Importance-Weighted Domain Adaptation for Sound Source Tracking

Bingxiang Zhong[1], *Thomas Dietzen*[1]

[1]KU Leuven, Dept. of Electrical Engineering (ESAT-PSI), EAVISE, Sint-Katelijne-Waver, Belgium.
{bingxiang.zhong, thomas.dietzen}@kuleuven.be

***Abstract*—In recent years, deep learning has significantly advanced sound source localization (SSL). However, training such models requires large labeled datasets, and real recordings are costly to annotate in particular if sources move. While synthetic data using simulated room impulse responses (RIRs) and noise offers a practical alternative, models trained on synthetic data suffer from domain shift in real environments. Unsupervised domain adaptation (UDA) can address this by aligning synthetic and real domains without relying on labels from the latter. The few existing UDA approaches however focus on static SSL and do not account for the problem of sound source tracking (SST), which presents two specific domain adaptation challenges. First, variable-length input sequences create mismatches in feature dimensionality across domains. Second, the angular coverages of the synthetic and the real data may not be well aligned either due to partial domain overlap or due to batch size constraints, which we refer to as directional diversity mismatch. To address these, we propose a novel UDA approach tailored for SST based on two key features. We employ the final hidden state of a recurrent neural network as a fixed-dimensional feature representation to handle variable-length sequences. Further, we use importance-weighted adversarial training to tackle directional diversity mismatch by prioritizing synthetic samples similar to the real domain. Experimental results demonstrate that our approach successfully adapts synthetic-trained models to real environments, improving SST performance.**

***Index Terms*—sound source tracking, sound source localization, unsupervised domain adaptation, importance weighting, adversarial training**

1. INTRODUCTION

Sound source localization (SSL), which refers to estimating the direction of arrivals (DoAs) of sound sources, is fundamental to applications requiring directional information, including speech enhancement [1], human-robot interaction [2], and automotive applications [3]. Traditional approaches exploit explicit models based on time difference of arrivals (TDoAs), such as steered response power with phase transform (SRP-PHAT) [4], [5], multiple signal classification (MUSIC) [6], and estimation of signal parameters via rotational invariance techniques (ESPRIT) [7]. However, these methods suffer significant performance degradation in challenging acoustic conditions with low signal-to-noise ratios (SNR) and high reverberation.

Deep learning has recently demonstrated substantial improvements [8]–[11] over traditional SSL methods, but their success critically depends on large-scale labeled datasets. Labeling data is time-consuming and costly, especially for moving sources, leading to a scarcity of real-world labeled data. As an alternative, room impulse response (RIR) simulation tools [12], [13] are used to generate extensive synthetic training data with diverse SNR and reverberation conditions, while small portions of real-world data serve for evaluation. However, models trained on synthetic data typically suffer performance drops due to simplified simulation models that do not fully capture the complexity of room acoustics and microphone characteristics, creating distribution mismatches between synthetic and real data, a phenomenon known as domain shift [14].

Unsupervised domain adaptation (UDA) attempts to mitigate the effect of domain shift by aligning features across source (synthetic) and target (real-world) domains without requiring target labels. While extensively investigated in computer vision and natural language processing [15], few studies have addressed this problem in SSL. Takeda et al. [16] proposed entropy minimization to regularize DoA classification probabilities, and later [17] introduced an eliminative posterior probability constraint that forces the probability of less likely candidates to become zero to eliminate incoherent errors. However, both approaches are limited to discrete classification formulations. Domain adversarial training [18] for SSL has shown mixed results. While [19] reported minimal improvements, [20] achieved significant gains using ensemble domain adversarial networks with multiple domain discriminators at different network layers. Notably, none of these approaches address domain adaptation specifically for sound source tracking (SST).

SST presents unique UDA challenges beyond static localization. First, tracking requires processing variable-length temporal sequences, while many existing domain adaptation methods [15], [18], [21] are designed for fixed feature dimensionality and cannot directly handle temporal variability. Second, directional diversity mismatch between source and target data creates feature alignment instability in domain adaptation through two effects. At the domain level, angular coverages may differ significantly. While source domains typically contain full directional diversity, target domains may cover only limited angular ranges depending on recording settings, creating partial domain overlap [22]. At the batch level, SST systems must handle long-duration audio sequences (typically > 10 seconds [8], [9]), necessitating small batch sizes due to memory constraints. Consequently, directional diversity within batches may differ significantly between source and target data, creating inconsistent feature distributions that hinder effective domain alignment. These challenges render standard adversarial training approaches [19], [20] ineffective, as they fundamentally assume complete label space alignment both across domains and within batches [22].

