applied sciences

MDPI

Review

A Review on Sound Source Localization in Robotics: Focusing on Deep Learning Methods

Reza Jalayer [1,*], Masoud Jalayer [2,3] and Amirali Baniasadi [4]

[1] Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 24/b, 20156 Milan, Italy
[2] Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland; masoud.jalayer@aalto.fi
[3] Department of Materials and Mechanical Engineering, University of Turku, 20014 Turku, Finland
[4] Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8P 5C2, Canada; amiralib@uvic.ca
* Correspondence: reza.jalayer@polimi.it

Abstract

Sound source localization (SSL) adds a spatial dimension to auditory perception, allowing a system to pinpoint the origin of speech, machinery noise, warning tones, or other acoustic events, capabilities that facilitate robot navigation, human–machine dialogue, and condition monitoring. While existing surveys provide valuable historical context, they typically address general audio applications and do not fully account for robotic constraints or the latest advancements in deep learning. This review addresses these gaps by offering a robotics-focused synthesis, emphasizing recent progress in deep learning methodologies. We start by reviewing classical methods such as time difference of arrival (TDOA), beamforming, steered-response power (SRP), and subspace analysis. Subsequently, we delve into modern machine learning (ML) and deep learning (DL) approaches, discussing traditional ML and neural networks (NNs), convolutional neural networks (CNNs), convolutional recurrent neural networks (CRNNs), and emerging attention-based architectures. The data and training strategy that are the two cornerstones of DL-based SSL are explored. Studies are further categorized by robot types and application domains to facilitate researchers in identifying relevant work for their specific contexts. Finally, we highlight the current challenges in SSL works in general, regarding environmental robustness, sound source multiplicity, and specific implementation constraints in robotics, as well as data and learning strategies in DL-based SSL. Also, we sketch promising directions to offer an actionable roadmap toward robust, adaptable, efficient, and explainable DL-based SSL for next-generation robots.

Keywords: sound source localization; auditory perception; speech recognition; human–robot interaction; deep learning

Academic Editors: Francesco Costantino, Silvia Colabianchi and Margherita Bernabei

Received: 3 July 2025
Revised: 17 August 2025
Accepted: 20 August 2025
Published: 26 August 2025

Citation: Jalayer, R.; Jalayer, M.; Baniasadi, A. A Review on Sound Source Localization in Robotics: Focusing on Deep Learning Methods. *Appl. Sci.* **2025**, *15*, 9354. https://doi.org/10.3390/app15179354

1. Introduction

Sound source localization (SSL) is the task of estimating the location or direction of a sound-emitting source relative to the sensor (microphone array). In robotic systems, SSL serves as a crucial component of robot audition, greatly enhancing a robot's perceptual capabilities [1]. An accurate SSL module enables a robot to orient its sensors towards the active speaker, disambiguate simultaneous talkers, or navigate autonomously towards an event that is audible but not visible. SSL serves various functional roles in robotics.

For instance, speech localization is critical for human–robot interaction (HRI), enabling a robot to accurately identify the location of a person giving a command, even in noisy environments [2,3]. While the domestic and service applications of SSL in HRI are well-documented, its role in industrial settings is crucial yet often understated, particularly for ensuring Occupational Health and Safety (OHS). Accurate SSL allows a robot to precisely identify and respond to human voice commands, including critical safety alerts, and to detect the presence of humans to avoid collisions on a noisy factory floor. Localizing machinery noise is also vital for fault detection and condition monitoring in industrial manufacturing. This allows a robot to detect the acoustic signature of a failing component and inspect the production line before it leads to a catastrophic and potentially dangerous system failure [4–6]. Furthermore, the detection of specific warning tones or the sudden appearance of new sound sources can be integrated into simultaneous localization and mapping (SLAM). This provides audio-goal seeking capabilities and can furnish acoustic landmarks or constraints when vision is occluded, enriching the robot's understanding of its environment. SSL is also a core component in other robotic applications, including autonomous vehicles [7], aerial robots [8,9], search-and-rescue systems [10–12], and security robots [13,14], where localizing auditory events is essential for effective decision making.

Alternative localization in robotics can be done by vision, light detection and ranging (LiDAR), Wi-Fi, Bluetooth, Global Positioning Systems (GPS), infrared (IR), and radio frequency identification (RFID). However, each of the aforementioned techniques has its pros and cons [15]. For example, cameras are susceptible to occlusion, darkness, glare, or privacy constraints [16,17]; LiDARs and Wi-Fi-based techniques are highly accurate in localization but expensive [18,19]; IR-based and Bluetooth localization are only accurate in specific conditions such that IR signals are easily obscured by physical obstacles and Bluetooth is range-limited; GPS is rendered unreliable by the presence of multiple building walls [20] and RFID-based localization suffers from interference between readers in the presence of multiple RFID tags and readers in the scene [21]. Acoustic waves, in contrast, propagate around obstacles and in complete darkness, furnishing a complementary channel that often operates beyond the line of sight of on-board cameras. SSL is therefore not a rival but a synergistic partner to these modalities, enriching multi-sensor fusion and improving the robustness of perceptual pipelines.

Despite these advantages, sound-based localization is notoriously sensitive to microphone geometry, environmental reverberation and noise, and the presence of multiple simultaneous sound emitters. Early robotic SSL systems, dating back to the Squirt robot in 1989 [22], relied on classical signal-processing algorithms [1]. These techniques model microphone spacing, speed of sound, and narrow-band free-field assumptions analytically; they achieve good performance in anechoic rooms or with a single talker but degrade quickly under strong reverberation, diffuse noise, and source motion. The last decade has witnessed a paradigm shift toward data-driven learning, mirroring breakthroughs in computer vision and speech recognition.

At the core of this shift lies deep learning, a class of machine learning techniques based on multi-layered neural networks capable of automatically extracting informative features from raw audio or spectrograms. Unlike classical signal-processing methods that rely on predefined rules or handcrafted features, deep learning models learn hierarchical representations directly from the data. This makes them especially effective in real-world SSL scenarios, where factors such as reverberation, background noise, and multiple overlapping sources present significant challenges. Over time, deep learning architectures have evolved to better capture both spatial and temporal information in acoustic signals, resulting in more robust and accurate localization systems.

Common deep learning-based SSL architectures include convolutional neural networks (CNNs) for spatial feature extraction from spectrograms, and convolutional–recurrent hybrids (CRNNs) that combine CNNs with recurrent layers to capture temporal dynamics. Further improvements such as the residual and densely connected networks facilitate the training of deeper models by improving gradient flow and encouraging feature reuse. More recently, attention-based transformers have demonstrated a remarkable ability to learn spatial features directly from raw waveforms or spectrograms [23]. These models are typically trained on large collections of synthetic or recorded room impulse responses, which help them generalize to unseen environments with varying noise and reverberation. The growing availability of low-cost microphone arrays and compact edge computing devices has also enabled the deployment of such models on mobile robotic platforms, opening up new opportunities for real-time, robust sound localization across a range of robotic applications.

Looking carefully to the existing reviews, we found many surveys on sound source localization in the past three decades; however, to better illustrate new studies, we summarize the ones published after 2017 in Table 1. A closer look at recent reviews indicates that no review has explored SSL in robotic platforms with a particular focus on deep learning models.

Table 1. Comparative summary of sound-source-localization (SSL) review papers.

Review Paper	Ref.	Year	Time Span	Main Focus	Robotics	DL-SSL on Robots
Localization of sound sources in robotics: A review	[1]	2017	Up to 2017	SSL in robotics; conventional techniques; SSL facets	Yes	No
Localization of sound sources: A systematic review	[24]	2021	2011–2021	SSL methods; influencing factors; practical constraints	No	No
Survey of sound source localization with deep learning methods	[23]	2022	2011–2021	DL-based SSL techniques, architectures, datasets	No	No
A review on sound source localization systems	[25]	2022	Up to 2021	Array types; classical vs. CNN-based; challenges (general SSL)	Partial	No
Nonverbal sound in human–robot interaction: A systematic review	[26]	2023	Up to April 2022	Non-verbal sound in HRI (broader than SSL)	Yes	No
A survey of sound source localization and detection methods and their applications	[27]	2023	Up to 2023	Classical + AI SSL methods and applications (general)	No	No
An overview of sound source localization based condition-monitoring robots	[5]	2024	Up to 2024	SSL in CMRs (domain-specific robotics)	Yes	Partial
A review on recent advances in sound source localization techniques, challenges, and applications	[28]	2025	Up to 2025	General SSL system architectures and types (multi-domain)	No	No

From the eight post-2017 SSL reviews in Table 1, only 3/8 (37.5%) have a robotics focus at all, and none of the three have a primary focus on deep learning methods. One review

is robotics-domain-specific (condition-monitoring robots) but does not center on DL-SSL. To make this explicit, we add a column "DL-SSL on robots?" in Table 1, which shows the absence of such a synthesis to date. This quantitative snapshot motivates our contribution: up-to-date literature bridging these two rapidly evolving fields. In this regard, Rascon and Meza's seminal survey [1], which could be the closest review to our review due to its robotic theme, predates the DL boom since it included papers before 2017. Liaghat et al. [24] had a broader focus to systematically review SSL works without focusing on specific applications (e.g., robotics), while reviewing the SSL methods between 2011 and 2021, overlooking the focus on DL models. In contrast, Grumiaux et al. [23] focused on new methods (from 2011 to 2021), especially in DL methods and their challenges. This review is very informative in the general domain of SSL and was very well cited, while it does not restrict its focus to robotics. The review by Desai and Mehendale [25], surveyed the SSL works based on the number of microphones, i.e., two microphones mimicking human auditory systems (binaural) and multiple microphones, and also based on the method (based on classical vs. convolutional neural networks). They also provided very informative information in general SSL, such as the challenges of each SSL systems, while they did not focus on robotics SSL. Zhang et al. [26], had a different focus such that it explored nonverbal sound in human–robot interaction by offering new taxonomies of function (perception vs. creation) and form (vocables, mechanical sounds, etc.). While not strictly about SSL, it identified how sound contributes to robot perception and communication, and emphasized underexplored aspects like robot-generated sound and shared datasets. Their focus was novel and important for understanding how SSL integrates into broader auditory HRI. Jekaterynczuk and Piotrowski [27] offered an extensive comparison of classical and AI-based SSL methods, with interesting classification by mic configuration, signal parameters, and neural architectures. They categorized the application of SSL works based on civil and military domains, and did not focus on robotics specifically. Lv et al. [5] recently wrote a very interesting and specific review on SSL in condition-monitoring robots (CMRs). They reviewed the diverse SSL techniques, including traditional and machine learning models. Their review was specifically narrowed to CMRs, and since SSL has not been extensively explored in there, they proposed a framework for future studies in this field. They also encourage future researchers in different condition monitoring tasks who use mobile robots to include SSL in addition to their existing monitoring systems, such as visual, infrared, etc. The most recent review on SSL was carried out by Khan et al. [28], where they explored traditional SSL as well as machine learning models. They also categorized works based on different applications, such as industrial domains, medical science, and speech enhancement.

This review fills the gap in other surveys by reviewing peer-reviewed literature from 2013 to 2024 in which different methods (especially machine learning and deep learning) in SSL are applied to, or evaluated on, robotic platforms. It also explained the fundamental SSL facets and traditional SSL as well as DL methods. We also explore the data, a pillar in deep learning models, together with different training strategies, then map SSL functions to concrete robotic tasks including service, social, search and rescue, and industrial applications and outline open challenges and future avenues. To better clarify our review structure, we depict an overall structure in Figure 1 explaining different aspects of our review.

The remainder of this review is organized accordingly, such that Section 2 details our literature-search and review. Section 3 revisits the acoustic foundations of SSL and the key environmental assumptions. Section 4 briefly outlines traditional SSL to contemporary deep learning architectures. Section 5 discusses the dataset and learning strategies underpinning DL-based SSL. Section 6 surveys how these techniques are deployed across the

different classes of robots and applications, and Section 7 identifies research challenges and future avenues.

Figure 1. Overall review structure.

2. Review Methodology

Our goal is to paint a clear picture of how different methods, especially deep learning approaches in sound-source localization (SSL) are currently being used, tested, and deployed on robotic platforms. To achieve this, we carried out a targeted, multi-stage literature search prioritizing the breadth of coverage over exhaustive enumeration.

2.1. Literature Search Strategy

Our search commenced by querying major engineering and robotics databases, including IEEE Xplore, ACM Digital Library, ScienceDirect, and SpringerLink. This was complemented by the extensive use of Google Scholar to capture publications from emerging workshops and arXiv preprints that may have subsequently undergone peer review. We also manually inspected conference proceedings from key robotics venues such as ICRA, IROS, RSS, and IEEE RO-MAN, alongside principal audio forums like ICASSP, WASPAA, Interspeech, DCASE, and EUSIPCO.

The core of our search strategy involved combining three conceptual blocks using Boolean operators. The primary block focused on sound localization terms, connected via "AND" to a block of robotics-related keywords. For instance, a common search string was:

(*"sound source locali*"* OR *"acoustic locali*"* OR *"sound source detection"* OR *"DOA estimation"*) AND (*robot** OR *"mobile robot*"* OR *"service robot*"* OR *"industrial robot*"* OR *cobot** OR *humanoid** OR *"legged robot*"* OR *drone** OR *UAV** OR *quadrotor** OR *multirotor**)

Variations such as *"speaker localisation"* , *"binaural CNN"*, or *"SELD robot"* were iteratively added as citation chaining uncovered new terminology and relevant keywords.

2.2. Inclusion and Exclusion Criteria

To ensure the relevance and quality of the reviewed literature, a strict set of inclusion and exclusion criteria was applied during the screening process. Papers were primarily included if they were

- Peer-reviewed publications, encompassing journal articles, full conference papers, or workshop papers with archival proceedings.
- Published within the window of 1 January 2013, to 1 May 2025, capturing the significant surge of deep learning applications in robotics.
- Relevant to a robotic context, meaning the work either (i) evaluated SSL on a physical or simulated robot, or (ii) explicitly targeted a specific robotic use-case (e.g., service, industrial, aerial, field, or social human–robot interaction). Some studies that have not directly implemented SSL on robotic platforms, such as simulations and conceptual frameworks, were also included if their findings were directly and explicitly transferable to the robotic context.
- Written in English.

Conversely, studies were excluded if they were patents, magazine tutorials, or non-archival extended abstracts. Articles limited to headphone spatial audio, hearing aids, pure speech recognition, or architectural acoustics that lacked direct robotic relevance were also discarded.

The screening procedure involved an initial scan of titles and abstracts to filter out papers clearly outside the defined scope. For the remaining records, a thorough full-text inspection was conducted. During this phase, particular attention was paid to the microphone configuration, the specific learning architecture employed, the evaluation protocol, and the presence of direct robotic experimentation or use-case targeting. Citation snowballing, both forward and backward, was applied to ensure that influential papers cited by the shortlisted works, or citing them, were not overlooked. This meticulous, iterative process ultimately converged on a corpus of 78 papers, which form the basis of this comprehensive review.

2.3. Publication Trends and Venues

Figure 2 illustrates the evolving landscape of SSL research in robotics through its annual publication trends. As observed, annual publications in SSL for robotics have consistently remained above five papers since 2014, reaching a peak of nine papers in 2019. The seemingly lower count for 2025, however, is attributed to our review's cut-off date of May 2025. Deep learning approaches, notably absent before 2015, began to gain significant traction that year. Since 2020, they have consistently accounted for approximately one-third of all SSL-for-robotics publications, contributing a steady 2–3 papers annually and underscoring their growing prominence in the field.

To look more closely at specific approaches, Figures 3 and 4 break down the yearly counts of robotic SSL papers by method for both traditional and DL/ML approaches.

Looking at Figure 3 shows that regarding traditional SSL, subspace analysis is the most frequently used approach across the years, with TDOA also popular—especially before 2020. However, SRP and beamforming appear less often, particularly in recent years. For DL/ML-based works (Figure 4), traditional ML and MLP models are more common before 2018; after that, most studies adopt CNNs, RNNs, or hybrid CRNNs. CNNs in particular account for most DL-based papers after 2019, indicating a shift from dense-layer MLPs toward convolutional architectures. Using recurrent neural layers after dense layers (RNNs) and convolutional layers (CRNNs) have also been frequently reported in SSL tasks in robotics in recent years.

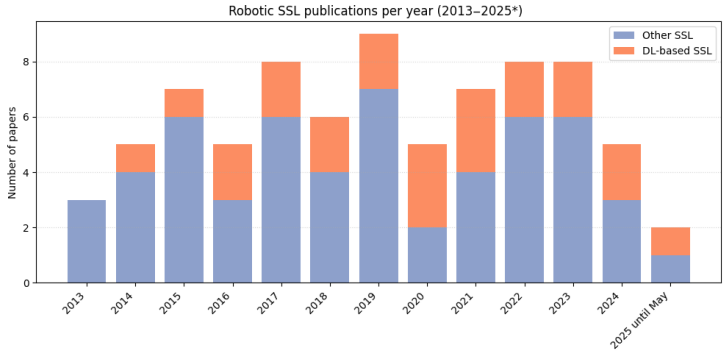


Figure 2. Number of robotics SSL publications with DL (orange) and non-DL approaches (blue).

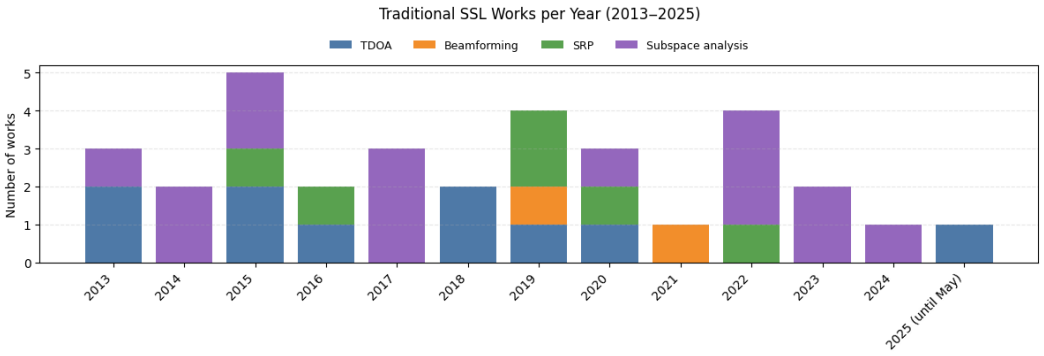


Figure 3. Number of robotics SSL publications using: time difference of arrival (TDOA) (blue column), beamforming (orange column), steered response power (SRP) (green column), and subspace analysis (purple column).

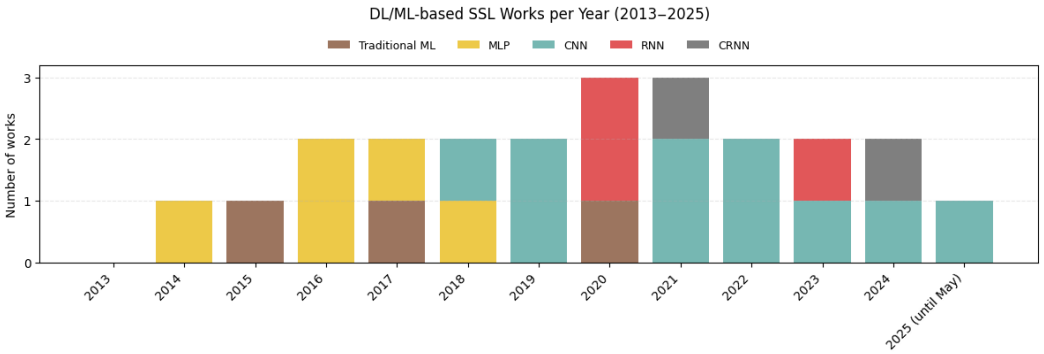


Figure 4. Number of robotics SSL papers traditional ML (yellow column), multilayer perceptron (MLP) (brown column), convolutional neural networks (CNNs) (cyan column), recurrent neural networks (RNNs) (red column), and convolutional–recurrent networks (CRNNs) (gray column).

To guide future research dissemination in this dynamic field, Table 2 presents the most common publication venues (those with two or more papers) from our corpus of 78 reviewed articles. Among these 78 papers, 36 were published as conference papers, 1 as a Ph.D. thesis [29], and 41 as journal papers. The table highlights the significant role of prominent conferences in SSL-for-robotics research. Notably, the two most prestigious robotics conferences, IROS (IEEE/RSJ International Conference on Intelligent Robots and Systems) and ICRA (IEEE International Conference on Robotics and Automation), show substantial contributions, with 11 and 3 papers, respectively. ICASSP (International Conference on Acoustics, Speech, and Signal Processing), a key conference venue in acoustic signal processing, also features prominently with 4 papers in our review. In terms of journals,

Table 2 indicates that only the IEEE Sensors Journal has contributed more than two papers to our review corpus; other journals are represented by fewer articles each.

Table 2. Most published venues among the papers in robotic SSL in our review. In the third column, "C" and "J" stand for conference and journal.

Venue	No. of Papers	C/J	Reference
IROS	11	C	[30–40]
ICASSP	4	C	[41–44]
ICRA	3	C	[45–47]
IEEE Sensors Journal	3	J	[10,48,49]
Robotics and Autonomous Systems	2	J	[1,50]
Drones	2	J	[51,52]
Applied Sciences	2	J	[53,54]
IEEE Transaction on Instrumentation and Measurement	2	J	[55,56]
IEEE/ACM Transactions on Audio, Speech, and Language Processing	2	J	[57,58]
ICDL	2	C	[59,60]

3. SSL Foundementals

Sound source localization (SSL) is the process of determining the spatial location of one or more sound sources based on measurements from acoustic sensors. This section provides an overview of the fundamental principles and terminology that form the foundation of SSL in robotics applications.

Sound propagates through air as pressure waves, traveling at approximately 343 m per second under standard conditions. When these waves encounter microphones or other acoustic sensors, they are converted into electrical signals that can be processed to extract spatial information. The core principle underlying SSL is that sound reaches different spatial positions at different times and with different intensities, creating patterns that can be analyzed to determine the source's location. In the context of robotics, SSL typically involves estimating direction of arrival (DOA) and distance or complete position. DOA is defined as an angle or vector pointing toward the sound source, often expressed in terms of azimuth (horizontal angle) and elevation (vertical angle). Distance is the range between the sound source and the receiver, and the position is the complete three-dimensional coordinates of the sound source in space. According to the objectives of each study, SSL tasks can include estimating DOA, distance, or position of the sound source.

