Robotics and Autonomous Systems 96 (2017) 184–210

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Localization of sound sources in robotics: A review

Caleb Rascon *, Ivan Meza

Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas, Universidad Nacional Autonoma de Mexico, Circuito Escolar S/N, Mexico 04510, Mexico

HIGHLIGHTS

- A highly detailed survey of sound source localization (SSL) used over robotic platforms.
- Classification of SSL techniques and description of the SSL problem.
- Description of the diverse facets of the SSL problem.
- Survey of the evaluation methodologies used to measure SSL performance in robotics.
- Discussion of current SSL challenges and research questions.

ARTICLE INFO

Article history:
Received 18 August 2016
Received in revised form 24 June 2017
Accepted 21 July 2017
Available online 5 August 2017

Keywords:
Robot audition
Sound source localization
Direction-of-arrival
Distance estimation
Tracking

ABSTRACT

Sound source localization (SSL) in a robotic platform has been essential in the overall scheme of robot audition. It allows a robot to locate a sound source by sound alone. It has an important impact on other robot audition modules, such as source separation, and it enriches human–robot interaction by complementing the robot's perceptual capabilities. The main objective of this review is to thoroughly map the current state of the SSL field for the reader and provide a starting point to SSL in robotics. To this effect, we present: the evolution and historical context of SSL in robotics; an extensive review and classification of SSL techniques and popular tracking methodologies; different facets of SSL as well as its state-of-the-art; evaluation methodologies used for SSL; and a set of challenges and research motivations.

1. Introduction

The goal of sound source localization (SSL) is to automatically estimate the position of sound sources. In robotics, this functionality is useful in several situations, for instance: to locate a human speaker in a waiter-type task, in a rescue scenario with no visual contact, or to map an unknown acoustic environment. Its performance is of paramount influence to the rest of a robot audition system since its estimations are frequently used in subsequent processing stages such as sound source separation, sound source classification and automatic speech recognition.

There are two components of a source position that can be estimated as part of SSL (in polar coordinates):

- Direction-of-arrival estimation (which can be in 1 or 2 dimensions)
- Distance estimation.

SSL in real-life scenarios needs to take into account that more than one sound source might be active in the environment. Therefore it is also necessary to estimate the position of multiple simultaneous sound sources. In addition, both the robot and the sound source are mobile, so it is important to track its position through time.

SSL has been substantially pushed forward by the robotics community by refining traditional techniques such as: single direction-of-arrival (DOA) estimation, learning-based approaches (such as neural network and manifold learning), beamforming-based approaches, subspace methods, source clustering through time and tracking techniques such as Kalman filters and particle filtering. While implementing these techniques onto robotics platforms, several facets relevant to SSL in robots have been made evident including: number and type of microphones used, number and mobility of sources, robustness against noise and reverberation, type of array geometry to be employed, type of robotic platforms to build upon, etc.

As it is shown in this review, the SSL field in robotics is quite mature, proof of which are the recent surveys in this topic. For instance, [1,2] present a survey on binaural robot audition, [3] offers a general survey of SSL in Chinese, [4] presents some SSL

* Corresponding author.
E-mail addresses: caleb.rascon@iimas.unam.mx (C. Rascon), ivanvladimir@turing.iimas.unam.mx (I. Meza).

http://dx.doi.org/10.1016/j.robot.2017.07.011

works based on binaural techniques and multiple-microphone arrays, and [5] presents an overview of the robot audition field as a whole. The aim of this work is to review the literature of SSL implemented over any type of robot, such as service, rescue, swarm, industrial, etc. We also review efforts that are targeted for an implementation in a robotic platform, even if they were not actually implemented in one. In addition, we review resources for SSL training or evaluation, including some that were not collected from a robotic perspective but could be applied to a robotic task. Finally, we incorporate research that uses only one microphone for SSL that, although not applied in a robotic platform, we believe has an interesting potential for the SSL robotic field.

In this work we present: the evolution of the field (Section 2); a definition of the SSL problem (Section 3); a classification of techniques used in SSL within the context of robot audition (Section 4); an overview of popular tracking techniques used for SSL (Section 5); several facets that describe the areas that SSL techniques are tackling (Section 6); a review of different evaluation methods that are currently being used for measuring the performance of SSL techniques (Section 7); and an insight on potentially interesting challenges for the community (Section 8). Finally, we highlight several motivations for future research questions in the robot audition community (Section 9).

2. The evolution of SSL

The surge of SSL in robotics is relatively new. To our knowledge, it started in 1989 with the robot *Squirt*, which was the first robot to have a SSL module [6,7]. *Squirt* was a tiny robot with two competing behaviors: hiding in a dark place and locating a sound source. The idea of using SSL as a behavior to drive interaction in a robot was later explored by Brook's own research team and it culminated with a SSL system for the *Cog* robot [8–11]. In the meantime, several Japanese researchers started to investigate the potential of SSL in a robot as well. In 1993, Takanashi et al. explored an anthropomorphic auditory system for a robot [12,13] (as described by [10]). This research was followed by notable advances in the field: Chiye robot [14], RWIB12-based robot [15–18], Jijo-2 [19,20], Robita [21] and Hadalay [22]. This first generation of robots tackled difficult scenarios such as human–robot interaction, integrating a complete auditory system (source separation feeding speech recognition), active localization, dealing with mobile sources and capture systems, and by exploring different methodologies for robust SSL.

At the turn of the 20th century, the binaural sub-field of robot audition started to become an important research effort, including SSL. Although robots from the first generation were technically binaural (e.g., Squirt, COG, Chiye, Hadalay), it is with the arrival of the SIG robot [23] that the field of binaural robot audition started to generate interest. SIG was built to promote audition as a basic skill for robots and was presented as an experimental platform for the RoboCup Humanoid Challenge 2000 [24]. This resulted in SIG becoming popular for researching robot perception. Binaural robot audition has been followed by other research teams and progress in the field has been constant [25–36].

During the 2000s, an important rift occurred in terms of the research motivations in the robot audition field, specifically in SSL techniques. Binaural audition cemented itself by the motivation to imitate nature: using only two ears/microphones. On the other hand, there was the motivation to increase performance (detailed in Section 4.3), which pushed for the use of more microphones. This opened the door for source localization techniques that use a high amount of sensors (such as MUSIC and beamformers) to carry out SSL in a robot. Subsequently, the facets of the SSL problem were broadened, which yielded a wide variety of solutions from the robot audition community.

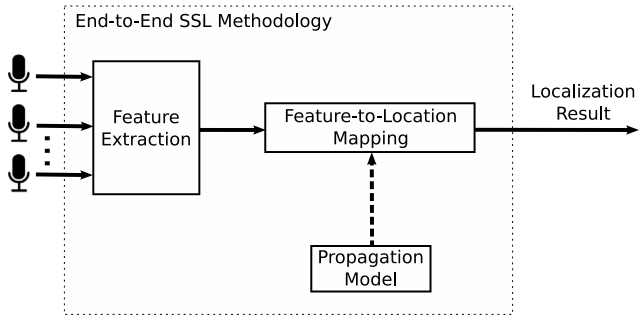


Fig. 1. The complete data pipeline of an end-to-end SSL methodology.

Throughout its history, a central goal for robots with a SSL system has been to support interaction with humans. In the first generations, an important contribution was to face the user, since it indicates that the robot is paying attention. One of the first robots to carry out this attention-based interaction was the Chiye robot [14] which has made its way into recent products such as the Paro robot [37]. Further on, SSL has been used in more complex settings in which other skills intertwine together to reach a specific goal, such as: playing the Marco-Polo game, acting as a waiter, taking assistance and finding its user when it visually lost him/her [38]; logging and detecting the origin of certain sounds while interacting with a caregiver [39]; playing a reduced version of hide and seek in which hand detection and SSL are used to guide the game [40]; providing visual clues from the sound sources as a complement of a telepresence scenario [41]; and directing a trivia-style game [42]. Given the evolution of SSL in robots, we are certain that the complexity of the scenarios will keep growing. In fact, we foresee that the challenges to come will definitely be more demanding (see Section 8 for further discussion).

3. Definition of the sound source localization problem

Sound source localization (SSL) tackles the issue of estimating the position of a source via audio data alone. This generally involves several stages of data processing. Its pipeline is summarized in Fig. 1.

Since this pipeline receives the data directly from the microphones and provides a SSL estimation, we consider a methodology that carries this out as *end-to-end*. Features are first extracted from the input signals. Then, a feature-to-location mapping is carried out, which usually relies on a sound propagation model. These three phases are referenced as such in the explanation of each methodology and their relevant variations in Section 4.

In this section a brief overview of these three phases is presented for ease of reference in the later detailed explanations.

3.1. Propagation models

The sound propagation model is proposed depending on: the positioning of the microphones, as there may be an object between them; the robotic application, as the user may be very close or far away from the microphone array; and the room characteristics, as they define how sound is reflected from the environment. In addition, the propagation model generally dictates the type of features to be used.

The most popular propagation model used is the free-field/far-field model, which assumes the following:

- *Free field:* The sound that is originated from each source reaches each microphone via a single, direct path. This

means that there are no objects between the sources and the microphones and that there are no objects between the microphones. In addition, it is assumed that there are no reflections from the environment (i.e., no reverberation).
- *Far field:* The relation between the inter-microphone distance and the distance of the sound source to the microphone array is such that the sound wave can be considered as being planar.

The second assumption greatly simplifies the mapping procedure between feature and location, as discussed in Section 4.1.

There are other type of propagation models that are relevant in SSL in robotics. The Woodworth–Schlosberg spherical head model [43, pp. 349–361] has been used extensively in binaural arrays placed on robotic heads [23,44] and is explained in Section 4.2. The near-field model [45] assumes that the user can be near the microphone array, which requires to consider the sound wave as being circular. There are a few robotic applications that use the near-field model, such as [46], however it is not as commonly used as the far-field model. In fact, there are approaches that use a modified far-field model successfully in near-field circumstances [47] or that modify the methodology design to consider the near-field case [48]. Nevertheless, as presented in [48], a far-field model directly used in near-field circumstances can decrease the SSL performance considerably. In addition, there are cases in which the propagation model is learned, such as the neural-network-based approaches in [49,50], manifold learning [33,51], linear regression [52] and as part of a multi-modal fusion [11,21].

3.2. Features

There are several acoustic features used throughout the reviewed methodologies. In this section, we provide a brief overview of the most popular:

Time difference of arrival (TDOA). It is the time-difference between two captured signals. In 2-microphone arrays (binaural arrays) that use external *pinnae*, this feature is also sometimes called the inter-aural time difference (ITD). There are several ways of calculating it, such as measuring the time difference between the moments of zero-level-crossings of the signals [18] or between the onset times calculated from each signal [6,7,14,17]. Another way to calculate the TDOA is by assuming the sound source signal is narrowband. Let us denote the phase difference of two signals at frequency f as $\Delta\varphi_f$. If f_m is the frequency with the highest energy, the TDOA for narrowband signals (which is equivalent to the inter-microphone phase difference, or IPD) can be obtained by $\frac{\Delta\varphi_{f_m}}{2\pi f_m}$ [23]. However, the most popular way of calculating the TDOA as of this writing is based on cross-correlation techniques, which are explained in detail in Section 4.1.

Inter-microphone intensity difference (IID). It is the difference of energy between two signals at a given time. This feature, when extracted from time-domain signals, can be useful to determine if the source is in the right, left or front of a 2-microphone array. To provide greater resolution, a many-microphone array is required [53] or a learning-based mapping procedure can be used [10]. The frequency-domain version of IID is the inter-microphone level difference (ILD) that is provided as the difference spectrum between the two short-time-frequency-transformed captured signals. This feature is also often used in conjunction with a learning-based mapping procedure [35].

A similar feature to the ILD are the set of differences of the outputs of a set of filters spaced logarithmically in the frequency domain (known as a filter bank). These set of features have shown more robustness against noise than the IID [9], while employing a feature vector with less dimensions than the ILD.

In [54], the ILD is calculated in the overtone domain. A frequency f_o is an overtone of another f when $f_o = rf$ (given that $r \in [2, 3, 4, \ldots]$) and their magnitudes are highly correlated through time. This approach has the potential of being more robust against interferences, since the correlation between the frequencies implies they belong to the same source.

Spectral notches. When using external *pinnae*[1] or inner-ear canals, there is a slight asymmetry between the microphone signals. Because of this, the result of their subtraction presents a reduction or amplification in certain frequencies, which depend on the direction of a sound source. These notches can be mapped against the direction of the sound source by experimentation [52]. However, because small changes to the external *pinnae* may hinder the results from these observations, it is advisable to use learning-based mapping when using these features [49].

Binaural/spectral cues. It is a popular term to refer to the feature set that is composed by the IPD and the ILD in conjunction. This feature set is often used with learning-based mapping [50,51]. They are often extracted on an onset to reduce the effect of reverberation [55]. It has been shown in practice that temporal smoothing of this feature set makes the resulting mapping more robust against moderate reverberation [56].

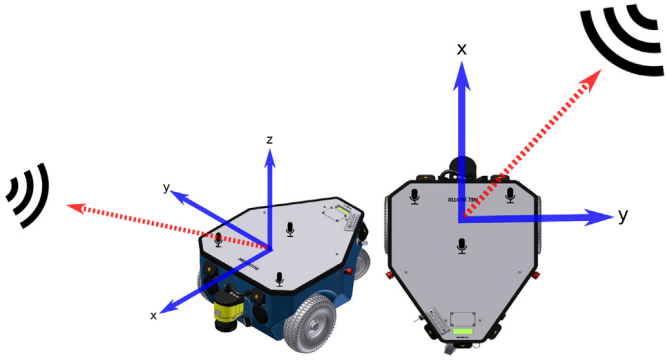
Besides these features, there are others that are also highly used, such as the MUSIC pseudo-spectrum and the beamformer steered-response. However, their application is bound to specific end-to-end methodologies. Because of this, their detailed explanation is given in Section 4.

3.3. Mapping procedures

A mapping procedure for SSL is expected to map a given extracted feature to a location. A typical manner to carry this out is by applying directly the propagation model, such as the free-field/far-field model or the Woodworth–Schlosberg spherical head model, both discussed in Section 3.1. However, there are some type of features (especially those used for multiple-source-location estimation) which require an exploration or optimization of the SSL solution space. A common approach is to carry out a *grid-search*, in which a mapping function is applied throughout the SSL space and the function output is recorded for each tested sound source location. This produces a solution spectrum in which peaks (or local maximums) are regarded as the SSL solutions. This is the most used type of mapping procedure for multiple-source-location estimation. Two important examples are the subspace orthogonality feature of MUSIC and the steered-response of a delay-and-sum beamformer. These are detailed further in Section 4.3.

There are types of mapping procedures other than *grid-search*. Their main focus is to train the mapping function based on recorded data of sources with known locations. As a result, the mapping function that was learned implicitly encodes the propagation model. In this survey, this type of mapping procedures are referred to as learning-based mapping. These are based in different training methodologies, such as neural networks [11,21,49], locally-linear regression [57], manifold learning [33,51], etc. Further details are given of each mapping procedure in the relevant branches of the methodology classification presented in Section 4.

[1] External ears.



(a) Full frame of reference. (b) Azimuth (top view).

(c) Elevation (side view).

Fig. 2. Graphical representation of the most used frame of reference. Based on the CAD views of the Neobotix MP-500 mobile robot [58].

4. Classification of sound source localization end-to-end methodologies

As mentioned before, the location of a sound source is usually considered as being composed of two parts: (1) the direction of arrival (DOA) of the source and (2) the distance of the source to the microphone array. The frame of reference that is used by most (if not all) of the reviewed works is exemplified in Fig. 2, which shows a robot mounted with a 3-microphone array.

As it can be seen, the center of the microphone array is generally considered as the origin. The *azimuth* plane (presented in Fig. 2b) is parallel to the horizon of the physical world and the *elevation* plane (presented in Fig. 2c) is orthogonal to it. This is the terminology and frame of reference used throughout the survey.

In this section, a classification of end-to-end methodologies used for SSL by the robotics community is presented. Because of the two-part composition of a sound source location, a type of divide-and-conquer philosophy of SSL has become popular: the estimation of the DOA and the distance is carried out separately. And, in most cases, the DOA is usually the only part of the location reported. Given this popularity, this section mostly reviews methodologies that estimate the DOA of the sound source. However, important advances in distance estimation have been made, which warrants their own branch.

The presented classification is summarized as follows:

- **1-dimensional single direction-of-arrival estimation.** In this branch, techniques that estimate the DOA in the azimuth plane of a single source are described.
- **2-dimensional single direction-of-arrival estimation.** In this branch, techniques that estimate the DOA in both the azimuth and the elevation plane of a single source are described.
- **Multiple direction-of-arrival estimation.** In this branch, techniques that estimate the DOA of multiple sources are described. These are mostly in the azimuth plane, but the process of how to generalize them into both planes is also described. This branch is further divided into three sub-branches:
  - *Beamforming-based.* Those that carry out spatial filtering towards several DOA candidates.
  - *Subspace methods.* Those that take advantage of the differentiation between signal and noise subspaces.
  - *Source clustering through time.* Those that carry out single DOA estimation throughout various time windows and provide a multiple-DOA solution by clustering these results.
- **Distance estimation.** In this branch, techniques that estimate the distance of the sound source to the microphone array are described.

The full list of the 188 reviewed works has been made available through an Excel file that is part of the additional external material that comes with this writing.

4.1. 1-dimensional single direction-of-arrival estimation

There are many works whose objective is to locate and track one sound source in the environment. A very commonly used feature to achieve this objective is the time-difference-of-arrival (TDOA) between a pair of sensors or microphones. The most popular way to estimate the TDOA as of this writing is based on calculating a cross-correlation vector (CCV) between two captured signals. One of the simplest way to calculate CCV is based on the Pearson correlation factor, as presented in Eq. (1):

$$CCV[\tau] = \frac{\sum_t (x_1[t] - \overline{x_1})(x_2[t-\tau] - \overline{x_2})}{\sqrt{\sum_t (x_1[t] - \overline{x_1})^2}\sqrt{\sum_t (x_2[t-\tau] - \overline{x_2})^2}} \quad (1)$$

where x_1 and x_2 are the two discrete signals being compared; τ is the point at which x_2 is being linearly shifted and the correlation is being calculated; and $\overline{x_1}$ and $\overline{x_2}$ are the mean values of x_1 and x_2, respectively. The TDOA of the sound source (τ_o) is the τ value that maximizes CCV. As mentioned before, the free-field/far-field propagation model is the most commonly used (presented in Eq. (2)) and it provides a simple feature-to-location mapping between τ_o and the DOA of the source (θ_o):

$$\theta_o = \arcsin\left(\frac{V_{sound} \cdot \tau_o}{f_{sample} \cdot d}\right) \quad (2)$$

where V_{sound} is the speed of sound ($\sim$343 m/s); f_{sample} is the sampling frequency in Hz; d is the distance between microphones in meters; and τ_o is the TDOA of the sound source in number of samples. To simplify SSL in real-life environments, the microphones are positioned such that the imaginary line between them is parallel to the azimuth plane, resulting in θ_o being the DOA in that plane. If elevation is required, the microphone pair can be positioned such that they cross the elevation plane. However, only the DOA in the plane of the microphone array is able to be estimated by this methodology.

Estimating the TDOA via CCV can be very sensitive to reverberations and other noise sources [59, pp. 213–215]. In these cases, correlation values are "spread" into other TDOAs [60], resulting in wide hills of correlation as well as TDOA estimation errors [61]. To counter this, a similar correlation vector can be calculated by an alternative approach. A frequency-domain-based cross-correlator (CC_F) [62] is presented in Eq. (3):

$$CC_F[f] = X_1[f]X_2[f]^* \quad (3)$$

where X_1 and X_2 are the Fourier transforms of x_1 and x_2 of Eq. (1) respectively; the $\{.\}^*$ operator stands for the complex conjugate

operation; and CC_F is a frequency-domain-based cross-correlator. It is important to state that the resulting $\mathcal{F}^{-1}(CC_F)$ presents the correlation information in a different manner than CCV. However, the peaks in $\mathcal{F}^{-1}(CC_F)$ are found in the same places of high correlation as in CCV in the range of $-\tau_{max} \leq \tau \leq \tau_{max}$, where τ_{max} is the maximum value of τ that can physically occur.[2]

By performing this operation in the frequency domain, a weighting function $\psi[f]$ can be applied as in Eq. (4), which is known as generalized cross-correlation (GCC) [60]:

$$GCC_F[f] = \psi[f]X_1[f]X_2[f]^*. \tag{4}$$

$\psi[f]$ varies depending on the objective of the correlation vector. If $\psi[f] = 1$, the resulting $\mathcal{F}^{-1}(GCC_F)$ equates to $\mathcal{F}^{-1}(CC_F)$ which suffers from sensitivity to reverberation, similar to CCV calculated by Eq. (1). Therefore, it is of interest that Dirac delta functions appear in the correlation vector only in places of high correlation. Since the Fourier transform of a Dirac delta function has all frequency magnitudes at 1, normalizing the magnitudes in GCC_F forces the presence of approximations of Dirac delta functions in places of high correlation. To normalize these magnitudes, $\psi[f]$ is equated to the inverse of the magnitude of the signal multiplication, as in Eq. (5):

$$\psi[f] = \frac{1}{|X_1[f]X_2[f]^*|}. \tag{5}$$

By applying $\psi[f]$ in Eq. (4), the phase information is left intact in GCC_F, thus $\psi[f]$ of Eq. (5) is known as the phase transform ($PHAT$). The generalized cross-correlation with phase transform [60] (GCC-PHAT) is presented in Eq. (6):

$$PHAT_F[f] = \frac{X_1[f]X_2[f]^*}{|X_1[f]X_2[f]^*|}. \tag{6}$$

Carrying out this normalization is equivalent to "whitening" the input signals, since all frequencies have a magnitude of 1. This has been shown to produce a "spikier" crosspower spectrum [63]. Because of this, interfering sources produced by either actual sound sources or environmental reflections (i.e., reverberation) tend to also "appear" as other peaks in the correlation vector. This offsets their effect on the correlation calculations in other TDOAs, which is not the case with the Pearson-based method described by Eq. (1). This provides GCC-PHAT robustness against reverberation, as shown via simulation in [61]. It also provides robustness against interfering sources in high signal-to-interference ratio (SIR) circumstances, as shown in a multiple source scenario in [64].