To address these challenges, we propose a novel UDA approach tailored for SST with two key contributions. First, adopting the SST architecture from [9], [23], we employ the final hidden state of a recurrent neural network (RNN) as a fixed-dimensional feature representation within the domain discriminator to handle variable-length sequences. Second, we leverage importance-weighted adversarial training [22] to tackle directional diversity mismatch by prioritizing synthetic samples similar to the real domain. Experimental results demonstrate that our approach successfully adapts synthetic-trained models to real environments, improving SST performance. The code is available at [24].

In section 2, we review a commonly used SST model adopted for domain adaptation. In section 3, we describe the proposed domain adaptation approach. Section 4 presents our experimental evaluation, and section 5 concludes the paper.

2. SOUND SOURCE TRACKING

We adopt the convolutional recurrent neural network (CRNN) architecture used in [9], [23] to perform single-source SST. The model

This project has received funding from KU Leuven internal funds STG/24/013.

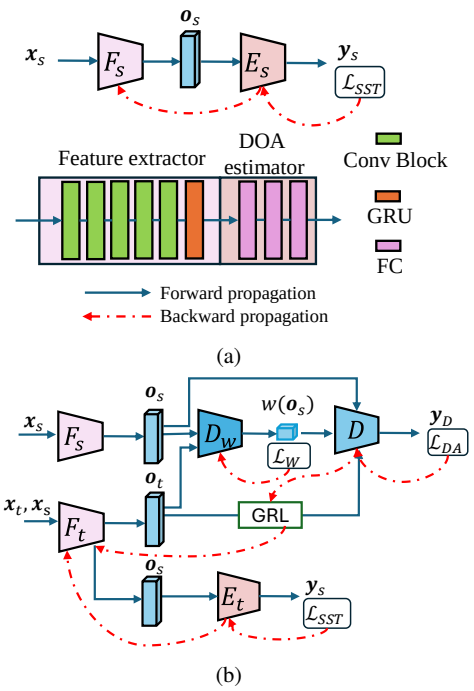


Fig. 1: The overview of the proposed method. (a) The CRNN model will first be trained on the source domain. (b) The pretrained model is adapted to the target domain using importance-weighted adversarial training.

is first trained on source domain data, considering tracking of sound sources in the azimuth direction spanning the range $[0, \pi)$. The input to the SST system concatenates the real and imaginary components of the short-time Fourier transform (STFT) of multi-channel acoustic signals, denoted as $\mathbf{x}_s$ of size $2M \times K \times L$, where M is the number of microphone channels, K is the number of frequency bins, and L is the number of time frames. Rather than directly regressing angular directions, we employ likelihood encoding as in [19], [23], [25], where the network output represents the probability of source existence across $J = 180$ candidate directions spanning the range $[0, \pi)$. The model is trained to regress a sequence of likelihood values for ground truth DoA values in the azimuth direction, denoted as $\mathbf{y}_s$ with dimensions $L \times J$. Given ground truth DoA values $\phi'(l)$ at each voice-active time frame l, and 180 candidate directions $\{\phi_i\}_{i=1}^{180}$, the desired output values are encoded via Gaussian functions

$$y_s(\phi_i, l) = \exp\Big\{-\frac{(\phi_i - \phi'(l))^2}{2\sigma^2}\Big\}, \tag{1}$$

where σ defines the beam width, set to 16° as in [23]. For non-voice-active frames, we set $y_s(\phi_i, l) = 0$. Voice-active frames are identified using the WebRTC voice activity detector (VAD) [26].