As noted by Rascon and Meza [1], the majority of SSL systems in robotics focus primarily on DOA estimation, as distance estimation presents additional challenges and is often less critical for many applications. One challenge in SSL is taking into consideration that the sound source (or sources) might be active, a source is active when emitting sound, or inactive during the localization task. Therefore, considering a source (or sources) as always active might be unrealistic in a practical setting. To deal with this challenge, as extensively pointed out in a survey by Grumiaux et al. [23], the source activity detection can be done either before (as a separate task) or simultaneously within the SSL task, for example, a neural network predicts both location and activity of a sound source [61]). It is important to note that sound source localization is broader than sound event detection (SED), where in SSL, the location of sound sources is obtained, as well as the detection of the presence of sound. Therefore, in many studies, SSL is referred to as sound event localization and detection (SELD) [23]. It is also worthwhile mentioning that sound source separation (in the presence of multiple active sounds) is another task that has to be done in SSL when dealing with more than one active sound source [62]. Therefore, some studies

are focusing on sound source counting as well as localization (such as [63]), while many assume we have prior knowledge of sound source numbers in the scene (e.g., [64]).

3.1. Acoustic Signal Propagation

Understanding acoustic signal propagation is essential for developing effective SSL systems. In ideal free-field conditions, sound waves propagate spherically from a point source, with amplitude decreasing proportionally to the distance from the source. However, real-world environments introduce several complexities:

Reverberation: Sound reflections from surfaces create multiple paths between the source and receiver, complicating the localization process.

Diffraction: This phenomenon happens where sound waves bend or spread around obstacles or through openings (like doorways or windows). It allows sound to be heard even when there is no direct line of sight to the source.

Refraction: This is the bending of sound waves as they pass from one medium into another, or as they travel through a medium where the speed of sound changes gradually. The speed of sound can vary due to changes in temperature, wind, or medium density.

Background noise (external noise) and sensor imperfection (internal noise): sounds from ambient, or noise generated from the robot operation itself [65], can mask or interfere with the target source; this noise can be referred to as external noise. Also, due to the imperfection of the receiver system, the recorded sound can be deviated from what it should be correctly recorded because of microphones or the audio acquisition system (e.g., analog-to-digital converters). This noise is inherent to the sensing hardware and its associated electronics.

These phenomena significantly impact the performance of SSL systems and have driven the development of increasingly sophisticated algorithms to address these challenges. To deal with these challenges, some studies considered some assumptions to simplify the SSL. For example, some early studies considered that there is no reverberation in the environment; this setting is called "anechoic". Despite not being realistic in most applications, the anechoic setting has been assumed in many SSL works [23,66,67]. To deal with noise, some studies used denoising techniques to overcome noise (e.g., [68]), while some considered the effect of noise levels, as defined by SNR (signal-to-noise ratio), in the localization of sound sources [69–72].

A core challenge in practical SSL is that the estimated DOA, denoted as $\hat{\theta}$, is not a perfect representation of the true DOA, θ_{true}. The difference between these two values is the localization error, ϵ_{loc}, which can be expressed as

$$\epsilon_{loc} = \hat{\theta} - \theta_{true}$$

This error arises from various sources, which can be modeled as additive terms to the ideal acoustic signal. For a simple DOA model based on the time difference of arrival (TDOA), the estimated time difference between two microphones, $\Delta\hat{t}$, is a function of the true time difference Δt_{true} and multiple error sources:

$$\Delta\hat{t} = \Delta t_{true} + \epsilon_{sync} + \epsilon_{prop} + \epsilon_{rev} + \epsilon_{noise}$$

where

- ϵ_{sync} is the synchronization error, which arises from imperfect clock synchronization between microphones or channels.
- ϵ_{prop} is the propagation speed error, caused by variations in the speed of sound due to changes in temperature, humidity, or wind.

- ϵ_{rev} represents the reverberation error, which results from multipath propagation, causing the direct sound to be masked or delayed by reflections.
- ϵ_{noise} is the signal-to-noise error, stemming from ambient noise and sensor imperfections.

In robotic SSL, several physical modeling assumptions are commonly made to simplify the problem, but they are particularly susceptible to failure in real-world, dynamic environments:

1. Free-field propagation: The assumption of a free-field environment (anechoic) is rarely valid in indoor robotics. As a robot navigates an office, factory, or home, reverberation artifacts are the most significant source of error. The direct path signal is often overshadowed by reflections, leading to inaccurate TDOA or phase-based estimates. To mitigate this, advanced methods should learn robust features that are invariant to reverberation.
2. Constant propagation speed: The speed of sound is assumed to be a constant 343 m/s. However, temperature and wind gradients in a real environment can cause this assumption to fail. While less impactful than reverberation in most indoor settings, spatial variations in temperature or air movement can introduce a non-negligible ϵ_{prop}. For precision applications or outdoor robots, this should be revised by taking into consideration temperature and wind factors to dynamically correct the speed of sound.
3. Point source model: The assumption that the sound originates from a single point is valid for small sources at a distance, but it fails for extended sound sources (e.g., a person speaking, a large machine). The phase and amplitude can vary across the source, which can confuse DOA algorithms that rely on simple time or phase differences.

Understanding these error sources and the conditions under which model assumptions fail is crucial for designing robust SSL systems for robotics. A successful system must either explicitly model these errors or learn to be invariant to their effects.

3.2. Coordinate Systems and Terminology

SSL systems typically employ either Cartesian or spherical coordinates in both 2D and 3D localization. In Cartesian coordinates, the positions of sound sources are obtained with respect to the X, Y, and Z axes in 3D (X and Y in 2D). In spherical coordinates, the location of sound sources are determined in terms of radius (distance), azimuth, and elevation angles. In 2D localization, many studies focused on azimuth and distance or azimuth (horizontal localization [73]) or azimuth and elevation (directional localization). Also, some works only restricted their objective to the azimuth angle (1D localization) relative to the microphone array position, and in some cases, they do the localization grid by dividing the 360 degrees azimuth angle into grid space, such as 8 sections [74] or narrower grid space (e.g., 360 sections [75]). This trick aims to perform SSL as a classification task in machine learning, where each class is devoted to a specific subregion for estimating sound sources [23].

In robotics applications, the choice of coordinate system often depends on the specific requirements of the task and the configuration of the robot's sensors. Also, it is worth mentioning that studies focusing on mobile robots might benefit from using the robot-centric coordinate (centered on the microphones mounted on the robot [76,77]). On the other hand, a fixed and static coordinate system defined by the environment itself (e.g., a corner of a room, a fixed marker) may be preferred for stationary applications, e.g., industrial robotic arms that have fixed locations in the workplace.

3.3. Microphone Numbers and Array Configurations

The arrangement of microphones plays a crucial role in SSL performance. Before explaining the different configurations, it is important to describe an open challenge

in this regard. As a general rule, a large number of microphones in SSL leads to high accuracy in localization [78]. However, including more microphones can cause higher computing, higher cost, and consequently higher latency in localization (not reaching real-time localization). Also, the variation in microphone number and arrangements adds other design hyperparameters (how many microphones? which arrangement is better?). These additional hyperparameters in each study results in not having a reference SSL design system as a benchmark. On the other hand, because of some constraints in each study, e.g., different objectives, different financial and computational budgets, and design constraints, it is not expected that each study follow the same microphone design. Various microphone configuration have been used in SSL studies in robotics, as shown in Figure 5. The variety in microphone array configurations can be categorized into the following configurations:

Binaural arrays: Mimicking human hearing with two microphones [79], often placed on a robot's head or within a dummy head structure [55,80,81], as represented in Figure 5a. These are particularly common in humanoid robots, or simple mobile robots [82,83], and offer natural spatial cues but may have limited resolution. The robot's head (or dummy head) can significantly impact sound waves through diffraction. This causes sound waves to bend around the head, leading to fluctuations in the time difference of arrival (TDoA), especially for sounds traveling around the front and back. The front–back ambiguity challenge in binaural SSL for robots, which means that the robot may struggle to differentiate between a sound coming from "ahead" or "behind" without additional cues or head movements. Binaural arrays are excellent at localizing sounds in the horizontal plane (azimuth) but offer very limited resolution for elevation (vertical direction). It is difficult for them to tell whether a sound is coming from above or below. These challenges lead to struggling with multiple simultaneous sound sources without using advanced separation techniques.

Linear arrays: Microphones are arranged in a straight line (see Figure 5b), providing good resolution in one dimension but suffering from front–back ambiguity and limited elevation estimation. The microphones are typically equally spaced, but non-uniform spacing can also be used to optimize performance [40]. In addition to the simplicity of design and implementation, this configuration is easier to calibrate than complex arrangements. A linear array mounted on the front or top of a robot can effectively localize sounds in the horizontal plane ahead of the robot [84,85]. This is useful for directional voice commands or detecting sounds from the front.

Circular arrays: As shown in Figure 5c, the microphones are arranged in a circle, offering 360-degree coverage in the horizontal plane and eliminating front–back ambiguity. Mounted on the "head" or "torso" of social robots to perceive sounds from any direction around them, crucial for engaging with multiple people in a room [50,86,87].

Spherical arrays: Microphones distributed over a spherical surface, as can be seen in Figure 5d, providing full three-dimensional coverage and the ability to decompose the sound field into spherical harmonics. The spherical arrays can enable a robot [88] or a drone [8,36,52] to precisely localize and track multiple speakers in 3D space, understanding who is speaking from where in complex social settings.

Arbitrary geometries: This refers to microphone configurations where the individual microphone elements are positioned without adherence to the particular geometric arrays. The placement might be dictated by factors external to optimal acoustic design, such as the physical constraints of a platform, aesthetic considerations, the repurposing of existing sensors, or opportunistic placement in an environment. In robotics, microphones could be placed in different parts of the body, as can be seen in Figure 5e, e.g., the head, torso, and limbs of a humanoid robot [41,57] or distributed inside the room [72,89]. The irregular microphone placements have also been implemented in non-humanoid such as a hose-shaped robot [12,90].

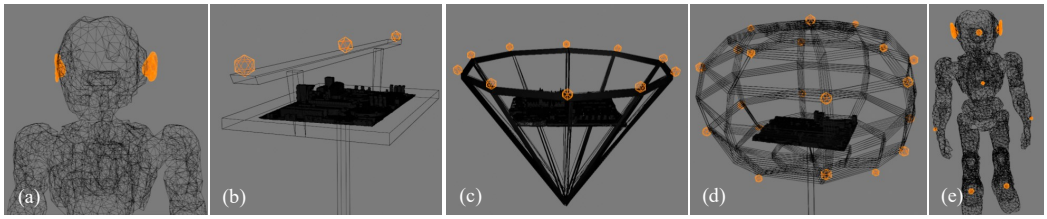


Figure 5. Different microphone array configurations used in SSL robotics studies: (**a**) Binaural microphone array; (**b**) Linear array; (**c**) Circular array; (**d**) Spherical array; and (**e**) Microphones distributed in robot parts.

The choice of the number of microphones and array configuration depends on the intention of researchers in each study. In robotics, many researchers want to mimic the human auditory system and chose binaural arrays. Also, it is noted that binaural is the most used configuration in SSL in robotics [1]. Interestingly, some researchers made some advancements by designing external pinnae for robots to mimic the human auditory system more than just putting two microphones on the head of the robot [55,91,92]. Regarding the number of microphones, the number of microphones varied a lot [1]. It includes some SSL studies having a single microphone [67,93,94], and some having a very dense microphone array (even 64 microphones [88]).

4. From Traditional SSL to DL Models

As robots are increasingly deployed in real-world settings, ranging from domestic assistants to industrial inspectors, the demands placed on SSL systems have grown considerably. These systems must now contend with complex, noisy, and reverberant environments, while maintaining low latency, minimal power consumption, and high spatial accuracy. Over the past decades, SSL methodologies have evolved in response to these challenges, beginning with analytically grounded, signal-processing techniques that exploit the physical properties of sound propagation and microphone arrays. While traditional methods such as time difference of arrival (TDOA), beamforming, and subspace analysis laid a robust foundation, their performance often degrades under non-ideal conditions and hardware constraints typical of mobile and embedded robotic systems. These limitations, in turn, have catalyzed a shift towards data-driven and learning-based paradigms, promising greater adaptability and robustness in real-world applications.

4.1. Traditional Sound Source Localization Methods

Before neural networks entered the scene, robot audition drew almost exclusively on classical array-signal-processing. Although most of today's deep models replace, or embed, those analytical blocks, an understanding of their logic remains essential because the same acoustic constraints (microphone geometry, reverberation, noise, and multipath) still apply. Below, we revisit the four pillars that have shaped the field, i.e., time-delay estimation, beamforming, steered-response power (SRP), and subspace analysisd. Throughout, we highlight why each method proved attractive for robots and where its weaknesses motivated the subsequent shift to learning-based pipelines.

4.1.1. Time Difference of Arrival (TDOA)

Time difference of arrival (TDOA) is one of the foundational techniques in sound source localization (SSL), widely applied due to its conceptual simplicity and computational efficiency. The method relies on the fundamental observation that a sound wave will arrive at different microphones in an array at slightly different times, depending on the location of the source relative to the array [23]. By measuring these inter-microphone

time delays, one can infer the direction or position of the sound source [95]. At the heart of TDOA-based SSL is the cross-correlation of audio signals recorded at different microphones. This process identifies the time lag at which the similarity between two signals is maximized, corresponding to the delay caused by the sound wave's path difference. To enhance the performance of this basic cross-correlation, especially in noisy or reverberant environments, researchers often employ generalized cross-correlation (GCC), which introduces a frequency-dependent weighting to the correlation process [96]. Among the variants of GCC, the phase transform (PHAT) weighting, commonly known as GCC-PHAT, has proven particularly effective in robotics contexts [97,98]. By emphasizing phase information and attenuating amplitude components, GCC-PHAT improves robustness to reverberation and environmental noise. Once the time delays are estimated, spatial localization is achieved through hyperbolic positioning. Each pairwise time difference corresponds to a hyperbolic curve (in 2D) or a hyperboloid surface (in 3D) upon which the sound source must lie. The intersection of these geometric constraints from multiple microphone pairs yields an estimate of the source's location.

Despite its popularity, the TDOA method is not without limitations. One major challenge is its sensitivity to noise and reverberation [99,100], particularly in enclosed environments with reflective surfaces. Although GCC-PHAT offers partial mitigation, its effectiveness diminishes in highly cluttered acoustic spaces. Additionally, the spatial resolution of TDOA-based systems is inherently limited by the sampling rate of the microphones and the inter-microphone distances, which constrain the granularity of detectable delays. Another well-documented drawback is the so-called front–back ambiguity, which arises primarily in linear microphone arrays. In such configurations, the system may struggle to distinguish whether a sound originates from in front of or behind the array due to symmetric time-delay profiles. Moreover, traditional TDOA techniques face difficulties in multi-source environments [101], where overlapping sound signals can interfere with accurate time-delay estimation, leading to degraded performance or incorrect localization. Nonetheless, TDOA remains a cornerstone of SSL research and application, especially in scenarios where computational simplicity and real-time performance are prioritized. Its principles have also served as a basis for hybrid systems that integrate TDOA with more advanced signal processing or learning-based techniques, expanding its utility in modern robotic auditory systems [33,37,45,47,102–107].

4.1.2. Beamforming

Beamforming is a spatial signal processing technique that plays a dual role in auditory systems: it not only facilitates sound source localization (SSL) but also enhances the quality of the captured audio by amplifying signals from a desired direction while attenuating noise from others. In robotic contexts, this capability is invaluable for tasks such as speech recognition, situational awareness, and interaction in acoustically complex environments [108–110]. At its core, beamforming involves the constructive combination of sound signals from multiple microphones, typically after delaying them to align with a hypothesized source direction. When the hypothesized direction matches the true direction of arrival (DOA), the signals reinforce, resulting in a high beam response. The simplest implementation of this principle is the delay-and-sum (DS) beamformer, which sums delayed microphone signals to estimate the DOA. More sophisticated approaches like the minimum variance distortionless response (MVDR) beamformer, also known as Capon's beamformer [111], improve robustness by minimizing output power from all directions except the desired one. The linearly constrained minimum variance (LCMV) beamformer allows multiple spatial constraints, making it suitable for multi-target environments [112]. Adaptive beamforming methods extend this paradigm by dynamically

adjusting parameters in response to changes in the environment, enhancing resilience to noise and interference.

In the survey by Lv et al. [5], some promising developments in beamforming were explained that could be implemented as beamforming-based SSL in robotics. For example, a novel beamforming technique by Yang et al. [113] tailored for far-field, large-scale environments, significantly improved localization accuracy across multiple sound sources. In low signal-to-noise ratio (SNR) conditions, Liu et al. [114] introduced a novel MVDR-based beamforming (MVDR-HBT) algorithm, which leverages statistical signal properties to boost robustness and precision. Zhang et al. [115] extended beamforming to periodic, steady-state sound sources by developing a high-resolution cyclic beamforming method that support both localization and fault diagnosis in machinery. While beamforming offers considerable advantages, it is not without its challenges. Traditional beamforming techniques tend to suffer from poor dynamic range and limited real-time performance, particularly in large environments or when precise temporal resolution is required. Moreover, accurate array calibration remains a critical prerequisite—small discrepancies in microphone positioning or gain can lead to substantial localization errors. The computational complexity of adaptive beamforming in real-time applications can also place significant demands on embedded processors typical of mobile robotic platforms and UAVs [51,116]. Nonetheless, beamforming remains a vital technique in SSL which could be used in many robotic applications. Its capacity for both directional enhancement and localization makes it particularly attractive for systems that require perceptual robustness in dynamic or noisy environments.

4.1.3. Steered-Response Power (SRP)

Steered response power (SRP) methods constitute a prominent category of traditional SSL techniques. These methods operate on the principle of beamforming: the microphone array is virtually "steered" in multiple candidate directions, and the response power, essentially the energy of the summed signals after delay alignment, is computed for each direction. The underlying assumption is that, when the beamformer is steered towards the true source direction, the accumulated energy or response power will be maximized. Among the various SRP techniques, the SRP-PHAT (phase transform) algorithm is widely recognized for its robustness and practical effectiveness. It incorporates the phase transform weighting strategy from GCC-PHAT into the SRP framework, attenuating amplitude information and emphasizing phase cues to improve reliability in reverberant environments [117].

One of the most valuable attributes of SRP methods (particularly SRP-PHAT) is their resilience in acoustically challenging settings. Unlike traditional cross-correlation techniques, SRP methods perform well even in the presence of significant reverberation, making them suitable for indoor or cluttered robotic applications. Another advantage is their capacity for multi-source localization. By analyzing the response power map across a spatial grid, systems can identify multiple peaks, each corresponding to a potential sound source. This ability makes SRP methods appealing in scenarios where robots must interact with multiple humans or monitor several machines simultaneously. However, SRP approaches are not without challenges. Chief among them is computational complexity. Evaluating the steered response across a dense spatial grid requires substantial processing, which can be burdensome for real-time systems or power-constrained platforms. Moreover, the spatial resolution of the localization is directly tied to the granularity of the search grid. Finer grids improve accuracy but exacerbate computational demands. Discretization also introduces errors, particularly when the true source lies between predefined grid points. Additionally, while SRP-PHAT improves robustness to reverberation, performance may still degrade in noisy environments or when multiple sources overlap in time and frequency,

leading to ambiguities in peak detection. Despite these limitations, SRP-based methods, especially when paired with optimization strategies or hierarchical search techniques, remain a powerful tool in the SSL toolbox. They bridge the gap between theoretical accuracy and real-world applicability and continue to be refined for emerging robotic use cases that demand high reliability in dynamic, reverberant, and multi-source auditory scenes [35,38,39,50,60,118].

4.1.4. Subspace Analysis

Subspace-based methods form one of the most analytically powerful approaches in the domain of sound source localization (SSL). These techniques, including the well-known multiple signal classification (MUSIC) and estimation of signal parameters via rotational invariance techniques (ESPRITs), operate by analyzing the eigenspectrum of the spatial covariance matrix derived from multichannel microphone signals. Their strength lies in the decomposition of this matrix into orthogonal signal and noise subspaces, which enables the extraction of direction-of-arrival (DOA) information with high angular precision. The MUSIC algorithm remains a prominent representative of this class. It estimates the covariance matrix of the received signals and then applies eigendecomposition to distinguish the dominant signal subspace from the residual noise subspace. A spatial pseudo-spectrum is then generated by scanning over candidate directions; peaks in this spectrum correspond to estimated DOAs, exploiting the principle that the array steering vectors of true sources are orthogonal to the noise subspace [119].

One of the significant limitations of traditional MUSIC is its dependency on accurate and noise-free estimation of the covariance matrix. In real-world scenarios with limited data or low signal-to-noise ratios (SNRs), these estimations can become unreliable. To address this, Zhang and Feng [120] proposed the use of a non-zero delay sample covariance matrix (SCM) combined with pre-projection techniques to filter out noise and improve signal subspace estimation. Similarly, Weng et al. [121] introduced the SHD-BMUSIC algorithm, which operates in the spherical harmonic domain and integrates wideband beamforming to enhance source discrimination, particularly in scenarios involving closely spaced or multiple sources. Despite their theoretical elegance and high spatial resolution, subspace methods are not without drawbacks. They are sensitive to coherent or correlated sound sources, a common condition in reverberant environments, which can collapse the signal subspace and compromise accuracy. Moreover, these methods demand significant computational resources [56], primarily due to eigendecomposition and exhaustive spatial scanning. They also typically require a substantial number of temporal snapshots to yield stable covariance estimates, reducing their responsiveness in rapidly changing acoustic scenes. Nonetheless, subspace methods remain an essential pillar of SSL research. Their precision in controlled environments and potential for multi-source resolution make them attractive for applications in collaborative mobile ground robotics [54,122–124] and UAVs [32,36,52,58,125,126] as well as humanoid audition [31,40,119,127–129]. Ongoing research aims to mitigate their limitations through techniques such as subspace smoothing, sparse array design, and integration with learning-based models, enabling their gradual transition from theoretical benchmarks to practical solutions in robotic auditory systems.