The PHAT weighting function is typically applied uniformly throughout the frequency bins which introduces sensitivity to broadband noise sources. To counter this, the PHAT weighting function can be modified such that additional weights are set depending on the "noisiness" of the frequency bins. Good examples of this are presented in [47,65,66], where an additional weighting term is added to the PHAT weighting function based on the frequency bin SNR, providing robustness against noise. However, applying non-zero weights to the frequency bin may produce noise leaking into other frequencies. To avoid this, an evolution of this approach is presented in [67], where a hard binary mask is applied instead. Meaning, only binary weights are added to the PHAT weighting function: 1 if the SNR is above a certain threshold, 0 otherwise. Unfortunately, using these hard masks results in leaks in the DOA estimation with unwanted dominant peaks. This issue is countered in [68], where a transition mask is used between noise and speech windows.

It is important to mention that these masking methods require on-line noise estimation to calculate the narrowband SNRs for each frequency. An alternative to this is to create a binary mask that only nullifies the frequency bins outside the frequency range used by the type of source the application calls for. In the case of [69], the authors aimed to track only speech sources, thus all frequency bins outside the frequency range of voice were nullified.

GCC-PHAT is probably the most commonly used TDOA estimation technique for single direction-of-arrival estimation in robot audition because of its robustness and its ease of implementation. For example, in [70,71], GCC-PHAT is used to carry out an acoustic map of the environment via an exploration carried out by a robot. Other works that use GCC-PHAT as part of their sound source localization systems for service robots can be found in [28,72–76].

Interestingly, the appearance of peaks in $PHAT$ bins other than the one representing the signal of interest may constitute other sources which are assumed as interfering. Because of this, some proposals use $PHAT$ as a simple way to estimate multiple directions-of-arrival [64,77]. However, even with the changes to $\psi[f]$ to make it applicable, the appearance of peaks is dependent on the ratio of power between the multiple sources [64]. As far as we know, this variation of the GCC-PHAT technique has not been applied in the context of robot audition.

As mentioned before, the free-field/far-field propagation model is the most commonly used for single-DOA estimation, however other sound propagation models can be used for 1-dimensional single-DOA estimation. In [78], a spherical head is assumed to be positioned between the microphones. For this purpose, the authors use the Woodworth–Schlosberg head model [43, pp. 349-361], shown in Eq. (7):

$$\tau(\theta) = \frac{d}{2V_{sound}}(\theta + \sin(\theta)) \tag{7}$$

where $d/2$ represents the radius of the head. Two propagation paths are then observed: one that propagates through the front of the head and another that propagates through the back. Although using the front propagation path should be enough for TDOA estimation, the back propagation path interferes with this estimation. To counter this, a multipath interference compensation factor is used. The resulting propagation model is presented in Eq. (8):

$$\tau(\theta) = \frac{d}{2V_{sound}}(\theta + \sin(\theta)) + \frac{d}{2V_{sound}}(sign(\theta)\pi - 2\theta)\,|\sin(\theta)| \tag{8}$$

where $sign(\theta)$ is described in Eq. (9):

$$sign(\theta) = \begin{cases} -1, & \theta < 0 \\ 1, & \theta \geq 0 \end{cases} \tag{9}$$

In [65], an addition to the propagation model in Eq. (8) is made to consider an attenuation factor β_m (typically set to 0.1, as suggested by the authors) as presented in Eq. (10):

$$\tau(\theta) = \frac{d}{2V_{sound}}(\theta + \sin(\theta)) + \frac{d}{2V_{sound}}(sign(\theta)\pi - 2\theta)\,|\beta_m \sin(\theta)|\,. \tag{10}$$

In [44], the authors reached the same models presented in Eqs. (2) and (7) from the point of view of auditory epipolar geometry (AEG) [23]. Epipolar geometry is popularly used in stereo computer vision to physically localize features extracted from two images simultaneously captured from two cameras with known locations [79]. The revision to AEG (RAEG) made in [44] is analogous to the Fourier transform of the model presented in Eq. (7). The authors applied the following grid-search mapping to carry out 1-D SSL: (1) using RAEG, a set of inter-microphone phase differences (IPD_f) are calculated for each possible f and DOA; (2) an $\widehat{IPD}_f$ is estimated from the incoming signals in the selected f's; and (3)

[2] Which happens when the sound source is placed in the imaginary line that crosses both microphones. That is to say, when $\theta = 90°$. Thus, it is calculated as $\tau_{max} = \frac{f_{sample} \cdot d}{V_{sound}} \sin(90°)$.

the final θ_o is the one associated with the IPD_f from the set that is the most similar to the estimated $\widehat{IPD}_f$.

Alternatively, machine learning approaches can be used to tackle SSL. These approaches do not define a sound propagation model, instead they learn a mapping function Ψ from a space of features κ of the sound signal to source locations θ.

$$\Psi_\phi : \kappa \mapsto \theta. \tag{11}$$

This map is modeled from recorded examples (training dataset) that are used to identify the parameters ϕ. Traditionally, additional examples are recorded to evaluate the trained model (testing dataset), and is expected to perform well in a real-world setting. Some types of training techniques that have been used for SSL are: recurrent neuronal networks [80–83]; bio-inspired spiking neural networks [84]; and deep learning architectures are recently being explored [85–87].

However, in order to train a model, training data diversity is important for generalization [88]. This is because the training process requires to "observe" a representative set of examples from which it will attempt to generalize the whole set of possible scenarios. If the circumstances in which the model is deployed in a real-world setting are vastly different from the circumstances in which it was trained, a basic assumption of most machine learning techniques is overlooked [89]. For instance, if the training data was recorded in a quiet office and the robot is deployed in a busy restaurant, the mismatch in noise levels will result in a poor SSL performance. To avoid this, a highly diverse training dataset is required, varying in terms such as: room response, number and location of sources, microphone placement, the use of external *pinnae* and/or inner-ear canals, head model, etc. In [90], a system is presented that both locates and identifies the users, and the authors proposed as future work to evaluate their SSL performance in mismatched training and testing conditions. An example of how this phenomenon impacts the performance of a SSL system can be found in [57], detailed in Section 4.2.

4.2. 2-dimensional single direction-of-arrival estimation

In a 3D environment, it is of interest to estimate the DOA of the source in both the azimuth and elevation plane. If the microphone array has a 3D geometry, the elevation angle can be calculated using the same TDOA-based techniques discussed in Section 4.1 along with the DOA in the azimuth plane, providing a 2-dimensional DOA estimation. This requires, however, additional microphones to be used.

Frequently in binaural hearing the microphones are positioned with a body between them (breaking the free-field assumption) and/or accompanied with external ears or with inner-ear canals. This is carried out so that the sound source signal is 'filtered' before it is captured by the microphones in a way that is dependent of its direction in both planes. Even though the free-field assumption is broken, this filtering effects can be used alongside non-free-field propagation models to carry out 2-dimensional DOA estimation.

For example, human beings are able to estimate both the azimuth and elevation of a sound source even when employing a two-microphone array [91]. When a sound source located on the side of a human head emits a sound wave, the ear farthest away from the source receives a modified version of the one received by the closest ear. This modification is carried out by several measurable physical phenomena when passing through and/or are reflected by[3] the human head, torso, external *pinnae* and inner-ear canals [92]. This set of phenomena can be measured by placing a microphone on each side of a specialized dummy head and capturing their impulse responses in an anechoic room (to avoid capturing the effects of the environment). These measurements can be used to calculate how the object between the microphones changes the properties of the signal received, depending on the location of the sound source (specifically, its azimuth and elevation). This set of properties can be used to create what is known as a head-related transfer function (HRTF), which is a type of filter that aims to emulate the physical phenomena that modify the sound source audio wave, given a pre-specified 2-dimensional DOA [93]. An HRTF can be used, among other applications, for spatial audio reproduction where a sound source is virtually "positioned" around a listener wearing headphones [94]. For the purpose of single-DOA estimation, a set of candidate 2D DOAs is proposed, with which a set of HRTFs are measured. Then, a database of filters that carry out the inverse function of an HRTF (referenced here as IHRTF) can be calculated from each HRTF-to-DOA association. These IHRTFs can then be applied to the incoming signals following a grid-search mapping. From each application, a proposed metric can be measured, such as the correlation between the output signals. The DOA associated with the IHRTF that maximizes such metric is then proposed as the estimated DOA [95]. It is important to consider that the external *pinnae* can also play an important role in the estimation of the HRTF, as shown in [96] where a spiral ear is used for 2-dimensional DOA estimation.

However, this type of approach requires measuring the effects of the database-IHRTFs (which are based on HRTFs usually estimated with measurements in low-noise anechoic rooms) to captured signals in real-world conditions (which usually include reverberation and noise). To counter this, room characteristics could be measured in advance and be considered as part of the calculations of the database-filters (philosophically different from the aforementioned IHRTFs), making them consistent with the real-world conditions. Unfortunately, deploying a robot into a real-world environment implies that the acoustic characteristics are not known in advance [97]. In addition, the room response is dependent of the position of the microphone array inside the room. This means that when a robot moves (either linearly or by rotation), a new room response must be measured and a new filter database must be calculated to account for this change [98]. A simulation of such a real-world environment can be carried out [99,100], from which room characteristics can be obtained to automatize the filter-database modification. However, carrying out such a simulation can be unfeasible since the dimensions and materials of the environment may not be known in advance. In the cases where they are known, such a simulation is time-consuming [44].

The revised auditory epipolar geometry (RAEG) [44], described in Section 4.1, can be used to surmount the issue of differences between real-world and training conditions. RAEG-based SSL does not require a set of IHRTFs to be estimated in an anechoic room, since the shape of the robotic head is already accounted for. However, it only provides a DOA in the azimuth plane. A generalization to 2D localization is possible via scattering theory.

Scattering theory is a field of physics which models how a wave or particle is perturbed from its path by some object [101,102]. In the case of acoustic or electromagnetic waves, this modeling process can be used for source localization [103]. As shown in [97], scattering theory, with a given IPD and IID, can be used to calculate both azimuth and elevation DOAs by assuming that the shape of the object between the microphones is spherical. A grid-search mapping (similar to the IHRTF-based and RAEG-based approaches already described) is used in [97] with a close-to-spherical robotic head between the microphones. The approach is summarized as follows: (1) a set of $[IPD_f, IID_f]$ tuples is calculated for each possible combination of azimuth DOA, elevation DOA and f's using the scattering theory equations; (2) an $\widehat{IPD}_f$ and $\widehat{IID}_f$ tuple is estimated

[3] The physical phenomenon observed when the audio signal is passing through the body are generally present in low frequencies, while the ones reflected are present in high frequencies.

from the incoming signals in the selected f's; and (3) the final 2-dimensional DOA is the one associated with the tuple from the set that is the most similar to the estimated tuple.

Another way to estimate a single DOA in 2 dimensions is presented in [104], called the space-domain distance (SDD) method. It relies on a distance metric that is applied between the captured time–frequency (TF) bin and a calculated TF bin that estimates what would have been captured if the signal would have been located in a given direction. This metric is based on the spherical Fourier transform (SFT) [105], which aims to investigate objects with rotational symmetry in terms of spherical coordinates, such as the perturbations of audio signals from a spherical robotic head. The SFT coefficients are estimated via measurements or simulations and can be used to pick the TF bins that contain information of the direct-path signal given a pre-defined DOA. The proposed SDD metric in [104] measures the distance between the direct-path TF bins that were calculated from the input signals and the same TF bin using the SFT coefficients given a pre-defined DOA. Given this metric, a grid-search is carried out to find the direction that minimizes it. This technique is robust against reverberation and invariant against the frequency range of the source. It is also used in [106] when testing the impact of a rotating head for single DOA estimation. The minimization of the SDD assumes only one source, however the authors believe that an extension to this approach for multiple-DOA estimation may be possible.

The elevation estimation approaches previously described use grid-search mapping. However, other type of modeling-based procedures can be used to map the feature space to an elevation estimation. For example, the approach proposed in [57] implicitly encodes the HRTF by applying a type of inverse regression called probabilistic piecewise-affine mapping (PPAM). The objective of PPAM in this case is to map a series of features to a location. A part of this set of features are the ILDs and IPDs of the whole frequency spectrum. Noise estimation is carried out via temporal averaging with which an activity measurement is calculated. This activity measurement is also used as a training feature to provide robustness to the mapping process against self noise, background noise and low reverberation. The azimuth accuracy of this technique is not affected when the microphones are moved, but the performance of its elevation estimation does decrease significantly when this happens. The authors state that to make their approach robust against these movements "would require combining training data from different real and/or simulated rooms". It is worthwhile mentioning that the approach described in [57] is able to locate two sources, making it a multiple-2D-DOA estimator. However, training data for such circumstances is required. This means that for it to be able to generically locate multiple sources it requires to be trained with vast amount of data. However, this is an important effort into making learning-based 2-dimensional SSL robust against changes in real-world and training conditions.

Further examples of learning-based 2D SSL are: in [49] a pinnae inspired by the human ear is used to extract spectral notches which are fed into a neural network; manifold learning is used to estimate both azimuth and elevation in a binaural system [33] and has been extended to include visual information [107]; and a linear regression approach is used as the mapping procedure in [52] to learn the propagation model using spectral notches extracted from the iCub robotic head with spiral ears.

4.3. Multiple direction-of-arrival estimation

For scenarios in which multiple sources are considered to be in the auditory scene, several techniques of multiple directions-of-arrival estimation can be used, which can be divided into three categories:

- Beamforming-based
- Subspace methods
- Source clustering through time.

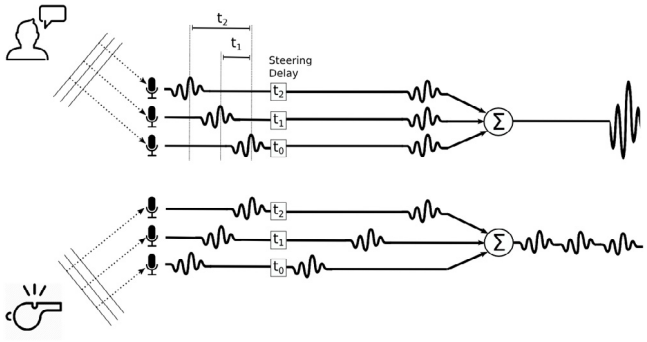


Fig. 3. Graphical representation of delay-and-sum beamforming.

For simplicity of their description, the 1-dimensional versions of these methodologies are presented. Thus, the estimated DOAs are proposed to be in the azimuth plane. However, it is important to consider that it is possible to extend them into 2 dimensions (azimuth and elevation) when using a 3D array geometry.

4.3.1. Based on beamforming

Beamforming is a filtering technique that is based on spatial weighting of the signals captured in an array of sensors such that its output is the signal approaching the sensor array from a pre-specified direction-of-arrival (DOA). It might be considered as counter-intuitive to use this technique for DOA estimation since it assumes that the DOA of the source is known. However, the overall steps of this approach are:

1. Propose an L-size set of candidate DOAs where sources are searched for.
2. Create a beamformer and *steer* it towards each candidate DOA.
3. Measure the response of each beamformer, usually measured as the energy of its output.
4. Create a steered-response spectrum.
5. Find peaks in the spectrum and propose their location as the DOA of a source.

A steered-response spectrum is a $1 \times L$ vector created from all the beamformer's responses ordered by their DOA. If the beamformer's response is measured as its output, the steered-response spectrum can be considered as an energy spectrum that shows how much energy is being received at the microphone array from each DOA. As it is apparent, the grid-search mapping is the most used for beamforming-based multiple-DOA estimation.

The simplest form of beamforming is know as the delay-and-sum beamforming (DAS), summarized in Fig. 3.

Assuming a free-field/far-field sound propagation model, a source's DOA has a direct relation to the TDOA of the captured signal between sensors, as shown in Eq. (2). Thus, a DAS beamformer aims to artificially shift the signals to counter such time difference and then add the shifted signals to obtain its output. The idea behind this is that the component of the captured signals that is received from the pre-defined DOA (or *steered direction*) is aligned in each shifted signal. This results in such a component being accentuated in the beamformer's output relative to other sources positioned in other directions (as shown in the upper part of Fig. 3). As it can be concluded, such accentuation is proportional to the number of sensors employed.

Assuming that the sensors are omnidirectional microphones, the DAS beamformer output can be represented by Eq. (12):

$$\hat{s_\theta}[t] = \sum_{n=1}^{N} x_n[t - \tau_n(\theta)] \tag{12}$$

where $\hat{s_\theta}$ is the beamformer's output, which is the estimation of the ground truth s_θ that reaches the array from the steered direction θ; t is the time bin; x_n is the signal received at microphone n; N is the number of microphones; and $\tau_n(\theta)$ is the TDOA of the source in microphone n related to the steered direction θ. It is important to mention that the TDOAs τ_n are generally calculated based on a reference microphone. If these are consistent to the array's geometry and propagation model, this approach can accommodate any array dimensionality. It can also accommodate several non-free-field/non-far-field propagation models, such as the near-field model [45] and the spherical head model [43, pp. 349-361].

The energy of the beamformer's output steered towards θ (E_θ) can be calculated by Eq. (13):

$$E_\theta = \sum_{t=1}^{T} \hat{s_\theta}[t]^2. \tag{13}$$

In Eq. (12), the time shift is carried out in the time domain. A time shift can also be performed by manipulating the captured signal in the frequency domain, as presented in Eq. (14):

$$X_{n_{\tau_n(\theta)}}[f] = X_n[f] e^{-2\pi f \tau_n(\theta)} \tag{14}$$

where X_n is the Fourier transform of x_n; f is the frequency bin; and $X_{n_{\tau_n(\theta)}}$ is the Fourier transform of $x_n(t - \tau_n(\theta))$. The arrangement of shifts related to a steered direction θ that are applied to the captured signals can be expressed as the complex-value $N \times F$ matrix W_θ, as shown in Eq. (15):

$$W_\theta = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{-2\pi f_1 \tau_2(\theta)} & e^{-2\pi f_2 \tau_2(\theta)} & \cdots & e^{-2\pi f_F \tau_2(\theta)} \\ e^{-2\pi f_1 \tau_3(\theta)} & e^{-2\pi f_2 \tau_3(\theta)} & \cdots & e^{-2\pi f_F \tau_3(\theta)} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-2\pi f_1 \tau_N(\theta)} & e^{-2\pi f_2 \tau_N(\theta)} & \cdots & e^{-2\pi f_F \tau_N(\theta)} \end{bmatrix} \tag{15}$$

where the f's are the frequency bins; N is the number of microphones; F is the frequency window size; and W_θ is the *broadband steering matrix*, where each of its columns represents a *narrowband steering vector*.

The Fourier transform of the output of the beamformer (S) can be constructed via Eq. (16):

$$\hat{S_\theta}[f] = W_\theta[f]^H X[f] \tag{16}$$

where the $\{.\}^H$ operator stands for the Hermitian transpose[4] ; X is a $N \times F$ complex-value matrix that holds all of the X_n's in its rows; $W_\theta[f]$ is the $N \times 1$ complex-value column of W_θ that holds all the beamforming weights for the frequency f; $X[f]$ is the $N \times 1$ column that holds the frequency information in f of all the captured signals; and $\hat{S}_\theta[f]$ is the beamformer's complex-value output steered towards θ in frequency f. $\hat{S_\theta}$ is the Fourier transform of $\hat{s_\theta}$, such that $\hat{s_\theta} = \mathcal{F}^{-1}(\hat{S_\theta})$.

Working in the frequency domain opens the door to additional refinements to W_θ to improve the beamformer's performance. For example, a set of weights ($A_{\theta_{MVDR}}$) that minimizes the energy of the beamformer while maintaining the same direction as W_θ is shown in Eq. (17):

$$A_{\theta_{MVDR}}[f] = \frac{R[f]^{-1} W_\theta[f]}{W_\theta[f]^H R[f]^{-1} W_\theta[f]} \tag{17}$$

where $R[f]$ is the $N \times N$ covariance matrix of $X[f]$. This is the well known Capon beamformer, also known as the minimum variance distortionless response (MVDR) [108]. An important consideration when implementing MVDR is the calculation of $R[f]$. A popular estimation method of $R[f]$ is the sample covariance matrix $\hat{R}[f]$, calculated from the average $X[f, t - T : t]X[f, t - T : t]^H$ over T time windows, where t is the current time window. This is an essential limitation of MVDR, since it does not provide reliable results until $\hat{R}[f] \approx R[f]$, which could take several time windows to do so.

The energy E_θ from Eq. (13) is based on the estimated $\hat{s_\theta}$, which means that the performance of the SSL relies on how well the sound source in the steered direction is accentuated. Thus, the signal-to-interference ratio (SIR) of the beamformer's output when steered towards the sound source has an effect on the relative height of its corresponding peak in the steered-response spectrum. And if the SIR is too low (i.e., the source is not accentuated enough), the peak can be difficult to find in the spectrum. Since this accentuation is related to the number of microphones used, a microphone array with a high number of microphones (8 or higher) is often employed (see Section 6.3.1 for more details).

An issue that arises when using DAS beamforming for multiple-DOA estimation is that the peaks that appear in the resulting steered-response spectrum are usually quite wide, making the resolution of the grid-search very poor. A way around this issue, as proposed in the ManyEars project [109], is to first rewrite the output of the beamformer in terms of cross-correlation vectors (CCV) as explained in Eq. (18):

$$\begin{aligned} E_\theta &= \sum_{t=1}^{T} \hat{s_\theta}[t]^2 = \sum_{t=1}^{T} \left(\sum_{n=1}^{N} x_n[t - \tau_n(\theta)] \right)^2 \\ &= \sum_{n=1}^{N} \sum_{t=1}^{T} x_n[t - \tau_n(\theta)]^2 \\ &\quad + 2 \sum_{n_1=1}^{N} \sum_{n_2=1}^{n_1-1} \sum_{t=1}^{T} x_{n_1}[t - \tau_{n_1}(\theta)] x_{n_2}[t - \tau_{n_2}(\theta)] \\ &= \sum_{n=1}^{N} \sum_{t=1}^{T} x_n[t - \tau_n(\theta)]^2 \\ &\quad + 2 \sum_{n_1=1}^{N} \sum_{n_2=1}^{n_1-1} CCV_{x_{n_1}, x_{n_2}}[\tau_{n_1}(\theta) - \tau_{n_2}(\theta)]. \end{aligned} \tag{18}$$

The last step is carried out by using the cross-correlation calculation in Eq. (1). As shown in [60], the GCC vector (as presented in Eq. (4)) can be used as a replacement for the CCV vector in Eq. (18). Thus the authors of [63] instead use the $PHAT$ vector from Eq. (6) to sharpen the peaks in the resulting steered-response spectrum.