The network architecture of the adopted SST model [9], [23] is shown in Fig. 1a. To facilitate the following explanation of domain adaptation, we split the model into two parts: a feature extractor F_s and a DoA estimator E_s, both specifically for the source domain. The feature extractor comprises five convolutional blocks followed by a gated recurrent unit (GRU). Each convolutional block contains two consecutive convolutional layers with 3×3 kernel size and 64 output channels, batch normalization, ReLU activation, and max pooling. Max pooling operations have kernel sizes of [4, 2, 2, 2, 2] along the frequency axis and [1, 1, 1, 1, 5] along the time axis. The resulting convolutional features are processed by a unidirectional GRU with 256 hidden units. The output of the GRU, denoted by $\mathbf{o}_s = F_s(\mathbf{x}_s)$, serves as the input to the DoA estimator E_s and also functions as a domain feature in the proposed domain adaptation approach. The DoA estimator consists of three fully connected (FC) layers with dimensions 512, 256, and 180, utilizing Tanh, ReLU, and sigmoid activations, respectively.

The model is trained to minimize the mean squared error (MSE) between the network predictions and the ground truth DoA likelihood representations. Given a batch of N input data samples $\mathbf{X}_s = \{\mathbf{x}_s^{(n)}\}$ with corresponding labels $\mathbf{Y}_s = \{\mathbf{y}_s^{(n)}\}$, $n = 1, ..., N$, the MSE loss for SST is computed as

$$\mathcal{L}_{SST}(F_s, E_s) = \frac{1}{N}\sum_{n=1}^{N} ||E_s(F_s(\mathbf{x}_s^{(n)})) - \mathbf{y}_s^{(n)}||_2^2, \tag{2}$$

where both F_s and E_s are updated during backward propagation. During inference, the estimated DoA is obtained by identifying the peak in the estimated encoded likelihood.

3. IMPORTANCE-WEIGHTED DOMAIN ADAPTATION FOR SOUND SOURCE TRACKING

Let $\mathcal{D}_s = \{\mathbf{X}_s, \mathbf{Y}_s\}$ represent the labeled source domain and $\mathcal{D}_t = \{\mathbf{X}_t\}$ represent the unlabeled target domain, where data samples are drawn from distributions p_s and p_t respectively. When deploying the model trained on the source domain to the target domain, the domain shift of the distribution mismatch between source and target domain data, i.e., $p_s \neq p_t$, leads to performance degradation on the target domain.

While domain adaptation approaches exist to address such distribution mismatch, SST presents two particular challenges: variable length features and diectional diversity mismatch between $\mathcal{D}_s$ and $\mathcal{D}_t$. To address these issues, we propose to use the final hidden state of the RNN to handle variable length sequences and adopt importance-weighted domain adversarial training [22] originally introduced for image classification and employ it for SST. The resulting architecture is shown in Fig. 1b.

The adaptation process begins by initializing the target domain feature extractor F_t and DoA estimator E_t with weights from the pretrained model. During domain adaptation, F_s remains frozen to maintain consistent source domain representations, while F_t adapts to extract domain-invariant features from both the source and target domain data. The approach incorporates two binary domain discriminators, D_w and D, which share an identical architecture. Source samples are labeled as 1 and target samples as 0. D_w facilitates the importance weighting mechanism, while D performs domain adversarial training. In the following, we describe the domain discriminator architecture, importance-weighted adversarial training, and importance weighting mechanism.

3.1. Domain discriminator architecture

To handle the variable-length input sequences, we design our discriminator architecture to effectively aggregate long and variable-length temporal dynamics. Unlike previous works [19], [20] that operate on fixed-dimensional features and can therefore use convolutional layers or simple FC layers for domain discrimination, our approach must address the challenge of variable-length temporal features. Therefore, we design our discriminator with a GRU and employ its final hidden state to encode the complete temporal dynamics into a fixed-dimensional representation independent of input length.

Both discriminators first process features $\mathbf{o}_s$ and $\mathbf{o}_t$ from the feature extractors through a 256-unit GRU layer. The final hidden state of this GRU then passes through two FC layers: a 128-unit layer with ReLU

activation followed by a single-output layer with sigmoid activation for domain classification.