Table 3 lists the information regarding the traditional SSL works in robotics that used the four typical localization families. As the table shows, a considerable number of papers in the last decade still rely on these classical SSL approaches (especially subspace analysis using MUSIC) in robotics. Interestingly, the vast majority of these studies target one stationary sound source in their localization experiments. This emphasis reflects an inherent limitation of traditional SSL techniques in tracking multiple or moving emitters with high accuracy.

Table 3. Studies on robotic platforms in our review that used the typical traditional SSL approaches (i.e., TDOA, beamforming, SRP, and subspace analysis). "S/M" denotes **s**tatic/**m**oving sound sources.

Paper	Method	Year	Robot Type	Max. Active Sources	S/M	No. of Mics (Geometry)	Performance (Accuracy or Error)
[106]	TDOA	2013	Mobile ground	1	S	4 (tetrahedral)	Error: 0.7° (azimuth), 0.1 m in 3 m distance
[102]	TDOA	2013	Humanoid	1	S	2 (binaural)	Error (RMSE): 1.96° (azimuth)
[107]	TDOA	2014	General	1	S	4	Above 77.5% accuracy within 30° (azimuth/elevation)
[103]	TDOA	2015	Humanoid	1	S	2 (binaural)	Error (RMSE): 1.99–4.53° (azimuth)
[33]	TDOA	2015	Mobile ground	1	S	2	—
[45]	TDOA	2016	Mobile ground	1	S	2	—
[47]	TDOA	2018	Mobile ground	1	S + M	8	Error: 0.8 m in a room $7 \times 7 \times 3\,m^3$
[37]	TDOA	2018	UAV	1	S	8	Error (RMSE): <2° (azimuth/elevation)
[104]	TDOA	2019	Mobile ground	1	S	5 (pyramid)	Eelative error: <1.5% (azimuth), 5–9% (distance)
[105]	TDOA	2020	Humanoid	1	M	2 (binaural)	error: 4.8° (azimuth), 0.86 m in open field $9 \times 6\,m^2$
[116]	Beamforming	2019	UAV	1	M	8 (circular)	Error: 6.03 m in $240 \times 160 \times 80\,m^3$
[51]	Beamforming	2021	UAV	1	S	32 (circular)	Error: 0.8–8.8° (azimuth), 1.4–10.4° (elevation) for 25.3–151.5 m range
[118]	SRP	2015	General	1	S	16 (cylindrical)	Error: 3.32° average 3D angle
[35]	SRP	2016	Humanoid	1	S	4	Error: 0.47–0.95° (azimuth), 1.47–2.12° (elevation)
[50]	SRP	2019	Mobile ground	4	S + M	8/16 (circular/closed cubic)	two sources: error (RMSE) 0.064–0.657° (azimuth)
[38]	SRP	2019	UAV	2	S	8 (circular)	Error: 3.5–8.4° 3D angle (2–6 m range)
[39]	SRP	2020	Mobile ground	1	S + M	16	Error: <0.3 m distance in 150 m^2 room
[60]	SRP	2022	Mobile ground	1	S	3	—
[126]	SRP/subspace	2024	UAV	1	S	6	—
[31]	Subspace analysis	2013	Humanoid	1	M	16 (circular)	Error: 6.5° avg. 3D angle (moving source)
[32]	Subspace analysis	2014	UAV	1	S	16 (circular and hemisphere)	error: <10° 3D angle (1–3 m, outdoor)
[128]	Subspace analysis	2014	Humanoid	1	S	7 (6 circular + 1 top)	relative error: 1.65% (azimuth), 3.8% (distance)
[129]	Subspace analysis	2015	Humanoid	1	S	2	—
[34]	Subspace analysis	2015	General	1	S + M	8 (planar)	Error: 4° avg. (azimuth) at 0.5–2 m in $7 \times 4\,m^2$ room
[125]	Subspace analysis	2017	UAV	1	S	12/16	—
[36]	Subspace analysis	2017	UAV	1	S	12 (spherical)	>80% accuracy (azimuth & elevation)
[122]	Subspace analysis	2017	Mobile ground	1	S	8 (circular)	Error: <10° (azimuth) at 1–5 m
[40]	Subspace analysis	2020	Humanoid	1	S + M	4 (linear)	74–95% within 2.5° (azimuth)
[58]	Subspace analysis	2022	UAV	2	S	30	Error: <2° (azimuth/elevation)
[123]	Subspace analysis	2022	Mobile ground	3	S	24 (rectangular)	Error (RMSE): 0.035 m in $2 \times 2\,m^2$ room

Table 3. *Cont.*

Paper	Method	Year	Robot Type	Max. Active Sources	S/M	No. of Mics (Geometry)	Performance (Accuracy or Error)
[124]	Subspace analysis	2022	Mobile ground	3	S	16 (planar)	—
[54]	Subspace analysis	2023	Mobile ground	3	S	4	82% accuracy (azimuth)
[52]	Subspace analysis	2023	UAV	2	M	16 (spherical)	Error (RMSE): 0.65 and 2.15 m for two moving sources (10 m range)

Acronyms: Mics—Microphones.

Regarding the localization accuracy of traditional methods, a unified comparison is challenging due to the varied metrics and experimental conditions found in the literature. As shown in the last column of Table 3, studies report accuracy using different metrics, such as angular or distance error, and across a wide range of setups. For example, in a large-scale outdoor experiment covering a field of 240 × 160 × 80 m^3, a mean distance error of 6.03 m [116] can be considered a successful outcome, aligning with the typical accuracy of open-field GPS. In contrast, very high precision can be reported in close-range indoor experiments, such as the 2–10 cm error achieved by [106] with the sound source 1.5–3 m away. The effect of distance on accuracy is also highlighted by [51], which reported a range of azimuth and elevation errors (0.8–8.8° and 1.4–10.4°, respectively) as the sound source moved from 25.3 to 151.5 m away. Noise is another critical factor. In [58], it was shown that while TDOA and subspace methods perform well in non-noisy conditions, they can lead to significant localization errors (e.g., a 100° azimuth error for TDOA-based GCC-PHAT) in extremely noisy environments with a signal-to-noise ratio (SNR) as low as −30. Beyond the model and environment, the microphone itself can influence accuracy; ref. [103] demonstrated that different pinnae designs on a humanoid's binaural microphone setup resulted in varying azimuth RMSE errors, even when using the same SSL model. Furthermore, some studies report accuracy as a success rate within a specific error threshold instead of a direct error value. For instance, ref. [40] reported that subspace analysis achieved a 74–95% success rate within a threshold of 2.5° for angle and 20 cm (at 5 m) for distance estimation.

Computational latency is a crucial factor for real-time robotic SSL, yet it has been largely overlooked in many studies. While this information is not widely available, some papers provide important comparisons. For example, ref. [37] noted that TDOA (GCC-PHAT) was twenty times faster than subspace analysis (MUSIC) on the same hardware, processing one second of signal in approximately 1.5 s. Similarly, ref. [105] found that a TDOA-based SSL was over 50% faster than SRP-PHAT for localizing a referee whistle, enabling a closer-to-real-time performance. A few researchers have directly addressed this challenge by developing computationally efficient models. For example, ref. [50] proposed a lightweight SRP model (SRP-PHAT-HSDA) that is up to 30 times more efficient than a typical SRP, allowing for real-time localization on low-cost embedded hardware. In a similar vein, ref. [31] introduced a modified MUSIC-based method called optimal hierarchical SSL (OH-SSL) that reduced the processing time from 1.073 s to a mere 0.028 s, making real-time processing feasible.

4.2. DL-Based SSL

Traditional SSL methods rely on explicit mathematical models of sound propagation and array geometry. While effective in controlled conditions, they often struggle in challenging real-world environments, such as high noise levels, reverberation, multiple sound sources, and unknown or changing array configurations, as well as moving sound sources. Machine learning (ML) and especially deep learning (DL) have emerged as transformative

approaches in SSL by enabling models to learn intricate patterns directly from data, with no or low need for pre-processing input data [23,29], often outperforming traditional methods in challenging scenarios. This section examines the evolution, methodologies, architectures, and innovations in deep learning-based SSL for robotics applications.

4.2.1. Traditional Machine Learning and Neural Networks

Before "deep" became fashionable, researchers already tried to map acoustic features to directions with classical machine learning. The support-vector machines (SVMs) model is a well-known ML that has been implemented in some SSL works as well as sound event classification [130–132]. Interestingly, in some SVM-based sound source localization, the data features from traditional SSL has been used, such as SVM using TDOA features [133] or SVM based on beamforming [134,135]. *K*-nearest neighbors (KNNs) is another classical ML model that has been used in sound source localization [136] and sound source classification [131,137].

Aside from traditional machine learning, which is primarily dependent on the quality and relevance of the hand-crafted features extracted from the raw audio signals (e.g., TDoA values from GCC [138]), neural networks were able to automatically learn optimal features directly from raw multi-channel audio data or low-level spectrograms. At the beginning, some research used simple neural networks (NNs) in the SSL task. This could be done in the simplest form using a shallow neural network (having a single hidden layer) or NNs with multiple hidden layers, called multi-layer perceptrons (MLPs) [139]. Shallow neural networks were not frequently used in SSL [140], since having only one layer as a hidden layer could not let the model learn the sound features for localization accurately. Song et al. [48], used a shallow neural network with fuzzy inference, called the fuzzy neural network (FNN), for fault detection from sound signals using a plant machinery diagnosis robot (MDR). In contrast, many SSL works used deeper neural networks, with multiple hidden layers, to effectively learn the sound features during the training phase. To the best of our knowledge, in the beginning of 2000, some early works used shallow MLPs in SSL, such as NNs with two hidden layers [141,142]. Kim et al. [143] used MLPs for both sound counting and localization, such that an MLP detected the number of sound sources, and then a different MLP localized each of the sound sources. Also, MLP was used in two studies by Davila-Chacon et al. [144,145] on a humanoid robot to estimate the azimuth angle and also to improve automatic speech recognition (ASR). In robotics, Youssef et al. [30] used an MLP to estimate the sound source azimuth angle by exploiting the binaural signal received in two microphones mounted on a humanoid robot. Another robotic study by Murray and Ervin [93] used an MLP for estimating the elevation angle while sounds were recorded through artificial pinnae, mimicking the human auditory system. In another robotic study by He et al. [46], they proposed an MLP-based model for simultaneously detecting and localizing multiple sound sources in human–robot interaction scenarios, and they showed that their model outperformed traditional SSL, such as MUSIC-based. Similarly, Takeda and Komatani used MLP in their works and showed that the MLP model outperforms MUSIC in SSL in studies with a robotic theme [42,43,146].

It is important to note that the neural network architecture in each study is different from each other and has many hyperparameters. For example, the activation function in each layer, the number of hidden layers and nodes in each layer, as well as the type of input data, i.e., could be raw signal, spectrogram, or sound features captured from traditional SSL, and could be different for each NN. Therefore, hyperparameter selection through an ablation study could help in achieving the best architecture. Also, another challenge in ML-based SSL is to ensure that the quantity and quality of data is sufficient in training,

such that the model could effectively learn in the training phase and avoid some common issues, such as overfitting [147].

4.2.2. Convolutional Neural Networks (CNNs)

A convolutional neural network is a famous DL architecture composed of learnable convolution kernels, interleaved with nonlinear activations, and, typically, pooling or normalization layers [139]. The defining feature is weight sharing: the same kernel is slid across the input, allowing the network to detect local patterns regardless of their absolute position. CNNs were first popularized for image analysis [148], yet their inductive biases, translation invariance, and local receptive fields map naturally to spectro-temporal audio representations. Compared to analytic TDOA or MUSIC pipelines, CNNs require no hand-crafted feature selection [25], can fuse information across multiple microphones and hundreds of frequency bins in a single forward pass, and can be trained to ignore ego-noise unique to a given platform. Their computation is a chain of dense tensor operations that map efficiently onto embedded GPUs, making real-time deployment feasible even on small service robots.

As pointed out in the survey by Grumiaux et al. [23], in 2015, the first CNN-based SSL was performed by Hirvonen [74]. In this study, a CNN model was trained to classify an audio signal containing one speech or musical source into spatial regions as a classification task. Some studies used raw audio waveforms as input for the CNN architecture [149–151]. This approach eliminates the need for hand-crafted features but may require larger models and more training data. However, other CNN-based SSL studies typically used the input in the form of a multichannel short-time Fourier transform (STFT) spectrogram [152,153]. These multichannel spectrograms are typically used as 3D tensors, with one dimension for time (or frames), one for frequency (bins), and one for channel. Some works used the feature of conventional SSL works, e.g., 2D images of DoA feature extracted by beamforming [154] or GCC features [44,155], as an input for the CNN architecture. The convolutional layers are regarded as an important hidden layer in CNNs, Conv1D, and Conv2D, and 3D convolutional layers have been used in SSL works. It is generally conceived that the higher dimension can lead to the better capture of the feature while it adds a computational cost [156]. Even though, some works suggest having hybrid convolutional dimensions in SSL [157], e.g., a set of Conv2D layers for frame-wise feature extraction followed by several Conv1D layers in the time dimension for temporal aggregation [158].

In robotics, several pioneering studies demonstrate the effectiveness of CNNs in SSL tasks, particularly in overcoming the limitations of traditional methods in complex, real-world scenarios. Notably, these approaches leverage CNNs to map binaural or multi-microphone audio features directly to sound source locations or directions. Nguyen et al. [159] explored an autonomous sensorimotor learning framework for a humanoid robot (iCub) to localize speech using a binaural hearing system. Their contribution involved an automated data collection and labeling procedure, and they successfully trained a CNN with white noise to map audio features to the relative angle of a sound source. Crucially, their experiments with speech signals showed the CNN's capability to localize sources even without explicit spectral feature handling, highlighting the network's inherent ability to learn robust mappings. Similarly, Boztas [160] focused on providing auditory perception for humanoid robots, emphasizing the importance of localizing moving sound sources in scenarios where cameras might be ineffective. By recording audio from four microphones on a robot's head and combining it with odometry data, the study demonstrated that a CNN could accurately estimate the location of a moving sound source, thus validating the robot's ability to sense sound positions with high accuracy, akin to living creatures. Furthermore, CNN-based SSL models are pushing the boundaries to

address more complex robotic challenges. He et al. [46] proposed using neural networks for the simultaneous detection and localization of multiple sound sources in human–robot interaction, moving beyond the single-source focus of many conventional methods. They introduced a likelihood-based output encoding for arbitrary numbers of sources and investigated sub-band cross-correlation features, demonstrating that their proposed methods significantly outperform traditional spatial spectrum-based approaches on real robot data. Jo et al. [2] presented a three-step deep learning-based SSL method tailored for human–robot interaction (HRI) in intelligent robot environments. Their approach prioritizes minimizing noise and reverberation by extracting excitation source information (ESI) using linear prediction residual to focus on the original sound components. Subsequently, GCC-PHAT is applied to cross-correlation signals from adjacent microphone pairs, with these processed single-channel, multi-input cross-correlation signals then fed into a CNN. This design, which avoids complex multi-channel inputs, allows the CNN architecture to independently learn TDOA information and classify the sound source's location. While not directly tested on a robot, Pang et al. [161] introduced a multitask learning approach using a time–frequency CNN (TF-CNN) for binaural SSL, aiming to simultaneously localize both azimuth and elevation under unknown acoustic conditions. Their method extracts interaural phase and level differences as input, mapping them to sound direction, and shows a promising localization performance, explicitly stating its usefulness for human–robot interaction. Similarly, Ko et al. [162] proposed a multi-stream CNN model for real-time SSL on low-power IoT devices, including its application to camera-based humanoid robots. Their model processes multi-channel acoustic data through parallel CNN layers to capture frequency-specific delay patterns, achieving high accuracy on noisy data with a low processing time, further emphasizing its applicability for robots mimicking human reactions in crowded environments. CNN-based SSL has also been tested in a multi-robot configuration. In this regard, Mjaid et al. [163] introduced AudioLocNet, a novel lightweight CNN that frames SSL as a classification task over a polar grid, letting each robot infer both azimuth and range to locate and communicate with each other.

4.2.3. Convolutional Recurrent Neural Networks (CRNNs)

A CNN treats each input window independently; it excels at learning spatial–spectral patterns, but it has no mechanism to remember how those patterns evolve over time. Sound localization in the wild, however, is inherently temporal. Robots rotate their heads, sources start and stop speaking, and early reflections arrive a few milliseconds after the direct path. To deal with temporal dependency, recurrent neural network (RNN) architectures are designed by adding state vectors that are updated at every time step, enabling the model to maintain a memory of past observations and to capture temporal dependencies that a vanilla CNN cannot represent. In the context of sound localization, an RNN ingests a sequence of time-–frequency frames, propagates information through gated recurrent units (GRUs) [164] or long short-term memory (LSTM) cells [165], and outputs either frame-wise direction estimates or a smoothed trajectory of source positions. Regarding RNNs in SSL, few works, such as those by Nguyen et al., [166] and Wang et al. [167], used RNNs without combining these with CNNs [23]. The former work implemented an RNN to match and fuse sound event detection (SED) and DOA predictions, and the latter one used a bidirectional LSTM to identify speech dominant time–frequency units for DOA estimation. Despite these successes, pure RNN solutions remained limited in SSL (e.g., two SSL works in robotics [89,168]) because the recurrent layers themselves could not learn spatial filters or phase relations from raw spectra. This bottleneck set the stage for convolutional–recurrent hybrids (CRNNs), which delegate spatial feature extraction to a front–end CNN and let the RNN specialize in temporal reasoning, thus combining the best of both worlds.

CRNNs have been regularly implemented for SSL since 2018 [23]. Having some convolutional layers following a recurrent unit, such as a bidirectional gated recurrent unit (BGRU) or LSTM, and a fully connected layer, is a typical CRNN architecture in SSL. This kind of model has been shown to be robustly able to detect overlapped sounds and a noisy environment [169,170]. Regarding the localization of multiple sound sources, some CRNN-based SSL works used a separate CRNN for sound source counting as well as one for localization [63], while some assumed the prior knowledge of sound sources and localize each with a trained model [72], and some simultaneously localize multiple sound sources within a single CRNN model [171,172].

In robotics, CRNNs have been recently (after 2020) used in SSL tasks. For example, Kim et al. [173] addressed the critical challenge of ego noise generated by mobile robots (robot vacuum cleaner), which significantly degrades SSL performance. Their approach proposed a multi-input–multi-output (MIMO) noise suppression algorithm built upon an extended time-domain audio separation network (TasNet [174]) architecture. They evaluated their work by the CNN (U-Net [175]) and CRNN (by adding an LSTM block) models across different noise levels. Aside from the works of Kim et al., other CRNN architectures in SSL have been used indirectly in robotic application. For instance, Mack et al. [176] implemented a Conv-BiLSTM model to estimate the DOA angles based on the spectrogram inputs of human voice and artificial noise with possible applications in robotics. Jalayer et al. [72] implemented CovLSTM-based SSL for the localization of humans and machines (CNC machines) using raw audio signals in an industrial noisy environment that can be applied for scenarios with industrial robots in the scene. Two studies by Altayeva et al. [177] and Akter et al. [178] implemented ConvLSTM-based sound detection and classification for various sound events in an urban environment, which could be applied in security robots in urban areas. Varnita et al. [179] similarly implemented ConvLSTM-based sound event localization and detection (SELD), which could be practically used for the accurate sound-based navigation and awareness of robots regarding the events.

4.2.4. Attention-Based SSL

The intuition behind *attention* is simple yet powerful: rather than processing an input sequence with a fixed, locality-biased kernel, a model can learn to *select* and *weight* the time–frequency regions that matter most for the task at hand. Introduced by Bahdanau et al. [180] for machine translation and generalized in the Transformer of Vaswani et al. [181], attention replaces recurrence with a data-driven affinity matrix that relates every element of the input to every other element. Each output representation is thus a context-aware combination of the entire input, enabling the network to capture long-range dependencies and intricate phase relationships across microphone channels that are essential for precise sound localization [23].

The first wave of attention-augmented SSL networks (after 2020) kept the overall CRNN structure and simply inserted a self-attention layer after the recurrent stack. Phan et al. [182] and Schymura et al. [183] reported consistent gains over the CRNN baseline: the localization error decreased while the models learned to focus automatically on time frames where sources were active. Mack et al. [176] placed two successive attention blocks, the first operating directly on phase spectrograms, and the second on convolutional feature maps, to generate dynamic masks that emphasize frequency bins dominated by the target speaker. Following the broader "*attention-is-all-you-need*" trend, fully Transformer-style encoders are now replacing recurrent back-ends altogether. Multi-head self-attention (MHSA) layers were coupled with lightweight CNN front ends by Emmanuel et al. [184] and Yalta et al. [185], cutting computation time while winning the SELD track. Many SSL research studies in recent years directly implemented Transformers in SSL [186–188]; for

instance, the most recent one by Zhang et al. [189] extends this trend by coupling a CNN front–end with a Transformer encoder that attends over sub-band spatial-spectrum features, enabling the network to identify true DOA peaks and suppress spurious ones, and thereby delivering state-of-the-art localization accuracy for multiple simultaneous sources under heavy noise and reverberation.

Although Transformer-style architectures already set the performance bar on SSL benchmarks, documented examples of their on-board use in mobile or humanoid robots remain scarce. Yet the very property that makes attention models excel in controlled evaluations, namely their ability to focus dynamically on the most informative time–frequency cues, addresses exactly the challenges faced by robots in the wild. In a bustling factory or a crowded café, an attentive SSL front–end can learn to highlight the spectral traces of a target speaker while suppressing background chatter, motor ego-noise, and sporadic impacts, thus preserving a stable acoustic focus for downstream speech recognition, navigation, or event detection. Because multi-head self-attention offers a global receptive field and permutation-invariant decoding, it naturally accommodates multi-source tracking without handcrafted heuristics, making it a promising cornerstone for robots that must localize critical events in real time and interact seamlessly with humans in complex soundscapes. Admittedly, full Transformers demand more computation and memory than lightweight CNNs, but recent engineering advances bring their inference cost within reach of embedded platforms. Sparse attention variants reduce the quadratic complexity of self-attention, slashing both latency and RAM with negligible loss in angular accuracy [190,191]. Further savings arise from weight-sharing across heads and post-training quantization [192], allowing multi-head self-attention to execute comfortably on Jetson-class GPUs or Edge-TPUs already found on many service, inspection, and aerial robots. These techniques suggest that attention-driven SSL is poised to become a practical and powerful component of next-generation robotic audition systems.