However, because of the magnitude normalization carried out by $PHAT$, each frequency bin contributes the same in the correlation calculation. This makes the process sensitive to noise. To counter this, as also presented in [109], spectral weighting can be applied to diminish the contribution of frequencies with low narrowband SNR. To do this, at every time window the mean power spectral density of all microphones is calculated, referred as the $1 \times F$ complex-value vector X_{mean}. Then, the frequency-domain noise signal (referred as the $1 \times F$ complex-value vector Y) is estimated by time-averaging X_{mean}. Then the weighting function $\psi[f]$ presented in Eq. (19) is applied to Eq. (4) for correlation calculation:

$$\psi[f] = \begin{cases} \psi_{PHAT}[f], & X_{mean}[f] \leq Y[f] \\ \psi_{PHAT}[f] \left(\dfrac{X_{mean}[f]}{Y[f]} \right)^{\gamma}, & X_{mean}[f] > Y[f] \end{cases} \tag{19}$$

[4] The Hermitian transpose is applied to negate the TDOAs in W_θ and, thus resulting in an aligned signal if it is steered in the DOA of a source signal.

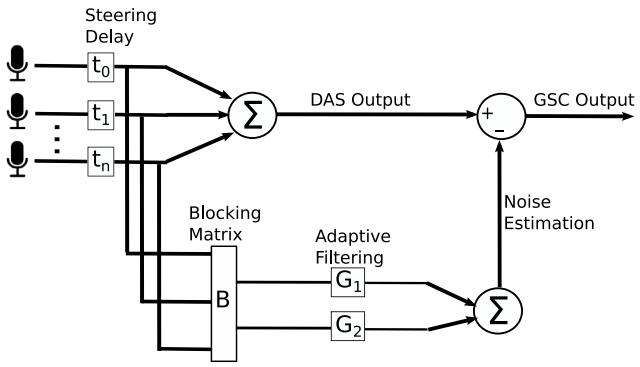


Fig. 4. Overview of the generalized sidelobe canceller (GSC).

where $\psi_{PHAT}[f] = \frac{1}{|X_1[f]X_2[f]^*|}$ is the phase transform; and $0 < \gamma < 2$ is a scalar which can be calibrated to define how much weight to give frequencies that have a high narrowband SNR. This results in an increase in robustness against noise.

One strong advantage of using beamforming for multiple-DOA estimation is that, since the beamformer is carrying out a type of source separation, features extracted from the separated sources can be used as metrics of the grid-search mapping. For example, in [110] the sound sources of interests are humans. Because of this, a voice similarity metric is employed to smooth the steered-response spectrum in such a way that the highest peaks can be assumed as being human sources, not noise. Human vowel sounds have an specific combination of spectral peaks or *peak signature* (J, a $1 \times F$ complex-value vector). This signature is distinct from other type of sounds and can be used for voice activity detection [111]. By capturing a corpus of vowel recordings from several speakers, a set of peak signatures templates can be obtained [110]. These templates only bear values of 1's in the frequency bins of the spectral peaks and 0's in all other bins. The similarity between the output of the beamform steered towards a proposed DOA ($\hat{S}_\theta[f]$) and a peak signature (J) is obtained by calculating the peak valley difference (PVD), described in Eq. (20):

$$PVD = \frac{\sum_{f=1}^{F} \hat{S}_\theta J[f]}{\sum_{f=1}^{F} J[f]} - \frac{\sum_{f=1}^{F} \hat{S}_\theta (1 - J[f])}{\sum_{f=1}^{F} J[f]}. \tag{20}$$

The PVD measures the distance of two spectra as the difference of the average energy around peaks that appear in both spectra and the average energy of other frequency bands.

Additionally in [110], another type of beamformer known as generalized sidelobe canceller (GSC) is applied in the directions of the n-best DOA candidates obtained from the smoothed steered-response spectrum. This is performed as a way to boost the signals in the "best" directions. GSC is used in [110] since it is designed to adapt to changes in noise/interference [112]. Fig. 4 presents an overview of GSC.

As it can be seen, GSC employs a delay-and-sum beamformer (DAS) in its top path. In the bottom path, it carries out noise estimation which it then subtracts from the DAS output. The noise estimation is carried out by applying a blocking matrix whose objective is to estimate the noise that is present in directions other than the proposed DOA by subtracting the delayed signals (basically, an anti-beamformer). The resulting signals are then filtered such that when summed together provide a noise estimation that is able to be subtracted directly from the DAS output. These filters are constantly optimized through time via least mean squares using the current estimated output and past noise estimations, resulting in adaptation to changes in the environment. The work presented in [110] is the only robot audition work we found that uses GSC as part of its multiple-DOA estimation method.

Another example is that of [113], where a frequency-based selection method is applied to the beamformer's output to remove the attenuated noise. The DAS-based steered-response spectrum is calculated, from which the highest-energy and the second-highest-energy sound source DOAs are obtained. The frequencies that are lower in the highest-energy beamformer's output than in the second-highest are filtered out. This filtered output is subtracted from the beamformers' output and the process is carried out again until only background noise is present. This results in high sensitivity to low energy sound sources which are not easily detectable with high energy interferences. However, the energy level of these low-energy sound sources is implicitly assumed to be higher than the background noise level.

An important drawback to using beamforming for multiple-DOA estimation is that it requires one beamformer per proposed DOA. Depending on the applied beamforming technique, a high amount of computational resources may be required to execute it [69].

4.3.2. Subspace methods

One of the most popular methods of this branch is multiple signal classification (MUSIC) [114]. It can be argued that it is the basis of all the other methodologies of this branch, thus its description is the main focus of this section.

The main concept behind MUSIC centers around the search of the DOAs that intersect the subspace that represents the signals of interest. Consider the captured signal model as presented in Eq. (21):

$$X = W_s S + V \tag{21}$$

where X is a $N \times F$ complex-value matrix that holds all of the X_n's in its rows; the X_n are the frequency-domain-transformed input discrete signals (x_n), each representing the captured signal at sensor n; S is a $D \times F$ complex-value matrix (where D is the number of source signals) that holds in its rows all of the source signals s_m's in the frequency domain; and V is a $N \times F$ complex-value matrix that holds in its rows a set of noise signals present at each sensor in the frequency domain. $W_s[f]$ is a $N \times D$ complex-value matrix which models the TDOAs ($\tau_{n:d}$) of each source signal (d) at each microphone (n) at a given frequency f, related to its DOA, as presented in (22):

$$W_s[f] = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{-2\pi f \tau_{2:1}} & e^{-2\pi f \tau_{2:2}} & \cdots & e^{-2\pi f \tau_{2:D}} \\ e^{-2\pi f \tau_{3:1}} & e^{-2\pi f \tau_{3:2}} & \cdots & e^{-2\pi f \tau_{3:D}} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-2\pi f \tau_{N:1}} & e^{-2\pi f \tau_{N:2}} & \cdots & e^{-2\pi f \tau_{N:D}} \end{bmatrix}. \tag{22}$$

It is important to mention that $W_s[f]$ is similar to the W_θ matrix of Eq. (15), but in this case it models the TDOAs of the received signals for each sound source. As it can be observed, this approach assumes the signal is narrowband, centered at f, which is not the case for speech. We will assume that the signal is narrowband for now, but examples are provided of broadband variations of MUSIC further on.

The objective of MUSIC is to estimate the subspace spanned by the columns of W_s. It carries this out by, first, performing the eigendecomposition of the $N \times N$ sample covariance matrix $\hat{R}[f]$ for the frequency f of the captured signals (calculated in the same manner as in the MVDR beamformer detailed in Section 4.3.1), as in Eq. (23):

$$\hat{R}[f] = Q[f] \Lambda[f] Q[f]^{-1} \tag{23}$$

where $\Lambda[f]$ is the $N \times N$ complex-value diagonal matrix holding the covariance's eigenvalues for the frequency f, sorted in descending order, as presented in Eq. (24):

$$\Lambda[f] = \begin{bmatrix} \lambda_1[f] & 0 & \cdots & 0 \\ 0 & \lambda_2[f] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N[f] \end{bmatrix} \tag{24}$$

$Q[f]$ is a $N \times N$ complex-value matrix which holds in its columns the set of the covariance's eigenvectors for frequency f sorted in the order established by Λ. It is presented in Eq. (25):

$$Q[f] = \begin{bmatrix} q_{1:1}[f] & q_{2:1}[f] & \cdots & q_{N:1}[f] \\ q_{1:2}[f] & q_{2:2}[f] & \cdots & q_{N:1}[f] \\ \vdots & \vdots & \ddots & \vdots \\ q_{1:N}[f] & q_{2:N}[f] & \cdots & q_{N:N}[f] \end{bmatrix} \tag{25}$$

$Q[f]$ can be divided into two subspaces, at the index λ_l, as shown in Eq. (26):

$$\begin{aligned} Q_s[f] &= Q[f][1 : \lambda_l] \\ Q_v[f] &= Q[f][\lambda_l + 1 : N] \end{aligned} \tag{26}$$

where $Q_s[f]$ is the $N \times \lambda_l$ matrix that holds in its columns the eigenvectors representing the signal subspace and $Q_v[f]$ is the $N \times (N - \lambda_l)$ matrix that holds in its columns the eigenvectors representing the noise subspace, both for frequency f. λ_l is typically set as the rank of $W_s[f]$. If this is difficult to estimate (as in noisy environments) λ_l can be set as the index of the eigenvalue that is the smallest of the set of high eigenvalues, meaning, those that are not close to zero. This follows the assumption that the eigenvalues of the noise subspace are considerably smaller than those in the signal subspace. Another approach is to whiten the noise before carrying out the eigendecomposition of the covariance matrix so that the eigenvalues of the noise subspace are values close to 1 and the eigenvalues of the signal subspace are greater than 1 [115,116]. This whitening is carried out by first estimating the $N \times N$ covariance matrix $K[f]$ of the noise $V[f]$ of Eq. (21) by capturing it when the sources are silent. Then, $K[f]$ is calculated by $K[f] = V[f]V[f]^H$ and the eigendecomposition presented in Eq. (27) is carried out:

$$K[f]^{-1}\hat{R}[f] = Q[f]\Lambda[f]Q[f]^{-1}. \tag{27}$$

Since $K[f]^{-1}$ can be any type of matrix that forces the $K[f]^{-1}\hat{R}[f]$ matrix to be square, this presents what is known as the generalized eigendecomposition (GEVD). When this decomposition is used as part of MUSIC (GEVD-MUSIC), it has been shown to be robust against non-correlated noise in the environment [115].

Having carried out the subspace division, a DOA search can be set in motion where several DOA candidates are proposed. To this effect, a $N \times 1$ vector $b[\theta, f]$ is calculated for each candidate θ, as in Eq. (28):

$$b[\theta, f] = \begin{bmatrix} 1 \\ e^{-2\pi f \tau_1(\theta)} \\ e^{-2\pi f \tau_2(\theta)} \\ \vdots \\ e^{-2\pi f \tau_N(\theta)} \end{bmatrix} \tag{28}$$

where N is the number of sensors; and, $\tau_n(\theta)$ is the TDOA of microphone n related to the candidate θ. The $\tau_n(\theta)$'s are calculated based on a given propagation model. As mentioned before, the free-field/far-field is the most commonly used, as is the case in this version of the methodology. But, in the same manner as the beamforming-based methodologies described in Section 4.3.1, other propagation models may be applied, such as the near-field model [45] and the spherical head model [43, pp. 349–361].

Because of the orthogonal nature between eigenvectors, when $b[\theta, f]$ is orthogonal to the eigenvectors in $Q_v[f]$, it represents one of the source signals in the signal subspace ($Q_s[f]$) and, thus points in the direction of a source signal. To test for this orthogonality, Eq. (29) is applied:

$$P_{MUSIC}[\theta, f] = \frac{1}{b[\theta, f]^H Q_v[f] Q_v[f]^H b[\theta, f]} \tag{29}$$

where the vector $P_{MUSIC}[f]$ is the MUSIC pseudo-spectrum at frequency f. The locations of its peaks represent the DOA of sound sources, similar to the steered-response spectrum of beamforming-based methods. $P_{MUSIC}[\theta, f]$ will not be defined when the denominator in Eq. (29) is 0, that is to say, when $b[\theta, f]$ is orthogonal to $Q_v[f]$. However, it is highly unusual for the noise to be completely non-correlated to the source signal, thus only values close to 0 are typically encountered. This does mean that the peaks in $P_{MUSIC}[f]$ are prone to have very large values.

It is important to mention that MUSIC requires at least one eigenvector spanning the noise subspace, if not $Q_v[f]$ is empty. Thus one essential requirement for MUSIC is that *there must be at most $N - 1$ sources present*. To work around this constraint, a microphone array with a high amount of microphones (8 or higher) is often employed (see Section 6.3.1 for more details).

Since MUSIC assumes that the source signal is narrowband, it is not directly applicable to speech signals which are broadband. One way to solve this issue is to calculate a narrowband MUSIC spectrum for each frequency f and propose their DOA-wise average as the broadband MUSIC spectrum [117,118]. Unfortunately, this results in requiring considerable computing resources to carry out MUSIC in real-time when using the noise-robust variation GEVD-MUSIC. To counter this, employing the generalized singular value decomposition (GSVD) instead of the generalized eigendecomposition (GEVD) provides the same orthogonality that is essential for MUSIC to be carried out, as shown in Eq. (30):

$$K[f]^{-1}\hat{R}[f] = Q_l[f]\Lambda[f]Q_r[f]^H \tag{30}$$

where $Q_l[f]$ and $Q_r[f]$ are left- and right-singular vectors for frequency f that are orthogonal to each other, as well as unitary. GSVD-MUSIC uses $Q_l[f]$ instead of $Q[f]$ from Eq. (27) onward. When it comes to computation cost, it is shown in [116] that using GSVD-MUSIC takes considerably less processing time than GEVD-MUSIC per time window. Additionally, in this same work, instead of carrying out a grid-search mapping throughout the whole set of possible DOAs, a coarse-search is carried out first and then a fine-search is carried out in the peaks of the coarse-search. The authors refer to this approach as Hierarchical SSL (H-SSL) and it can provide a resolution below 1° in real-time.

In [119], MUSIC's robustness against noisy environments is increased by (1) decontaminating the covariance matrix by rectifying noisy phases by linear regression between frequency bins and by (2) neglecting noisy frequency bins by carrying out subspace-based SNR estimation. It is important to remember that the phase data is present in the sample covariance matrix $\hat{R}[f]$. The proposed method uses a two-microphone array, which results in $X[f]$ only having two rows, thus, only one source can be located. However, as it is shown further on, this method can be generalized to a multiple-DOA estimator.

The row $X_1[f]$ is considered as the signal received at the reference microphone, thus, $X_2[f] = X_1[f]e^{-i2\pi f \tau}$. In this scenario, the phase data can be extracted from the covariance matrix as it is shown in Eq. (31):

$$\begin{aligned} \hat{R}[f] &= X[f]X[f]^H \\ &= \begin{bmatrix} X_1[f]X_1[f]^H & X_1[f]X_2[f]^H \\ X_2[f]X_1[f]^H & X_2[f]X_2[f]^H \end{bmatrix} \\ &= \begin{bmatrix} X_1[f]^2 & X_1[f]^2 e^{i2\pi f \tau} \\ X_1[f]^2 e^{-i2\pi f \tau} & X_1[f]^2 \end{bmatrix} \end{aligned} \tag{31}$$

where $\zeta(f,\tau) = e^{-i2\pi f\tau}$ represents the phase data, which is a type of mapping between f and τ. Traditionally, in a noise-less environment the relationship between these two is close to being linear. However, the authors found that this linearity is lost in noisy environments, which produces errors in the DOA estimations. To this effect, the authors propose to use the following procedure to rectify the phase data:

1. Calculate the sample covariance matrix $\hat{R}[f]$.
2. Extract the phase data $\zeta(f,\tau)$ from $\hat{R}[f]$ at each frequency f.
3. Regression stage (repeat until error reaches a pre-defined minimum):
   (a) Carry out a first-order regression of the phase data such that it approaches what the authors refer to as *phase line*. These are obtained by calculating the phase data of observed signals from different directions and should be straight in noiseless environments.
   (b) Update the phase data that is lower than its corresponding value in the estimated phase line.
   (c) Calculate the error between updated phase data and the estimated phase line.
4. Reconstruct the sample covariance matrix with the updated $\zeta(f,\tau)$.

Additionally, frequency bin selection is carried out based on the metric shown in Eq. (32):

$$\epsilon[f] = \log\left(\frac{\lambda_1\hat{R}[f]}{\lambda_2\hat{R}[f]}\right) \tag{32}$$

where $\lambda_1[f]$ is the eigenvalue of the first eigenvector of $\hat{R}[f]$ and $\lambda_2[f]$ is the eigenvalue of the second. Since in this case scenario only one source is able to be estimated, $\lambda_1[f]$ represents the eigenvalue of the signal subspace and $\lambda_2[f]$ represents that of the noise subspace. Thus, $\epsilon[f]$ represents the narrowband subspace signal-to-noise ratio (subspace SNR) of the frequency f from where $\hat{R}[f]$ is calculated. A pre-specified threshold can be applied to ignore the frequency bins with a low subspace SNR [119].

Although the authors do not specify it, it may be possible to generalize this method to use an array with more than two microphones (and carry out multiple-DOA estimation). The subspace SNR can be calculated by finding the ratio between the sum of the eigenvalues in the signal subspace and the sum of the eigenvalues of the noise subspace, as presented in Eq. (33):

$$\epsilon[f] = \log\left(\frac{\sum_{n=1}^{\lambda_l}\lambda_n[f]}{\sum_{n=\lambda_l+1}^{N}\lambda_n[f]}\right). \tag{33}$$

Further examples of the application of MUSIC as part of SSL for a service robot are found in [120–123].

4.3.3. DOA clustering through time

Because of the nature of this type of multiple direction-of-arrival estimators, it is important to first define what is meant by "time window", since it is an essential part of SSL methods based on DOA clustering through time (DOA-cluster). The techniques described in the previous sections assume that all of the information provided to the estimator is from a certain portion of time, defined by a past number of samples (aka time window). Given this, the estimators previously explained provide a multi-DOA result using only that data.

The techniques described here estimate a single[5] DOA throughout several time windows and group up the different DOAs into clusters, each representing a sound source. A good example of this is [15], where a DOA is assigned to a cluster of DOAs (called a "peak track") if it is close to its last DOA.

The authors of [69] carry out redundant single DOA estimation by calculating the TDOAs from the three pairs of a triangular array [124] using a variation of the GCC-PHAT method. By taking advantage of its redundancy measures as well as the non-overlapping nature of simultaneous speech, the single DOA estimator is able to estimate a single DOA from one source even in multiple-speaker situations. A DOA calculated from time window t may be of a different source from the one estimated from time window $t+1$. To this effect, a clustering method[6] based on radar tracking techniques [69] is applied to create tracks out of the consecutive estimated single DOAs, each track representing a sound source. At each time window, all the DOAs of each track are fed into a Kalman filter. The proposed system provides a multiple-DOA estimation by carrying out the DOA prediction of all tracks. A similar Kalman-based approach is used in [122], although just one source was located.

Another method, proposed by [126], carries out multiple-DOA estimation through a tracker based on a Gaussian mixture model. It uses a probability hypothesis density filter (GM-PHD) to constantly estimate the probability density function of all the possible positions of all the current speakers (referred to as the "intensity function"). At each time window, when being fed of new DOA estimations, the intensity function is "refreshed" by considering both the new estimations as well as the past ones. In addition, the authors also used a process to automatically introduce new speakers into the intensity function, referred to as a "birth process". Interestingly, this tracking system can also provide a rough approximation of the distance of each speaker by accompanying each DOA estimation with a distance component and evaluating its divergence through time.

In [127] a method is presented that uses an adaptive variation of the *K-Means++* clustering method. The *K-Means* algorithm is a clustering method that starts from a given set of centers. The *K-Means++* algorithm pre-chooses those centers given a certain set of probabilities calculated from the data [128]. The work in [127] proposes an adaptive version of the *K-Means++* algorithm that does not assume to know the number of clusters beforehand. It iteratively tests with different number of clusters, until the resulting cluster centers are close to the pre-chosen ones.

Learning-based mapping procedures can also be employed in this category of multiple-DOA estimators. For example, an expectation–maximization framework (described in Section 5.3) can be incorporated into the learning process, as seen in [129,130]. This has also been carried out for regression [51], manifold learning [33] and probabilistic models [131–133]. All these examples are mentioned elsewhere in this survey, but their multiple-DOA estimation capabilities are worth noticing, specially when considering that they employ binaural arrays.

Although DOA-cluster methods are not as frequently used as beamforming-based or subspace-based methods, they provide some advantages. As discussed in Sections 4.3.1 and 4.3.2, both types of methods require a high number of microphones for good SSL performance. However, if a clustering-based method relies solely on single-DOA estimations through time, only one sound source is estimated at each time window. This means that these techniques only require the amount of microphones employed by a single-DOA estimator, which is generally lower than the employed by the techniques in other multiple-DOA estimation categories. For example, the microphone-array used in the clustering-based multiple-DOA estimator in [69] employed only 3 microphones, while the subspace-based methods used in [19,134,135] and the

[5] A multi-DOA estimator can also be applied at each time window if need be.