3.2. Importance-weighted adversarial training

The importance-weighted adversarial training [22] follows the same training approach as standard adversarial training [18]. The core principle is to align feature representations across domains through a weighted adversarial game between a feature extractor and a domain discriminator. In our implementation, the features $\mathbf{o}_s$ and $\mathbf{o}_t$ extracted from source and target domain data serve as input to the domain discriminator D. The loss function is given by [22],

$$\begin{aligned}\mathcal{L}_{DA}(D, F_t) = \frac{1}{N}\sum_{n=1}^{N}[w(\mathbf{o}_s^{(n)}) \log D(\mathbf{o}_s^{(n)}) \\ + \log(1 - D(\mathbf{o}_t^{(n)}))],\end{aligned} \tag{3}$$

where $\mathbf{o}_s^{(n)} = F_s(\mathbf{x}_s^{(n)})$ and $\mathbf{o}_t^{(n)} = F_t(\mathbf{x}_t^{(n)})$. For standard adversarial training [18], $w(\mathbf{o}_s^{(n)}) = 1$. The objective is to solve a minimax game, $\min_{F_t} \max_D \mathcal{L}_{DA}(D, F_t)$. The discriminator D maximizes the objective by correctly classifying source domain features and target domain features. Simultaneously, the target feature extractor F_t minimizes the objective by generating features that fool the discriminator. Through this adversarial process, the feature extractor learns domain-invariant representations that enable effective knowledge transfer from the labeled source domain to the unlabeled target domain. A gradient reversal layer (GRL) enables simultaneous training of the minmax game by reversing gradients for the feature extractor.

3.3. Importance weighting mechanism

To obtain the importance weights w, a second domain discriminator D_w is employed. The goal is to assign larger weights to source data samples that share similar feature characteristics with the target data, while assigning smaller weights to those with less overlap. This discriminator is first trained to distinguish between source and target domains using the binary cross entropy loss,

$$\mathcal{L}_W(D_w) = -\frac{1}{N}\sum_{n=1}^{N}[\log D_w(\mathbf{o}_s^{(n)}) + \log(1 - D_w(\mathbf{o}_t^{(n)}))]. \tag{4}$$

Since the final activation function in D_w is the logistic sigmoid function, the output $D_w(\mathbf{o}_s)$ represents the probability that a source sample belongs to the source domain. When $D_w(\mathbf{o}_s) \approx 1$, source samples are highly distinguishable from target samples and should receive lower weights in the adversarial training. Conversely, when $D_w(\mathbf{o}_s) \approx 0$, the source samples are similar to target samples and should receive higher weights. The importance weight is therefore inversely related to $D_w(\mathbf{o}_s)$,

$$w(\mathbf{o}_s) = 1 - D_w(\mathbf{o}_s). \tag{5}$$

These weights are normalized by the mean weight value per batch [22]. Note that the domain discriminator D_w is used solely to calculate the importance weights and does not participate in the adversarial training; therefore, gradients from D_w do not back propagate to update the feature extractor F_t. The obtained weights are then incorporated into the loss function in (3) to prioritize source samples that are most similar to the target domain during adversarial training.

3.4. Overall training objective

To preserve source domain performance while adapting to the target domain, we minimize the combined loss function,

$$\begin{aligned}\mathcal{L}(D, D_w, F_t, E_t) = \mathcal{L}_{SST}(F_t, E_t) \\ + \lambda\mathcal{L}_{DA}(D, F_t) + \mathcal{L}_W(D_w).\end{aligned} \tag{6}$$

Here, $\mathcal{L}_{SST}$ is computed exclusively using source domain data to retain performance on the original task. The trade-off parameter λ controls the influence of the domain adversarial loss. When λ is too small, domain alignment is weak, and the model behaves similarly to source-only training. When λ is too large, the adversarial objective dominates, encouraging excessive domain invariance and causing the learned features to lose essential task-relevant characteristics.

4. EXPERIMENTS

4.1. Experimental setups

4.1.1. Datasets: **Target Domain (Real-world Data)**: we use the RealMAN dataset [23], which contains 35.4 hours of moving speaker recordings across 32 different acoustic scenes with background noise, sampled at 48 kHz. The dataset employs a 32-channel microphone array, from which we extract a 9-channel subarray consisting of an 8-channel uniform circular array (3 cm radius) plus a center microphone. In the experiments, we consider single-source moving scenarios in three distinct acoustic scenes spanning indoor to outdoor environments: OfficeRoom3, OfficeLobby, and BasketballCourt1. These scenes provide training, validation, and test data with durations of (81, 7, 8), (51, 28, 8), and (40, 8, 42) minutes, respectively. We select these three scenes because they provide complete training, validation, and testing splits along with scene-specific background noise recordings. During adversarial training, the training data is augmented with background noise at SNR levels ranging from -10 dB to 15 dB, matching the SNR range found in the validation and test data [23].