Table 4 summarizes the studies that apply machine learning and deep learning techniques to SSL on robotic platforms. Early work (2013–2017) relied mainly on shallow neural networks and MLPs or on classical feature-reduction methods and machine learning (e.g., PCA and SVM). From 2017 onward, however, the field has shifted decisively toward convolutional, recurrent, and hybrid (CRNN) architectures. These models can jointly exploit the spatial–spectral structure of microphone signals and their temporal dynamics, enabling, for example, the localization of several simultaneous active sound sources [72] or the tracking of moving emitters [72,89,160]. Notably, to the best of our knowledge, Transformer-based networks—now state-of-the-art in many audio tasks—have not yet been adopted for robotic SSL, leaving a promising avenue for future research.

Regarding performance accuracy, we observed that studies report a variety of metrics, making a direct comparison challenging. For example, some works provide an angular (either azimuth or elevation or both) or distance error, while others report localization accuracy based on a specific prediction threshold. This inconsistency stems from diverse experimental setups, including different environmental conditions and sound source configurations. Despite this, several studies have directly compared traditional and ML/DL-based SSL methods, often demonstrating the latter's superiority in challenging scenarios. For instance, Ref. [55] reported that a proposed ML-based model significantly outperformed traditional SRP in high-noise (SNR = -10) and highly reverberant environments, achieving 40–50% lower angular error and 50% higher localization accuracy. Similarly, two studies by Takeda et al. [42,146] showed that MLP-based methods were more accurate than subspace analysis (MUSIC) in reverberant conditions. Importantly, DL-based approaches (MLPs and CNNs) have proven more effective in multi-source scenarios. Moreover, Ref. [46] found that DL-based models maintained a low angular error (below 10°) when localizing multiple

sources, whereas SRP and MUSIC errors increased to 20–40°. More recently, ref. [2] demonstrated that a CNN-based model achieved a 98.75% accuracy for single-source localization (within a 15° threshold), a significant improvement over TDOA-based methods (GCC and GCC-PHAT), which achieved around 70% accuracy.

Table 4. Machine and deep learning based SSL studies on robotic platforms.

Paper	ML/DL	Year	Robot	Max. Active Sources	S/M	No. of Mics (Geometry)	Performance (Accuracy or Error)
[30]	MLP	2013	Humanoid	1	S	2 (binaural)	Error: 2–5° (azimuth)
[55]	PCA + DFE	2014	Humanoid	1	S	2 (binaural)	Error: 1.5° avg. (azimuth)
[144]	MLP	2014	Humanoid	1	S	2 (binaural)	—
[80]	PPAM	2015	Humanoid	3	S	2 (binaural)	Error: 2.1° avg. (azimuth), 1.1° avg. (elevation)
[146]	MLP	2016	General	2	S	2 (binaural)	Above 74.8% azimuth accuracy
[42]	MLP	2016	General	1	S	2 (binaural)	Above 74.2% azimuth accuracy
[138]	SVM	2017	General	1	S	6	Error: 4.6° avg. (azimuth), 3.1° avg. (elevation)
[43]	MLP	2017	Humanoid	1	S	2 (binaural)	—
[46]	CNN	2018	Humanoid	2	S	4	Error: <5° (azimuth)
[145]	MLP	2018	Humanoid	2	S	2 (binaural)	Error: <5° (azimuth)
[161]	CNN	2019	General	1	S	2 (binaural)	Above 90% azimuth & elevation accuracy
[48]	FNN	2020	Mobile ground	1	S	8	Error: average 0.2 m (distance)
[168]	LSTM	2020	Mobile ground	1	S	6	Above 85% for azimuth & elevation accuracy
[89]	RNN	2020	Mobile ground	1	M	20	—
[59]	CNN	2021	Humanoid	1	S	2 (binaural)	97% azimuth directional accuracy (left/right/front)
[173]	CRNN	2021	Mobile ground	1	S	4 (tetrahedral)	—
[155]	CNN	2022	Humanoid	1	S	2 (binaural)	96.21% azimuth localization accuracy
[160]	MLP/CNN/LSTM	2023	Humanoid	1	M	4	Above 92% for LSTM/Bi-LSTM, 89% for CNN (distance accuracy)
[72]	CNN + LSTM	2024	General	5	S+M	3	Above 95% single-source localization (classification)
[2]	CNN	2025	Mobile ground	1	S	3	98.75% within 15° threshold (azimuth)

Acronyms: Mics—Microphones; PCA—pincipal component analysis; DFE—diffused field equalization; PPAM—probabilistic piecewise affine mapping; FNN—feedforward neural network.

In addition to comparing against traditional methods, some studies have also compared different DL architectures. Ref. [161] reported that a CNN-based model outperformed an MLP for both azimuth and elevation detection in noisy conditions (SNR = −5), showing more than a 10% gain in accuracy. For tracking moving sources, ref. [160] showed that LSTM and CNN models were considerably more accurate than MLPs, with localization errors (MAE) worth approximately 25% of the MLP's error. Hybrid architectures also show promise, as [72] demonstrated that a combined CNN and LSTM model (ConvLSTM) surpassed a pure CNN in localizing both human speech and machine-generated sounds. However, the performance of these models can still be impacted by the number of sources,

as shown in [80], where adding a second and third simultaneous source increased the azimuth error from 2.1° to 4.7° and 12°, respectively. Ref. [72] reported a similar effect, noting that more sources reduced accuracy, especially in noisy conditions. The distance between the sound source and the microphone array is another critical factor similar to what was reported for traditional SSL (Table 3). The authors in [30] found that an MLP's azimuth error slightly decreased as the source moved from 1 m to 2.8 m, while [168] showed that an LSTM's accuracy dropped by about 7% when the source distance increased from 1 m to 2 m.

Similarly to traditional methods, most ML/DL-based studies have overlooked the critical factor of computational latency and real-time performance. This lack of reporting is a notable gap in the literature. However, a few works have provided this information. For example, ref. [55] reported that their ML-based model (PCA + DFE) had a latency of just 220 ms, enabling real-time SSL. Ref. [138] also highlighted that, while SVM-based SSL requires a long training time, its localization processing time is faster than traditional methods like TDOA, facilitating closer-to-real-time performance.

5. Data and Learning Strategies

Deep learning sound-source localization requires extensive and diverse corpora as well as training paradigms that bridge the gap between laboratory conditions and everyday robotics. Mobile robots face motor ego-noise, rapidly changing geometries, overlapping talkers, and strong reverberation. We review (i) how training data are generated or collected, (ii) the augmentation schemes that mitigate over-fitting, and (iii) supervised, semi-/weakly-supervised, self-supervised, and transfer-learning paradigms that turn those data into robust models.

5.1. Data

Simulation pipelines: Most systems bootstrap with synthetic data because collecting ground-truth directions for every time frame is costly. The standard recipe convolves dry audio (a reverberation-free data) with room-impulse responses (RIRs) generated by an image-source method (ISM) [193] or its GPU-accelerated variants [194]. Open source libraries such as Habets' RIR generator [195], Pyroomacoustics [196], and ROOMSIM [197,198] can render millions of multichannel RIRs spanning room sizes, reverberation times and source–array poses [23]. Models trained on such corpora generalize surprisingly well [172,199], e.g., the VAST dataset [200] covers a vast variety of different geometries simulated in ROOMSIM and showed that virtually learned mappings on this dataset generalize to real test data. Although they have some limitations, for example, a simulated acoustic room (e.g., generated by Pyroomacoustics) is inherently unable to generate external diffuse noise and simulate obstacles or separating walls within the simulated room [72,201]. Dry-signal choice matters: mixing speech, noise, and sound events outperforms noise-only training (Krause et al. [202]). For robotics, to narrow the gap between simulation and reality, specific considerations such as ego-noise (generated by the robot itself), simulating recording from moving microphones [203], in case microphones are mounted on a mobile robot, should be taken into account. Achieving a high-fidelity simulation of SSL in robotics can avoid risky and costly field trials.

Recorded datasets: Real corpora remain indispensable for evaluation. Regarding the data collection, the emergence of some worldwide challenges organized in recent years has motivated the public sharing of the datasets. One of the most widely used real-recorded benchmarks for modern localization networks are the sound-event localization and detection (SELD) datasets released by the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [204]. Open science is a guiding principle of DCASE: every

task ships an openly licensed dataset, a baseline system, and a fixed evaluation protocol, enabling reproducible comparison across systems and years. For SSL, the relevant tasks included the localization of static sound sources [205] as well as moving ones [206–208]. The latest development of the SELD task include using audio-visual input [209], and, in 2024, distance estimation [210]. These datasets were recorded in the real-world reverberant and noisy environments containing different sound types (e.g., human speech and barking dog sounds). As highlighted in the survey by Grumiaux et al. [23], the datasets of the DCASE challenges have become the benchmark for deep learning-based SSL, e.g., [211,212]. Although the recordings are not captured on mobile robots, the datasets' noisy, reverberant, multi-source scenarios mirror many robotic deployments and could therefore serve as an invaluable data source for some robotic research. The acoustic source localization and tracking (LOCATA) challenge [213] provides another comprehensive data corpus encompassing scenarios from a single static source to multiple moving speakers, using various microphone arrays (from a 15-microphone planar array to a 12-microphone robot head array and even binaural hearing aids). This dataset could be very beneficial for robotic SSL since it includes the data captured from a pseudo-spherical array with 12 microphones integrated into a prototype head for the humanoid robot NAO.

Aside from those well-established datasets in acoustic challenges, some robotics-specific datasets have also been published. One of the examples is the SSLR data introduced by He et al. [46]. In this dataset, all recordings were made with the four head-mounted microphones of a Pepper humanoid, so every recording includes the robot's own fan noise. Two subsets are available: a loudspeaker subset in which Pepper automatically pans its head while speech is played from random positions, and a human-interaction subset that captures single and overlapping utterances during live dialogues. Both subsets provide motion-capture ground truth and voice-activity labels, making SSLR particularly suitable for human–robot interaction studies. Another dataset by [35], designed for robot audition, was recorded across four real-world environments with varying reverberation times, utilizing an NAO humanoid robot equipped with microphones on its head. It comprises audio-only data, i.e., speech emitted via loudspeaker in the lab, fixed distances, 360° azimuth/elevation range, and audio-visual data, i.e., speech within the robot's camera field-of-view, both predominantly affected by the robot's self-generated fan noise. For aerial robotics, the most complete resource is DREGON, published by Strauss et al. [37]. Here, an eight-microphone cube is suspended beneath a quadrotor; flights are tracked by a Vicon system and accompanied by synchronized IMU and motor-speed logs. The corpus combines challenging in-flight speech or noise at negative SNRs with separate ego-noise flights and semi-anechoic "clean" loud-speaker recordings, so researchers can test localization, noise suppression, and data synthesis with the same material. Wang et al. [38] extended the aerial scenario with an audio-visual quadcopter (AVQ) dataset. The 50 min collection is divided into two subsets. In the first, up to two talkers stand at nine predefined locations between two and six meters from the drone while speech and rotor noise are recorded separately, allowing controlled SNR studies. In the second subset, a loud-speaker is carried along three-minute trajectories, giving moving-source material at varying thrust levels. Video frames at 30 fps provide the ground-truth angles for both subsets. Most recently, Jekaterynczuk et al. [214] presented UaVirBASE, a publicly downloadable database aimed at ground-based drone monitoring. It contains the high-quality recordings of a single UAV captured from many distances, heights, and azimuths, and includes detailed metadata such as array coordinates and environmental conditions. A baseline mel-spectrogram DNN trained on UaVirBASE achieves roughly half-meter mean error in range and about one degree in azimuth, demonstrating the dataset's suitability for acoustic surveillance applications. Taken together, SSLR, DREGON, AVQ, and UaVirBASE provide complementary test

beds that include ego-noise, platform motion, and realistic microphone layouts, conditions that conventional laboratory corpora and purely simulated data rarely reproduce in full.

Data augmentation. The performance of deep learning-based SSL models, especially in complex robotic applications, is highly dependent on the quantity and diversity of training data. To overcome limitations in real-world data availability and mitigate the potential for models to overfit to simulated conditions, data augmentation techniques are indispensable. These methods create new training examples from existing data, enriching the dataset without requiring additional physical recordings, and ultimately leading to more robust and generalizable SSL models for robots. Various data augmentation strategies have been explored, many inspired by their success in other domains. For instance, techniques like SpecAugment [215], originally used for speech recognition [216], involve masking certain time frames or frequency bands within spectrograms. This teaches the model to focus on the more fundamental spatial cues rather than relying on specific spectral content, a valuable trait for robots encountering diverse sound characteristics. Other methods leverage the properties of multi-channel audio formats, such as first-order ambisonics (FOA). Mazzon et al. [217] proposed techniques like channel swapping or inversion and label-oriented rotation, which effectively generate new sound source locations and orientations by transforming the FOA channels and their corresponding labels. This allows a robot to learn how sound propagates and presents itself from a wider range of directions than might be practically recordable. More broadly, methods like Mixup [218] generate new training samples by creating convex combinations of existing data pairs, while techniques like pitch shifting [219] and the random mixing of multiple training signals [220,221] create novel mixtures, simulating scenarios with varying speaker characteristics or multiple overlapping sound sources. For robotics, these data augmentation techniques translate directly into significant practical advantages. By exposing SSL models to a vastly expanded and diversified set of synthetic acoustic environments, robots can develop robust "acoustic perception" that is less susceptible to the unpredictability of real-world deployment. Robots equipped with models trained on augmented data will be better at localizing sources even in highly reverberant rooms, amidst unforeseen background noises, or when facing multiple simultaneous speakers.

5.2. Learning Paradigms

The effectiveness of deep neural networks in sound source localization heavily depends on the chosen learning strategy, which dictates how the model acquires knowledge from data. While most SSL systems rely on supervised learning [23], the limitations of labeled data in complex robot environments necessitate exploring semi-supervised, weakly supervised, and transfer learning approaches.

Supervised learning: Supervised learning is the predominant paradigm in DL-based SSL. It involves training a neural network on a dataset where each input (e.g., multi-channel audio features) is paired with a corresponding ground truth output (the "label"), the known position, or the direction of the sound source. The network learns by minimizing a cost function (or loss function), which quantifies the discrepancy between its predicted output and the true label.

For robotic SSL, supervised learning typically manifests in two primary ways:

- Classification: When the sound source's location is discretized into angular bins (e.g., five-degree sectors around the robot). Here, a softmax activation function is often used at the output, and the model minimizes the categorical cross-entropy loss. This is suitable for tasks where a robot needs to identify which general direction a sound is coming from [2].

- Regression: When the goal is to predict continuous values for the sound source's azimuth, elevation, or 3D Cartesian coordinates. In this case, the mean square error (MSE) is the most common choice for the cost function [222]. This enables robots to pinpoint locations with greater precision. While MSE is prevalent, other metrics like angular error or L1-norm are sometimes employed to capture specific aspects of localization accuracy [223].

The primary limitation of supervised learning for robotic applications is its insatiable demand for large amounts of accurately labeled training data. Collecting such data in real-world robotic environments is incredibly resource-intensive, as it requires meticulously tracking sound sources and robot poses. Existing real-world SSL datasets for robotics are often limited in size and variety, making them insufficient for robust DL model training. To address this, supervised learning is often augmented with extensive data simulation and data augmentation techniques, which synthetically expand the dataset's diversity.

Beyond full supervision: To cope with the scarcity of labeled real-world data and enhance model robustness to unseen conditions, researchers leverage strategies that go beyond purely supervised learning:

- Semi-supervised learning: This approach combines both labeled and unlabeled data for training [23]. The core idea is to perform part of the learning in a supervised manner (using available labels) and another part in an unsupervised manner (learning from unlabeled data) [224]. For robotics, this is highly valuable because robots can continuously collect vast amounts of unlabeled audio data during their operation. Semi-supervised learning methods can fine-tune a network pre-trained on labeled data (often simulated data), adapting it to real-world conditions without requiring exhaustive manual labeling. Techniques such as minimizing overall entropy (e.g., SSL work by Takeda et al. [225]) or employing generative models (e.g., SSL work by Bianco et al. [226]) that learn underlying data distributions have been proposed. Adversarial training (e.g., SSL work by Le Moing et al. [227]) is another example, where a discriminator network tries to distinguish real from simulated data, while the SSL network learns to "fool" it by producing realistic outputs from simulated inputs, thus adapting to real-world acoustic characteristics. This allows a robot to refine its SSL capabilities based on its own experiences in the operational environment, even if precise ground truth is unavailable.
- Weakly supervised learning: In this paradigm, the training data comes with "weak" or imprecise labels, rather than detailed ground truth in many domains when labeling is costly or challenging [228]. For SSL, this might mean only knowing the number of sound sources present, or having a rough idea of their general location without the exact coordinates. Models are designed with specialized cost functions that can account for these less precise labels. For instance, He et al. [229] fine-tuned networks using weak labels representing the known number of sources, which helped regularize predictions. Other approaches, like those using triplet loss functions by Opochinsky et al. [230], involve training the network to correctly infer the relative positions of sound sources (e.g., that one source is closer than another), even if their absolute coordinates are not provided. For robotics, weakly supervised learning offers a pragmatic solution when collecting precise ground truth labels is impractical, allowing robots to learn from more easily obtainable, albeit less granular, supervisory signals. This means a robot could learn to localize effectively just by being told, for example, "there is a sound coming from that general direction" rather than requiring exact coordinates, making deployment and ongoing learning more feasible.

Transfer learning: Transfer learning is a pivotal strategy that addresses the common challenge of data scarcity in specific target domains [231], a particularly pertinent issue for

robotic SSL. Rather than training a deep neural network from scratch on a limited robot-specific dataset, transfer learning involves leveraging the knowledge acquired by a model pre-trained on a related, typically larger, source dataset. This strategy offers significant advantages for robotic applications.

The typical workflow for transfer learning in SSL involves two main phases. First, a base model is pre-trained on a vast dataset, which consists of general acoustic scenes. In this regard, some works, such as SSL studies by Nguyen et al. and Park et al. [166,232], pre-trained their model on a large realistic dataset (e.g., audio datasets [233]), and some (e.g., Zhang et al. [234]) used simulation to generate the training data for pre-training. This initial training allows the network to learn rich, generalized representations of sound features and spatial cues. Second, the pre-trained model is then adapted to the specific SSL task through fine-tuning. This fine-tuning process involves the further training of the model on a comparatively smaller dataset, such as a robot-specific dataset, often with a reduced learning rate to preserve the learned general knowledge. During fine-tuning, the selected layers of the pre-trained network might be frozen (acting as fixed feature extractors), or the entire network might be updated, allowing it to specialize in the specific domain, for example, the robot's acoustic environment, its ego-noise characteristics, and the unique geometry of its microphone array.

For robotics, transfer learning is highly advantageous. It significantly reduces the amount of labeled robot-specific data required, thereby accelerating development and deployment cycles. A robot can benefit from the models pre-trained on generic acoustic datasets or other robotic datasets, such as SSLR, DREGON, AVQ, and UaVirBASE datasets which are detailed before, then quickly adapt to its unique operational environment with minimal new data. This strategy also improves the model generalization capabilities, as the network benefits from the broader knowledge base of the pre-training data, making it more robust to unforeseen acoustic conditions, various sound sources, and environmental variations that a robot might encounter in real-world scenarios. Consequently, transfer learning stands as a powerful enabler for rapidly deploying accurate and robust SSL capabilities on diverse robotic platforms.

6. Applications of SSL in Robotics

Sound source localization (SSL) is a fundamental capability that significantly enhances the autonomy, perceptual awareness, and interactive abilities of robotic systems. Its applications span a wide array of robot types and operational domains, enabling robust and versatile perception in complex, real-world environments. To better illustrate where SSL has already proven its worth, we categorize the papers reviewed in this study first by the robotic platform employed, namely humanoid robots, mobile ground robotic platforms, and UAVs. Subsequently, we delineate their applications based on motivating domains.

6.1. Categorization by Robot Type

The integration of SSL capabilities varies significantly depending on the physical characteristics and primary functions of different robotic platforms. From the 78 robotic SSL papers analyzed, the predominant robot types fall into three categories: mobile ground robotic platforms, humanoid robots, and UAVs. Additionally, a notable portion of studies proposed SSL methods generalizable across platforms without specifying a particular robot type for implementation, which we categorize as "general robotic platforms". The distribution of these categories among the reviewed papers is depicted in Figure 6.

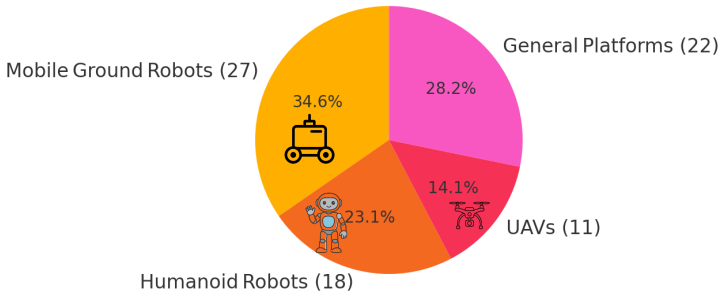


Figure 6. Distribution of reviewed SSL studies by robot types.

Mobile ground robots: Representing the largest category in our review with 27 papers, mobile ground robots are a cornerstone of robotic auditioning systems. These platforms, encompassing wheeled, legged, and tracked designs, leverage SSL to navigate, monitor, and interact with their environment, particularly in scenarios where visual information may be limited or unreliable. The TurtleBot was a commonly utilized mobile ground robot in these studies [39,47,81,85,160], while many researchers also employed customized robotic platforms tailored to their specific research objectives, such as ANYmal legged robot [53], a pet robot [122], pipe inspector robot [235], a hose-shaped robot [12], and a bio-robot [10]. Figure 7 represents these mobile robots.