[6] Its first iteration is based on a simple DOA distance from a cluster center [125].

beamforming-based methods used in [136–138] used arrays with 8 microphones or more. Another advantage is due to the use of filtering techniques in clustering-based methods (such as the Kalman filter in [69] or the PHD filter in [126]), which reduce the influence of "noisy" DOAs in the end result.

However, an important disadvantage of DOA-cluster methods is that they rely heavily in *several* DOA calculations through time (i.e., several time windows). As mentioned before, subspace-based and beamforming-based methods only require one time window to provide a complete multiple-DOA estimation. This means that DOA-cluster methods may suffer from low responsiveness compared to their counterparts. However, it is important to point out that the techniques that require the calculation of a covariance matrix (such as MUSIC and MVDR-based beamforming) only provide a *reliable* multi-DOA result after several time windows in which the covariance matrix is iteratively calculated.

It is important to mention that some of the grouping techniques employed by DOA-cluster methods are in fact tracking techniques. Examples of which are the Kalman filter and particle filtering, which are described in Section 5 in more detail.

4.4. Distance estimation

After the direction-of-arrival, distance is the remaining component of a sound source location. To this effect, a sound source is located inside the hyperbola determined by the TDOA between the captured signals from a pair of microphones. If several pairs of microphones are employed, the source can be located by calculating the intersection of the hyperbolic curves of each microphone pair in the array. This method is a variation of the triangulation method usually used for source localization in antenna-based systems but, instead of having the source inside the sensor array, the source is located outside. However, it has been shown that the dynamic range of the TDOA caused by distance variation is very small in far distances and that it is non-linear in close distances [18]. What this implies is that if a source is far away, any changes to its distance to the microphone array is not reported by a change in the TDOA, and thus neither by the intersections of the hyperbolic curves. This can produce a high amount of distance estimation errors. In addition, factors like timbre, loudness and reflections can affect the distance estimation much more severely than the DOA estimation [18].

An alternative approach can be used that is similar to the learning-based 2-dimensional DOA estimators described in Section 4.2. Features, such as the ones described in Section 3.2, can be extracted from captured signals of sound sources at different distances and be used as training data for distance estimation [139]. However, in the case of the IID feature, the variability of intensity between different sound sources is much greater than of the same sound source recorded at different distances. This means that the IID is not as sensitive to distance changes in the ranges relevant to robotic applications. To work around this issue, a correction factor based on the azimuth of the sound source can be used to account for such variability [139]. However, the same issues of elevation estimation via training (as described in Section 4.2) are still present in this type of learning-based mapping.

Another way to estimate the distance of the sound source is to take advantage of the robot's mobility and carry out the typical method of triangulation. The DOA of the sound source is estimated at different known positions in the environment and, with simple trigonometry, its distance can be estimated [32,53,140]. In [32], however, the distance estimations varied considerably (1 m in standard deviation).

If the array is large enough it can be divided in several spatially separated sub-arrays, each estimating a DOA, and the distance can be estimated by employing triangulation [141]. But, sound separation is required to achieve good performance. A more general way is to build an evidence grid of the environment based on the DOA of a sound source when the robot is in different positions [70,71]. A similar grid-based method is employed in [71,121], where a grid is built for the objective of getting close to the source.

Distance can also be estimated by employing optimization methods over a Cartesian plane [74]. The GCC-PHAT-based TDOAs can be used to propose a position model where a least-squares solution is found using Lagrange multipliers. By considering that the distance is a redundant variable, a linear correction can be carried out exploiting the least-squares solution in a second phase. Or like in [142], where a particle-filter-based tracker provides estimations mapped directly onto the Cartesian-plane.

The relative inter-microphone intensity difference (RIID) can be used as the input of a parameter-less self-organizing map (PL-SOM) such that one of its outputs is mapped to the value of the distance [143,144]. However, the performance is reduced dramatically when the source is in front of the array, since the RIID is reduced to a basic IID. As mentioned before, the IID is not sensitive to distance variation.

A useful phenomenon that can be used for distance estimation is how the source signal is reflected while propagating through the environment. When the distance of the sound source changes, the energy from the reflections (i.e., the reverberant diffuse sound field) is assumed to remain constant while the energy from the direct-path varies [145]. The ratio between these two energies is known as the direct-to-reverberant ratio (DRR), and is related to the distance of the sound source. Unfortunately, most methods that estimate the DRR require *a-priori* measurements of the room response. In [146], a DRR-estimation technique is proposed that relied on an equalization–cancellation method that, in turn, relied on DOA localization to appropriately select the signal samples with which to estimate the direct-path energy. The reverberant energy is calculated as the subtraction between the signal energy and the estimated direct-path energy. Unfortunately, in practice, the reverberant energy does not remain constant when the distance varies. To overcome this, a Gaussian mixture model (GMM) can be used to map the relationship between the DRR and the sound source distance [146], however this technique has not been carried out in a robotic platform.

It is important to mention that a similar concept to DRR is used to increase the robustness against reverberation of a DOA estimator in a robotic platform [16,18]. It relies on a generalized pattern of an impulse response in a reverberant environment for onset detection. However, this approach does not provide a distance estimation. This brings up an interesting tendency in the surveyed works. It is not unusual that features related to reverberation (DRR, reverberation time, room response, etc.) are extracted for the benefit of SSL. However, as presented in Section 6.2.2, these features are extracted mainly to increase robustness against reverberation for DOA estimation, not to carry out distance estimation. At the moment of this writing, it seems from the surveyed works that there has not been a robotic application where reverberation characteristics have been used as features for distance estimation. This tendency notwithstanding, we believe the previously described work in [146] could be a good step forward in that direction.

Learning based approaches can also be used for distance estimation. However, these systems do not estimate the distance directly but it is a byproduct of estimating the position in the Cartesian plane. The neural network architecture in [147] estimates the Cartesian coordinates using only one microphone. The hybrid deterministic/probabilistic method proposed in [131,132] is able to not only integrate visual information but to estimate a 3D position of the sound sources. In these approaches, the mapping procedures are based on a learned model similar to the 2-dimensional DOA estimation approach in [57] explained in Section 4.2.

5. Tracking

Most of the single-DOA and multi-DOA estimators described in Section 4 require the audio information contained in only one time window[7] to provide a DOA estimation result (exceptions to this are described in Section 4.3.3). The DOA estimation results of several time windows can be used as data with which a sound source can be tracked.

The tracking of a sound source uses this data to propose candidate scenarios in which it may be present. The tracking result is a location calculated using several single-window location estimations and a movement model. Additionally, it can provide a prediction of the source's movement in future time windows, depending on the movement model the tracking technique employs.

The tracking result can also be considered as a "refinement" or "filtered version" of the single-window location estimation, which is important in noisy environments and in presence of interferences. If the proposed movement model is robust enough, such interferences and/or noise sources can be considered as part of the environment. This not only improves SSL performance, but can also lead to the tracking of multiple sound sources.

Although there is a large amount of tracking techniques in literature, the ones usually employed for SSL in robot audition are based on:

- Kalman filtering
- Particle filtering.

However, there are other tracking techniques that are sparsely used which are worth mentioning and are briefly described in Section 5.3.

5.1. Kalman filters

A Kalman filter is a type of Bayes filter whose aim is to estimate the state of a system through noisy measurements over time, assuming that the noise satisfies a normal distribution [148]. Because of its consideration of noise in its calculations, it is frequently used for smoothing the trajectory of a mobile target [149].

A Kalman filter relies on a state-space paradigm which implies that: (1) time is assumed to be discrete and (2) a new state of the system is estimated at every time step. Although this would also imply an input–output system in which the inputs are being controlled by a known agent, this may not always be the case.

Generally, the Kalman filter carries out two stages to estimate the system state:

- **Predictor stage:** the system state is estimated based on previous states, as well as the current inputs being fed into the system (if there are any).
- **Corrector stage:** the estimated state is updated given the current observed measurements.

The output of the *corrector stage* is what is typically provided as the output of the Kalman filter. It is important to acknowledge the optional nature of the inputs during the predictor stage. As any typical state-space system, the input to the system (c) can be part of its model, as presented in Eq. (34):

$$\mathbf{p}_t = A\mathbf{p}_{t-1} + B\mathbf{c}_{t-1} + v_{t-1} \tag{34}$$

and the measurement is modeled as in Eq. (35):

$$z_t = H\mathbf{p}_t + u_t \tag{35}$$

[7] As explained in Section 4.3.3, a time window is a certain portion of time defined by a past number of samples.

where t is the time index; $\mathbf{p}_t$ is the state of the system at time t; c_t is the input vector received at time t; z_t is the measurement observed at time t; A is the transition matrix; H is the measurement matrix; and v_t and u_t are the process and measurement noise respectively at time t.

However, in the case of sound source tracking, the input vector c is rarely available since it would involve knowledge of the motivation behind the movement of the sound source. This leaves only the current system state to be used, eliminating the $B\mathbf{c}_{t-1}$ component in Eq. (34), resulting in Eq. (36):

$$\mathbf{p}_t = A\mathbf{p}_{t-1} + v_{t-1} \tag{36}$$

This, however, may produce stagnation issues in the system prediction, since there is no new information to be used except for the process noise. A possible way around this issue is to use past observations to calculate the difference of position that has been observed between time steps [150].

It is also worth mentioning that it is common to define the system state as a state vector $\mathbf{p}_t = [\theta_t, r_t, \dot{\theta}_t, \dot{r}_t]^T$ containing the Cartesian coordinates and velocity of the source. If only directional data is available (meaning, distance estimation was not carried out), to obtain these Cartesian coordinates it can be assumed that the sound source is on the unit circumference (with a radius of 1 m) [69].

Some important assumptions are that both noises (process and measurement) are zero-mean, independent of each other and are distributed normally with covariances R_v and R_u, as shown in Eq. (37):

$$\mathbf{v} \sim \mathcal{N}(0, R_v), \qquad \mathbf{u} \sim \mathcal{N}(0, R_u) \tag{37}$$

This greatly simplifies the equations used in both stages. Instead of using the whole probability density function (PDF) as in a typical Bayes filter, only means and covariances are used.

There are two different types of states that are used in the calculations in both predictor and corrector stages:

- **The *a-priori* estimate of the system:** $\hat{\mathbf{p}}_t^{\,-}$, calculated using all the information prior to time t.
- **The *a-posteriori* estimate of the system:** $\hat{\mathbf{p}}_t$, calculated at time t using the current measurement z_t.

The overall processing arch of the Kalman filter is as follow:

1. In the predictor stage, $\hat{\mathbf{p}}_{t-1}$ is used to calculate $\hat{\mathbf{p}}_t^{\,-}$.
2. In the subsequent corrector stage, $\hat{\mathbf{p}}_t^{\,-}$ is used to calculate $\hat{\mathbf{p}}_t$, which is provided as the result of the Kalman filter for the current time step t.

The manner in which $\hat{\mathbf{p}}_t^{\,-}$ is calculated is presented in Eq. (38):

$$\hat{\mathbf{p}}_t^- = A\hat{\mathbf{p}}_{t-1}. \tag{38}$$

While in the predictor stage, the covariance of the *a-priori* estimate error at time t ($R_{e_t}^-$) is predicted using Eq. (39):

$$R_{e_t}^- = AR_{e_{t-1}}A^T + R_v \tag{39}$$

where $R_{e_{t-1}}$ is the covariance of the *a-posteriori* estimate error at time $t-1$.

Then, in the corrector stage, the *a-posteriori* estimate (the output of the Kalman filter) is calculated using Eq. (40):

$$\hat{\mathbf{p}}_t = \hat{\mathbf{p}}_t^- + K(z_t - H\hat{\mathbf{p}}_t^-). \tag{40}$$

The term $(z_t - H\hat{\mathbf{p}}_t^-)$ is referred to as the *residual* between the estimated measurement $H\hat{\mathbf{p}}_t^-$ and the actual measurement z_t. The weighting of this residual is the main basis of which of these two measurements is "trusted" more. The weight of the residual

is K, which is usually referred to as the *filter gain*, and its value can fluctuate in the range of $[0, H^{-1}]$. K can be calculated using Eq. (41):

$$K = \frac{R_{e_t}^- H^T}{H R_{e_t}^- H^T + R_u} \tag{41}$$

where $R_{e_t}^-$ is the covariance of the *a-priori* estimate error e_t^- at time t, both calculated using Eq. (42):

$$R_{e_t}^- = e_t^- e_t^{-T}, \qquad e_t^- = \mathbf{p}_t - \hat{\mathbf{p}}_t^-. \tag{42}$$

As it can be seen, K is based on both the covariance of the *a-priori* estimate error $R_{e_t}^-$ and the covariance of the noise of the measurements R_u. Because of this, when the estimate error $R_{e_t}^-$ is low, the estimated measurement $H\hat{\mathbf{p}}_t^-$ is "trusted" more; when the measurement noise R_u is low, the actual measurement z_t is "trusted" more [149]. This balancing process makes the Kalman filter a very attractive technique for noisy environments.

Finally, while in the corrector stage, R_{e_t} is calculated via Eq. (43):

$$R_{e_t} = (I - KH) R_{e_t}^- \tag{43}$$

which is used to calculate the covariance of the *a-priori* estimate error in the following time step $t + 1$.

Given Eqs. (34) and (36), it can be deduced that Kalman filters are only applicable to linear systems. Unfortunately, many SSL systems are not linear. To this effect, the extended Kalman filter (EKF) can be used. Its assumption is that the system is differentiable and uses its Jacobian derivatives instead of transition matrices. This results in a linearization of the model around the current working point. Unfortunately, because of this linearization, the EKF does not assure an optimal solution in the same way the linear Kalman filter does. Regardless, EKF has been applied successfully for SSL in robot audition [74,151].

Another important notion is that there is an underlying assumption that all the data that is fed to a Kalman filter is of only one target. Thus, it cannot be used alone for multiple sound source tracking. However, Kalman-based multiple-DOA estimation can be carried out by clustering initial location estimations based on their similarity to past estimations and employing one Kalman filter per cluster, as in [69,122].

As mentioned before, a Kalman filter is a type of Bayes filter, but modified to be used with multivariate normal distributions. For this reason, SSL approaches that rely on a Bayes filter, such as [109,152], can be considered as part of this branch of tracking techniques.

5.2. Particle filters

A particle filter has a similar objective as a Kalman filter and can even be presented using the same state-space model. A particle filter does not assume that both the system and measurement models are linear nor that they have a Gaussian distribution [153]. It instead employs a sampling method to obtain an estimate of the probability distribution functions (PDF) of the system state. One sampling method that is widely used is the sequential importance sampling (SIS) algorithm. It is a type of Monte Carlo method that randomly generates a set of weighted samples (or *particles*) from a given probability density, which are then used to estimate the PDF of the system state. Because of this random sampling, the solution provided by a particle filter is not optimal. However, it is flexible in the types of systems it can be applied to.

The expected value of the posterior PDF of the system state is usually provided as the end result of the particle filter,[8] and its variance can be used as a type of confidence measurement.

[8] If the PDF is very sparse, the value of the bin with the highest probability can also be used.

The posterior PDF of the system state can be approximated by applying Eq. (44):

$$P(\mathbf{p}_t \mid z_{1:t}) \approx \sum_{n=1}^{N_p} w_t^n \delta(\mathbf{p}_t - \mathbf{p}_t^n) \tag{44}$$

where $P(\mathbf{p}_t \mid z_{1:t})$ is the posterior PDF of the system state $\mathbf{p}_t$ at time t; $\{\mathbf{p}_{0:t}^n, w_t^n\}_{n=1}^{N_p}$ is a set of randomly-chosen weighted samples known as *particles* that are used to characterize $P(\mathbf{p}_t \mid z_{1:t})$; the weights w_t^n are normalized such that $\sum_{n=1}^{N_p} w_t^n = 1$; N_p is the number of particles; and $\delta(i)$ is the Dirac function centered in i. The Dirac function formalizes w_t^n as the value of the bin $[\mathbf{p}_t - \mathbf{p}_t^n]$ of the PDF.

To calculate the weights w_t^n the SIS algorithm is used. An *importance density* is required to generate particles. The choice of this *importance density* is one of the most important design steps of a particle filter. Specifically, it can be used to minimize an issue where, after some iterations, only one particle has a non-negligible weight. This issue is known as the *degeneracy phenomenon* and, since the variance of the weights increases over time, such phenomenon is bound to happen. A frequently used *importance density* [154] is the prior PDF of the system state $P(\mathbf{p}_t \mid \mathbf{p}_{t-1}^n)$; weights can be calculated by applying it to the Bayes rule, as shown in Eq. (45):

$$w_t^n = w_{t-1}^n P(z_t \mid \mathbf{p}_t^n) \tag{45}$$

In addition, to reduce the *degeneracy phenomenon*, a supplementary step known as *resampling* is carried out to "refresh" the weight values whenever a threshold[9] is passed. This step involves generating a new set of particles from the newly calculated weighted ones $\{\mathbf{p}_{0:t}^n, w_t^n\}_{n=1}^{N_p}$ by discarding those that have small weights. To complete the set of N_p particles, new particles are generated from the ones that are left.

There are different ways to carry out the resampling step, but one of the most popular [155] is known as *systematic resampling*, where the new particles are copies of the old particles. As part of this process, the number of times that each old nth particle is copied (N_n) needs to be calculated at each iteration. To do this, an ordered set of N_p numbers is proposed as U, where its ith element is $U_i = U_1 + \frac{i-1}{N_p}$ and U_1 is a uniform-randomly chosen number in the range of $[0, N_p)$. Systematic resampling proposes the value of N_n as the number of elements in U whose value is inside the range $[\sum_{k=1}^{n-1} w_t^k, \sum_{k=1}^{n} w_t^k)$.

The complete set of steps to carry out this version of a particle filter at any time $t > 1$ is as follows:

1. Build the current *importance density* given the set of particles of time step $t - 1$.
2. Randomly choose N_p particles from the current *importance density*.
3. Calculate the weights of the particles w_t^n using Eq. (45).
4. Normalize w_t^n such that $\sum_{n=1}^{N_p} w_t^n = 1$.
5. Calculate the degeneracy of the weights.
6. If the degeneracy is below a threshold, resample using *systematic resampling*.
7. Estimate the posterior PDF of the system state $P(\mathbf{p}_t \mid z_{1:t})$ using Eq. (44).
8. Calculate the expected value and variance of $P(\mathbf{p}_t \mid z_{1:t})$ and present them as the tracking result with a confidence value.

As it can be gathered, particle filters are quite flexible in the types of models that can be employed, since $P(\mathbf{p}_t \mid z_{1:t})$ can be the PDF of any model. This is relevant for SSL, as several models have been proven useful:

[9] Usually measured as $1/\sum_{n=1}^{N_p} (w_t^n)^2$.

- Multiple simultaneous sound sources [120].
- Sound sources that cross each other by including an inertia effect [138,156].
- Human tracking by fusing different perceptual modalities, such as sound source localization (audio) and person visual detection (vision) [66,73,157].

Although the whole set of particles is acting as one Bayes filter (since its result is the estimation of a PDF), the calculation steps employed in a Bayes filter are actually carried out for each particle when updating its weights. Thus, it can be argued that in terms of computational complexity, a particle filter is actually implementing several Bayes filters concurrently (one per particle). Additionally, since the particles are being combined or copied during the resampling step, parallelization opportunities are limited [154]. This means that the high performance and model flexibility of particle filtering comes at the expense of it being computationally expensive [69]. This will have an impact in the resources that are left for other modules in the robot, such as navigation, vision, manipulation, planning, etc.

5.3. Other approaches

There are other approaches that are encountered in the literature that are worth mentioning that can be employed for tracking purposes.

Log-likelihood. In [70,71], the objective is to create a grid that represents the space surrounding the robot in order to provide in each cell the log-likelihood of a source being located there. This requires the robot to be listening and moving at the same time, as this auditory evidence grid is created using both audio and self-localization data. It has shown poor performance when having more than 2 sources present in its first iteration [70], but with an additional refinement during exploration, its performance can improve substantially [71].

Expectation–maximization. It is based on two iterative steps. The maximization step computes several parameters that maximizes the likelihood of the data given a pre-calculated expected value. The expectation step then calculates an expected value given the parameters of the last maximization step. This method can be used for tracking humans using both SSL and face localization [30].

Recurrent neural networks (RNN). They are a type of artificial neural networks that have dynamic internal states which serve as historical information that can be relevant for the recognition process. This makes them able to handle temporal signals proficiently. In the works of Murray et. al. [82,83], estimated DOAs (calculated via a cross-correlation technique) are fed into an RNN to estimate the next location of the source. The RNN represents the current and past locations of the source as internal states.

Random sample consensus (RANSAC) It is a type of model estimation algorithm that iteratively learns the parameters of a model by randomly sampling observed data, with the assumption that such data has "inliers" and "outliers". Via a voting scheme of several possible models, it assumes that the "outliers" do not vote in a consistent manner, while the "inliers" do. In [113], having acquired several DOAs (via beamforming) while the robot is moving, RANSAC is used to track three sources in the environment.

6. SSL facets

In the previous section, a review and classification of techniques used for SSL in robotics is presented. As a part of this work, the 188 surveyed techniques were compiled in an external Excel file where each column represents a different facet inherent to SSL. These facets are the dimensions of the SSL problem that the reviewed techniques are aiming to solve. In this section, each facet is described and discussed in detail.

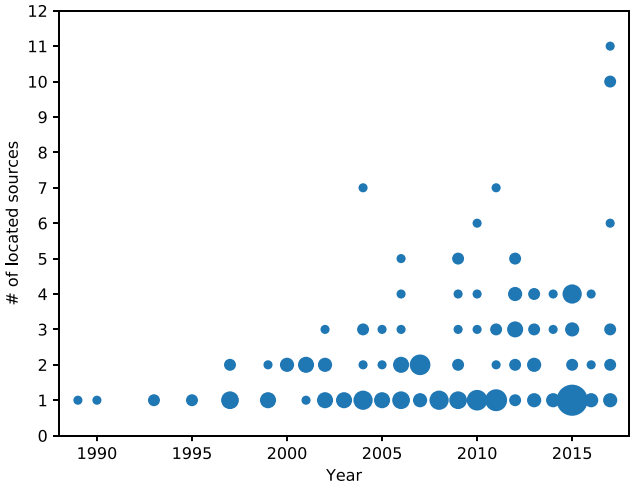


Fig. 5. Evolution of number of located sources in SSL.