Source Domain (Synthetic Data): we generate synthetic data using the same 9-channel microphone configuration as the target domain to ensure acoustic consistency. Following established protocols [8], we simulate rooms with dimensions randomly selected from 3×3×2.5 m to 10×8×6 m and reverberation times from 0.2 s to 1.0 s. RIRs are generated using gpuRIR [13]. Speech signals are randomly selected from the LibriSpeech corpus [27] and convolved with the generated RIRs, then corrupted with diffused noise at SNRs ranging from -10 dB to 15 dB. The diffused noise is generated using the toolbox [28] applied to the NOISEX-92 corpus [29]. Speech segments vary from 2 s to 10 s in duration. To enhance training diversity, each sample is generated on-the-fly with randomized acoustic conditions including source trajectories, microphone positions, source signals, noise characteristics, reverberation times, and SNRs.

Both target and source domain data is downsampled to 16 kHz for computational efficiency. STFT analysis uses a 32 ms window with 20 ms frame shift to extract time-frequency representations.

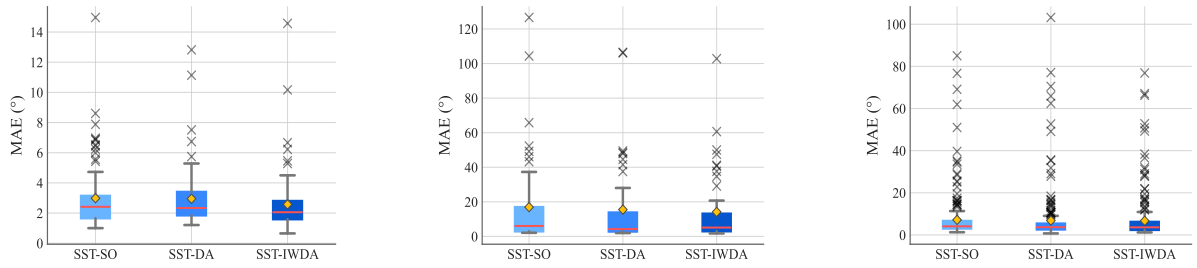
4.1.2. Evaluation metrics: Performance is evaluated during voice-active segments using mean absolute error (MAE) of azimuth estimates for across all time frames, and localization accuracy (Acc), defined as the percentage of frames with MAE below 5°.

4.1.3. Training configuration: The training process comprises two phases. In the first phase, the model is trained on the source domain for 150 epochs with a batch size of 16, using the Adam optimizer [30] and a learning rate of 0.0001. The model checkpoint is selected based on its MAE performance on simulated testing data.

Subsequently, the model undergoes domain adaptation for an additional 60 epochs with the same learning rate, and early stopping. To mitigate training instability in early stages, we employ a gradual schedule for the adversarial trade-off parameter following [18],

$$\lambda = \frac{2u}{1 + \exp(-p)} - u, \tag{7}$$

where p gradually increases over training steps, allowing λ to smoothly approach its upper bound u. The model checkpoint selection is



(a) OfficeRoom3 (b) OfficeLobby (c) BasketballCourt1

Fig. 2: Boxplots comparing the MAE distributions of different methods across various acoustic scenes. The boxes span the interquartile range (IQR, 25th to 75th percentiles), with the red line marking the median. Whiskers extend to the minimum and maximum values within 1.5×IQR from the quartiles. Diamond markers indicate mean MAE values, while crosses denote outliers beyond the whiskers.

also based on the model's MAE performance on a small validation set. Although this may appear to conflict with the principles of unsupervised learning, validation-based model selection is widely accepted in domain adaptation for computer vision tasks [18], [21], [22]. While alternative strategies for checkpoint selection without validation sets exist [31], they are typically tailored to classification problems. Moreover, collecting and annotating a small validation set is far more feasible than creating large annotated training datasets.