Microphone arrays are typically mounted on their chassis or integrated into dedicated sensor heads. Their applications include

- Acoustic navigation and mapping: Mobile robots can utilize SSL to identify and localize fixed sound sources (e.g., specific machinery hums in an industrial setting, ventilation systems, or public address announcements) as acoustic landmarks for simultaneous localization and mapping (SLAM) or precise navigation in GPS-denied or visually ambiguous areas.
- Hazard detection and avoidance: For service robots in public spaces or industrial robots on factory floors, SSL enables the early detection and localization of unexpected or dangerous sounds (e.g., a car horn, breaking glass, a falling object, abnormal machinery sounds). This facilitates proactive collision avoidance or emergency response.
- Sound-guided exploration and search: In search and rescue scenarios (e.g., navigating collapsed buildings or smoke-filled areas), mobile robots can rely on SSL to precisely locate sounds like human voices or whistles, effectively guiding their exploration towards potential survivors.
- Security and surveillance: Mobile robots can patrol large areas, employing SSL to detect and track intruders or suspicious acoustic events (e.g., footsteps, suspicious voices) in various service or security contexts.

For mobile ground robots, the choice of SSL method is often driven by the need for robust performance in diverse and sometimes noisy environments (e.g., industrial settings, outdoor search-and-rescue). TDOA and SRP are commonly used for their computational efficiency, which is important for on-the-go processing. As seen in Table 3, studies on mobile ground robots have successfully used TDOA [47,104] and SRP [39,50] to localize both static and moving sources, making them suitable for dynamic applications like sound-guided exploration. For more complex scenarios involving multiple sources or high noise, DL methods are increasingly preferred. As shown in Table 4, approaches like FNN [48], RNN [89], and CRNN [173] have been applied to mobile ground robots, offering improved accuracy for tasks like navigation and object detection in noisy environments. The

combination of CNN and LSTM, as in [72], proves highly effective for localizing multiple static and moving sources, making it a powerful tool for complex acoustic mapping tasks.

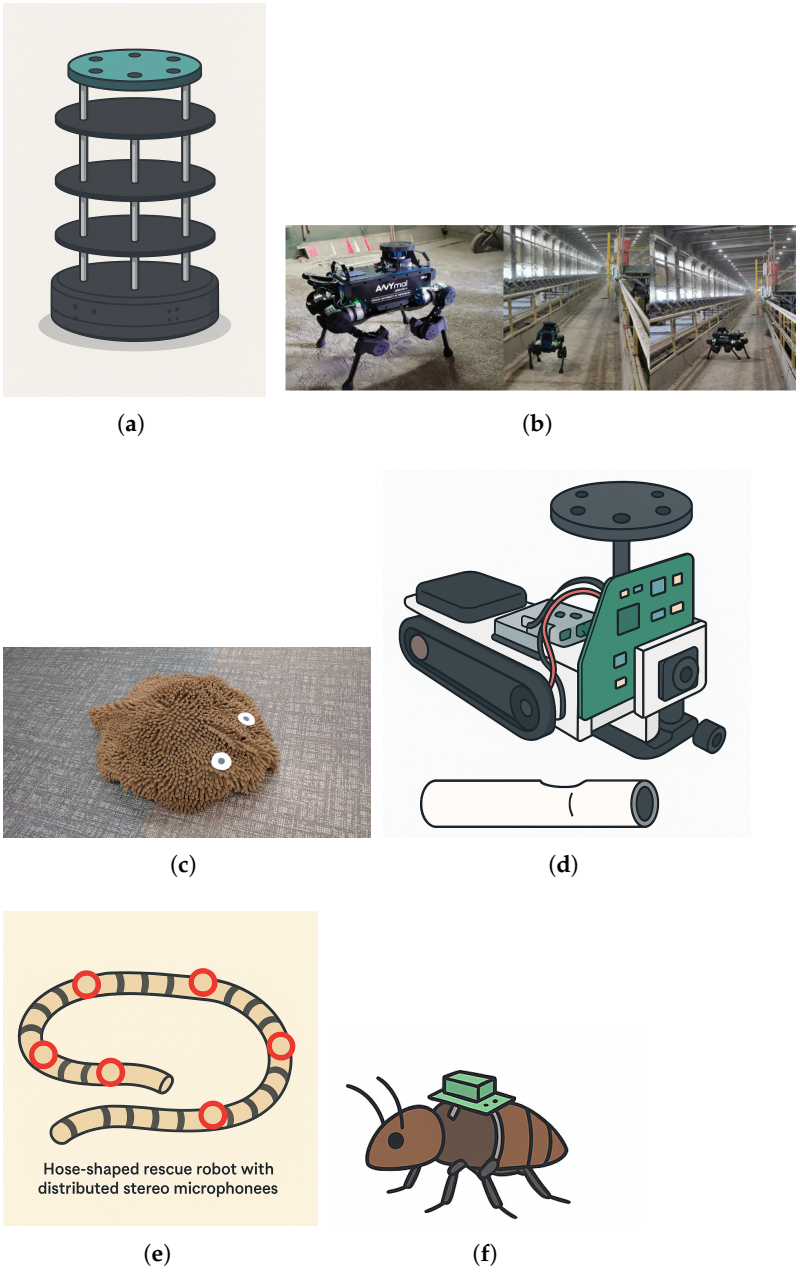
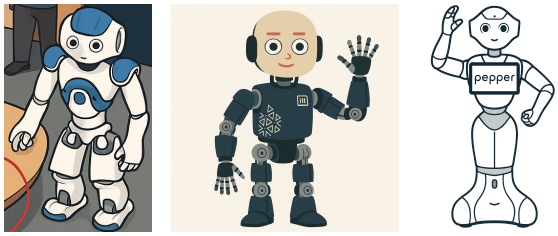


Figure 7. Different mobile robot types employed in SSL studies. (**a**) TurtleBot2 with a circular microphone array mounted on top; (**b**) ANYmal quadruped using SSL to inspect a mining belt conveyor [53]. Reproduced without modifications from [53], article distributed under the Creative Commons Attribution 4.0 International licence (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/ (accessed on 15 June 2025)). (**c**) Fur-covered "pet" robot that localizes humans via SSL for social interaction [122]. Reproduced without modifications from [122], article distributed under the Creative Commons Attribution-NoDerivatives 4.0 International licence (CC BY-ND 4.0, https://creativecommons.org/licenses/by-nd/4.0/ (accessed on 15 June 2025)). (**d**) Schematic of the tracked pipe-inspection robot with a four-microphone array used for sound-source-localization-based defect detection (redrawn and adapted from [235]); (**e**) Hose-shaped rescue robot equipped with distributed stereo microphones for sound-source localisation (redrawn and adapted from [12]); (**f**) Insect-scale bio-robot equipped with a lightweight microphone array for SSL in search-and-rescue missions (redrawn and adapted from [10]).

Humanoid robots: Eighteen papers among the 78 reviewed implemented SSL on humanoid robots or platforms specifically designed with a dummy head configuration, e.g., Knowles Electronics Mannequin for Acoustic Research (KEMAR), widely used in SSL [55]. Figure 8 shows the typical humanoid robots used in SSL. The NAO robot (Figure 8a) [35,43,105,160] and iCub (Figure 8b) [59,144,145,236] were frequently employed platforms in these studies, alongside others such as Pepper (Figure 8c) [29,46], Hearbo [128] and HRP-2 [129].



(**a**) (**b**) (**c**)

Figure 8. Common humanoids in SSL studies. (**a**) NAO used in [35,43,105,160]; (**b**) iCub used in [59,144,145,236]; and (**c**) Pepper robot used in [29,46].

These humanoid robots, designed to interact with and operate in human-centric environments, benefit immensely from sophisticated auditory perception. Equipped with microphone arrays, often integrated into their heads (mimicking binaural hearing) or across their torsos, they utilize SSL for

- Natural human–robot interaction (HRI): Localizing the direction of human speech allows humanoids to naturally orient their head and gaze towards a speaker, facilitating engaging conversations and conveying attentiveness. This is crucial for social and companion robots aiming to build rapport with users.
- Multi-speaker tracking: In dynamic social settings, humanoids can employ advanced SSL techniques (e.g., using deep learning or subspace methods) to identify and track multiple simultaneous speakers, enabling complex multi-party conversations and selective listening.
- Enhanced situational awareness: Beyond direct interaction, SSL enables humanoids to detect and localize various environmental sounds, such as alarms, doorbells, or footsteps, contributing to their overall understanding of the surrounding service environment.

Humanoid robots, with their emphasis on natural human–robot interaction (HRI), benefit from methods that can handle complex speech signals and multiple speakers. Binaural SSL and DL methods are particularly well-suited due to their head-like microphone configuration. MLP-based SSL [30,80] has been successfully implemented in humanoid robots to provide directional cues in experiments. More recently, as shown in Table 4, DL methods like CNNs [46,59] and RNNs like LSTM and Bi-LSTM [160] have shown superior performance over MLPs in experiments using humanoids. These models are ideal for processing the temporal and spectral features of speech, which are essential for multi-speaker tracking and enhanced HRI. The LSTM model, in particular, has demonstrated high accuracy in localizing moving sources, which is crucial for dynamic interactions.

Unmanned aerial vehicles (UAVs): Eleven papers in our review specifically utilized UAVs for their SSL investigations. UAVs present unique challenges and opportunities for SSL due to their inherent ego-noise (from propellers) and often limited payload capacity. Despite these hurdles, SSL offers them extended perception capabilities:

- Remote monitoring and surveillance: UAVs equipped with SSL can perform long-range acoustic monitoring for environmental applications (e.g., detecting illegal logging, tracking wildlife based on vocalizations) or security tasks (e.g., localizing gunshots or human activity over large, inaccessible terrains).
- Search and rescue in challenging terrains: In search and rescue operations over vast or difficult-to-traverse areas (e.g., dense forests, mountainous regions), UAVs can use SSL to pinpoint distress calls or specific human sounds, directing ground teams more efficiently.
- Hazard identification: Detecting and localizing critical sounds like explosions, gas leaks (through associated sounds), or structural failures from a safe distance in industrial or disaster zones.
- Traffic monitoring: In service or urban planning contexts, UAVs can localize vehicles based on their sounds, contributing to traffic flow analysis.

For UAVs, the primary challenge is the overwhelming propeller ego-noise. Therefore, SSL methods must be highly robust to this broadband noise. Subspace analysis methods, such as MUSIC, are particularly suitable as they can effectively separate the sound source signal from the robot's self-noise based on the structure of the signal's covariance matrix [32,52,58]. However, as highlighted by [125], there is a trade-off between noise tolerance and real-time performance within subspace analysis. For instance, SEVD-MUSIC offers a low computational cost and short delay but has poor noise tolerance, making the UAV's own noise visible in the processed spectrum. Conversely, iGSVD-MUSIC includes noise whitening, which provides excellent noise tolerance at the cost of a longer delay (2–3 s longer than SEVD-MUSIC), as seen in [125]'s findings. This suggests that the choice of subspace analysis method for UAVs depends on whether real-time performance or noise robustness is the priority for a given application. As seen in Table 3, this approach has been successfully applied to UAVs with excellent results, with some studies achieving an average 3D angle error of less than 2° in real-world scenarios [58]. Beamforming is another highly suitable method for UAVs. By steering a directional beam and forming a spatial filter, it effectively enhances the signal from a specific direction while suppressing noise from other directions, including the robot's own noise. Studies like [51,116] demonstrate that beamforming is an effective solution for long-range acoustic monitoring and surveillance, where the sound of interest is often faint and distant. TDOA-based methods can be less robust to ego-noise without significant preprocessing. However, with a sufficient number of microphones and careful array design, they have been successfully applied, as demonstrated by the use of an eight-microphone array for an indoor UAV in [37]. The choice of method, therefore, depends on the level of noise and the available computational resources.

6.2. Categorization by Application Domain

While the robot type defines the platform, the application domain dictates the specific tasks and environmental challenges that SSL must address. Based on our review, four primary domains have emerged: social and companion, search and rescue, service, and industrial robotics. We also introduce a fifth category, "General" for studies that do not restrict their applications to a single specific domain. The proportion of reviewed papers falling into each domain is depicted in a pie chart in Figure 9.

Social and companion: Among the reviewed literature, fourteen studies focused on enhancing social interaction and companionship through robotic systems. These works primarily employed humanoid robots [35,43,46,54,103,105,144,145,161], although some mobile ground robots were also utilized [3,54]. For robots designed to interact with and assist humans in social contexts, SSL is paramount for fostering natural and engaging relationships. It enables accurate speech localization for fluent conversations, allows the

robot to discern who is speaking in group settings, and facilitates attentive behaviors by orienting towards auditory cues, thereby enhancing the robot's perceived intelligence and empathy (through possible facial analysis as an additional modality [3,54]). A central theme in these studies is the detection and localization of human speech, often considering variations such as male and female speakers due to the human-centric nature of the sound sources [43,161]. Notably, an interesting research by Chen et al. [105] explored localizing the referee whistle in RoboCup competitions for NAO robots, requiring distinction from other ambient sounds like human speech and claps.

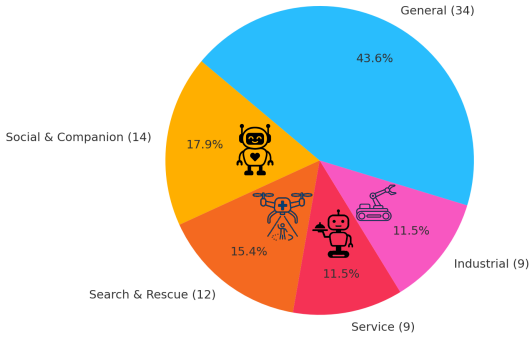


Figure 9. Distribution of SSL applications in robotics across different domains in our review.

Search and rescue: Twelve papers in our review addressed the theme of search and rescue, a domain where SSL serves as a vital, often life-saving, capability for robots deployed in hazardous or disaster-stricken areas where visibility is severely compromised. Sound separation and detection are critical factors in this domain, as the robot must reliably distinguish target sounds from its own ego-noise and other non-pertinent ambient noises. These robots leverage SSL to pinpoint the exact locations of trapped survivors by detecting specific sounds, such as human speech and chirps [37,58] or whistles [36,125], thereby guiding rescue teams to precise areas within debris fields or collapsed structures. Most of these study UAVs in both outdoor [9,36,49,52,125] and indoor settings [37,58]. Beyond UAVs, some research also incorporated mobile ground robots [11] and humanoids [129], as well as custom-designed robots like hose-shaped or bio-inspired platforms for specialized search and rescue experiments [10,12].

Service: Nine studies among the reviewed papers focused on robotic applications within the service domain. SSL empowers service robots to operate effectively in diverse human environments, ranging from smart homes to public spaces. Key applications include speaker localization for intuitive interaction, the detection of specific household sounds (e.g., door knocks, appliance alarms, and hairdryers [45,173]) or public place acoustics (e.g., in classrooms [118]), and the monitoring for unusual acoustic events in residential or commercial settings, including security applications [14]. These studies primarily focused on indoor environments, predominantly utilizing ground mobile robots [2,45,60,123,124,173].

Industrial: Among the reviewed articles, nine papers specifically addressed the industrial domain. Six of these studies directly focused on industrial condition monitoring for fault detection using SSL with robotic platforms [5,6,48,53,235,237]. Complementing these, three other papers, while not directly implementing robots in their experimental setups, explored SSL within contexts highly relevant to future robotic integration in industrial environments. Jalayer et al. [4,72] investigated applications in manufacturing settings, considering both machinery sounds and human speech, thereby laying groundwork for robots operating alongside human workers. With the same spirit, Sun et al. [138] discussed the potential for industrial SSL implementation by utilizing common sounds in such environ-

ments, including human speech, machinery sounds, and telephone rings in indoor settings. Although SSL remains relatively underexplored in this domain compared to others, manufacturing plants, warehouses, and other industrial settings hold immense potential for robots equipped with advanced SSL techniques to significantly enhance automation and safety. One of the great use cases for SSL in this context could be directly tied to occupational health and safety. Robots can leverage SSL for predictive maintenance by localizing anomalous machinery sounds (e.g., grinding, squealing) to identify failing components, preventing equipment malfunctions that could lead to worker injury. Furthermore, SSL could be a critical component of industrial collaborative robot (cobot) systems, enabling the real-time detection of human presence and voice commands for collision avoidance and emergency shutdowns. By monitoring for critical safety alarms or the sound of dropped tools and other auditory anomalies, SSL provides a complementary safety layer that is not susceptible to the line-of-sight limitations of cameras or LiDAR, thereby offering a robust, multi-modal approach to ensuring a safer work environment.

7. Challenges and Future Avenues

Despite significant advances in SSL for robotics, particularly with the integration of deep learning methods, numerous challenges remain. This section examines the current limitations and emerging research directions that aim to address these challenges and further advance the field.

7.1. Current Challenges

7.1.1. Environmental Robustness

Although compared to traditional signal processing, DL models performed better in complex environments, achieving robust and reliable performance across diverse and highly dynamic acoustic environments is still a hurdle for robot audition systems. These challenges can be categorized as follows:

- **Extreme reverberation:** In large indoor spaces with reflective surfaces, extreme reverberation severely degrades a robot's ability to precisely map its acoustic surroundings. This phenomenon continues to be a significant obstacle to SSL accuracy in multi-source and noisy environments, causing sound reflections to confuse the robot's directional cues. In highly reflective environments, the multi-path distorts the inter-microphone cross-correlation, producing spurious TDOA peaks and front–back/elevation ambiguities for linear or near-planar arrays. Classical SRP-PHAT/MUSIC degrade as the direct-path coherence drops and DL-based models trained on mild reverberation overfit to room responses and suffer under unseen room geometry. This degradation has been empirically confirmed; for example, Keyrouz [55] reported significant performance drops in extreme reverberation for both ML-based methods and, more severely, for SRP-PHAT-based SSL.
- **Different noise types:** Real-world scenarios are characterized by complex, unpredictable background noise, such as human chatter, machinery, or general environmental sounds, which can mask critical acoustic cues. Furthermore, a robot's own self-generated noise (ego-noise) from motors, actuators, and movement significantly complicates its internal auditory focus. Although many studies considered ego-noise in their studies and datasets [37], considering multiple noise types and their cumulative effect that can happen in realistic applications is still a challenge. Also, considering the effect of a microphone's internal noise (caused by imperfection [72]) could add complexity, and this should be addressed since microphones do not always work perfectly. To the best of our knowledge, still no robotic studies have systematically and comprehensively examined the simultaneous presence of multiple noise types during

experiments. Nevertheless, prior work has demonstrated that high-noise conditions (SNR below 0 dB) considerably impair localization accuracy in both traditional SSL methods (e.g., TDOA and subspace approaches [103]) and DL-based models (e.g., MLPs and, to a lesser degree, CNNs [161]). Despite these findings, the cumulative effects of diverse and concurrent noise sources remain largely unaddressed in the literature, representing an important gap for realistic robotic applications.

- **Dynamic acoustic conditions and unstructured environments:** As robots navigate, acoustic conditions continuously change due to varying listener positions relative to sources and reflections. Even under controlled experimental conditions, source–listener distance alone can significantly influence localization and tracking accuracy. This effect has been reported for both ML/DL-based methods (e.g., MLP [30], LSTM [168]) and traditional SSL approaches [51]. When combined with changes in acoustic conditions, such variations present even greater challenges. Despite their practical relevance, a comprehensive analysis of these combined effects remains unexplored in robotic experiments. Adapting to these dynamic changes without requiring re-calibration or extensive retraining presents a significant challenge for both traditional and deep learning methods. Moreover, extending SSL approaches to outdoor (especially for UAVs) or highly unstructured environments introduces additional complexities related to wind, varying atmospheric conditions, and unpredictable sound propagation.

7.1.2. Multi-Source Scenarios

Accurately localizing and distinguishing multiple simultaneous sound sources, particularly in cluttered acoustic environments, poses a significant scene analysis problem for robotic systems. This challenge is highlighted in many SSL studies in general [23], including robotics for both traditional and ML/DL-based methods. The main difficulties can be detailed as follows:

- **Source separation vs. localization interdependence:** A fundamental dilemma for a robot's auditory processing pipeline is the interdependency between source separation and localization. The effective separation of individual sound streams often requires prior knowledge of source locations, while accurate localization may require separated or enhanced source signals, creating a "chicken-and-egg" problem for the robot's acoustic intelligence. In this regard, some previous SSL studies implemented separate DL models for sound counting, classification and localization tasks [143], while some did all within one model [46].
- **Overlapping sources:** When multiple sound sources extensively overlap in time and frequency, it profoundly challenges a robot's ability to differentiate and pinpoint distinct acoustic events. This is particularly difficult for speech sources with similar spectral characteristics or when multiple human speakers are closely spaced (a cocktail party scenario [54]), affecting a robot's ability to focus on a specific speaker. Performance degradation in the presence of multiple simultaneous sound-emitting sources has been observed in practice. For instance, Yamada et al. [52] reported increased localization errors when two sources emitted sounds simultaneously using a traditional SSL method (MUSIC). Similarly, Jalayer et al. [72] documented reduced accuracy for a DL-based method (ConvLSTM) when multiple sound sources overlapped in time, underscoring the persistence of this challenge across methodological paradigms.
- **Variable source numbers:** Real-world scenarios involve a fluctuating number of active sound sources over time. Developing robotic auditory systems that can dynamically adapt to changing numbers of sources without explicit prior knowledge remains a complex task, requiring flexible and scalable processing architectures.

- **Source tracking and identity maintenance:** Maintaining consistent the identity tracking of multiple moving sound sources over extended periods [56], especially through instances of silence, occlusion, or signal degradation, presents significant difficulties that current robotic audition methods have not fully resolved [23]. This capability is essential for a robot to maintain situational awareness and interact intelligently with dynamic entities.

7.1.3. Practical Implementation Constraints

Deploying robust SSL systems on real robotic platforms introduces several practical constraints that must be overcome:

- **Computational efficiency and power consumption:** Real-time processing requirements coupled with the limited computational resources and strict power budgets on many robotic platforms directly impact a robot's operational endurance and real-time responsiveness. There is a growing interest in "energy-efficient wake-up technologies" [238] for SSL, as highlighted by Khan et al. [28], to enable long-duration missions for battery-powered robots by minimizing energy expenditure on continuous acoustic monitoring.
- **Microphone array limitations and flexibility:** Physical constraints on microphone placement, quality, and array geometry across diverse robotic platforms (e.g., humanoid heads, mobile robot chassis, UAVs) impose inherent sensory limitations on a robot's acoustic perception and can significantly impact SSL performance. The type of input features and microphone array configurations heavily influence the effectiveness of deep learning approaches, often requiring model retraining for different robot setups. Moreover, supporting dynamic microphone array configurations (e.g., on articulated or reconfigurable robots) remains a challenge.
- **Calibration and maintenance:** Ensuring consistent acoustic performance over a robot's operational lifespan demands robust self-calibration routines within its perceptual architecture to address issues such as microphone drift, physical damage, or changes in robot configuration that may affect the acoustic properties of the system. This is crucial for maintaining the integrity of the robot's auditory data.