6.1. Source characteristics

There are several characteristics that locatable sound sources possess, such as: number of sources, mobility of sources and distance of sources.

It is important to mention that most of the works discussed do not impose any constraints on the types of sound sources that are to be located. Given their application in a robot that interacts with humans, it is tempting to presume that such methodologies implicitly assume that the sound sources are human speech. There are some works that do constrain the type of sound sources to be of human origin, such as those that corroborate acoustic activity with mouth movements [158], carry out face recognition in parallel [135], or take advantage of speech characteristics for the benefit of SSL [68,69,159]. Moreover, there are works that aim to locate non-speech sources, such as cricket sounds [160] or generic broadband signals [161]. However, the reader is invited to assume that the reviewed works aim to locate any type of sound source.

6.1.1. Number of sources

There is a wide variety of number of sources located by the reviewed works. However, the vast majority aim to locate one source, such as [40,52,80,81,137,139,162]. Locating two sources is also common, as in the works presented in [44,71,98,134,140,163,164]. There are works that aim to locate 3 sources [115,120,141,152,165–167] and 4 sources [42,69,118,125,138], but it seems that this is the 'soft' limit for the robot audition community, since very few works carry out SSL for more than that amount. Exceptions to this limit are [113,117,127,168] which locate 5 and 6 sources, the work in [109,169] which locate up to 7 sources, and the work of [90,170] which locate 10 and 11 sources, the maximum we found in the literature.

As it is seen in Fig. 5, there is a high amount of works that locate a low amount of sources. The reverse is also apparent: there is a low amount of works that locate a high amount of sources. This is expected since increasing the number of sources also increases the complexity of the SSL technique. However, as it is also seen in Fig. 5, the amount of sources to be located has increased throughout the evolution of SSL in robotics.

It is important to mention that most of the reviewed works locating more than one source assume that the sources to be located are active simultaneously. This can be considered unnecessary since it is infrequent that two users interrupt each other in

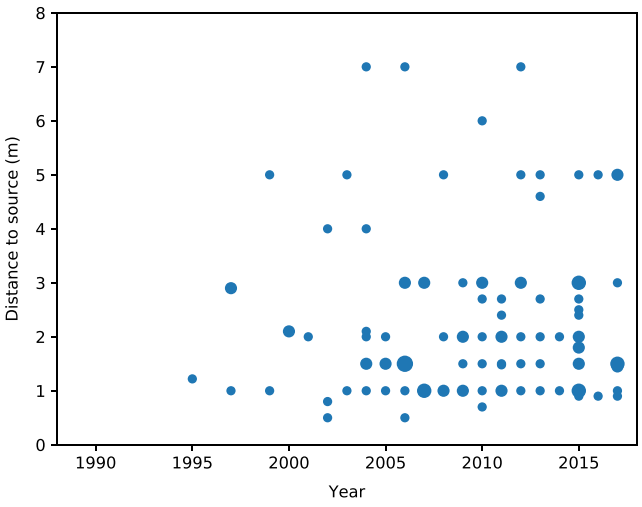


Fig. 6. Evolution of distance of sources in SSL.

a conversation [171]. However, it is frequent to have interfering sources from other conversations in settings such as restaurants or parties. In addition, users talking over each other is encouraged in circumstances such as having the robot act as a quizmaster in a trivia-type contest [42], take a food order from multiple clients [170] or just as an experimental setting [90].

6.1.2. Mobility of sources

In the reviewed works, the mobility of the sources is heavily considered, which has encouraged the use of well-known tracking techniques. Some of these techniques are described in Section 5. Examples of these efforts are presented in [73,109,113,115,120,138,172].

It is important to note that the mobility assumption is important since the robot audition module usually resides in a robot that is expected to move around in the environment. This implies that, even if the source is static, its relative position to the robot is dynamic while the robot is navigating or moving its robotic head.

6.1.3. Distance of sources

In Fig. 6, it can be seen that the majority of the works locate sound sources no farther away than 5 m from the microphone array. This approximately corresponds to the size of a room inside a house or an office.

As discussed in Section 4.4, there are several issues that arise when estimating the distance of the sound source by the analysis of sound signals alone. However, such information is useful in several acoustic scenarios where a robot is expected to perform. For example, sources that are too far away from the microphone array can be considered irrelevant to the interaction and thus ignored in the rest of the audio signal processing stages. Furthermore, using the source's distance and some clever manipulation of the room response, important characteristics of the environment can be extracted, such as the number of walls, the room size and shape, and the ceiling height [173]. To this effect, some works that aim to estimate the distance of the source are [18,74,139,140,143,144].

6.2. Environment

The characteristics of the environment play an important part of the complexity that SSL methodologies face. The ones that are more prevalent in the reviewed literature are: noise and reverberation.

6.2.1. Noise

We consider "noise" as audio data that is present in all microphones in a random manner. For the sake of this section, we differ "noise" from "interference", since the latter can be localized. In fact, an interference can be considered as a sound source in the environment, while there are other means that create noise in the input data that cannot be localized (such as electrical line current, residual environmental noise, etc.). Since noise is expected to happen in robot audition systems, it is important that SSL techniques are robust against it. Methods such as MUSIC (explained in Section 4.3.2) or the Kalman filter (detailed in Section 5.1) consider noise as part their signal models. Currently, examples of works that aim to tackle noise robustness are [69,71,117,140,141,166,174].

It is important to mention that the amount of noise varied greatly from one work to another: from an office-type noise (computer fans, inter-cubicle shatter, etc.) [69] to a very noisy exhibition hall [141].

6.2.2. Reverberation

Reverberation is the accumulation of reflections of the sound source signal from the physical elements in the environment, such as walls, ceilings, floor, even objects. A reflected signal has similar characteristics as the original sound source, having only suffered a decrease of energy (depending on the material from which it was reflected and the distance from the microphone). Because of this, such a reflection can be considered as an additional sound source in the environment. In practice, these reflections occur from all directions, which results in substantial decreases in performance for those techniques that are not robust against it, such as the original version of MUSIC. This can happen even with moderately reverberant environments.

As mentioned in Section 4.1, generalized cross-correlation with the phase transform (GCC-PHAT) provides a moderate amount of robustness against reverberation and is used in robotic-based works, such as [69,71,117,125,166,175].

Another reverberation-robust approach is presented in [16,18], where a SSL technique based on onset-detection is proposed. The pattern of an impulse response in a reverberant environment can be generalized into two parts: (1) the initial peak produced by the direct capture of the sound source (onset) and (2) a decreasing exponential decay (echoes). Since a signal has a higher energy when captured directly at its onset than its echoes, onset detection can be carried out by applying a threshold to the sound-to-echo ratio (similar to the direct-to-reverberant ratio of [146] discussed in Section 4.4). A feedback algorithm can be used to obtain these high sound-to-echo-ratio sections of the signal. This method shows a small amount of error (4°) while being able to locate two sources in practice. However, the generalized pattern requires to be calibrated to the room.

Another approach that aims to tackle reverberation-robustness is presented in [176], which is an extended version of the clustering-based method presented in [127] (detailed in Section 4.3.3). This extension is based on estimating the TDOA with a method that combines GCC-PHAT and MUSIC. The correlation vector is calculated only using the first principal component vector of the signal subspace, reducing the noise sensitivity of the reverberation-robust GCC-PHAT. Up to 6 sources are able to be located, but its performance decreases as the amount of sources increases: from an error of 2° with one source, to 18° with 6 sources.

A reverberation-robust approach is presented in [104] and discussed in Section 4.1. It carries out the SDD method and defines the signal subspace solely as the eigenvector with the largest eigenvalue. This proposal is based on the assumption that the direct-path component has the highest energy. Thus, it is understandable that this method locates one sound source at a time, however it shows a very small amount of error (1°).

A binaural particle-filter-based approach that is robust against reverberation is presented in [36]. The method is based on measuring the interaural coherence (IC), or linear dependency, between the two captured signals. The IC is used to ensure that a signal feature (such as TDOA or IID) is calculated only using the time–frequency bins of the captured signal that originated directly from the sound source and not from a reflection. When locating one source, its 3-dimensional localization has an average error of 0.157 m. However, permutation issues when calculating the joint multi-target probability density are present when locating two sources at the same time. The authors propose as future work to use more elaborate measures with simultaneous sources.

As it can be seen, these methods provide a high SSL performance in low to moderate reverberant rooms. This is important since their testing conditions were similar to the conditions where a robot is expected to perform in: houses, offices, etc. In addition, the amount of sources being located are similar to those discussed in Section 6.1.1, which is promising. Furthermore, SSL is carried out with no prior knowledge of the environment and small amounts of calibration. However, no distance estimation is carried out, which is possible by using reverberation information, as shown in [146] (discussed in Section 4.4).

6.3. Hardware

There are several aspects that the hardware side of the SSL problem include: capture equipment, number of microphones, array geometry and the robotic platforms used.

6.3.1. Capture equipment

There are two essential parts of the audio capture hardware system usually employed for SSL and audio signal processing in general.

Microphones. These convert the changes in air pressure caused by a sound source to an electric signal. There is an important tendency in the reviewed SSL works to use omnidirectional microphones. These aim to have a close-to-circular polar pattern in the azimuth plane. This tendency is understandable since a sound source can be located anywhere in the angular range of the microphone array and an omnidirectional microphone can capture it uniformly regardless of its direction.

Because of the nature of omnidirectional microphones, these tend to also capture signals with a substantial presence of noise and interferences. Thus, there are some efforts to use cardioid microphones which have a semi-circular polar pattern and can significantly reduce the presence of interferences located in the back of the microphone. However, this can cause issues if the source of interest is located there. Some works that use cardioid microphones are [28,50,84,177]. An important mention is the effort presented in [177] where a hybrid solution (using both omnidirectional and cardioid microphones) is used for SSL in a robot.

Audio interfaces. These convert the electrical signal captured by the microphones and modify it such that it can be analyzed and processed by a computer. This process includes digitization and amplification. In the earliest reviewed efforts, generic digital signal processors were adapted to be used with audio signals such as the cases of [9,178]; in other circumstances, they were built in-house [15,19]. In more recent efforts, off-the-shelf audio interfaces are more commonly used and a case can be made that this is the norm [47,70,179,180]. However, important endeavors to build many-microphone portable interfaces are worth mentioning, such as the 8SoundsUSB project [156], the EAR sensor [181] and the 32-microphone board presented in [182].

There is a wide variety of number of microphones used in the reviewed techniques, ranging from only just one microphone [49,183], up to 32 [113,140,184] and 64 microphones [185].

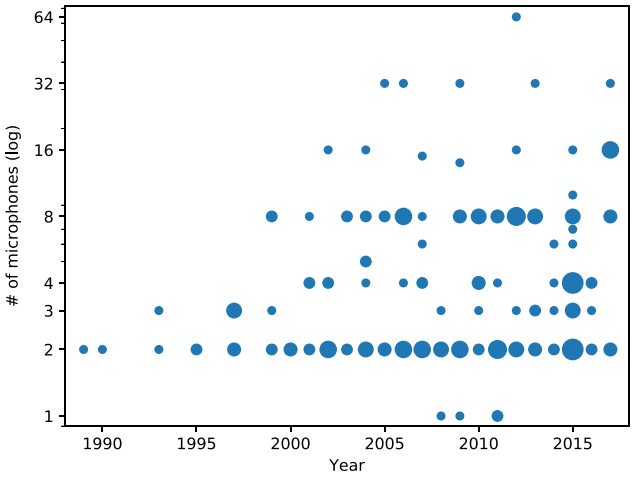


Fig. 7. Evolution of number of microphones in SSL (log scale in vertical axis).

However, as it can be seen in Fig. 7, the vast majority of the surveyed works aim to tackle SSL with a binaural approach (two microphones), such as [95,186–191]. In addition, there are also a considerable amount of works using eight microphones [42,109,115,118,120,127,134,152,165,192].

It is important to mention that there are approaches that use an intermediate amount of microphones, like three [69,75,124,125], four [40,70,71,122,193,194], five [74], six [73,104,141] and seven [106]. It is also important to mention that there are works that use a much larger amount of microphones, such as 14 [117], 15 [195] and 16 [87,158,166,196].

The motivation behind choosing how many microphones to employ in a SSL solution is highly dependent on the interests of the research group that is developing it. There is a wide range of research interests that motivate the community, thus it is important to discuss this topic in their context:

Human imitation. This is one of the main motivations (if not, *the* main motivation) behind the development of binaural methodologies, since humans use two ear canals for SSL. Many of these methodologies are accompanied by the use of external *pinnae*. These not only bring them closer to the human baseline, but are also useful in tackling the mirroring issue prevalent in binaural techniques [96,197] as well as estimating elevation [55,198–200]. In this same vein, there are other human-like approaches used for SSL, such as: multi-modal techniques that complement SSL via vision (which is close to what humans do) [54,201–204]; the use of epipolar geometry [205]; and moving the head to address the mirroring issue [206]. However, a case can be made that, because of their mechanical nature, the hair cells that are part of the inner ear sensory system can each be considered as a microphone [207,208]. In addition, a link has been found between the tactile sensory system and the auditory system in Macaque monkeys [209,210] which can mean that the human skin may also play a part in human audition abilities. This makes sense since sound is a change in air pressure and skin may be able to perceive it in a minor way. Thus, the human auditory system is more complex than two receptors and to imitate it requires a biological-based robotic development effort.

Austerity Some research groups aim to use the least amount of microphones possible. The amount of data analyzed is reduced in such cases, and thus the computational requirements can also be reduced. This is important since the robot audition modules are usually run in the same computer in parallel with other software modules, such as vision, navigation, etc. An extreme solution to the

austerity motivation is to use just one microphone. Current examples of 1-microphone SSL generally use learning-based mapping procedures, accompanied with the use of external *pinnae* and/or inner-ear canals. The 1-microphone approach presented in [183] can only locate the types of sources for which it is trained for, therefore its performance might be affected by unknown sources. A way around this issue is presented in [147], where a neural network is trained with recorded sound sources in known locations from different array positions. However, its distance estimation suffers from poor performance. A binaural array is a popular compromise for austerity, and some techniques are even able to carry out multiple-DOA estimation with it. Arrays of 3 and 4 microphones are able to carry out multiple-DOA estimation via clustering-through-time (described in Section 4.3.3), but suffer from poor response time. To this effect, a balance requires to be struck between hardware austerity (number of microphones) and software austerity (dataset size, response time, etc.). Such balance seems to be in the range between 2 and 8 microphones and it would seem that austere SSL approaches tend to employ a number of microphones in the lower side of that range.

Performance As discussed in Section 4, when carrying out multiple-DOA estimation it is important to consider the amount of microphones being employed, as it can be argued that it is directly correlated to SSL performance. As discussed in Section 4.3.3, beamforming-based and subspace-based methods tend to have better performance when using a large amount of microphones. Traditionally, the limiting factor for the amount of microphones to employ has been the space occupied by the body of the robot. Thus, what "large amount" means is up to the miniaturization techniques that are available at the time. In the case of condenser microphones,[10] the size of the microphone diaphragm is somewhat correlated to the microphone's sensitivity. Larger diaphragms tend to be more sensitive and do not require much amplification (i.e., have a high signal-to-noise ratio, SNR), resulting in the reduction of amplification noise that usually accompany small-diaphragm condenser microphones [211]. Thus, when using condenser microphones, a trade-off usually emerges. Using large-diaphragm condenser microphone array comes with a relatively high space requirement and high SNR. While using a small-diaphragm condenser microphone array (such as the 8SoundsUSB project [156]) comes with a low space requirement and relatively low SNR. More recently, microelectromechanical system (MEMS) microphones [212] have been gaining interest, primarily because of their small size. The first iterations of this technology provided moderately low SNR in comparison to condenser microphones [213]. However, current iterations (such as the STMicroelectronics MP33AB01H [214] and the InvenSense ICS-40618 [215]) have achieved SNR levels comparable to those of large-diaphragm condenser microphones. Thus, we suspect that this type of microphones will be used more and more in future SSL approaches, minimizing the impact of the issue of physical space in relation to performance, as done in [216–219]. Nonetheless, as it can be appreciated in Fig. 7, there is a high concentration of approaches that use 8 microphones, hinting that this may be the minimum that the community has established for a good performance.

6.3.2. Array geometry

There is a wide variety of array geometries used for SSL, for which we propose the following categorization:

- **Symmetric.** In this category, three sub-categories emerge, based on their dimensionality:
  - **1-dimensional.** They are also known as linear arrays, from which the most popular are the binaural arrays [220,221]. Linear arrays with more than two microphones can also be used in a robot for SSL [48,222].
  - **2-dimensional.** These include triangular [223,224], square [40,225] and circular [142,158] geometries.
  - **3-dimensional.** These include cubic [109,226], column-based [169,176], pyramidal [141], hemispherical [219] and spherical [104,185] geometries.
- **Irregular.** In this category, the microphones are scattered throughout the robot's body [136,227,228] and, in some cases, the array does not even have a static structure, such as:
  - A reconfigurable microphone array geometry that optimizes its performance dynamically [229].
  - A robot that moves its microphones to test its SSL performance [230].
  - A robot that accounts for its movements (and changes in array geometry) to enhance its SSL performance [231].
  - The impact to SSL performance of head rotation and limb movements of a humanoid robot with an array distributed over its body [106].
  - A hose-shaped robot that estimates its pose (and its array geometry) by reversing the SSL problem [232].

The most used array geometry is the binaural array. Between the symmetrical and irregular arrays, the symmetrical are much more popular since they are simpler to configure.[11] However, it is important to mention that irregular arrays present a more integrated solution to the robot, since the robot's physicality defines its array complexity not the other way around.

6.3.3. Robotic platforms

There is a wide variety of robotic platforms with a SSL system installed. We propose the following categorization of such robots, based on their physical appearance as well as their overall functionality:

- **Robotic heads.** These systems present only the head [51,56] or the upper torso of a humanoid, like Cog [8–10] or SIG [30,137,206]. They usually involve the use of a binaural array, each microphone in place of an ear on each side of the head. An important exception to this is the SIG2 robotic head which employs an additional binaural array inside the head for ego-noise cancellation [98,172]. In these cases, it is common that HRTFs [95,233] or epipolar geometry [44] are used to overcome the breaking of the free-field assumption.
- **Mobile bases.** These systems are usually used for outdoor/urban navigation and/or mapping [234]. They consist of a mobile robotic base with a microphone array installed [136,227,235]. In this category, unmanned aerial vehicles [216,218,219] (UAV, which are equivalent to flying mobile bases) are included. They employ microphone arrays for the purpose of carrying out SSL in outdoor environments [53,236,237].
- **Service robots.** These systems employ several functionalities in conjunction (vision, manipulation, navigation, speech recognition, etc.) for the benefit of servicing a human. There are complete service robotic systems that carry out SSL as part of these functionalities, such as: ASIMO [238], Jijo-2 [19], Golem [239,240], Hadaly and

[10] Only a small amount of approaches report the type of microphones employed but, from the ones that do, it seems that condenser microphones are a popular choice.

[11] Basically, configure one half and multiply by -1 the other half.

Hadaly-2 [22,241], Spartacus [138,228], HRP-2 [135], RobotAssist [159], HEARBO [158,242] and the works presented in [39,40,243]. It is also important to mention that the RoboCup@Home service robotics competition has been evaluating the SSL functionality in the participating robots since 2015 [244,245], thus such service robots are included in this category.

- **Complete commercial systems.** These systems are commercially available off-the-shelf and include SSL as one of their features. Examples of this category are: NAO [106,230,231,246], AIBO [143,144] and Paro [37].

6.4. Software frameworks

Complete SSL software frameworks are those that are provided to the community so that it can be downloaded, configured and employed directly into a robotic system with a given microphone array. As of this writing, there are two major SSL frameworks reportedly available for robotic systems: HARK [165,170] and ManyEars [156]. Both provide other robot audition functionalities as well.

It is important to emphasize that these frameworks are the ones that have been reported in the literature, but should not be considered as the only ones available. For instance, a SSL framework could have been applied but not reported; or only part of the framework is presented. Other efforts choose to report the low-level audio capture architectures such as: the JACK Connection Toolkit [247] in [69,240]; and RtAudio [248] in [66] and the ManyEars framework [156]. However, this is an important issue in the SSL literature: SSL software is seldom reported and rarely openly provided to the community.

7. Evaluation methodologies

SSL researchers use the experimentation–evaluation paradigm to report the progress in the field. Generally, when evaluating performances, the SSL system is considered as one whole unit, which includes the end-to-end methodology and hardware configuration. Moreover, most evaluations focus on three levels.

In the first level, the evaluation aims to characterize the performance of the SSL system for a certain facet. That is to say, it aims to measure the precision of the estimated positions. In this level, the evaluation is *one-to-many*, in the sense that one approach is tested under many acoustic environments. One can vary the position of the source [65], its distance [139], the SNR [226] or the environmental conditions [117].

In the second level, the evaluation aims to establish a common ground for the comparison of different methodologies. Meaning, it aims to compare system *A* versus system *B*. In this level, the evaluation is *many-to-many*, in the sense that many approaches are tested under many acoustic settings. The challenge in these evaluations is that all of the approaches have to "experience" the same input. One option is to use sound speakers as sources and play recordings through them using the same configuration [65]. A second option is to use databases of recordings [69]. And a third option is to use a simulation of sources and acoustic environments [231].

In the third and final level, the evaluation aims to measure the impact of a SSL system on an specific task. For example, it can aim to answer: "What is the benefit of locating a sound source in a waiter robot?" In this level, measuring the impact of the SSL system in a task is not straightforward. The best option is an *on/off* evaluation, in the sense that a task performed by the robot is tested with and without using the SSL functionality. However, since the tasks heavily depend on SSL, it is almost impossible to compare both runs. In addition, this type of evaluation only measures the impact of the functionality in a task; more subtle aspects of the interaction are not explored. Because of this, other strategies are used for evaluation at this level, such as measuring specific facets of the SSL system during the task [40], reporting if the robot completed the task or not [165], or carrying out user questionnaires as the evaluation metric [240].