4.1.4. Baselines: We denote the proposed method as SST-IWDA. To evaluate its effectiveness, we compare it with two baseline models. The first baseline, SST-SO, is a source-only model trained exclusively on source domain data. For a fair comparison, the pretrained model is further trained for 60 additional epochs, with model checkpoint selection based on validation performance on the target domain data for each acoustic scene. The second baseline, SST-DA, implements standard adversarial domain adaptation [18] without the importance weighting mechanism.

4.2. Results and discussion

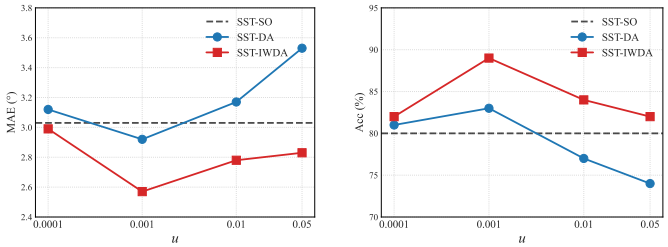
In the first experiment, we adapt the pretrained model using target data from various acoustic environments. The results using $u = 0.001$, presented in Table 1, show that SST-DA yields modest performance gains over the source-only baseline SST-SO. However, the proposed approach SST-IWDA consistently achieves the lowest MAE and highest accuracy among all methods. For example, in OfficeRoom3, SST-IWDA reduces the MAE from 3.03° to 2.57° and boosts accuracy from 80% to 89%. Similarly, in BasketballCourt1, we observe a reduction in MAE from 6.26° to 5.62°, with accuracy improving from 68% to 76%.

While SST-SO performs reasonably well in OfficeRoom3 and BasketballCourt1, both characterized by relatively low MAEs, its performance deteriorates substantially in OfficeLobby, where the estimation error reaches 17.12°. This degradation is likely due to a pronounced mismatch in spatial correlation coefficients between the real-world office lobby noise and the diffuse noise assumption employed in our simulated training data [23]. Despite this domain gap, SST-IWDA improves the MAE by nearly 2° (from 17.12° to 15.39°) and increases accuracy by 5 percentage points.

The boxplot analysis in Fig. 2 reveals additional insights into the performance of the proposed approach. Domain adaptation can reduce both the number of outliers and their magnitude, which correlates with the observed increase in accuracy. SST-IWDA consistently reduces the interquartile range (IQR) and whisker lengths, as well as the median values. In contrast, SST-DA exhibits inconsistent behavior: it sometimes increases the quartile values and whisker lengths (as observed in OfficeRoom3), while in other cases it reduces the IQR effectively but introduces outliers with larger magnitudes (as seen in BasketballCourt1).

Table 1: Performance comparison across different acoustic scenes (lower MAE and higher Acc indicate better performance).

	OfficeRoom3		OfficeLobby		BasketballCourt1	
Method	MAE (°)	Acc (%)	MAE (°)	Acc (%)	MAE (°)	Acc (%)
SST-SO	3.03	80	17.12	51	6.26	68
SST-DA	2.92	83	16.19	**56**	5.93	71
SST-IWDA	**2.57**	**89**	**15.39**	**56**	**5.62**	**76**



(a) MAE (b) Acc

Fig. 3: Domain adaptation performance across different values of parameter u for the OfficeRoom3 acoustic scene.

In the second experiment, we evaluate domain adaptation performance in OfficeRoom3 under varying values of the upper bound parameter u from (7), which controls the strength of adversarial training. We test u of values 0.0001, 0.001, 0.01, 0.05 for both SST-DA and SST-IWDA. As shown in Fig. 3, small u values yield performance similar to SST-SO. Increasing u initially improves results, but excessive values degrade both MAE and accuracy. SST-IWDA consistently outperforms SST-SO across all tested values, whereas SST-DA rapidly surpasses SST-SO's MAE (3.03°) and drops below its Acc (80%) as u rises. These findings confirm the robustness of the proposed importance-weighted adversarial training in balancing adaptation and task-specific learning.