7.1.4. Data and Learning Challenges

Deep learning approaches, while powerful, face specific challenges related to data management and learning processes for SSL in robotics:

- **Limited Training Data and Annotation Complexity:** Collecting diverse, high-quality labeled acoustic data for SSL is inherently time-consuming and expensive for robotic applications. Unlike vision datasets (where images can often be manually labeled), obtaining the true direction or position of a sound source at each time requires specialized equipment (e.g., motion tracking systems as used in LOCATA [213]) or careful calibration with known source positions. This becomes even harder if either the sound source or the robot (or both) are moving. As a result, truly comprehensive spatial audio datasets for robotics are scarce. Furthermore, a significant impediment is that most research studies often do not share their collected data, hindering broader research progress and comparative analysis.
- **Lack of comprehensive benchmarks in SSL:** Despite some efforts (e.g., CASE challenges in the last decade [204]), there is a notable lack of comprehensive, widely adopted benchmark datasets (like ImageNet in visual object recognition) in the field of SSL. This absence makes it difficult for researchers to uniformly test and compare the performance of different models, impeding the standardized evaluation and progress towards common goals.

- **Sim-to-real gap and generalization:** While simulated environments, such as Pyroomacoustics [196] and ROOMSIM [197,198], can generate large amounts of training data, bridging the fidelity gap between simulated and real-world acoustics remains a critical barrier to a robot's transition from simulated training to autonomous real-world operation. Consequently, models trained on specific environments or conditions often fail to generalize robustly to new, unseen scenarios, limiting their real-world applicability for robotic deployment.
- **Interpretability:** The "black box" nature of many deep learning models complicates debugging, understanding their decision-making processes, and ensuring their reliability. This lack of interpretability can be particularly problematic for safety-critical robotics (e.g., rescue robots [36] or condition monitoring robots [6]) applications where explainability and a human operator's trust in the robot's auditory decisions are paramount.
- **Accurate 3D localization and distance estimation:** While 2D direction of arrival (DoA) is commonly addressed, accurately estimating the distance to a sound source, especially in reverberant environments, remains a more complex task for a robot's spatial awareness. As highlighted by Rascon et al. [1], reliable 3D localization (azimuth, elevation, and distance) with high resolution is essential for a robot's precise navigation, manipulation, and interaction within a volumetric space, but it is not yet consistently achievable in real-time.

7.2. Future Opportunities and Avenues

Building upon the current challenges, the field of robotic sound source localization presents numerous exciting opportunities for research and development. These avenues aim to elevate a robot's auditory intelligence, enabling more autonomous, perceptive, and adaptable machines for a wide range of applications.

7.2.1. Enhancing Robustness and Adaptability in Robot Audition

Future work will focus on equipping robots with auditory systems capable of navigating the most challenging acoustic environments with unprecedented reliability:

- **Adaptive noise and reverberation suppression:** Developing advanced signal processing and deep learning techniques that can adaptively suppress diverse non-stationary noise and mitigate extreme reverberation in real-time. This includes research into robust ego-noise cancellation specific to robotic platforms, ensuring that a robot can maintain auditory focus even during rapid movement or noisy operations.
- **Outdoor and unstructured acoustic modeling:** Expanding SSL research beyond the controlled indoor environments of the labs to truly unstructured outdoor settings. This involves developing new models for sound propagation in open air, by introducing novel outdoor acoustic simulators, accounting for environmental factors like wind, and leveraging multi-modal fusion with visual or inertial sensors to enhance robustness where audio cues might be ambiguous.
- **Dynamic acoustic scene understanding:** Moving beyond static snapshot localization to the continuous, real-time comprehension of evolving acoustic scenes. This includes rapid adaptation to changing sound source characteristics, varying background noise, and dynamic room acoustics as the robot moves, fostering a more fluid and context-aware auditory perception.

7.2.2. Advanced Multi-Source Auditory Scene Analysis

Opportunities abound in enabling robots to dissect complex auditory environments with multiple, interacting sound sources:

- **Integrated sound event localization and detection (SELD) with source separation:** Developing holistic systems that simultaneously detect, localize, and separate multiple overlapping sound events. This integrated approach, often framed as a multi-task learning problem, promises to break the current "chicken-and-egg" dilemma between separation and localization, allowing robots to extract distinct auditory streams for specific analysis or interaction.
- **Robust multi-object acoustic tracking:** Advancing algorithms for the reliable and persistent tracking of multiple moving sound sources, including handling occlusions, disappearances, and re-appearances. This is crucial for collaborative robots interacting with multiple humans or for inspection robots monitoring various machinery components in parallel.
- **Dynamic source counting and characterization:** Equipping robots with the ability to dynamically estimate the number of active sound sources in real-time, along with their types (e.g., speech, machinery, alarms). This capability enhances a robot's contextual awareness, allowing it to prioritize relevant acoustic information within a complex environment.

7.2.3. Efficient and Flexible Robotic System Integration

Future research will drive the development of SSL solutions optimized for practical deployment on diverse robotic hardware:

- **Lightweight and energy-efficient deep learning models:** Designing novel, compact deep learning architectures and employing techniques like model quantization, pruning, and knowledge distillation tailored for execution on resource-constrained edge AI platforms. Also, advancements in power-saving techniques (e.g., the wake-up strategy described in Khan et al. survey [28]) promise to keep the auditory system off until an acoustic event of interest occurs, allowing the robot to reserve precious energy for localization and task execution. This will enable robots to perform complex SSL tasks while maintaining long operational durations and reducing power consumption.
- **Flexible and self-calibrating microphone arrays:** Investigating SSL methods that are robust to variations in microphone array geometry, supporting dynamic reconfigurations inherent to mobile or articulated robots. Furthermore, developing autonomous, self-calibrating routines for microphone arrays will reduce deployment complexity and ensure consistent performance over a robot's lifespan, compensating for sensor drift or minor physical changes. Techniques like non-synchronous measurement (NSM) technology, which allows a small moving microphone array to emulate a larger static one, offer promising avenues for achieving high-resolution localization with fewer microphones while managing the challenges of dynamic array configurations [6].
- **Hardware-software co-design for robot audition:** Fostering a holistic approach where microphone array design, sensor placement on the robot, and signal processing algorithms are jointly optimized. This co-design can lead to highly efficient and purpose-built auditory systems that maximize performance within a robot's physical and computational constraints.

7.2.4. Advancements in Data-Driven Learning and Robot Intelligence

Significant opportunities lie in overcoming data limitations and enhancing the intelligence and transparency of a robot's acoustic learning processes:

- **Large-scale, diverse, and shared datasets:** A critical opportunity is the creation and broad dissemination of large-scale, diverse, and meticulously annotated benchmark datasets specifically for robotic SSL. These datasets should include challenging real-world scenarios, diverse robot platforms, varying microphone array configurations

(including dynamic ones), and precise ground truth for moving sources and robots. Initiatives to encourage data sharing across research institutions are vital to foster collaborative progress and provide standardized benchmarks for model comparison.

- **Unsupervised, self-supervised, and reinforcement learning for SSL:** Exploring learning paradigms that minimize the reliance on labor-intensive labeled data. This includes techniques that allow robots to learn spatial acoustic features from vast amounts of unlabeled audio data gathered during operation, as well as reinforcement learning approaches where a robot can optimize its SSL performance through interaction with its environment [23,27].
- **Bridging the sim-to-real gap with domain adaptation:** Developing robust domain adaptation and transfer learning techniques to effectively transfer knowledge from simulated acoustic environments to real-world robotic deployment. This will involve more sophisticated acoustic simulations that accurately model complex reverberation and noise, coupled with techniques that make deep learning models more resilient to the inevitable discrepancies between simulated and real data [234].
- **Foundation models for semantic interpretation:** A promising avenue involves integrating large language models (LLMs) into the robot's auditory processing pipeline, particularly after successful sound source localization and speech recognition. Once an SSL system pinpoints the origin of human speech, and an automatic speech recognition (ASR) system transcribes it, LLMs can be employed to provide a deeper contextual understanding and semantic interpretation of verbal commands, queries, or intentions. This would enable robots to engage in more natural, nuanced, and multi-turn dialogues, disambiguate vague instructions based on conversational history or common sense, and ground abstract concepts in the robot's physical environment, moving beyond mere utterance processing to true linguistic comprehension.
- **Hybrid models and multi-modal fusion for holistic perception:** Future research will increasingly focus on hybrid models that intelligently combine deep learning with traditional signal processing techniques, and crucially, on fusing SSL outputs with other sensory modalities. This includes the tightly coupled integration of acoustic data with visual information from cameras, e.g., for audio-visual speaker tracking, identifying sounding objects, human facial expression in HRI [3,54], and haptic feedback (e.g., for contact localization on robot limbs). Such multi-modal fusion allows robots to build a more comprehensive and robust perception of their environment, compensating for limitations in any single modality and enabling a richer semantic understanding of the auditory scene.
- **Joining "hearing" and "touch" for safer human–robot teamwork:** Beyond vision, force/torque and impedance signals captured at a robot's joints offer information. Recent impedance-learning methods already let a robot adjust its stiffness and damping by feeling the force that a person applies [239]. A natural next step is to let the robot listen. If the robot knows where a voice command or warning sound is coming from, it can quickly relax its grip, steer away, or steady its tool in that direction. Achieving this will require (i) shared audio-and-force datasets collected while people guide a robot and talk to it, and (ii) simple learning rules that map those combined signals to clear impedance changes. Such "audio-haptic" control could make everyday tasks—like power-tool assistance, bedside help, or shared carrying— both easier for the user and safer for everyone nearby.
- **Explainable AI for robot audition:** Research into methods that allow deep learning-based SSL systems to provide transparent explanations for their localization decisions. This will enhance human operators' trust in autonomous robots and facilitate debugging and the refinement of auditory perception systems in safety-critical applications.

8. Conclusions

Sound source localization (SSL) is a pivotal capability for enhancing robot autonomy, facilitating seamless human–robot interaction, and enabling intelligent environmental awareness. This review has provided a comprehensive synthesis of SSL in robotics, with a particular focus on the transformative impact of deep learning methods over the past decade. We began by revisiting the fundamental principles of SSL and the traditional methods that formed its bedrock, such as time difference of arrival (TDOA), beamforming, steered-response power (SRP), and subspace analysis methods, highlighting their core mechanisms and inherent limitations. We then transitioned into a comprehensive exploration of deep learning architectures, from foundational machine learning, shallow neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) to sophisticated attention-based models. We examined how these data-driven approaches implicitly handle complex acoustic phenomena like reverberation and noise—challenges that traditionally limited classical methods. The critical roles of data and training strategies, including synthetic data generation and various learning paradigms, were also thoroughly discussed. Furthermore, we mapped the diverse applications of SSL in robotics across various robotic types (i.e., mobile ground robots, humanoids, UAVs) and domains, including social and companion, search and rescue, service, and industrial. This provided a practical overview of how SSL is being integrated into real-world robotic systems, ranging from speech command recognition to anomaly detection and situational awareness.

Despite significant advancements, particularly with deep learning, several key challenges remain. Achieving environmental robustness in highly reverberant, noisy, and dynamic conditions, especially in unstructured outdoor environments, continues to demand innovative solutions. The accurate localization and tracking of multiple, overlapping sound sources remain complex, requiring better source separation and identity maintenance capabilities. Practical implementation constraints such as computational efficiency, power consumption, and the inherent limitations of microphone arrays on robotic platforms necessitate the development of lightweight, energy-efficient algorithms and flexible hardware designs. Lastly, data and learning challenges, including the scarcity of diverse, labeled datasets, the "sim-to-real" gap, and the current lack of comprehensive benchmarks, hinder the generalization and real-world deployability of deep learning models. The black-box nature of many deep learning approaches also poses interpretability challenges, particularly for safety-critical robotic applications.

Looking forward, the future of robotic SSL is rich with promising opportunities. Developing adaptive and robust auditory systems that can autonomously cope with extreme acoustic conditions will be paramount. Advancements in multi-source auditory scene analysis, combining localization with intelligent source separation and tracking, will unlock new levels of robot perception in complex social and industrial settings. Crucially, the integration of foundation models, such as large language models, with SSL outputs presents a transformative avenue for robots to not only pinpoint sound sources but also semantically interpret human speech, enabling more natural and intelligent human–robot interaction. Similarly, hybrid models and multi-modal fusion with visual and haptic promise to yield a more holistic and robust understanding of the environment. Overcoming data limitations through unsupervised and self-supervised learning, coupled with the creation of large-scale, and shared benchmark datasets for robotic contexts, will accelerate progress. Finally, innovations in efficient hardware–software co-design and explainable AI will pave the way for deployable, trustworthy, and energy-efficient SSL solutions for the next generation of autonomous robots. By continuing to bridge the gap between acoustic signal processing, advanced deep learning, and robotic system integration, the field is poised to equip robots with auditory capabilities that rival, and in some aspects even surpass, human hearing in

specific operational contexts. This will enable robots to become more perceptive, interactive, and intelligent agents in our increasingly complex world.

Author Contributions: Conceptualization, R.J. and M.J.; methodology, R.J. and M.J.; software, R.J.; validation, R.J., M.J. and A.B.; investigation, R.J.; data curation, R.J.; writing—original draft preparation, R.J. and M.J.; writing—review and editing, R.J. and M.J.; visualization, R.J. and M.J.; supervision, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rascon, C.; Meza, I. Localization of sound sources in robotics: A review. *Robot. Auton. Syst.* **2017**, *96*, 184–210. [CrossRef]
2. Jo, H.M.; Kim, T.W.; Kwak, K.C. Sound Source Localization Using Deep Learning for Human–Robot Interaction Under Intelligent Robot Environments. *Electronics* **2025**, *14*, 1043. [CrossRef]
3. Korayem, M.; Azargoshasb, S.; Korayem, A.; Tabibian, S. Design and implementation of the voice command recognition and the sound source localization system for human–robot interaction. *Robotica* **2021**, *39*, 1779–1790. [CrossRef]
4. Jalayer, R.; Jalayer, M.; Orsenigo, C.; Vercellis, C. A Conceptual Framework for Localization of Active Sound Sources in Manufacturing Environment Based on Artificial Intelligence. In Proceedings of the 33rd International Conference on Flexible Automation and Intelligent Manufacturing (FAIM 2023), Porto, Portugal, 18–22 June 2023; pp. 699–707.
5. Lv, D.; Tang, W.; Feng, G.; Zhen, D.; Gu, F.; Ball, A.D. An Overview of Sound Source Localization based Condition Monitoring Robots. *ISA Trans.* **2024**, *158*, 537–555. [CrossRef] [PubMed]
6. Lv, D.; Feng, G.; Zhen, D.; Liang, X.; Sun, G.; Gu, F. Motor Bearing Fault Source Localization Based on Sound and Robot Movement Characteristics. In Proceedings of the 2024 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD), Huangshan, China, 31 October 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
7. Marques, I.; Sousa, J.; Sá, B.; Costa, D.; Sousa, P.; Pereira, S.; Santos, A.; Lima, C.; Hammerschmidt, N.; Pinto, S.; et al. Microphone array for speaker localization and identification in shared autonomous vehicles. *Electronics* **2022**, *11*, 766. [CrossRef]
8. Yamada, T.; Itoyama, K.; Nishida, K.; Nakadai, K. Sound source tracking by drones with microphone arrays. In Proceedings of the 2020 IEEE/SICE International Symposium on System Integration (SII), Honolulu, HI, USA, 12–15 January 2020; IEEE: New York, NY, USA, 2020; pp. 796–801.
9. Yamamoto, T.; Hoshiba, K.; Yen, B.; Nakadai, K. Implementation of a Robot Operation System-based network for sound source localization using multiple drones. In Proceedings of the 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Macau, China, 3–6 December 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
10. Latif, T.; Whitmire, E.; Novak, T.; Bozkurt, A. Sound localization sensors for search and rescue biobots. *IEEE Sens. J.* **2015**, *16*, 3444–3453. [CrossRef]
11. Zhang, B.; Masahide, K.; Lim, H. Sound source localization and interaction based human searching robot under disaster environment. In Proceedings of the 2019 SICE International Symposium on Control Systems (SICE ISCS), Kumamoto, Japan, 7–9 March 2019; IEEE: New York, NY, USA, 2019; pp. 16–20.
12. Mae, N.; Mitsui, Y.; Makino, S.; Kitamura, D.; Ono, N.; Yamada, T.; Saruwatari, H. Sound source localization using binaural difference for hose-shaped rescue robot. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; IEEE: New York, NY, USA, 2017; pp. 1621–1627.
13. Park, J.H.; Sim, K.B. A design of mobile robot based on Network Camera and sound source localization for intelligent surveillance system. In Proceedings of the 2008 International Conference on Control, Automation and Systems, Seoul, Republic of Korea, 14–17 October 2008; IEEE: New York, NY, USA, 2008; pp. 674–678.
14. Han, Z.; Li, T. Research on sound source localization and real-time facial expression recognition for security robot. In Proceedings of the Journal of Physics: Conference Series, Hangzhou, China, 25–26 July 2020; IOP Publishing: Bristol, UK, 2020; Volume 1621, p. 012045.
15. Obeidat, H.; Shuaieb, W.; Obeidat, O.; Abd-Alhameed, R. A review of indoor localization techniques and wireless technologies. *Wirel. Pers. Commun.* **2021**, *119*, 289–327. [CrossRef]
16. Tarokh, M.; Merloti, P. Vision-based robotic person following under light variations and difficult walking maneuvers. *J. Field Robot.* **2010**, *27*, 387–398. [CrossRef]
17. Hall, D.; Talbot, B.; Bista, S.R.; Zhang, H.; Smith, R.; Dayoub, F.; Sünderhauf, N. The robotic vision scene understanding challenge. *arXiv* **2020**, arXiv:2009.05246. [CrossRef]

18. Belkin, I.; Abramenko, A.; Yudin, D. Real-time lidar-based localization of mobile ground robot. *Procedia Comput. Sci.* **2021**, *186*, 440–448. [CrossRef]
19. Yu, Z. A WiFi Indoor Localization System Based on Robot Data Acquisition and Deep Learning Model. In Proceedings of the 2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI), Guangzhou, China, 26–28 July 2024; IEEE: New York, NY, USA, 2024; pp. 367–371.
20. Wahab, N.H.A.; Sunar, N.; Ariffin, S.H.; Wong, K.Y.; Aun, Y. Indoor positioning system: A review. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 477–490. [CrossRef]
21. Alfurati, I.S.; Rashid, A.T. Performance comparison of three types of sensor matrices for indoor multi-robot localization. *Int. J. Comput. Appl.* **2018**, *181*, 22–29. [CrossRef]
22. Flynn, A.M.; Brooks, R.A.; Wells, W.M., III; Barrett, D.S. *Squirt: The Prototypical Mobile Robot for Autonomous Graduate Students*; Massachusetts Institute of Technology: Cambridge, MA, USA, 1989 .
23. Grumiaux, P.A.; Kitić, S.; Girin, L.; Guérin, A. A survey of sound source localization with deep learning methods. *J. Acoust. Soc. Am.* **2022**, *152*, 107–151. [CrossRef] [PubMed]
24. Liaquat, M.U.; Munawar, H.S.; Rahman, A.; Qadir, Z.; Kouzani, A.Z.; Mahmud, M.P. Localization of sound sources: A systematic review. *Energies* **2021**, *14*, 3910. [CrossRef]
25. Desai, D.; Mehendale, N. A review on sound source localization systems. *Arch. Comput. Methods Eng.* **2022**, *29*, 4631–4642. [CrossRef]
26. Zhang, B.J.; Fitter, N.T. Nonverbal sound in human–robot interaction: A systematic review. *ACM Trans. Hum.-Robot. Interact.* **2023**, *12*, 1–46. [CrossRef]
27. Jekateryńczuk, G.; Piotrowski, Z. A survey of sound source localization and detection methods and their applications. *Sensors* **2023**, *24*, 68. [CrossRef]
28. Khan, A.; Waqar, A.; Kim, B.; Park, D. A Review on Recent Advances in Sound Source Localization Techniques, Challenges, and Applications. *Sens. Actuators Rep.* **2025**, *9*, 100313. [CrossRef]
29. He, W. Deep Learning Approaches for Auditory Perception in Robotics. Ph.D. Thesis, EPFL, Lausanne, Switzerland, 2021.
30. Youssef, K.; Argentieri, S.; Zarader, J.L. A learning-based approach to robust binaural sound localization. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; IEEE: New York, NY, USA, 2013; pp. 2927–2932.
31. Nakamura, K.; Gomez, R.; Nakadai, K. Real-time super-resolution three-dimensional sound source localization for robots. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; IEEE: New York, NY, USA, 2013; pp. 3949–3954.
32. Ohata, T.; Nakamura, K.; Mizumoto, T.; Taiki, T.; Nakadai, K. Improvement in outdoor sound source detection using a quadrotor-embedded microphone array. In Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; IEEE: New York, NY, USA, 2014; pp. 1902–1907.
33. Grondin, F.; Michaud, F. Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; IEEE: New York, NY, USA, 2015; pp. 6149–6154.
34. Nakamura, K.; Sinapayen, L.; Nakadai, K. Interactive sound source localization using robot audition for tablet devices. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; IEEE: New York, NY, USA, 2015; pp. 6137–6142.
35. Li, X.; Girin, L.; Badeig, F.; Horaud, R. Reverberant sound localization with a robot head based on direct-path relative transfer function. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Republic of Korea, 9–14 October 2016; IEEE: New York, NY, USA, 2016; pp. 2819–2826.
36. Nakadai, K.; Kumon, M.; Okuno, H.G.; Hoshiba, K.; Wakabayashi, M.; Washizaki, K.; Ishiki, T.; Gabriel, D.; Bando, Y.; Morito, T.; et al. Development of microphone-array-embedded UAV for search and rescue task. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; IEEE: New York, NY, USA, 2017; pp. 5985–5990.
37. Strauss, M.; Mordel, P.; Miguet, V.; Deleforge, A. DREGON: Dataset and methods for UAV-embedded sound source localization. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; IEEE: New York, NY, USA, 2018; pp. 1–8.
38. Wang, L.; Sanchez-Matilla, R.; Cavallaro, A. Audio-visual sensing from a quadcopter: Dataset and baselines for source localization and sound enhancement. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; IEEE: New York, NY, USA, 2019; pp. 5320–5325.