The evaluation methodologies rely on metrics to measure the performance of the SSL system. The following are some of the common metrics we have found that are used in the field:

Average error. It measures the error of the estimation. This metric is commonly used for azimuth [187], elevation [47] and distance [140]. A set of estimations are compared against the true position of the source. The average difference between both is reported. There is the *discrete* version of this metric, in which a final stable estimation is compared with the source position [187]. The other is the *continuous* version, in which a stream of estimations through time are compared against a stream of true positions [239]. In particular, this second version is used when the robot or the source are mobile. Other similar metrics are *absolute error* [178], *maximum error* [18], *mean square error* [74] and *root mean square error* [66]. Other works also include the standard deviation of the error [238].

Average accuracy. This is the complement of the *error* in the sense that it measures the correctness of the estimations [81].

Correct detection. This metric is commonly used for DOA estimations, although it can be adapted to be used with distance estimations. It measures how many times the detection is correct, deemed so if it is within a given range of error from the true position of the source. Discrete [95] and continuous [230] versions of this metric are found in the literature.

Precision, recall and F_1-score. These metrics are based upon the *correct detections* metric. Although less commonly used, these metrics are able to measure the recovery capabilities of the SSL system. If a position is correct it is considered a *true positive*; if it is incorrect, a *false positive*; and, if no source is detected but it should have, it is considered a *false negative*. These three cases can be used to calculate the *precision* and *recall* metrics [181,237]. The *precision* metric measures how well the system is at predicting true positives. The *recall* metric compares between the predicted true positives and those that should have been predicted. These metrics, in turn, are used to calculate the F_1-score [249], which provides a balance between the two.

Number of sources. The previous metrics can be modified to consider the estimation of the number of sources instead of the estimation of their location. In this manner, these metrics measure the capability of counting the sources in an acoustic environment, rather than the quality of the estimated positions [176,239].

Word error rate (WER). It is a metric for speech recognition and it is used to indirectly measure the quality of SSL. Some sound source separation techniques depend on an accurate sound source localization. Therefore, a good SSL performance has a positive impact in separation quality, resulting in an intelligible signal that can be evaluated using speech recognition [54].

In addition to the previous metrics, it is common to find graphs to illustrate the performance of the system. We have identified two types of graphs.

Characterization graphs plot the full range of values for a variable. The simplest characterization graph plots the estimated variable versus the real value variable. In the case of DOA estimation, this can be the estimated angle versus the ground-truth angle (for both azimuth or elevation [27]). The characterization graph of an ideal SSL system plots a diagonal. Additionally, it is possible to characterize the performance of the system by plotting

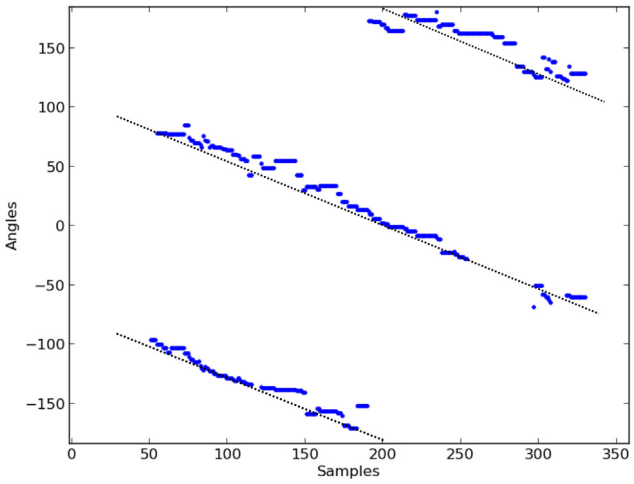


Fig. 8. Example of a graph path, taken from [69].

different variables against each other, such as: estimated angle vs error [30]; estimated angle vs. distance [139]; correct localization vs. SNR [118]; WER vs. SNR [165]; RMSE vs. number of sources [176]; average error vs. number of speakers [168]; and F_1-score vs. SNR [230]. Although usually these graphs are given in Cartesian coordinates, there are versions that use polar coordinates [168]. All these graphs have the goal to show the performance of the system under different circumstances and they are used to show its robustness or disadvantages.

The second type of graph are track graphs that show the localization of the sources through time. In these graphs, the actual source location and the estimated source location are drawn over each other. This is effective at providing a good visualization of the system's performance when the robot or sources are mobile. An example of a graph path is shown in Fig. 8 where two mobile sources are moving around a robot [69]. The source true position is shown as dashed lines and the estimated source position is shown as scattered points. Additionally, it is possible to plot the track in a 2-dimensional map with an initial and a final point [142].

7.1. SSL datasets

For the purpose of *many-to-many* evaluations, different acoustic datasets can be used. In particular, we have identified eight datasets that are freely available and that are appropriate for SSL evaluation:

- **RWCP database**[12] [250] This is one of the first datasets collected for scene understanding. It contains positions of several types of audio sources which were moved using a mechanical device. Recordings were made using a lineal array (14 microphones) and semi-spherical array (54 microphones).
- **V16.3 audio visual corpus**[13] [251] This is a dataset of video and audio recordings acquired in meeting rooms with one or more speakers, using two 16-microphone arrays. The mobility of each speaker varies from one recording to another. In this dataset, the audio capture system is static.
- **CAVA dataset**[14] [252] This dataset consists of recordings using a helmet mounted in a human head. Audio was acquired using two microphones and video was acquired using a camera mounted directly on the helmet. The recorded scenes are of meetings with multiple mobile speakers and noise sources.
- **CAMIL dataset**[15] [33] This dataset consists of recordings of one source with a realistic dummy head equipped with a binaural array and mounted on a pan/tilt robotic neck. The source is static, but the recordings were acquired during and after head movements. The acoustic environment is that of an office.
- **2nd CHiME dataset**[16] [253] This dataset collected with the purpose of sound separation and speech recognition. It can be re-purposed for SSL since information of the speaker positions are available.
- **RAVEL corpus**[17] [254] In this dataset, the primary data consists of video and audio recordings using a binaural head. It includes scenes with static and mobile speakers. The audio recording system is static.
- **AVASM dataset**[18] [255] This dataset provides information of a position of a moving source in the visual perception field. The audio of these video recordings are from a dummy head equipped with a binaural array. The dummy head is static but the sound position of the source varies in the video recordings.
- **AIRA corpus**[19] [125] This dataset consists of audio recordings from a 3-microphone array. There are three settings: anechoic, office and hall. The scenes include mobile speakers and a mobile audio acquisition system installed over a robotic platform.

Unfortunately this list is quite short and we attribute that to the fact that the creation of a dataset is challenging, mainly because of the heterogeneous diversity of acquisition hardware. The configuration used in a robot can be different from the one used to gather the dataset in terms of: the array geometry, the number and/or type of microphones, the microphone spacing, the audio interface, etc. In addition, to collect a dataset is time consuming, however [256] presents a proposal to expedite its collection using several robots. Each robot carries a microphone array and an electronic speaker while moving around, in this way it can collect several recordings with varying positions in a realistic environment.

8. Challenges

In this section a discussion is put forward on the SSL challenges that are of interest to be solved by the robot audition community.

8.1. Increase resolution in real-time

The concept of performance is an over-arching issue that is in continuous improvement, but the manner in which it is improved heavily depends on the scenario, constraints and testing conditions. We believe that a way to improve performance in a generic

12 Information about RWCP is available from: http://research.nii.ac.jp/src/en/RWCP-SSD.html.

13 V16.3 is available from: http://www.glat.info/ma/av16.3/.

14 Extra information for CAVA: https://team.inria.fr/perception/cava/.

15 CAMIL is available from: http://perception.inrialpes.fr/~Deleforge/CAMIL_Dataset/.

16 2nd CHiME Challengue dataset is available from: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2013.

17 RAVEL is available from: http://perception.inrialpes.fr/datasets/Ravel/.

18 AVASM is available from: https://team.inria.fr/perception/the-avasm-dataset/.

19 AIRA is available from: http://golem.iimas.unam.mx/proj_viewer.php?lang=en{&}sec=projects{&}proj=aira.

manner is to increase the localization resolution. The surveyed approaches can currently provide a high-resolution result, but the size of the 3D search space (azimuth, elevation and distance) is too large to provide results in real-time. That is to say, the challenge is to carry out 3D localization with high resolution ($<1^\circ$ in angles, <1 cm in distance) in real-time. There are already some works that have tackled this issue in the azimuth plane. The authors of [118] proposed a fine high-resolution search seeded by the results of a broad search. In [116], a super-resolution technique helped increase the robustness of simultaneous speech recognition.

8.2. Dynamic acoustic environment

A typical acoustic environment in which a robot is expected to work is flooded by a level of dynamism that is seldom tackled in conjunction. Examples of issues that come with this dynamism are: ever-changing acoustic environments (noise and reverberation), multiple mobile sources, non-stationary mobile interferences, complications that arise from the robot's mobility, etc. Although there are many techniques that aim to solve these issues (as discussed in Section 6), these are usually tackled in isolation and/or in a controlled environment. The challenge is to carry out SSL in real-life scenarios. The following are examples of research currently tackling this challenge: the HARK framework [170,186] can be used in very noisy environments with multiple users [141]; the ManyEars framework [156] uses noise masks to improve its robustness in dynamic environments [68]; and, the lightweight SSL solution used in the Golem robot [69] is able to carry out SSL as part of a waiter-type task in a noisy RoboCup@Home arena [240].

8.3. Off-the-Shelf ssl

Since SSL is of interest in the robotics community, it is also of interest to have SSL solutions that are easily integrated to already established platforms. Unfortunately, the requirements for an off-the-shelf SSL solution are challenging.

To begin with, the microphone array geometry presents important obstacles. Since different robots have different bodies, the microphone array that would accompany them differs in the same manner. In addition, it is also of interest to quickly change the number of microphones, either to remove faulty ones or add more for performance improvements. This means that an off-the-shelf SSL solution should be able to be configured according to the robot's microphone array in a trivial manner.

Moreover, the spacing between microphones can change during the task execution, as in the case of the hose-shaped robot used in [232] and the UAV swarm solution used in [53]. Thus, it should also be able to handle online changes in the array geometry.

Examples of systems which aim to tackle this challenge are HARK [165] and ManyEars [156], which provide an end-to-end framework that is available in their respective websites. It is also important to mention that one-microphone approaches are also being used, such as [49,147,183,199], which should simplify the initial configuration of an off-the-shelf SSL solution.

8.4. Datasets for machine learning methodologies

As presented in Section 7.1, there are few datasets devoted to sound localization. The rise of machine learning methodologies, in particular of Deep Learning [257], has been in part possible because of the accessibility of large scale datasets. For instance, in the field of image recognition there exists ImageNet [258], a dataset with one million annotated images; in the field of speech recognition, datasets can contain hundreds or thousands of speech hours with their transcriptions [259]. A way to promote the progress of the field would be to have a vast collection of recordings of sound sources in various positions and in multiple settings. In fact, this effort could be expanded to other robot audition tasks, such as sound source separation and classification. Towards this goal, [256] proposes a methodology to collect this large scale dataset using robots.

8.5. Integration of SSL in robotic tasks

The SSL functionality can be very useful in the overall task of a robot. For example, the work in [42] uses SSL as part of a task in which the robot acts as a quizmaster, where users try to answer as fast as possible a trivia question given by the robot. The robot is the judge of not only who said what answer (which is solved by SSL), but also who answered first. Another example is presented in [40], where the robot plays 'hide and seek' using SSL to find the user. Robots, such as the one presented in [41], can use SSL as part of a teleoperation task by complementing the representation of the scene to the operating user. Service robots for elderly care, as in [39], use SSL to complement the interaction with the user. Other examples of the applications of SSL in robotic tasks can be: the robot playing 'Marco Polo' with the user, taking attendance in a class and following a person that it has lost visually [38].

In all the aforementioned examples, SSL contributed to the overall behavior. This contribution can be expanded to other types of robots such as social, rescue, cognitive, industrial, as well as to many other types of service robotics tasks.

9. Discussion

In the last the three decades, there has been a lot of progress in the SSL field and it has become clear that there are several parts of the SSL problem that have been addressed profoundly from different vantage points. A clear example is the fact that robustness against noise and reverberation, a problem that was deemed near impossible half a century ago, is now being tackled in a regular basis by the community (see Sections 6.2.1 and 6.2.2). Moreover, multiple-DOA estimation and tracking techniques are now able to perform well in circumstances that are in par with typical acoustic environments. This, in turn, has provided SSL results that are more than suitable to be used in the next processing stages of a robot audition system (see Sections 4.3 and 5). Furthermore, even though there is still work to be done to achieve a good distance estimation relying solely in audio data, the current tendencies indicate that this research question is close to being solved (see Section 4.4). Moreover, even though an actual standard to compare SSL systems at a framework level is still missing, the field has at its disposal many methodologies and metrics to evaluate the performance of SSL systems (see Section 7).

To push this progress further, new research questions need to be formed and addressed. We have identified three motivations that can guide the robot audition community in putting together these questions:

- **The quest for the ideal ear.** There is a wide variety of hardware configurations currently being used, but we suspect that there are plenty more hardware settings and acoustic scenarios to explore. For example, the omnidirectional microphone is still widely used even when there are several techniques in which its polar pattern is altered by using external *pinnae*. This suggests that other acquisition hardware models are worthy of being considered to be built for robot audition, such as MEMS microphone-arrays. Biology-based systems, in which audio data is acquired using more than one type of sensor, might be another source of inspiration for these new models.

- **More extreme acoustic situations.** Currently, SSL scenarios have been mostly focused on laboratory settings. However, as with any research field, new areas of application should also be investigated. For instance, in scenarios such as inside a cave or a badly acoustically-designed classroom, the reverberation is much more extreme than in a household. In autonomous warehouses and self-driving cars, the level of noise is expected to be much higher than in an office. In addition, the dynamism of the acoustic environment, which includes the number, mobility and intensity of sound sources, can be much more complex in a setting such as a cocktail party, restaurant, public transport hub and busy streets. Finally, there are types of sources other than human speech that can be localized, such as dog barks, security alarms, whistles, body movements, stride pacing, etc. These are all dimensions of a fertile landscape in which the SSL field can grow.
- **New horizon.** Although the goal of SSL is well bounded, the standard with which SSL techniques are measured by has been evolving. It can be argued that, as of this writing, we are reaching human equivalent capabilities in terms of SSL, and that some techniques have the potential to go even beyond these limits. The question now is what is the new horizon of what we want to achieve in terms of, for example: an increasing number of located sources, 3-dimensional localization of the source, faster mobile sources, real-time processing, higher localization resolution, etc.

These motivations can fuel current and future research questions of SSL in the robotics field. We believe that answering these questions will not only benefit the field on its own, but will have an impact on the later processing stages of robot audition. In addition, it will widen the range of robotic systems where SSL can be applied, as well as open the door to richer types of robotic tasks.

Since 1989, when the Squirt robot was given a way to locate the sound sources surrounding it, the community has made steady progress to provide answers to SSL challenges. There are still frontiers to conquer in the field but, as it can be seen from this survey, these are well within the community's grasp and will surely motivate more progress that will transcend the field, improving robot audition in general.

Acknowledgments

The authors thank the support of CONACYT through the projects 81965, 178673 and 251319, PAPIIT-UNAM through the project IN107513 and ICYTDF through the project PICCO12-024. The authors would also like to thank Dr. Gibran Fuentes for his input in the description of tracking techniques, as well as Carmen Valle for her editorial support.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.robot.2017.07.011.

References

[1] S. Argentieri, A. Portello, M. Bernard, P. Danès, B. Gas, Binaural systems in robotics, in: J. Blauert (Ed.), The Technology of Binaural Listening, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 225–253.
[2] S. Argentieri, P. Danès, P. Souères, A survey on sound source localization in robotics: From binaural to array processing methods, Comput. Speech Lang. 34 (1) (2015) 87–112.
[3] L. Xiaofei, L. Hong, A survey of sound source localization for robot audition, CAAI Trans. Intell. Syst. 7 (1) (2012) 9–20.
[4] K. Nakadai, K. Nakamura, Sound source localization and separation, Wiley Encyclopedia of Electrical and Electronics Engineering.
[5] H.G. Okuno, K. Nakadai, Robot audition: Its rise and perspectives, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2015, pp. 5610–5614.
[6] A.M. Flynn, R.A. Brooks, W.M. Wells III, D.S. Barrett, Squirt: The prototypical mobile robot for autonomous graduate students, Tech. rep., DTIC Document, 1989.
[7] R.A. Brooks, Elephants don't play chess, Robot. Auton. Syst. 6 (1) (1990) 3–15.
[8] R.A. Brooks, L.A. Stein, Building brains for bodies, Auton. Robots 1 (1) (1994) 7–25.
[9] R.E. Irie, Robust sound localization: An application of an auditory perception system for a humanoid robot, Ph.D. thesis, MIT, 1995.
[10] R.E. Irie, Multimodal sensory integration for localization in a humanoid robot, in: Proceedings of IJCAI Workshop on Computational Auditory Scene Analysis, (CASA), Morgan Kaufmann Publishers, Inc., 1997, pp. 54–58.
[11] R.A. Brooks, C. Breazeal, M. Marjanović, B. Scassellati, M.M. Williamson, The cog project: Building a humanoid robot, in: Proceedings of Computation for Metaphors, Analogy, and Agents, Springer, 1999, pp. 52–87.
[12] A. Takanishi, S. Masukawa, Y. Mori, T. Ogawa, Study on anthropomorphic auditory robot continuous localization of a sound source in horizontal plane, in: Proceedings of Japan Robot Society Arts and Science Lecture Series, RSJ, 1993, pp. 793–796, (in Japanese).
[13] A. Takanishi, S. Masukawa, Y. Mori, T. Ogawa, Development of an anthropomorphic auditory robot that localizes a sound direction, Bull. Centre Inform. 20 (1995) 24–32.
[14] K. Nagashima, T. Yoshiike, A. Konno, M. Inaba, H. Inoue, Attention-based interaction between human and the robot chiye, in: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN, 1997, pp. 100–105.
[15] J. Huang, N. Ohnishi, N. Sugie, Building ears for robots: Sound localization and separation, Artif. Life Robot. 1 (4) (1997) 157–163.
[16] J. Huang, N. Ohnishi, N. Sugie, Sound localization in reverberant environment based on the model of the precedence effect, IEEE Trans. Instrum. Meas. 46 (4) (1997) 842–846.
[17] F. Wang, Y. Takeuchi, N. Ohnishi, N. Sugie, A mobile robot with active localization and disrimination of a sound source, J. Robot. Soc. Jpn. 15 (1997) 61–67.
[18] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, N. Sugie, A model-based sound localization system and its application to robot navigation, Robot. Auton. Syst. 27 (4) (1999) 199–209.
[19] F. Asono, H. Asoh, T. Matsui, Sound source localization and signal separation for office robot Jijo-2, in: Proceedings of IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems, MFI, 1999, 243–248.
[20] T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, N. Otsu, Integrated natural spoken dialogue system of "Jijo-2" mobile robot for office services, in: Proceedings of the National Conference on Artificial Intelligence and the Innovative Applications of Artificial Intelligence, American Association for Artificial Intelligence, 1999, pp. 621–627.
[21] Y. Matsusaka, T. Tojo, S. Kubota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, T. Kobayashi, Multi-person conversation via multi-modal interface - a robot who communicate with multi-user, in: Proceedings of European Conference on Speech Communication and Technology, EUROSPEECH, Vol. 99, 1999, pp. 1723–1726.
[22] S. Hashimoto, S. Narita, H. Kasahara, A. Takanishi, S. Sugano, K. Shirai, T. Kobayashi, H. Takanobu, T. Kurata, K. Fujiwara, et al., Humanoid robot - development of an information assistant robot Hadaly, in: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, (RO-MAN), IEEE, 1997, pp. 106–111.
[23] K. Nakadai, T. Lourens, H.G. Okuno, H. Kitano, Active audition for humanoid, in: Proceedings of National Conference on Artificial Intelligence, (AAAI), AAAI, 2000, pp. 832–839.
[24] H. Kitano, H.G. Okuno, K. Nakadai, T. Sabisch, T. Matsu, Design and architecture of sig the humanoid: an experimental platform for integrated perception in robocup humanoid challenge, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, Vol. 1, (IROS), IEEE, 2000, pp. 181–190.
[25] H.G. Okuno, K. Nakadai, T. Lourens, H. Kitano, Sound and visual tracking for humanoid robot, Appl. Intell. 20 (3) (2004) 253–266.
[26] M. Kumon, T. Shimoda, R. Kohzawa, I. Mizumoto, Z. Iwai, Audio servo for robotic systems with pinnae, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS), IEEE, 2005, pp. 1881–1886.
[27] F. Keyrouz, W. Maier, K. Diepold, A novel humanoid binaural 3d sound localization and separation algorithm, in: Proceedings of IEEE-RAS International Conference on Humanoid Robots, IEEE, 2006, pp. 296–301.
[28] V.M. Trifa, A. Koene, J. Morén, G. Cheng, Real-time acoustic source localization in noisy environments for human–robot multimodal interaction, in: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN, 2007, pp. 393–398.
[29] A. Kulaib, M. Al-Mualla, D. Vernon, 2d binaural sound localization: for urban search and rescue robotics, in: Proceedings of the International Conference on Climbing and Walking Robots, 2009, pp. 9–11.