5. CONCLUSION

This paper proposes a novel domain adaptation approach for SST with two key contributions. First, we utilize GRU's final hidden state to encode variable-length sequences into fixed-dimensional representations within the domain discriminator. Second, we use importance-weighted adversarial training to mitigate directional diversity mismatch between synthetic and real data distributions. Experimental results demonstrate that the proposed method outperforms baseline approaches and exhibits enhanced robustness across different scenes.

REFERENCES

[1] H.-Y. Lee, J.-W. Cho, M. Kim, and H.-M. Park, "DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1091–1095, 2016.

[2] S. Argentieri, P. Danes, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech & Language*, vol. 34, no. 1, pp. 87–112, 2015.

[3] Y. Schulz, A. K. Mattar, T. M. Hehn, and J. F. Kooij, "Hearing what you cannot see: Acoustic vehicle detection around corners," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2587–2594, 2021.

[4] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: signal processing techniques and applications*. Springer, 2001, pp. 157–180.

[5] E. Grinstein, E. Tengan, B. Çakmak, T. Dietzen, L. Nunes, T. van Waterschoot, M. Brookes, and P. A. Naylor, "Steered response power for sound source localization: A tutorial review," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 59, 2024.

[6] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[7] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 2002.

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

[9] B. Yang, H. Liu, and X. Li, "SRP-DNN: Learning direct-path phase difference for multiple moving sound source localization," in *Proceedings 2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2022)*, Singapore, Singapore, May 2022, pp. 721–725.

[10] E. Grinstein, C. M. Hicks, T. van Waterschoot, M. Brookes, and P. A. Naylor, "The neural-SRP method for universal robust multi-source tracking," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 19–28, 2023.

[11] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

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

[13] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[14] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, pp. 151–175, 2010.

[15] P. Singhal, R. Walambe, S. Ramanna, and K. Kotecha, "Domain adaptation: challenges, methods, datasets, and applications," *IEEE access*, vol. 11, pp. 6973–7020, 2023.

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

[17] 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 *Proceedings 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, Apr. 2018, pp. 3514–3518.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[19] W. He, P. Motlicek, and J.-M. Odobez, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *Proceedings 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 770–774.

[20] G. Le Moing, P. Vinayavekhin, D. J. Agravante, T. Inoue, J. Vongkulbhisal, A. Munawar, and R. Tachibana, "Data-efficient framework for real-world multiple sound source 2D localization," in *Proceedings 2021 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2021)*, Toronto, ON, Canada, Jun. 2021, pp. 3425–3429.

[21] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proceedings Computer Vision–ECCV 2016 Workshops (ECCV 2016)*, Amsterdam, the Netherlands, Oct. 2016, pp. 443–450.

[22] J. Zhang, Z. Ding, W. Li, and P. Ogunbona, "Importance weighted adversarial nets for partial domain adaptation," in *Proceedings 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, Salt Lake City, UT, USA, Jun. 2018, pp. 8156–8164.

[23] B. Yang, C. Quan, Y. Wang, P. Wang, Y. Yang, Y. Fang, N. Shao, H. Bu, X. Xu, and X. Li, "RealMAN: A real-recorded and annotated microphone array dataset for dynamic speech enhancement and localization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 105 997–106 019, 2024.

[24] B. Zhong and T. Dietzen, "Importance-weighted domain adaptation for sound source tracking," GitHub repository, 2025. [Online]. Available: https://github.com/bingxiang-zhong/SST-IWDA

[25] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proceedings 2018 IEEE International Conference on Robotics and Automation (ICRA 2018)*, Brisbane, QLD, Australia, May 2018, pp. 74–79.

[26] J. Wiseman, "py-webrtcvad: Python interface to the WebRTC voice activity detector," GitHub repository, Nov. 2019. [Online]. Available: https://github.com/wiseman/py-webrtcvad

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proceedings 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5206–5210.

[28] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[29] A. Varga and H. Steeneken, "II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings 2015 International Conference on Learning Representations (ICLR 2015)*, San Diego, USA, May 2015.

[31] J. Yang, H. Qian, Y. Xu, K. Wang, and L. Xie, "Can we evaluate domain adaptation models without target-domain labels?" *arXiv preprint arXiv:2305.18712*, 2023.