39. Michaud, S.; Faucher, S.; Grondin, F.; Lauzon, J.S.; Labbé, M.; Létourneau, D.; Ferland, F.; Michaud, F. 3D localization of a sound source using mobile microphone arrays referenced by SLAM. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; IEEE: New York, NY, USA, 2020; pp. 10402–10407.
40. Sewtz, M.; Bodenmüller, T.; Triebel, R. Robust MUSIC-based sound source localization in reverberant and echoic environments. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; IEEE: New York, NY, USA, 2020; pp. 2474–2480.
41. Tourbabin, V.; Barfuss, H.; Rafaely, B.; Kellermann, W. Enhanced robot audition by dynamic acoustic sensing in moving humanoids. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; IEEE: New York, NY, USA, 2015; pp. 5625–5629.
42. Takeda, R.; Komatani, K. Sound source localization based on deep neural networks with directional activate function exploiting phase information. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; IEEE: New York, NY, USA, 2016; pp. 405–409.
43. Takeda, R.; Komatani, K. Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: New York, NY, USA, 2017; pp. 2217–2221.
44. Ferguson, E.L.; Williams, S.B.; Jin, C.T. Sound source localization in a multipath environment using convolutional neural networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: New York, NY, USA, 2018; pp. 2386–2390.
45. Grondin, F.; Michaud, F. Noise mask for TDOA sound source localization of speech on mobile robots in noisy environments. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; IEEE: New York, NY, USA, 2016; pp. 4530–4535.
46. He, W.; Motlicek, P.; Odobez, J.M. Deep neural networks for multiple speaker detection and localization. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; IEEE: New York, NY, USA, 2018; pp. 74–79.
47. An, I.; Son, M.; Manocha, D.; Yoon, S.E. Reflection-aware sound source localization. In Proceedings of the 2018 IEEE international conference on robotics and automation (ICRA), Brisbane, Australia, 21–25 May 2018; IEEE: New York, NY, USA, 2018; pp. 66–73.
48. Song, L.; Wang, H.; Chen, P. Automatic patrol and inspection method for machinery diagnosis robot—Sound signal-based fuzzy search approach. *IEEE Sens. J.* **2020**, *20*, 8276–8286. [CrossRef]
49. Clayton, M.; Wang, L.; McPherson, A.; Cavallaro, A. An embedded multichannel sound acquisition system for drone audition. *IEEE Sens. J.* **2023**, *23*, 13377–13386. [CrossRef]
50. Grondin, F.; Michaud, F. Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. *Robot. Auton. Syst.* **2019**, *113*, 63–80. [CrossRef]
51. Go, Y.J.; Choi, J.S. An acoustic source localization method using a drone-mounted phased microphone array. *Drones* **2021**, *5*, 75. [CrossRef]
52. Yamada, T.; Itoyama, K.; Nishida, K.; Nakadai, K. Placement planning for sound source tracking in active drone audition. *Drones* **2023**, *7*, 405. [CrossRef]
53. Skoczylas, A.; Stefaniak, P.; Anufriiev, S.; Jachnik, B. Belt conveyors rollers diagnostics based on acoustic signal collected using autonomous legged inspection robot. *Appl. Sci.* **2021**, *11*, 2299. [CrossRef]
54. Shi, Z.; Zhang, L.; Wang, D. Audio–visual sound source localization and tracking based on mobile robot for the cocktail party problem. *Appl. Sci.* **2023**, *13*, 6056. [CrossRef]
55. Keyrouz, F. Advanced binaural sound localization in 3-D for humanoid robots. *IEEE Trans. Instrum. Meas.* **2014**, *63*, 2098–2107. [CrossRef]
56. Wang, Z.; Zou, W.; Su, H.; Guo, Y.; Li, D. Multiple sound source localization exploiting robot motion and approaching control. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 7505316. [CrossRef]
57. Tourbabin, V.; Rafaely, B. Theoretical framework for the optimization of microphone array configuration for humanoid robot audition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1803–1814. [CrossRef]
58. Manamperi, W.; Abhayapala, T.D.; Zhang, J.; Samarasinghe, P.N. Drone audition: Sound source localization using on-board microphones. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 508–519. [CrossRef]
59. Gonzalez-Billandon, J.; Belgiovine, G.; Tata, M.; Sciutti, A.; Sandini, G.; Rea, F. Self-supervised learning framework for speaker localisation with a humanoid robot. In Proceedings of the 2021 IEEE International Conference on Development and Learning (ICDL), Beijing, China, 23–26 August 2021; IEEE: New York, NY, USA, 2021; pp. 1–7.
60. Gamboa-Montero, J.J.; Basiri, M.; Castillo, J.C.; Marques-Villarroya, S.; Salichs, M.A. Real-Time Acoustic Touch Localization in Human-Robot Interaction based on Steered Response Power. In Proceedings of the 2022 IEEE International Conference on Development and Learning (ICDL), London, UK, 12–15 September 2022; IEEE: New York, NY, USA, 2022; pp. 101–106.

61. Yalta, N.; Nakadai, K.; Ogata, T. Sound source localization using deep learning models. *J. Robot. Mechatronics* **2017**, *29*, 37–48. [CrossRef]
62. Chen, L.; Chen, G.; Huang, L.; Choy, Y.S.; Sun, W. Multiple sound source localization, separation, and reconstruction by microphone array: A dnn-based approach. *Appl. Sci.* **2022**, *12*, 3428. [CrossRef]
63. Tian, C. Multiple CRNN for SELD. *Parameters* **2020**, *488211*, 490326.
64. Bohlender, A.; Spriet, A.; Tirry, W.; Madhu, N. Exploiting temporal context in CNN based multisource DOA estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1594–1608. [CrossRef]
65. Li, X.; Liu, H.; Yang, X. Sound source localization for mobile robot based on time difference feature and space grid matching. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 25–30 September 2011; IEEE: New York, NY, USA, 2011; pp. 2879–2886.
66. El Zooghby, A.H.; Christodoulou, C.G.; Georgiopoulos, M. A neural network-based smart antenna for multiple source tracking. *IEEE Trans. Antennas Propag.* **2000**, *48*, 768–776. [CrossRef]
67. Ishfaque, A.; Kim, B. Real-time sound source localization in robots using fly Ormia ochracea inspired MEMS directional microphone. *IEEE Sens. Lett.* **2022**, *7*, 6000204. [CrossRef]
68. Athanasopoulos, G.; Dekens, T.; Brouckxon, H.; Verhelst, W. The effect of speech denoising algorithms on sound source localization for humanoid robots. In Proceedings of the 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), Montreal, QC, Canada, 2–5 July 2012; IEEE: New York, NY, USA 2012; pp. 327–332.
69. Subramanian, A.S.; Weng, C.; Watanabe, S.; Yu, M.; Yu, D. Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition. *Comput. Speech Lang.* **2022**, *75*, 101360. [CrossRef]
70. Goli, P.; van de Par, S. Deep learning-based speech specific source localization by using binaural and monaural microphone arrays in hearing aids. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 1652–1666. [CrossRef]
71. Wu, S.; Zheng, Y.; Ye, K.; Cao, H.; Zhang, X.; Sun, H. Sound source localization for unmanned aerial vehicles in low signal-to-noise ratio environments. *Remote Sens.* **2024**, *16*, 1847. [CrossRef]
72. Jalayer, R.; Jalayer, M.; Mor, A.; Orsenigo, C.; Vercellis, C. ConvLSTM-based Sound Source Localization in a manufacturing workplace. *Comput. Ind. Eng.* **2024**, *192*, 110213. [CrossRef]
73. Risoud, M.; Hanson, J.N.; Gauvrit, F.; Renard, C.; Lemesre, P.E.; Bonne, N.X.; Vincent, C. Sound source localization. *Eur. Ann. Otorhinolaryngol. Head Neck Dis.* **2018**, *135*, 259–264. [CrossRef]
74. Hirvonen, T. Classification of spatial audio location and content using convolutional neural networks. In Proceedings of the Audio Engineering Society Convention 138, Warsaw, Poland, 7–10 May 2015; Audio Engineering Society: New York, NY, USA, 2015.
75. Xiao, X.; Zhao, S.; Zhong, X.; Jones, D.L.; Chng, E.S.; Li, H. A learning-based approach to direction of arrival estimation in noisy and reverberant environments. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; IEEE: New York, NY, USA, 2015; pp. 2814–2818.
76. Geng, Y.; Jung, J.; Seol, D. Sound-source localization system based on neural network for mobile robots. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; IEEE: New York, NY, USA 2008; pp. 3126–3130.
77. Liu, G.; Yuan, S.; Wu, J.; Zhang, R. A sound source localization method based on microphone array for mobile robot. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; IEEE: New York, NY, USA 2018; pp. 1621–1625.
78. Lee, S.Y.; Chang, J.; Lee, S. Deep learning-based method for multiple sound source localization with high resolution and accuracy. *Mech. Syst. Signal Process.* **2021**, *161*, 107959. [CrossRef]
79. Acosta, O.; Hermida, L.; Herrera, M.; Montenegro, C.; Gaona, E.; Bejarano, M.; Gordillo, K.; Pavón, I.; Asensio, C. Remote Binaural System (RBS) for Noise Acoustic Monitoring. *J. Sens. Actuator Netw.* **2023**, *12*, 63. [CrossRef]
80. Deleforge, A.; Forbes, F.; Horaud, R. Acoustic space learning for sound-source separation and localization on binaural manifolds. *Int. J. Neural Syst.* **2015**, *25*, 1440003. [CrossRef] [PubMed]
81. Gala, D.; Lindsay, N.; Sun, L. Realtime active sound source localization for unmanned ground robots using a self-rotational bi-microphone array. *J. Intell. Robot. Syst.* **2019**, *95*, 935–954. [CrossRef]
82. Baxendale, M.D.; Nibouche, M.; Secco, E.L.; Pipe, A.G.; Pearson, M.J. Feed-forward selection of cerebellar models for calibration of robot sound source localization. In Proceedings of the Conference on Biomimetic and Biohybrid Systems, Nara, Japan, 9–12 July 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 3–14.
83. Gala, D.; Sun, L. Moving sound source localization and tracking for an autonomous robot equipped with a self-rotating bi-microphone array. *J. Acoust. Soc. Am.* **2023**, *154*, 1261–1273. [CrossRef]
84. Mumolo, E.; Nolich, M.; Vercelli, G. Algorithms for acoustic localization based on microphone array in service robotics. *Robot. Auton. Syst.* **2003**, *42*, 69–88. [CrossRef]

85. Nguyen, Q.V.; Colas, F.; Vincent, E.; Charpillet, F. Long-term robot motion planning for active sound source localization with Monte Carlo tree search. In Proceedings of the 2017 Hands-free Speech Communications and Microphone Arrays (HSCMA), San Francisco, CA, USA, 1–3 March 2017; IEEE: New York, NY, USA, 2017; pp. 61–65.
86. Tamai, Y.; Kagami, S.; Amemiya, Y.; Sasaki, Y.; Mizoguchi, H.; Takano, T. Circular microphone array for robot's audition. In Proceedings of the SENSORS, 2004 IEEE, Vienna, Austria, 24–27 October 2004; IEEE: New York, NY, USA, 2004; pp. 565–570.
87. Choi, C.; Kong, D.; Kim, J.; Bang, S. Speech enhancement and recognition using circular microphone array for service robots. In Proceedings of the Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No. 03CH37453), Las Vegas, NV, USA, 27–31 October 2003; IEEE: New York, NY, USA, 2003; Volume 4, pp. 3516–3521.
88. Sasaki, Y.; Kabasawa, M.; Thompson, S.; Kagami, S.; Oro, K. Spherical microphone array for spatial sound localization for a mobile robot. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal, 7–12 October 2012; IEEE: New York, NY, USA, 2012; pp. 713–718.
89. Jin, L.; Yan, J.; Du, X.; Xiao, X.; Fu, D. RNN for solving time-variant generalized Sylvester equation with applications to robots and acoustic source localization. *IEEE Trans. Ind. Inform.* **2020**, *16*, 6359–6369. [CrossRef]
90. Bando, Y.; Mizumoto, T.; Itoyama, K.; Nakadai, K.; Okuno, H.G. Posture estimation of hose-shaped robot using microphone array localization. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; IEEE: New York, NY, USA, 2013; pp. 3446–3451.
91. Kim, U.H. Improvement of Sound Source Localization for a Binaural Robot of Spherical Head with Pinnae. Ph.D. Thesis, Kyoto University, Kyoto, Japan, 2013.
92. Kumon, M.; Noda, Y. Active soft pinnae for robots. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 25–30 September 2011; IEEE: New York, NY, USA, 2011; pp. 112–117.
93. Murray, J.C.; Erwin, H.R. A neural network classifier for notch filter classification of sound-source elevation in a mobile robot. In Proceedings of the The 2011 International Joint Conference on Neural Networks, San Jose, CA, USA, 16 January 2012; IEEE: New York, NY, USA, 2011; pp. 763–769.
94. Zhang, Y.; Weng, J. Grounded auditory development by a developmental robot. In Proceedings of the IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222), Washington, DC, USA, 15–19 July 2001; IEEE: New York, NY, USA, 2001; Volume 2, pp. 1059–1064.
95. Xu, B.; Sun, G.; Yu, R.; Yang, Z. High-accuracy TDOA-based localization without time synchronization. *IEEE Trans. Parallel Distrib. Syst.* **2012**, *24*, 1567–1576. [CrossRef]
96. Knapp, C.; Carter, G. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.* **2003**, *24*, 320–327. [CrossRef]
97. Park, B.C.; Ban, K.D.; Kwak, K.C.; Yoon, H.S. Performance analysis of GCC-PHAT-based sound source localization for intelligent robots. *J. Korea Robot. Soc.* **2007**, *2*, 270–274.
98. Wang, J.; Qian, X.; Pan, Z.; Zhang, M.; Li, H. GCC-PHAT with speech-oriented attention for robotic sound source localization. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; IEEE: New York, NY, USA, 2021; pp. 5876–5883.
99. Lombard, A.; Zheng, Y.; Buchner, H.; Kellermann, W. TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 1490–1503. [CrossRef]
100. Huang, B.; Xie, L.; Yang, Z. TDOA-based source localization with distance-dependent noises. *IEEE Trans. Wirel. Commun.* **2014**, *14*, 468–480. [CrossRef]
101. Scheuing, J.; Yang, B. Disambiguation of TDOA estimation for multiple sources in reverberant environments. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 1479–1489. [CrossRef]
102. Kim, U.H.; Nakadai, K.; Okuno, H.G. Improved sound source localization and front-back disambiguation for humanoid robots with two ears. In Proceedings of the Recent Trends in Applied Artificial Intelligence: 26th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2013, Amsterdam, The Netherlands, 17–21 June 2013; Proceedings 26; Springer: Berlin/Heidelberg, Germany, 2013; pp. 282–291.
103. Kim, U.H.; Nakadai, K.; Okuno, H.G. Improved sound source localization in horizontal plane for binaural robot audition. *Appl. Intell.* **2015**, *42*, 63–74. [CrossRef]
104. Chen, G.; Xu, Y. A sound source localization device based on rectangular pyramid structure for mobile robot. *J. Sens.* **2019**, *2019*, 4639850. [CrossRef]
105. Chen, H.; Liu, C.; Chen, Q. Efficient and robust approaches for three-dimensional sound source recognition and localization using humanoid robots sensor arrays. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420941357. [CrossRef]
106. Xu, Q.; Yang, P. Sound Source Localization Strategy Based on Mobile Robot. In Proceedings of the 2013 Chinese Intelligent Automation Conference: Intelligent Automation & Intelligent Technology and Systems, Yangzhou, China, 26–28 August 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 469–478.

107. Alameda-Pineda, X.; Horaud, R. A geometric approach to sound source localization from time-delay estimates. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1082–1095. [CrossRef]
108. Valin, J.M.; Michaud, F.; Rouat, J. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robot. Auton. Syst.* **2007**, *55*, 216–228. [CrossRef]
109. Luzanto, A.; Bohmer, N.; Mahu, R.; Alvarado, E.; Stern, R.M.; Becerra Yoma, N. Effective Acoustic Model-Based Beamforming Training for Static and Dynamic Hri Applications. *Sensors* **2024**, *24*, 6644. [CrossRef]
110. Kagami, S.; Thompson, S.; Sasaki, Y.; Mizoguchi, H.; Enomoto, T. 2D sound source mapping from mobile robot using beamforming and particle filtering. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; IEEE: New York, NY, USA, 2009; pp. 3689–3692.
111. Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* **2005**, *57*, 1408–1418. [CrossRef]
112. Dmochowski, J.; Benesty, J.; Affes, S. Linearly constrained minimum variance source localization and spectral estimation. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 1490–1502. [CrossRef]
113. Yang, C.; Wang, Y.; Wang, Y.; Hu, D.; Guo, H. An improved functional beamforming algorithm for far-field multi-sound source localization based on Hilbert curve. *Appl. Acoust.* **2022**, *192*, 108729. [CrossRef]
114. Liu, M.; Qu, S.; Zhao, X. Minimum Variance Distortionless Response—Hanbury Brown and Twiss Sound Source Localization. *Appl. Sci.* **2023**, *13*, 6013. [CrossRef]
115. Zhang, C.; Wang, R.; Yu, L.; Xiao, Y.; Guo, Q.; Ji, H. Localization of cyclostationary acoustic sources via cyclostationary beamforming and its high spatial resolution implementation. *Mech. Syst. Signal Process.* **2023**, *204*, 110718. [CrossRef]
116. Faraji, M.M.; Shouraki, S.B.; Iranmehr, E.; Linares-Barranco, B. Sound source localization in wide-range outdoor environment using distributed sensor network. *IEEE Sens. J.* **2019**, *20*, 2234–2246. [CrossRef]
117. DiBiase, J.H.; Silverman, H.F.; Brandstein, M.S. Robust localization in reverberant rooms. In *Microphone Arrays: Signal Processing Techniques and Applications*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 157–180.
118. Yook, D.; Lee, T.; Cho, Y. Fast sound source localization using two-level search space clustering. *IEEE Trans. Cybern.* **2015**, *46*, 20–26. [CrossRef]
119. Ishi, C.T.; Chatot, O.; Ishiguro, H.; Hagita, N. Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 10–15 October 2009; IEEE: New York, NY, USA, 2009; pp. 2027–2032.
120. Zhang, X.; Feng, D. An efficient MUSIC algorithm enhanced by iteratively estimating signal subspace and its applications in spatial colored noise. *Remote Sens.* **2022**, *14*, 4260. [CrossRef]
121. Weng, L.; Song, X.; Liu, Z.; Liu, X.; Zhou, H.; Qiu, H.; Wang, M. DOA estimation of indoor sound sources based on spherical harmonic domain beam-space MUSIC. *Symmetry* **2023**, *15*, 187. [CrossRef]
122. Suzuki, R.; Takahashi, T.; Okuno, H.G. Development of a robotic pet using sound source localization with the hark robot audition system. *J. Robot. Mechatronics* **2017**, *29*, 146–153. [CrossRef]
123. Chen, L.; Sun, W.; Huang, L.; Yu, L. Broadband sound source localisation via non-synchronous measurements for service robots: A tensor completion approach. *IEEE Robot. Autom. Lett.* **2022**, *7*, 12193–12200. [CrossRef]
124. Chen, L.; Huang, L.; Chen, G.; Sun, W. A large scale 3d sound source localisation approach achieved via small size microphone array for service robots. In Proceedings of the 2022 5th International Conference on Information Communication and Signal Processing (ICICSP), Shenzhen, China, 26–28 November 2022; IEEE: New York, NY, USA, 2022; pp. 589–594.
125. Hoshiba, K.; Washizaki, K.; Wakabayashi, M.; Ishiki, T.; Kumon, M.; Bando, Y.; Gabriel, D.; Nakadai, K.; Okuno, H.G. Design of UAV-embedded microphone array system for sound source localization in outdoor environments. *Sensors* **2017**, *17*, 2535. [CrossRef]
126. Azrad, S.; Salman, A.; Al-Haddad, S.A.R. Performance of DOA Estimation Algorithms for Acoustic Localization of Indoor Flying Drones Using Artificial Sound Source. *J. Aeronaut. Astronaut. Aviat.* **2024**, *56*, 469–476.
127. Nakamura, K.; Nakadai, K.; Asano, F.; Hasegawa, Y.; Tsujino, H. Intelligent sound source localization for dynamic environments. In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 10–15 October 2009; IEEE: New York, NY, USA, 2009; pp. 664–669.
128. Narang, G.; Nakamura, K.; Nakadai, K. Auditory-aware navigation for mobile robots based on reflection-robust sound source localization and visual slam. In Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA, 5–8 October 2014; IEEE: New York, NY, USA, 2014; pp. 4021–4026.
129. Asano, F.; Morisawa, M.; Kaneko, K.; Yokoi, K. Sound source localization using a single-point stereo microphone for robots. In Proceedings of the 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Hong Kong, China, 16–19 December 2015; IEEE: New York, NY, USA, 2015, pp. 76–85.
130. Tran, H.D.; Li, H. Sound event recognition with probabilistic distance SVMs. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 1556–1568. [CrossRef]