[30] H.-D. Kim, K. Komatani, T. Ogata, H.G. Okuno, Human tracking system integrating sound and face localization using an expectation-maximization algorithm in real environments, Adv. Robot. 23 (6) (2009) 629–653.
[31] K. Youssef, S. Argentieri, J.-L. Zarader, Multimodal sound localization for humanoid robots based on visio-auditive learning, in: Proceeding of IEEE International Conference on Robotics and Biomimetics, ROBIO, 2011, pp. 2517–2522.
[32] A. Portello, P. Danès, S. Argentieri, Acoustic models and kalman filtering strategies for active binaural sound localization, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2011, pp. 137–142.
[33] A. Deleforge, R. Horaud, Learning the direction of a sound source using head motions and spectral features, Tech. rep. Institut National Polytechnique de Grenoble, 2011.
[34] A. Portello, P. Danès, S. Argentieri, Active binaural localization of intermittent moving sources in the presence of false measurements, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2012, pp. 3294–3299.
[35] B. Garcia, M. Bernard, S. Argentieri, B. Gas, S. Argentieri, A. Portello, M. Bernard, P. Danès, B. Gas, M. Bernard, et al., Sensorimotor learning of sound localization for an autonomous robot, in: Proceedings of EAA Congress on Acoustics, Forum Acusticum, Springer, 2012, pp. 188–198.
[36] I. Kossyk, M. Neumann, Z.-C. Marton, Binaural bearing only tracking of stationary sound sources in reverberant environment, in: Proceedings of IEEE-RAS International Conference on Humanoid Robots, IEEE, 2015, pp. 53–60.
[37] K. Wada, T. Shibata, T. Saito, K. Sakamoto, K. Tanie, Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, 2005, pp. 2785–2790.
[38] I. Meza, C. Rascon, G. Fuentes, L.A. Pineda, On indexicality, direction of arrival of sound sources, and human–robot interaction, J. Robot. (2016).
[39] H.M. Do, W. Sheng, M. Liu, An open platform of auditory perception for home service robots, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2015, pp. 6161–6166.
[40] H. Liu, M. Shen, Continuous sound source localization based on microphone array for mobile robots, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2010, pp. 4332–4339.
[41] A. Reveleau, F. Ferland, M. Labbe, D. Letourneau, F. Michaud, Visual representation of interaction force and sound source in a teleoperation user interface for a mobile robot, J. Hum.-Robot Inter. 4 (2) (2015) 1–23.
[42] I. Nishimuta, K. Itoyama, K. Yoshii, H.G. Okuno, Toward a quizmaster robot for speech-based multiparty interaction, Adv. Robot. 29 (18) (2015) 1205–1219.
[43] R.S. Woodworth, H. Schlosberg, Experimental Psychology, Oxford and IBH Publishing, 1954.
[44] K. Nakadai, H. Okuno, H. Kitano, Epipolar geometry based sound localization and extraction for humanoid audition, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, Vol. 3, 2001, pp. 1395–1401.
[45] J.C. Chen, K. Yao, R.E. Hudson, Acoustic source localization and beamforming: Theory and practice, EURASIP J. Adv. Signal Process. 2003 (4) (2003) 926837.
[46] M. Sato, A. Sugiyama, O. Hoshuyama, N. Yamashita, Y. Fujita, Near-field sound-source localization based on a signed binary code, IEICE Trans. Fundam. Electron. Commun. Comput. Sci. E88-A (8) (2005) 2078–2086.
[47] J. Valin, F. Michaud, J. Rouat, D. Letourneau, Robust sound source localization using a microphone array on a mobile robot, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, Vol. 2, 2003, pp. 1228–1233.
[48] S. Argentieri, P. Danès, P. Soueres, Modal analysis based beamforming for nearfield or farfield speaker localization in robotics, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2006, pp. 866–871.
[49] J.C. Murray, H.R. Erwin, A neural network classifier for notch filter classification of sound-source elevation in a mobile robot, in: Proceedings of International Joint Conference on Neural Networks, IJCNN, 2011, pp. 763–769.
[50] E. Saffari, A. Meghdari, B. Vazirnezhad, M. Alemi, Ava (a social robot): Design and performance of a robotic hearing apparatus, in: Social Robotics, Springer, 2015, pp. 440–450.
[51] A. Deleforge, F. Forbes, R. Horaud, Acoustic space learning for sound-source separation and localization on binaural manifolds, Int. J. Neural Syst. 25 (1) (2015) 1–19.
[52] J. Hornstein, M. Lopes, J.S. Victor, F. Lacerda, Sound localization for humanoid robots - building audio-motor maps based on the hrtf, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS), IEEE, 2006, pp. 1170–1176.
[53] S. Lana, K.N.K.N.H. Takahashi, T. Kinoshita, Consensus-based sound source localization using a swarm of micro-quadrocopters, in: Proceedings of the Conference of the Robotics Society of Japan, 2015, pp. 1–4.
[54] K. Nakadai, K. Hidai, H.G. Okuno, H. Kitano, Real-time speaker localization and speech separation by audio-visual integration, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, Vol. 1, 2002, pp. 1043–1049.
[55] T. Rodemann, G. Ince, F. Joublin, C. Goerick, Using binaural and spectral cues for azimuth and elevation localization, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2008, pp. 2185–2190.
[56] K. Youssef, S. Argentieri, J.L. Zarader, A learning-based approach to robust binaural sound localization, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2013, pp. 2927–2932.
[57] A. Deleforge, R. Horaud, Y.Y. Schechner, L. Girin, Co-localization of audio sources in images using binaural features and locally-linear regression, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (4) (2015) 718–731.
[58] Neobotix. Mp-500 - neobotix http://www.neobotix-robots.com/mobile-robot-mp-500.html [online, cited 25.05.17].
[59] D. Wang, G.J. Brown (Eds.), Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, IEEE Press/Wiley-Interscience, 2006.
[60] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. Acoust. Speech Signal Process. 24 (4) (1976) 320–327.
[61] M. Brandstein, H. Silverman, A robust method for speech signal time-delay estimation in reverberant rooms, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Vol. 1, 1997, pp. 375–378.
[62] J. Hassab, R. Boucher, Optimum estimation of time delay by a generalized correlator, IEEE Trans. Acoust. Speech Signal Process. 27 (4) (1979) 373–380.
[63] M. Omologo, P. Svaizer, Acoustic event localization using a crosspower-spectrum phase based technique, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Vol. 2, 1994, pp. II/273–II/276.
[64] B. Kwon, Y. Park, Y.-s. Park, Analysis of the GCC-PHAT technique for multiple sources, in: Proceedings of International Conference on Control Automation and Systems, ICCAS, 2010, pp. 2070–2073.
[65] U.-H. Kim, K. Nakadai, H. Okuno, Improved sound source localization in horizontal plane for binaural robot audition, Appl. Intell. 42 (1) (2015) 63–74.
[66] I. Markovic, I. Petrovic, Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering, Robot. Auton. Syst. 58 (11) (2010) 1185–1196.
[67] F. Grondin, F. Michaud, Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2015, pp. 6149–6154.
[68] F. Grondin, F. Michaud, Noise mask for tdoa sound source localization of speech on mobile robots in noisy environments, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, 2016, pp. 1–6.
[69] C. Rascon, G. Fuentes, I. Meza, Lightweight multi-DOA tracking of mobile speech sources, EURASIP J. Audio Speech Music Proces. (11) (2015).
[70] E. Martinson, A. Schultz, Auditory evidence grids, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2006, pp. 1139–1144.
[71] E. Martinson, A. Schultz, Robotic discovery of the auditory scene, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, 2007, pp. 435–440.
[72] E.B. Reuven, Y. Singer, Discriminative binaural sound localization, in: Proceedings of Advances in Neural Information Processing Systems, Vol. 15, 2002, pp. 1229–1236.
[73] R. Stiefelhagen, H. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, A. Waibel, Enabling multimodal human–robot interaction for the karlsruhe humanoid robot, IEEE Trans. Robot. 23 (5) (2007) 840–851.
[74] D. Bechler, M. Schlosser, K. Kroschel, System for robust 3d speaker tracking using microphone array measurements, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, Vol. 3, 2004, pp. 2117–2122.
[75] K.-C. Kwak, An enhanced multimodal sound localization with humanlike auditory system for intelligent service robots, Int. J. Latest Res. Sci. Technol. 2 (6) (2013) 26–31.
[76] G.I. Parisi, J. Bauer, E. Strahl, S. Wermter, A multi-modal approach for assistive humanoid robots, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2015, pp. 10–15.
[77] A. Clifford, J. Reiss, Calculating time delays of multiple active sources in live sound, in: Proceedings of Convention of the Audio Engineering Society, 2010, pp. 8157.1–8157.8.
[78] U.-H. Kim, T. Mizumoto, T. Ogata, H. Okuno, Improvement of speaker localization by considering multipath interference of sound wave for binaural robot audition, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2011, pp. 2910–2915.
[79] O. Faugeras, Three-Dimensional Computer Vision: A Geometric Viewpoint, MIT Press, Cambridge, MA, USA, 1993.
[80] J.C. Murray, H. Erwin, S. Wermter, Robotics sound-source localization and tracking using interaural time difference and cross-correlation, in: Proceedings of AI Workshop on NeuroBotics, 2004, pp. 89–97.

[81] J. Murray, S. Wermter, H. Erwin, Auditory robotic tracking of sound sources using hybrid cross-correlation and recurrent networks, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS), IEEE, 2005, pp. 3554–3559.

[82] J. Murray, S. Wermter, H. Erwin, Bioinspired auditory sound localisation for improving the signal to noise ratio of socially interactive robots, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2006, pp. 1206–1211.

[83] J.C. Murray, H.R. Erwin, S. Wermter, Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks, Neural Netw. 22 (2) (2009) 173–189.

[84] J. Liu, H. Erwin, S. Wermter, Mobile robot broadband sound localisation using a biologically inspired spiking neural network, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2008, pp. 2191–2196.

[85] R. Takeda, K. Komatani, Discriminative multiple sound source localization based on deep neural networks using independent location model, in: Proceedings of the IEEE Spoken Language Technology Workshop, (SLT), IEEE, 2016, pp. 603–609.

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

[87] N. Yalta, K. Nakadai, T. Ogata, Sound source localization using deep learning models, J. Robot. Mechatronics 29 (1) (2017) 37–48.

[88] L. Yu, S. Wang, K. Keung Lai, Testing of diversity strategy and ensemble strategy in svm-based multiagent ensemble learning, in: J. Mehnen, M. Koeppen, A. Saad, A. Tiwari (Eds.), Applications of Soft Computing: From Theory To Praxis, Springer-Verlag Berlin Heidelberg, Berlin, 2009, pp. 431–440 (Chapter 47).

[89] M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of Machine Learning, The MIT Press, US, 2012.

[90] K. Youssef, K. Itoyama, K. Yoshii, Simultaneous identification and localization of still and mobile speakers based on binaural robot audition, J. Robot. Mechatronics 29 (1) (2017) 59–71.

[91] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, MIT Press, 1997.

[92] G.S. Kendall, A 3-d sound primer: directional hearing and stereo reproduction, Comput. Music J. 19 (4) (1995) 23–46.

[93] C.I. Cheng, G.H. Wakefield, Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space, J. Audio Eng. Soc. 49 (4) (2001) 231–249.

[94] D. Pralong, S. Carlile, Generation and validation of virtual auditory space, in: Virtual Auditory Space: Generation and Applications, Springer, 1996, pp. 109–151.

[95] F. Keyrouz, Y. Naous, K. Diepold, A new method for binaural 3-d localization based on hrtfs, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Vol. 5, 2006, pp. V341–V344.

[96] S. Hwang, Y. Park, Y. sik Park, Sound direction estimation using an artificial ear for robots, Robot. Auton. Syst. 59 (3–4) (2011) 208–217.

[97] K. Nakadai, D. Matsuura, H. Okuno, H. Kitano, Applying scattering theory to robot audition system: robust sound source localization and extraction, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, Vol. 2, 2003, pp. 1147–1152.

[98] K. Nakadai, H.G. Okuno, H. Kitano, H.G. Okuno, H. Kitano, Real-time sound source localization and separation for robot audition, in: Proceedings of IEEE International Conference on Spoken Language Processing, ICSLP, 2002, pp. 193–196.

[99] L. Savioja, J. Huopaniemi, T. Lokki, R. Väänänen, Virtual environment simulation - advances in the DIVA project. in: Proceedings of the International Conference on Auditory Display, ICAD, 1997, pp. 43–46.

[100] Siemens, Lms sysnoise, 2016. https://www.plm.automation.siemens.com/en_us/products/lms/virtual-lab/legacy-applications/sysnoise.shtml.

[101] M. Reed, B. Simon, Methods of Modern Mathematical Physics III: Scattering Theory, Academic Press, Inc., 1979.

[102] P.D. Lax, R.S. Phillips, Scattering Theory, Vol. 26, Academic Press, 1990.

[103] D. Colton, R. Kress, Inverse Acoustic and Electromagnetic Scattering Theory, Vol. 93, Springer Science & Business Media, 2012.

[104] V. Tourbabin, B. Rafaely, Speaker localization by humanoid robots in reverberant environments, in: Proceedings of IEEE Convention of Electrical Electronics Engineers in Israel, IEEEI, 2014, pp. 1–5.

[105] Q. Wang, O. Ronneberger, H. Burkhardt, Fourier analysis in polar and spherical coordinates, Tech. rep., University of Freiburg, internal Report 1/08, 2008.

[106] V. Tourbabin, H. Barfuss, B. Rafaely, W. Kellermann, Enhanced robot audition by dynamic acoustic sensing in moving humanoids, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2015, pp. 5625–5629.

[107] A. Deleforge, Acoustic space mapping: A machine learning approach to sound source separation and localization, Ph.D. thesis, Universite de Grenoble, 2013.

[108] J. Capon, High-resolution frequency–wavenumber spectrum analysis, Proc. IEEE 57 (8) (1969) 1408–1418.

[109] J.-M. Valin, F. Michaud, B. Hadjou, J. Rouat, Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, Vol. 1, 2004, pp. 1033–1038.

[110] H. Lim, I.-C. Yoo, Y. Cho, D. Yook, Speaker localization in noisy environments using steered response voice power, IEEE Trans. Consum. Electron. 61 (1) (2015) 112–118.

[111] I.-C. Yoo, D. Yook, Robust voice activity detection using the spectral peaks of vowel sounds, ETRI J. 31 (4) (2009) 451–453.

[112] L. Griffiths, C. Jim, An alternative approach to linearly constrained adaptive beamforming, IEEE Trans. Antennas and Propagation 30 (1) (1982) 27–34.

[113] Y. Sasaki, S. Kagami, H. Mizoguchi, Multiple sound source mapping for a mobile robot by self-motion triangulation, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2006, pp. 380–385.

[114] R. Schmidt, Multiple emitter location and signal parameter estimation, IEEE Trans. Antennas and Propagation 34 (3) (1986) 276–280.

[115] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, H. Tsujino, Intelligent sound source localization for dynamic environments, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2009, pp. 664–669.

[116] K. Nakamura, K. Nakadai, H.G. Okuno, A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition, Adv. Robot. 27 (12) (2013) 933–945.

[117] C. Ishi, O. Chatot, H. Ishiguro, N. Hagita, Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2009, pp. 2027–2032.

[118] K. Nakamura, K. Nakadai, G. Ince, Real-time super-resolution sound source localization for robots, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2012, pp. 694–699.

[119] S.-C. Lee, B.-W. Chen, J.-F. Wang, M.-J. Liao, W. Ji, Subspace-based doa with linear phase approximation and frequency bin selection preprocessing for interactive robots in noisy environments, Comput. Speech Lang. 34 (1) (2015) 113–128.

[120] T. Otsuka, K. Nakadai, T. Ogata, H.G. Okuno, Bayesian extension of music for sound source localization and tracking, in: Proceedings of Annual Conference of the International Speech Communication Association, INTERSPEECH, 2011, pp. 3109–3112.

[121] E. Vincent, A. Sini, F. Charpille, Audio source localization by optimal control of a mobile robot, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2015, pp. 5630–5634.

[122] S. Pourmehr, J. Bruce, J. Wawerla, R.T. Vaughan, A sensor fusion framework for finding an hri partner in crowd, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2015, pp. 1–6.

[123] T. Suzuki, H. Otsuka, W. Akahori, Y. Bando, H.G. Okuno, Influence of different impulse response measurement signals on music-based sound source localization, J. Robot. Mechatronics 29 (1) (2017) 72–82.

[124] C. Rascon, H. Aviles, L. Pineda, Robotic orientation towards speaker for human–robot interaction, in: Lecture Notes in Advances in Artificial Intelligence, IBERAMIA, Vol. 6433, 2010, pp. 10–19.

[125] C. Rascon, L. Pineda, Multiple direction-of-arrival estimation for a mobile robotic platform with small hardware setup, in: H.K. Kim, S.-I. Ao, M.A. Amouzegar, B.B. Rieger (Eds.), IAENG Transactions on Engineering Technologies, in: Lecture Notes in Electrical Engineering, vol. 247, Springer, Netherlands, 2014, pp. 209–223.

[126] C. Evers, A. Moore, P. Naylor, J. Sheaffer, B. Rafaely, earing-only acoustic tracking of moving speakers for robot audition, in: Proceedings of IEEE International Conference on Digital Signal Processing, DSP, 2015, pp. 1206–1210.

[127] J.S. Hu, C.H. Yang, C.K. Wang, Estimation of sound source number and directions under a multi-source environment, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2009, pp. 181–186.

[128] D. Arthur, S. Vassilvitskii, K-means++: The advantages of careful seeding, in: Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms, (SODA), Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007, pp. 1027–1035.

[129] A. Deleforge, R. Horaud, 2d sound-source localization on the binaural manifold, in: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, (MLSP), IEEE, 2012, pp. 1–6.

[130] A. Deleforge, F. Forbes, R. Horaud, Variational em for binaural sound-source separation and localization, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), IEEE, 2013, pp. 76–80.

[131] J. Sanchez-Riera, X. Alameda-Pineda, J. Wienke, A. Deleforge, S. Arias, J. Čech, S. Wrede, R. Horaud, Online multimodal speaker detection for humanoid robots, in: Proceedings of IEEE-RAS International Conference on Humanoid Robots, IEEE, 2012, pp. 126–133.

