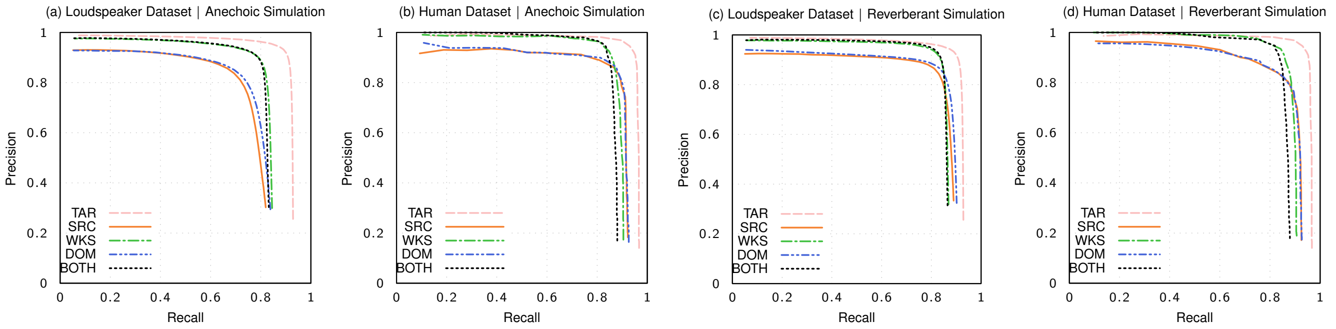

Fig. 4: Localization performance on different simulation conditions and different test sets (indicated by the titles). The DOA estimations are considered correct if their error is less than 5°. The curves are generated by varying the threshold ξ in Eq. 4.

array on the robot, and the source locations are randomly chosen. We generate the impulse responses with the RIR generator [17] and convolve them with clean speech data from the AMI corpus [18]. We simulate two sets of source domain data with different room conditions: anechoic and reverberant (RT60 up to 800ms). Frames of single sources are mixed, so that we generate overlapping sources with diverse location combinations. Finally, the sources are added with the real background noise (robot fan noise) recorded by the robot. We control the SNR (with respect to the background noise) between 0 and 20dB. For each room condition, we generate 1 million samples.

4.2. Network Training

We pre-train the network with the source domain data, according to the two-stage training scheme [16]. The network is trained with supervision on the intermediate time-frequency local prediction in four epochs, before it is supervised on the final output in ten epochs. Then, the network is adapted for ten epochs with both the source and target domain data. We choose $\mu = 1$, and λ varying from 0 to 10^{-3} during the training:

$$\lambda(p) = \left(\frac{2}{1 + \exp(-10p)} - 1 \right) \times 10^{-3}, \qquad (10)$$

where $p \in [0, 1]$ is the adaptation progress. During both the pre-training and adaptation, we use Adam optimizer [19], and mini-batches of size 100.

4.3. Methods

We include the following methods for comparison.

TAR Model trained with fully labelled target domain data.

SRC Unadapted model trained with simulation data.

WKS Adapted with weak supervision ($\lambda = 0$).

DOM Adapted with domain-adversarial training ($\mu = 0$).

BOTH Adapted with both weak supervision and domain-adversarial training.

4.4. Results

We evaluate the localization performance in term of precision and recall, using the same evaluation criteria in [7]. The DOA estimations are considered correct if their error is less than 5°. We generate the precision-recall plots for each individual source domains and different test sets (Fig. 4).

The results show that the weak supervision increases the performance significantly in all the conditions. The absolute precision improvement is roughly 10% while keeping the same recall. This confirms that the number of sources indeed provide useful information for the weakly supervised adaptation.

The adaptation with domain-adversarial training only shows insignificant improvement on the loudspeaker test set (Fig 4a,c). Moreover, combining domain-adversarial training with weak supervision does not improve the result with respect to using only weak supervision. We have explored different values of λ, different architecture of the domain classifier, and different layers of features extractor, however further improvements were not obtained. In practice, introducing domain-invariance suffers the risk of reducing the discriminative power of the features, because the feature extractor may produce irrelevant features in order to fool the domain classifier. Furthermore, finding the balance between domain-invariance and discriminative power is difficult.

5. CONCLUSION

In conclusion, we have studied two methods and their combination for domain adaptation of multiple sound source localization DNNs: weak supervision and domain-adversarial training. The weak supervision regularizes the network output, making it closer to the possible output space based on the inexact labels, such as the number of sources. The domain-adversarial training aims to find domain-invariant features. The experiments have shown significant improvement of models adapted with weak supervision, however, the combination of domain-adversarial training does not further improve the performance according to our experiments.

6. REFERENCES

[1] 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.

[2] Ning Ma, Guy J. Brown, and Tobias 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.

[3] 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.

[4] 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.

[5] Nelson Yalta, Kazuhiro Nakadai, and Tetsuya Ogata, “Sound source localization using deep learning models,” *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, Feb. 2017.

[6] S. Chakrabarty and E. A. P. Habets, “Broadband DOA estimation using convolutional neural networks trained with noise signals,” in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2017, pp. 136–140.

[7] Weipeng He, Petr Motlicek, and Jean-Marc Odobez, “Deep neural networks for multiple speaker detection and localization,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, May 2018, pp. 74–79.

[8] Sharath Adavanne, Archontis Politis, and Tuomas 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, Sept. 2018, arXiv: 1710.10059.

[9] 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.

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

[11] 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*, Apr. 1997, vol. 1, pp. 375–378 vol.1.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Franois Laviolette, Mario Marchand, and Victor Lempitsky, “Domain-adversarial training of neural networks,” *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[13] R. Takeda and K. Komatani, “Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 2217–2221.

[14] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 3514–3518.

[15] 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.

[16] Weipeng He, Petr Motlicek, and Jean-Marc Odobez, “Joint localization and classification of multiple sound sources using a multi-task neural network,” in *Proc. Interspeech 2018*, Hyderabad, India, Sept. 2018, pp. 312–316.

[17] Emanuel AP Habets, “Room impulse response generator,” *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

[18] Iain McCowan, Jean 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*, 2005, vol. 88.

[19] Diederik P. Kingma and Jimmy Ba, “Adam: a method for stochastic optimization,” in *International Conference on Learning Representations (ICLR)*, San Diego, May 2015.