131. Wang, J.C.; Wang, J.F.; He, K.W.; Hsu, C.S. Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptor. In Proceedings of the 2006 IEEE International Joint Conference on Neural Network Proceedings, Vancouver, BC, Canada, 16–21 July 2006; IEEE: New York, NY, USA, 2006; pp. 1731–1735.
132. Yussif, A.M.; Sadeghi, H.; Zayed, T. Application of machine learning for leak localization in water supply networks. *Buildings* **2023**, *13*, 849. [CrossRef]
133. Chen, H.; Ser, W. Sound source DOA estimation and localization in noisy reverberant environments using least-squares support vector machines. *J. Signal Process. Syst.* **2011**, *63*, 287–300. [CrossRef]
134. Salvati, D.; Drioli, C.; Foresti, G.L. A weighted MVDR beamformer based on SVM learning for sound source localization. *Pattern Recognit. Lett.* **2016**, *84*, 15–21. [CrossRef]
135. Salvati, D.; Drioli, C.; Foresti, G.L. On the use of machine learning in microphone array beamforming for far-field sound source localization. In Proceedings of the 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), Vietri sul Mare, Italy, 13–16 September 2016; IEEE: New York, NY, USA, 2016; pp. 1–6.
136. Gadre, C.M.; Patole, R.K.; Metkar, S.P. Comparative analysis of KNN and CNN for Localization of Single Sound Source. In Proceedings of the 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), Bengaluru, India, 1–2 September 2023; IEEE: New York, NY, USA, 2023; pp. 1–6.
137. Nando, P.; Putrada, A.G.; Abdurohman, M. Increasing the precision of noise source detection system using KNN method. *Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control* **2019**, *4*, 157–168. [CrossRef]
138. Sun, Y.; Chen, J.; Yuen, C.; Rahardja, S. Indoor sound source localization with probabilistic neural network. *IEEE Trans. Ind. Electron.* **2017**, *65*, 6403–6413. [CrossRef]
139. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
140. Fu, T.; Zhang, Z.; Liu, Y.; Leng, J. Development of an artificial neural network for source localization using a fiber optic acoustic emission sensor array. *Struct. Health Monit.* **2015**, *14*, 168–177. [CrossRef]
141. Jin, C.; Schenkel, M.; Carlile, S. Neural system identification model of human sound localization. *J. Acoust. Soc. Am.* **2000**, *108*, 1215–1235. [CrossRef] [PubMed]
142. Pu, C.J.; Harris, J.G.; Principe, J.C. A neuromorphic microphone for sound localization. In Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics, Computational Cybernetics and Simulation, Orlando, FL, USA, 12–15 October 1997; IEEE: New York, NY, USA, 1997; Volume 2, pp. 1469–1474.
143. Kim, Y.; Ling, H. Direction of arrival estimation of humans with a small sensor array using an artificial neural network. *Prog. Electromagn. Res. B* **2011**, *27*, 127–149. [CrossRef]
144. Davila-Chacon, J.; Twiefel, J.; Liu, J.; Wermter, S. Improving Humanoid Robot Speech Recognition with Sound Source Localisation. In Proceedings of the Artificial Neural Networks and Machine Learning–ICANN 2014: 24th International Conference on Artificial Neural Networks, Hamburg, Germany, 15–19 September 2014; Proceedings 24; Springer: Berlin/Heidelberg, Germany, 2014; pp. 619–626.
145. Dávila-Chacón, J.; Liu, J.; Wermter, S. Enhanced robot speech recognition using biomimetic binaural sound source localization. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 138–150. [CrossRef] [PubMed]
146. Takeda, R.; Komatani, K. Discriminative multiple sound source localization based on deep neural networks using independent location model. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13–16 December 2016; IEEE: New York, NY, USA, 2016; pp. 603–609.
147. Bianco, M.J.; Gerstoft, P.; Traer, J.; Ozanich, E.; Roch, M.A.; Gannot, S.; Deledalle, C.A. Machine learning in acoustics: Theory and applications. *J. Acoust. Soc. Am.* **2019**, *146*, 3590–3628. [CrossRef]
148. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
149. Vera-Diaz, J.M.; Pizarro, D.; Macias-Guarasa, J. Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates. *Sensors* **2018**, *18*, 3418. [CrossRef]
150. Suvorov, D.; Dong, G.; Zhukov, R. Deep residual network for sound source localization in the time domain. *arXiv* **2018**, arXiv:1808.06429. [CrossRef]
151. Huang, D.; Perez, R.F. Sseldnet: A fully end-to-end sample-level framework for sound event localization and detection. In Proceedings of the DCASE, Online, 15–19 November 2021.
152. Vincent, E.; Virtanen, T.; Gannot, S. *Audio Source Separation and Speech Enhancement*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
153. Chakrabarty, S.; Habets, E.A. Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 8–21. [CrossRef]
154. Wu, Y.; Ayyalasomayajula, R.; Bianco, M.J.; Bharadia, D.; Gerstoft, P. Sound source localization based on multi-task learning and image translation network. *J. Acoust. Soc. Am.* **2021**, *150*, 3374–3386. [CrossRef]
155. Butt, S.S.; Fatima, M.; Asghar, A.; Muhammad, W. Active Binaural Auditory Perceptual System for a Socially Interactive Humanoid Robot. *Eng. Proc.* **2022**, *12*, 83.

156. Krause, D.; Politis, A.; Kowalczyk, K. Comparison of convolution types in CNN-based feature extraction for sound source localization. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 18–21 January 2021; IEEE: New York, NY, USA, 2021; pp. 820–824.
157. Diaz-Guerra, D.; Miguel, A.; Beltran, J.R. Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *29*, 300–311. [CrossRef]
158. Bologni, G.; Heusdens, R.; Martinez, J. Acoustic reflectors localization from stereo recordings using neural networks. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: New York, NY, USA, 2021; pp. 1–5.
159. Nguyen, Q.; Girin, L.; Bailly, G.; Elisei, F.; Nguyen, D.C. Autonomous sensorimotor learning for sound source localization by a humanoid robot. In Proceedings of the IROS 2018-Workshop on Crossmodal Learning for Intelligent Robotics in Conjunction with IEEE/RSJ IROS, Madrid, Spain, 1–5 October 2018.
160. Boztas, G. Sound source localization for auditory perception of a humanoid robot using deep neural networks. *Neural Comput. Appl.* **2023**, *35*, 6801–6811. [CrossRef]
161. Pang, C.; Liu, H.; Li, X. Multitask learning of time-frequency CNN for sound source localization. *IEEE Access* **2019**, *7*, 40725–40737. [CrossRef]
162. Ko, J.; Kim, H.; Kim, J. Real-time sound source localization for low-power IoT devices based on multi-stream CNN. *Sensors* **2022**, *22*, 4650. [CrossRef] [PubMed]
163. Mjaid, A.Y.; Prasad, V.; Jonker, M.; Van Der Horst, C.; De Groot, L.; Narayana, S. Ai-based simultaneous audio localization and communication for robots. In Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation, San Antonio, TX, USA, 9–12 May 2023; pp. 172–183.
164. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078. [CrossRef]
165. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
166. Nguyen, T.N.T.; Nguyen, N.K.; Phan, H.; Pham, L.; Ooi, K.; Jones, D.L.; Gan, W.S. A general network architecture for sound event localization and detection using transfer learning and recurrent neural network. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: New York, NY, USA, 2021; pp. 935–939.
167. Wang, Z.Q.; Zhang, X.; Wang, D. Robust speaker localization guided by deep learning-based time-frequency masking. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *27*, 178–188. [CrossRef]
168. Andra, M.B.; Usagawa, T. Portable keyword spotting and sound source detection system design on mobile robot with mini microphone array. In Proceedings of the 2020 6th International Conference on Control, Automation and Robotics (ICCAR), Singapore, 20–23 April 2020; IEEE: New York, NY, USA, 2020; pp. 170–174.
169. Adavanne, S.; Politis, A.; Nikunen, J.; Virtanen, T. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE J. Sel. Top. Signal Process.* **2018**, *13*, 34–48. [CrossRef]
170. Lu, Z. Sound event detection and localization based on CNN and LSTM. Technical Report. In Proceedings of the Detection Classification Acoustic Scenes Events Challenge, New York, NY, USA, 25–26 October 2019.
171. Perotin, L.; Serizel, R.; Vincent, E.; Guérin, A. CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 22–33. [CrossRef]
172. Grumiaux, P.A.; Kitić, S.; Girin, L.; Guérin, A. Improved feature extraction for CRNN-based multiple sound source localization. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 23–27 August 2021; IEEE: New York, NY, USA, 2021; pp. 231–235.
173. Kim, J.H.; Choi, J.; Son, J.; Kim, G.S.; Park, J.; Chang, J.H. MIMO noise suppression preserving spatial cues for sound source localization in mobile robot. In Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Republic of Korea, 22–28 May 2021; IEEE: New York, NY, USA, 2021; pp. 1–5.
174. Han, C.; Luo, Y.; Mesgarani, N. Real-time binaural speech separation with preserved spatial cues. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: New York, NY, USA, 2020; pp. 6404–6408.
175. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
176. Mack, W.; Bharadwaj, U.; Chakrabarty, S.; Habets, E.A. Signal-aware broadband DOA estimation using attention mechanisms. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: New York, NY, USA, 2020; pp. 4930–4934.
177. Altayeva, A.; Omarov, N.; Tileubay, S.; Zhaksylyk, A.; Bazhikov, K.; Kambarov, D. Convolutional LSTM Network for Real-Time Impulsive Sound Detection and Classification in Urban Environments. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*. [CrossRef]

178. Akter, R.; Islam, M.R.; Debnath, S.K.; Sarker, P.K.; Uddin, M.K. A hybrid CNN-LSTM model for environmental sound classification: Leveraging feature engineering and transfer learning. *Digit. Signal Process.* **2025**, *163*, 105234. [CrossRef]
179. Varnita, L.S.S.; Subramanyam, K.; Ananya, M.; Mathilakath, P.; Krishnan, M.; Tiwari, S.; Shankarappa, R.T. Precision in Audio: CNN+ LSTM-Based 3D Sound Event Localization and Detection in Real-world Environments. In Proceedings of the 2024 2nd International Conference on Networking and Communications (ICNWC), Chennai, India, 2–4 April 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
180. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
181. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems, Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): La Jolla, CA, USA, 2017; Volume 30.
182. Phan, H.; Pham, L.; Koch, P.; Duong, N.Q.; McLoughlin, I.; Mertins, A. On multitask loss function for audio event detection and localization. *arXiv* **2020**, arXiv:2009.05527. [CrossRef]
183. Schymura, C.; Ochiai, T.; Delcroix, M.; Kinoshita, K.; Nakatani, T.; Araki, S.; Kolossa, D. Exploiting attention-based sequence-to-sequence architectures for sound event localization. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 18–21 January 2021; IEEE: New York, NY, USA, 2021; pp. 231–235.
184. Emmanuel, P.; Parrish, N.; Horton, M. Multi-scale network for sound event localization and detection. In Proceedings of the Technol Report of DCASE Challenge, Online, 15–19 November 2021.
185. Yalta, N.; Sumiyoshi, Y.; Kawaguchi, Y. The Hitachi DCASE 2021 Task 3 system: Handling directive interference with self attention layers. Technical Report. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021, Online, 15–19 November 2021; Volume 29.
186. Zhang, G.; Geng, L.; Xie, F.; He, C.D. A dynamic convolution-transformer neural network for multiple sound source localization based on functional beamforming. *Mech. Syst. Signal Process.* **2024**, *211*, 111272. [CrossRef]
187. Zhang, R.; Shen, X. A Novel Sound Source Localization Method Using Transformer. In Proceedings of the 2024 9th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 21–23 November 2024; IEEE: New York, NY, USA, 2024; Volume 9, pp. 361–366.
188. Chen, X.; Zhao, L.; Cui, J.; Li, H.; Wang, X. Hybrid Convolutional Neural Network-Transformer Model for End-to-End Binaural Sound Source Localization in Reverberant Environments. *IEEE Access* **2025**, *13*, 36701–36713. [CrossRef]
189. Zhang, D.; Chen, J.; Bai, J.; Wang, M.; Ayub, M.S.; Yan, Q.; Shi, D.; Gan, W.S. Multiple sound sources localization using sub-band spatial features and attention mechanism. *Circuits Syst. Signal Process.* **2025**, *44*, 2592–2620. [CrossRef]
190. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *AI Open* **2022**, *3*, 111–132. [CrossRef]
191. Mu, H.; Xia, W.; Che, W. Improving domain generalization for sound classification with sparse frequency-regularized transformer. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 10–14 July 2023; IEEE: New York, NY, USA, 2023; pp. 1104–1108.
192. Liu, Z.; Wang, Y.; Han, K.; Zhang, W.; Ma, S.; Gao, W. Post-training quantization for vision transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 28092–28103.
193. Allen, J.B.; Berkley, D.A. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **1979**, *65*, 943–950. [CrossRef]
194. Diaz-Guerra, D.; Miguel, A.; Beltran, J.R. gpuRIR: A python library for room impulse response simulation with GPU acceleration. *Multimed. Tools Appl.* **2021**, *80*, 5653–5671. [CrossRef]
195. Habets, E.A. *Room Impulse Response Generator*; Tech. Rep.; Technische Universiteit Eindhoven: Eindhoven, The Netherlands, 2006; Volume 2, p. 1.
196. Scheibler, R.; Bezzam, E.; Dokmanić, I. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: New York, NY, USA, 2018; pp. 351–355.
197. Campbell, D.; Palomaki, K.; Brown, G. A Matlab simulation of "shoebox" room acoustics for use in research and teaching. *Comput. Inf. Syst.* **2005**, *9*, 48.
198. Schimmel, S.M.; Muller, M.F.; Dillier, N. A fast and accurate "shoebox" room acoustics simulator. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; IEEE: New York, NY, USA, 2009; pp. 241–244.
199. Varanasi, V.; Gupta, H.; Hegde, R.M. A deep learning framework for robust DOA estimation using spherical harmonic decomposition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1248–1259. [CrossRef]

200. Gaultier, C.; Kataria, S.; Deleforge, A. VAST: The virtual acoustic space traveler dataset. In Proceedings of the Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, 21–23 February 2017; Proceedings 13; Springer: Berlin/Heidelberg, Germany, 2017; pp. 68–79.
201. Cheng, R.; Bao, C.; Cui, Z. Mass: Microphone array speech simulator in room acoustic environment for multi-channel speech coding and enhancement. *Appl. Sci.* **2020**, *10*, 1484. [CrossRef]
202. Krause, D.; Politis, A.; Kowalczyk, K. Data diversity for improving DNN-based localization of concurrent sound events. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; IEEE: New York, NY, USA, 2021; pp. 236–240.
203. Evers, C.; Moore, A.H.; Naylor, P.A. Localization of moving microphone arrays from moving sound sources for robot audition. In Proceedings of the 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 28 August–2 September 2016; IEEE: New York, NY, USA, 2016; pp. 1008–1012.
204. Mesaros, A.; Serizel, R.; Heittola, T.; Virtanen, T.; Plumbley, M.D. A decade of DCASE: Achievements, practices, evaluations and future challenges. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: New York, NY, USA, 2025; pp. 1–5.
205. Adavanne, S.; Politis, A.; Virtanen, T. A multi-room reverberant dataset for sound event localization and detection. *arXiv* **2019**, arXiv:1905.08546. [CrossRef]
206. Politis, A.; Mesaros, A.; Adavanne, S.; Heittola, T.; Virtanen, T. Overview and evaluation of sound event localization and detection in DCASE 2019. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *29*, 684–698. [CrossRef]
207. Politis, A.; Adavanne, S.; Krause, D.; Deleforge, A.; Srivastava, P.; Virtanen, T. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection. *arXiv* **2021**, arXiv:2106.06999. [CrossRef]
208. Politis, A.; Shimada, K.; Sudarsanam, P.; Adavanne, S.; Krause, D.; Koyama, Y.; Takahashi, N.; Takahashi, S.; Mitsufuji, Y.; Virtanen, T. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *arXiv* **2022**, arXiv:2206.01948. [CrossRef]
209. Shimada, K.; Politis, A.; Sudarsanam, P.; Krause, D.A.; Uchida, K.; Adavanne, S.; Hakala, A.; Koyama, Y.; Takahashi, N.; Takahashi, S.; et al. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 72931–72957.
210. Krause, D.A.; Politis, A.; Mesaros, A. Sound event detection and localization with distance estimation. In Proceedings of the 2024 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 26–30 August 2024; IEEE: New York, NY, USA, 2024; pp. 286–290.
211. Yeow, J.W.; Tan, E.L.; Bai, J.; Peksi, S.; Gan, W.S. Squeeze-and-Excite ResNet-Conformers for Sound Event Localization, Detection, and Distance Estimation for DCASE 2024 Challenge. *arXiv* **2024**, arXiv:2407.09021.
212. Shimada, K.; Takahashi, N.; Takahashi, S.; Mitsufuji, Y. Sound event localization and detection using activity-coupled cartesian DOA vector and RD3Net. *arXiv* **2020**, arXiv:2006.12014. [CrossRef]
213. Evers, C.; Löllmann, H.W.; Mellmann, H.; Schmidt, A.; Barfuss, H.; Naylor, P.A.; Kellermann, W. The LOCATA challenge: Acoustic source localization and tracking. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1620–1643. [CrossRef]
214. Jekateryńczuk, G.; Szadkowski, R.; Piotrowski, Z. UaVirBASE: A Public-Access Unmanned Aerial Vehicle Sound Source Localization Dataset. *Appl. Sci.* **2025**, *15*, 5378. [CrossRef]
215. Zhang, J.; Ding, W.; He, L. Data augmentation and prior knowledge-based regularization for sound event localization and detection. In Proceedings of the DCASE 2019 Detection and Classification of Acoustic Scenes and Events 2019 Challenge, New York, NY, USA, 25–26 October 2019.
216. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779. [CrossRef]
217. Mazzon, L.; Koizumi, Y.; Yasuda, M.; Harada, N. First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation. *arXiv* **2019**, arXiv:1910.04388. [CrossRef]
218. Pratik, P.; Jee, W.J.; Nagisetty, S.; Mars, R.; Lim, C. Sound event localization and detection using CRNN architecture with Mixup for model generalization. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), New York, NY, USA, 25–26 October 2019; pp. 199–203.
219. Noh, K.; Jeong-Hwan, C.; Dongyeop, J.; Joon-Hyuk, C. Three-stage approach for sound event localization and detection. Technical Report. In Proceedings of the Detection Classification Acoustic Scenes Events Challenge, New York, NY, USA, 25–26 October 2019.
220. Wang, Q.; Du, J.; Wu, H.X.; Pan, J.; Ma, F.; Lee, C.H. A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 1251–1264. [CrossRef]
221. Takahashi, N.; Gygli, M.; Pfister, B.; Van Gool, L. Deep convolutional neural networks and data augmentation for acoustic event detection. *arXiv* **2016**, arXiv:1604.07160. [CrossRef]

222. He, W.; Motlicek, P.; Odobez, J.M. Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1303–1317. [CrossRef]
223. Jenrungrot, T.; Jayaram, V.; Seitz, S.; Kemelmacher-Shlizerman, I. The cone of silence: Speech separation by localization. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 20925–20938.
224. Hu, Y.; Samarasinghe, P.N.; Gannot, S.; Abhayapala, T.D. Semi-supervised multiple source localization using relative harmonic coefficients under noisy and reverberant environments. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 3108–3123. [CrossRef]
225. Takeda, R.; Kudo, Y.; Takashima, K.; Kitamura, Y.; Komatani, K. Unsupervised adaptation of neural networks for discriminative sound source localization with eliminative constraint. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: New York, NY, USA, 2018; pp. 3514–3518.
226. Bianco, M.J.; Gannot, S.; Gerstoft, P. Semi-supervised source localization with deep generative modeling. In Proceedings of the 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP), Espoo, Finland, 21–24 September 2020; IEEE: New York, NY, USA, 2020; pp. 1–6.
227. Le Moing, G.; Vinayavekhin, P.; Agravante, D.J.; Inoue, T.; Vongkulbhisal, J.; Munawar, A.; Tachibana, R. Data-efficient framework for real-world multiple sound source 2D localization. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: New York, NY, USA, 2021; pp. 3425–3429.
228. Ren, Z.; Wang, S.; Zhang, Y. Weakly supervised machine learning. *CAAI Trans. Intell. Technol.* **2023**, *8*, 549–580. [CrossRef]
229. He, W.; Motlicek, P.; Odobez, J.M. Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: New York, NY, USA, 2019; pp. 770–774.
230. Opochinsky, R.; Laufer-Goldshtein, B.; Gannot, S.; Chechik, G. Deep ranking-based sound source localization. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; IEEE: New York, NY, USA, 2019; pp. 283–287.
231. Niu, S.; Liu, Y.; Wang, J.; Song, H. A decade survey of transfer learning (2010–2020). *IEEE Trans. Artif. Intell.* **2021**, *1*, 151–166. [CrossRef]
232. Park, S.; Jeong, Y.; Lee, T. Many-to-Many Audio Spectrogram Tansformer: Transformer for Sound Event Localization and Detection. In Proceedings of the DCASE, Online, 15–19 November 2021; pp. 105–109.
233. Gemmeke, J.F.; Ellis, D.P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: New York, NY, USA, 2017; pp. 776–780.
234. Zhang, M.; Yu, S.; Hu, Z.; Xia, K.; Wang, J.; Zhu, H. Sound source localization with sparse Bayesian-based feature matching via deep transfer learning in shallow sea. *Measurement* **2025**, *253*, 117873. [CrossRef]
235. Yu, Y.; Horoshenkov, K.V.; Sailor, G.; Tait, S. Sparse representation for artefact/defect localization with an acoustic array on a mobile pipe inspection robot. *Appl. Acoust.* **2025**, *231*, 110545. [CrossRef]
236. Kothig, A.; Ilievski, M.; Grasse, L.; Rea, F.; Tata, M. A bayesian system for noise-robust binaural sound localisation for humanoid robots. In Proceedings of the 2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE), Ottawa, ON, Canada, 17–18 June 2019; IEEE: New York, NY, USA, 2019; pp. 1–7.
237. Shiri, H.; Wodecki, J.; Ziętek, B.; Zimroz, R. Inspection robotic UGV platform and the procedure for an acoustic signal-based fault detection in belt conveyor idler. *Energies* **2021**, *14*, 7646. [CrossRef]
238. Yang, D.; Zhao, J. Acoustic wake-up technology for microsystems: A review. *Micromachines* **2023**, *14*, 129. [CrossRef]
239. Xing, X.; Burdet, E.; Si, W.; Yang, C.; Li, Y. Impedance learning for human-guided robots in contact with unknown environments. *IEEE Trans. Robot.* **2023**, *39*, 3705–3721. [CrossRef]