[132] X. Alameda-Pineda, Egocentric audio-visual scene analysis. a machine learning and signal processing approach, Ph.D. thesis, Université Joseph-Fourier-Grenoble I, 2013.
[133] A. Deleforge, R. Horaud, The cocktail party robot: Sound source separation and localisation with an active binaural head, in: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, (HRI), ACM, 2012, pp. 431–438.
[134] F. Asano, M. Goto, K. Itou, H. Asoh, Real-time sound source localization and separation system and its application to automatic speech recognition, in: Proceedings of Annual Conference of the International Speech Communication Association, INTERSPEECH, 2001, pp. 1013–1016.
[135] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, K. Yamamoto, Robust speech interface based on audio and video information fusion for humanoid HRP-2, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, Vol. 3, 2004, pp. 2404–2410.
[136] L. Mattos, E. Grant, Passive sonar applications: target tracking and navigation of an autonomous robot, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, Vol. 5, 2004, pp. 4265–4270.
[137] M. Murase, S. Yamamoto, J.-M. Valin, K. Nakadai, K. Yamada, K. Komatani, T. Ogata, H.G. Okuno, Multiple moving speaker tracking by microphone array on mobile robot, in: Proceedings of Annual Conference of the International Speech Communication Association, INTERSPEECH, 2005, pp. 249–252.
[138] J.-M. Valin, F. Michaud, J. Rouat, Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering, Robot. Auton. Syst. 55 (3) (2007) 216–228.
[139] T. Rodemann, A study on distance estimation in binaural sound localization, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS), IEEE, 2010, pp. 425–430.
[140] Y. Tamai, Y. Sasaki, S. Kagami, H. Mizoguchi, Three ring microphone array for 3d sound localization and separation for mobile robot audition, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2005, pp. 4172–4177.
[141] Q. Nguyen, J. Choi, Selection of the closest sound source for robot auditory attention in multi-source scenarios, J. Intell. Robot. Syst. (2015) 1–13.
[142] J.M. Valin, F. Michaud, J. Rouat, Robust 3d localization and tracking of sound sources using beamforming and particle filtering, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Vol. 4, 2006, pp. IV–841–IV–844.
[143] E. Berglund, J. Sitte, G. Wyeth, Active audition using the parameter-less self-organising map, Auton. Robots 24 (4) (2008) 401–417.
[144] E. Berglund, J. Sitte, Sound source localisation through active audition, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2005, pp. 653–658.
[145] P. Zahorik, Direct-to-reverberant energy ratio sensitivity, J. Acoust. Soc. Am. 112 (5) (2002) 2110–2117.
[146] Y.C. Lu, M. Cooke, Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources, IEEE Trans. Audio Speech Lang. Process. 18 (7) (2010) 1793–1805.
[147] P. Kumarakulasingam, A. Agah, Neural network-based single sensor sound localization using a mobile robot, Intell. Autom. Soft Comput. 14 (1) (2008) 89–103.
[148] R.E. Kalman, A new approach to linear filtering and prediction problems, ASME J. Basic Eng. 82 (1) (1960) 35–45.
[149] G. Welch, G. Bishop, An introduction to the kalman filter, Tech. rep. University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1995.
[150] Z. Liang, X. Ma, X. Dai, Robust tracking of moving sound source using multiple model Kalman filter, Appl. Acoust. 69 (12) (2008) 1350–1355.
[151] S. Ogiso, T. Kawagishi, K. Mizutani, N. Wakatsuki, K. Zempo, Self-localization method for mobile robot using acoustic beacons, Robomech. J. 2 (12) (2015).
[152] Y. Bando, T. Otsuka, K. Itoyama, K. Yoshii, Y. Sasaki, S. Kagami, H. Okuno, Challenges in deploying a microphone array to localize and separate sound sources in real auditory scenes, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2015, pp. 723–727.
[153] N. Gordon, D. Salmond, A. Smith, Novel approach to nonlinear/non-gaussian bayesian state estimation, IEE Proc. F Radar Signal Process. 140 (2) (1993) 107–113.
[154] M. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking, IEEE Trans. Signal Process. 50 (2) (2002) 174–188.
[155] A. Doucet, A.M. Johansen, A tutorial on particle filtering and smoothing: Fifteen years later, in: Handbook of Nonlinear Filtering, Vol. 12, 2009, pp. 656–704.
[156] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, F. Michaud, The ManyEars open framework, Auton. Robots 34 (3) (2013) 217–232.
[157] H. Asoh, F. Asano, T. Yoshimura, K. Yamamoto, Y. Motomura, N. Ichimura, I. Hara, J. Ogata, An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion, in: Proceedings of Information Fusion, IF, 2004, pp. 805–812.
[158] R. Gomez, L. Ivanchuk, K. Nakamura, T. Mizumoto, K. Nakadai, Utilizing visual cues in robot audition for sound source discrimination in speech-based human–robot communication, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2015, pp. 4216–4222.
[159] B. Hilsenbeck, N. Kirchner, Listening for people: Exploiting the spectral structure of speech to robustly perceive the presence of people, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2011, pp. 2903–2909.
[160] A.D. Horchler, R.E. Reeve, B. Webb, R.D. Quinn, Robot phonotaxis in the wild: a biologically inspired approach to outdoor sound localization, Adv. Robot. 18 (8) (2004) 801–816.
[161] P. Danès, J. Bonnal, Information-theoretic detection of broadband sources in a coherent beamspace music scheme, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2010, pp. 1976–1981.
[162] R. Liu, Y. Wang, Azimuthal source localization using interaural coherence in a robotic dog: Modeling and application, Robotica 28 (7) (2010) 1013–1020.
[163] L. Calmes, G. Lakemeyer, H. Wagner, Azimuthal sound localization using coincidence of timing across frequency on a robotic platform, J. Acoust. Soc. Am. 121 (4) (2007) 2034–2048.
[164] S. Argentieri, P. Danès, Broadband variations of the music high-resolution method for sound source localization in robotics, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS), IEEE, 2007, pp. 2009–2014.
[165] K. Nakadai, T. Takahashi, H.G. Okuno, H. Nakajima, Y. Hasegawa, H. Tsujino, Design and implementation of robot audition system 'HARK' open source software for listening to three simultaneous speakers, Adv. Robot. 24 (5–6) (2010) 739–761.
[166] J. Even, N. Kallakuri, Y. Morales, C. Ishi, N. Hagita, Multi-modal sound localization from a mobile platform, in: JSAI Technical Report SIG-Challenge-B202–10, 2012, pp. 58–63.
[167] X. Alameda-Pineda, R. Horaud, Vision-guided robot hearing, Int. J. Robot. Res. 34 (4–5) (2015) 437–456.
[168] M. Đurković, Localization, tracking, and separation of sound sources for cognitive robots, Ph.D. thesis, Technische Universität München, 2012.
[169] J.S. Hu, C.Y. Chan, C.K. Wang, C.C. Wang, Simultaneous localization of mobile robot and multiple sound sources using microphone array, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, 2009, pp. 29–34.
[170] K. Nakadai, H.G. Okuno, T. Mizumoto, Development, deployment and applications of robot audition open source software hark, J. Robot. Mechatronics (2017) 16–25.
[171] T. Stivers, N.J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J.P. de Ruiter, K.-E. Yoon, S.C. Levinson, Universals and cultural variation in turn-taking in conversation, Proc. Natl. Acad. Sci. 106 (26) (2009) 10587–10592.
[172] K. Nakadai, T. Matsui, H.G. Okuno, H. Kitano, Active audition system and humanoid exterior design, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, Vol. 2, 2000, pp. 1453–1461.
[173] B. Günel, Room shape and size estimation using directional impulse response measurements, in: Proceedings of EAA Congress on Acoustics, Forum Acusticum, 2002, pp. 1–7.
[174] S. Argentieri, P. Danès, P. Souères, Prototyping filter-sum beamformers for sound source localization in mobile robotics, in: Proceedings of IEEE International Conference on Robotics and Automation, (ICRA), IEEE, 2005, pp. 3551–3556.
[175] T. Otsuka, K. Ishiguro, H. Sawada, H.G. Okuno, Unified auditory functions based on bayesian topic model, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2012, pp. 2370–2376.
[176] J.-S. Hu, C.-H. Yang, Estimation of sound source number and directions under a multisource reverberant environment, EURASIP J. Adv. Signal Process. (63) (2010).
[177] F. Asano, M. Morisawa, K. Kaneko, K. Yokoi, source localization using a single-point stereo microphone for robots, in: Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA, 2015, pp. 76–85.
[178] J. Huang, T. Supaongprapa, I. Terakura, N. Ohnishi, N. Sugie, Mobile robot and sound localization, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, Vol. 2, 1997, pp. 683–689.
[179] O. Deniz, J. Cabrera, M. Hernandez, Building a sound localization system for a robot head, Rev. Iberoam. Inteligencia Artif. 2003 (18) (2003) 17–24.
[180] H.s. Kim, J. Choi, Binaural sound localization based on sparse coding and som, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2009, pp. 2557–2562.
[181] V. Lunati, J. Manhès, P. Danès, A versatile system-on-a-programmable-chip for array processing and binaural robot audition, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2012, pp. 998–1003.
[182] Y. Sasaki, S. Masunaga, S. Thompson, S. Kagami, H. Mizoguchi, Sound localization and separation for mobile robot tele-operation by tri-concentric microphone array, J. Robot. Mechatronics 19 (3) (2007) 281.

[183] A. Saxena, A.Y. Ng, Learning sound location from a single microphone, in: Proceedings of IEEE International Conference on Robotics and Automation, (ICRA), IEEE Press, 2009, pp. 4310–4315.
[184] Y. Sasaki, N. Hatao, K. Yoshii, S. Kagami, Nested igmm recognition and multiple hypothesis tracking of moving sound sources for mobile robot audition, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2013, pp. 3930–3936.
[185] Y. Sasaki, M. Kabasawa, S. Thompson, S. Kagami, K. Oro, Spherical microphone array for spatial sound localization for a mobile robot, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2012, pp. 713–718.
[186] H. Okuno, K. Nakadai, Computational auditory scene analysis and its application to robot audition, in: Proceedings of Hands-Free Speech Communication and Microphone Arrays, HSCMA, 2008, pp. 124–127.
[187] K. Nakadai, H.G. Okuno, T. Laurens, H. Kitano, Humanoid active audition system, in: Proceedings of IEEE-RAS International Conference on Humanoid Robots, 2000, pp. 1–15.
[188] F. Keyrouz, W. Maier, K. Diepold, Robotic localization and separation of concurrent sound sources using self-splitting competitive learning, in: Proceedings of Computational Intelligence in Image and Signal Processing, CIISP, 2007, pp. 340–345.
[189] F. Keyrouz, W. Maier, K. Diepold, Robotic binaural localization and separation of more than two concurrent sound sources, in: Proceedings of Signal Processing and Its Applications, ISSPA, 2007, pp. 1–4.
[190] L. Calmes, H. Wagner, S. Schiffer, G. Lakemeyer, Combining sound localization and laser-based object recognition, in: Proceedings of AAAI Spring Symposium, 2007,pp. 1–6.
[191] L. Calmes, Biologically inspired binaural sound source localization and tracking for mobile robots, Ph.D. thesis, Aachen University 2009.
[192] K. Nakadai, H. Nakajima, M. Murase, S. Kaijiri, K. Yamada, T. Nakamura, Y. Hasegawa, H.G. Okuno, H. Tsujino, Robust tracking of multiple sound sources by spatial integration of room and robot microphone arrays, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 4, (ICASSP), IEEE, 2006, pp. IV929–IV932.
[193] Aldebaran Robotics, NAO Key Feature: Sound Source Localization.
[194] H. Li, T. Yosiara, Q. Zhao, T. Watanabe, J. Huang, A spatial sound localization system for mobile robots, in: Proceedings of IEEE Instrumentation and Measurement Technology Conference Proceedings, IMTC, 2007, pp. 1–6.
[195] S. Argentieri, P. Danès, Convex optimization and modal analysis for beamforming in robotics: Theoretical and implementation issues, in: Proceedings of European Signal Processing Conference, EUSIPCO, 2007, pp. 773–777.
[196] T. Nishiura, M. Nakamura, A. Lee, H. Saruwatari, K. Shikano, Talker tracking display on autonomous mobile robot with a moving microphone array, in: Proceedings of the International Conference on Auditory Display, ICAD, 2002, pp. ICAD02–1–ICAD02–4.
[197] M. Bernard, S. N'Guyen, P. Pirim, B. Gas, J.-A. Meyer, Phonotaxis behavior in the artificial rat psikharpax, in: Proceedings of International Symposium on Robotics and Intelligent Sensors, IRIS, 2010, pp. 118–122.
[198] T. Shimoda, T. Nakashima, M. Kumon, R. Kohzawa, I. Mizumoto, Z. Iwai, Spectral cues for robust sound localization with pinnae, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2006, pp. 386–391.
[199] M. Kumon, Y. Noda, Active soft pinnae for robots, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2011, pp. 112–117.
[200] W. Odo, D. Kimoto, M. Kumon, T. Furukawa, Active sound source localization by pinnae with recursive bayesian estimation, J. Robot. Mechatronics 29 (1) (2017) 49–58.
[201] H.G. Okuno, K. Nakadai, K.I. Hidai, H. Mizoguchi, H. Kitano, Human-robot interaction through real-time auditory and visual multiple-talker tracking, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, Vol. 3, 2001, pp. 1402–1409.
[202] K. Nakadai, K. ichi Hidai, H.G. Okuno, H. Kitano, Real-time multiple speaker tracking by multi-modal integration for mobile robots, in: Proceedings of European Conference on Speech Communication and Technology, EUROSPEECH, 2001, pp. 1193–1196.
[203] K. Nakadai, K.-i. Hidai, H. Mizoguchi, H.G. Okuno, H. Kitano, Real-time auditory and visual multiple-object tracking for humanoids, in: Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI, 2001, pp. 1425–1432.
[204] H.G. Okuno, K. Nakadai, H. Kitano, Social interaction of humanoid robot based on audio-visual tracking, in: T. Hendtlass, M. Ali (Eds.), Proceedings of International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 725–735.
[205] K. Nakadai, H.G. Okuno, H. Kitano, Exploiting auditory fovea in humanoid-human interaction, in: Proceedings of National Conference on Artificial Intelligence, AAAI, 2002, pp. 431–438.
[206] H.D. Kim, K. Komatani, T. Ogata, H.G. Okuno, Design and evaluation of two-channel-based sound source localization over entire azimuth range for moving talkers, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2008, pp. 2197–2203.
[207] A. Lelli, P. Kazmierczak, Y. Kawashima, U. Müller, J.R. Holt, Development and regeneration of sensory transduction in auditory hair cells requires functional interaction between cadherin-23 and protocadherin-15, J. Neurosci. 30 (34) (2010) 11259–11269.
[208] A.W. Peng, F.T. Salles, B. Pan, A.J. Ricci, Integrating the biophysical and molecular mechanisms of auditory hair cell mechanotransduction, Nature Commun. 2 (2011) 523.
[209] C.E. Schroeder, R.W. Lindsley, C. Specht, A. Marcovici, J.F. Smiley, D.C. Javitt, Somatosensory input to auditory association cortex in the macaque monkey, J. Neurophysiol. 85 (3) (2001) 1322–1327.
[210] K.-M.G. Fu, T.A. Johnston, A.S. Shah, L. Arnold, J. Smiley, T.A. Hackett, P.E. Garraghty, C.E. Schroeder, Auditory cortical neurons respond to somatosensory stimulation, J. Neurosci. 23 (20) (2003) 7510–7515.
[211] J. Eargle, The Microphone Book: From Mono To Stereo To Surround - A Guide To Microphone Design and Application, Focal Press, 2004.
[212] J. Lewis, Analog and digital mems microphone design considerations, Tech. rep. Analog Devices, Inc. no. MS-2472 2013.
[213] J. Lewis, Low self noise: The first step to high-performance mems microphone applications, 2012. http://www.eetimes.com/document.asp?doc_id=1280170.
[214] STMicroelectronics, Mems audio surface-mount bottom-port silicon microphone with analog output, Tech. rep. STMicroelectronics, 2013. http://www.st.com/content/ccc/resource/technical/document/datasheet/d2/06/84/85/f3/19/44/12/DM00075180.pdf/files/DM00075180.pdf/jcr:content/translations/en.DM00075180.pdf.
[215] InvenSense, High snr microphone with differential output and low-power mode, Tech. rep. InvenSense, 2016. https://www.invensense.com/wp-content/uploads/2016/02/DS-000044-ICS-40618-v1.0.pdf.
[216] K. Hoshiba, O. Sugiyama, A. Nagamine, R. Kojima, M. Kumon, K. Nakadai, Design and assessment of sound source localization system with a uav-embedded microphone array, J. Robot. Mechatronics 29 (1) (2017) 154–167.
[217] R. Suzuki, T. Takahashi, H.G. Okuno, Development of a robotic pet using sound source localization with the hark robot audition system, J. Robot. Mechatronics 29 (1) (2017) 146–153.
[218] T. Ishiki, K. Washizaki, M. Kumon, Evaluation of microphone array for multirotor helicopters, J. Robot. Mechatronics 29 (1) (2017) 168–176.
[219] T. Ohata, K. Nakamura, A. Nagamine, T. Mizumoto, T. Ishizaki, R. Kojima, O. Sugiyama, K. Nakadai, Outdoor sound source detection using a quadcopter with microphone array, J. Robot. Mechatronics 29 (1) (2017) 177–187.
[220] D. Li, S.E. Levinson, A linear phase unwrapping method for binaural sound source localization on a robot, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, Vol. 1, 2002, pp. 19–23.
[221] S.B. Andersson, A.A. Handzel, V. Shah, P.S. Krishnaprasad, Robot phonotaxis with dynamic sound-source localization, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, Vol. 5, 2004, pp. 4833–4838.
[222] J. Bonnal, S. Argentieri, P. Danès, J. Manhès, P. Souères, M. Renaud, The EAR project, J. Robot. Soc. Japan 28 (1) (2010) 10–13.
[223] C.-T. Kim, T.-Y. Choi, B. Choi, J.-J. Lee, Robust estimation of sound direction for robot interface, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, 2008, pp. 3475–3480.
[224] N. Mahadev, K.B. Austin, Sound localization by robot using inter-aural time differences, J. Comput. Sci. Coll. 30 (4) (2015) 50–56.
[225] R.C. Luo, C.H. Huang, C.Y. Huang, Search and track power charge docking station based on sound source for autonomous mobile robot applications, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2010, pp. 1347–1352.
[226] A. Badali, J.M. Valin, F. Michaud, P. Aarabi, Evaluating real-time audio localization algorithms for artificial audition in robotics, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2009, pp. 2033–2038.
[227] F. Ferland, D. Létourneau, A. Aumont, J. Frémy, M.-A. Legault, M. Lauria, F. Michaud, Natural interaction design of a humanoid robot, J. Hum.-Robot Inter. 1 (2) (2012) 118–134.
[228] M. Fréchette, D. Létourneau, J.M. Valin, F. Michaud, Integration of sound source localization and separation to improve dialogue management on a robot, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2012, pp. 2358–2363.
[229] E. Martinson, T. Apker, M. Bugajska, Optimizing a reconfigurable robotic microphone array, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2011, pp. 125–130.
[230] R. Takeda, K. Komatani, Performance comparison of music-based sound localization methods on small humanoid under low snr conditions, in: Proceedings of IEEE-RAS International Conference on Humanoid Robots, 2015, pp. 859–865.

[231] V. Tourbabin, B. Rafaely, Direction of arrival estimation using microphone array processing for moving humanoid robots, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (11) (2015) 2046–2058.
[232] Y. Bando, T. Mizumoto, K. Itoyama, K. Nakadai, H.G. Okuno, Posture estimation of hose-shaped robot using microphone array localization, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2013, pp. 3446–3451.
[233] M.Z.S. Ahmed, R. Lobo, C.R. Somaiah, Sound localization used in robotics, in: Proceedings of IRF International Conference, 2015, pp. 18–24.
[234] R. Tanabe, Y. Sasaki, H. Takemura, Probabilistic 3d sound source mapping system based on monte carlo localization using microphone array and lidar, J. Robot. Mechatronics 29 (1) (2017) 94–104.
[235] S.H. Young, M.V. Scanlon, Detection and localization with an acoustic array on a small robotic platform in urban environments, Tech. rep. DTIC Document 2003.
[236] K. Okutani, T. Yoshida, K. Nakamura, K. Nakadai, Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2012, pp. 3288–3293.
[237] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, H.G. Okuno, Noise correlation matrix estimation for improving sound source localization by multirotor uav, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2013, pp. 3943–3948.
[238] K. Nakadaij, H. Nakajima, M. Murase, H.G. Okuno, Y. Hasegawa, H. Tsujino, Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2006, pp. 852–859.
[239] L.A. Pineda, L. Salinas, I. Meza, C. Rascon, G. Fuentes, SitLog: A programming language for service robot tasks, Int. J. Adv. Robot. Syst. 10 (358) (2013).
[240] C. Rascon, I. Meza, G. Fuentes, L. Salinas, L. Pineda, Integration of the multi-DOA estimation functionality to human-robot interaction, Int. J. Adv. Robot. Syst. 12 (8) (2015).
[241] S. Hashimoto, S. Narita, H. Kasahara, K. Shirai, T. Kobayashi, A. Takanishi, S. Sugano, J. Yamaguchi, H. Sawada, H. Takanobu, K. Shibuya, T. Morita, T. Kurata, N. Onoe, K. Ouchi, T. Noguchi, Y. Niwa, S. Nagayama, H. Tabayashi, I. Matsui, M. Obata, H. Matsuzaki, A. Murasugi, S. Haruyama, T. Okada, Y. Hidaki, Y. Taguchi, K. Hoashi, E. Morikawa, Y. Iwano, D. Araki, J. Suzuki, M. Yokoyama, I. Dawa, D. Nishino, S. Inoue, T. Hirano, E. Soga, S. Gen, T. Yanada, K. Kato, S. Sakamoto, Y. Ishii, S. Matsuo, Y. Yamamoto, K. Sato, T. Hagiwara, T. Ueda, N. Honda, K. Hashimoto, T. Hanamoto, S. Kayaba, T. Kojima, H. Iwata, H. Kubodera, R. Matsuki, T. Nakajima, K. Nitto, D. Yamamoto, Y. Kamizaki, S. Nagaike, Y. Kunitake, S. Morita, Humanoid robots in Waseda University–Hadaly-2 and WABIAN, Auto. Robot 12 (1) (2002) 25–38.
[242] K. Nakamura, K. Nakadai, F. Asano, G. Ince, Intelligent sound source localization and its application to multimodal human tracking, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2011, pp. 143–148.
[243] K. Teachasrisaksakul, N. Iemcha-od, S. Thiemjarus, C. Polprasert, Speaker tracking module for indoor robot navigation, in: Proceedings of Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON, 2012, pp. 1–4.
[244] L. van Beek, K. Chen, D. Holz, M. Matamoros, C. Rascon, M. Rudinac, J.R. des Solar, S. Wachsmuth, RoboCup@Home 2015: Rule and regulations, 2015. http://www.robocupathome.org/rules/2015_rulebook.pdf.
[245] L. van Beek, K. Chen, D. Holz, L.L. Sanchez, M.M.A. Nagano, C. Rascon, J. de Souza, M. Rudinac, SvenWachsmuth, RoboCup@Home 2016: Rules Regulations, 2016. http://www.robocupathome.org/rules/2016_rulebook.pdf.
[246] R. Takeda, K. Komatani, Noise-robust music-based sound source localization using steering vector transformation for small humanoids, J. Robot. Mechatronics 29 (1) (2017) 26–36.
[247] P. Davis, JACK Connecting a World of Audio. http://jackaudio.org [online, cited 13.04.16].
[248] G.P. Scavone, The rtaudio home page. https://www.music.mcgill.ca/~gary/rtaudio/ [online, cited 13.04.16].
[249] C.D. Manning, P. Raghavan, H. Schutze, Introduction To Information Retrieval, Cambridge University Press, US, 2008.
[250] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, T. Yamada, Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, in: Proceedings of the International Conference on Language Resources and Evaluation, LREC, 2000, pp. 965–968.
[251] G. Lathoud, J.-M. Odobez, D. Gatica-Perez, Av16. 3: an audio-visual corpus for speaker localization and tracking, in: Machine Learning for Multimodal Interaction, Springer, 2004, pp. 182–195.
[252] E. Arnaud, H. Christensen, Y.-C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant, et al., The cava corpus: synchronised stereoscopic and binaural datasets with head movements, in: Proceedings of the International Conference on Multimodal Interfaces, ACM, 2008, pp. 109–116.
[253] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, M. Matassoni, The second chime speech separation and recognition challenge: Datasets, tasks and baselines, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), IEEE, 2013, pp. 126–130.
[254] X. Alameda-Pineda, J. Sanchez-Riera, J. Wienke, V. Franc, J. Čech, K. Kulkarni, A. Deleforge, R. Horaud, Ravel: An annotated corpus for training robots with audiovisual abilities, J. Multimodal User Interfaces 7 (1–2) (2013) 79–91.
[255] A. Deleforge, V. Drouard, L. Girin, R. Horaud, Mapping sounds onto images using binaural spectrograms, in: Proceedings of European Signal Processing Conference, (EUSIPCO), IEEE, 2014, pp. 2470–2474.
[256] J. Le Roux, E. Vincent, J.R. Hershey, D.P. Ellis, Micbots: collecting large realistic datasets for speech and audio research using mobile robots, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), IEEE, 2015, pp. 5635–5639.
[257] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[258] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), IEEE, 2009, pp. 248–255.
[259] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, J. Chen, J. Chen, Z. Chen, M. Chrzanowski, A. Coates, G. Diamos, K. Ding, N. Du, E. Elsen, J. Engel, W. Fang, L. Fan, C. Fougner, L. Gao, C. Gong, A. Hannun, T. Han, L. Johannes, B. Jiang, C. Ju, B. Jun, P. LeGresley, L. Lin, J. Liu, Y. Liu, W. Li, X. Li, D. Ma, S. Narang, A. Ng, S. Ozair, Y. Peng, R. Prenger, S. Qian, Z. Quan, J. Raiman, V. Rao, S. Satheesh, D. Seetapun, S. Sengupta, K. Srinet, A. Sriram, H. Tang, L. Tang, C. Wang, J. Wang, K. Wang, Y. Wang, Z. Wang, Z. Wang, S. Wu, L. Wei, B. Xiao, W. Xie, Y. Xie, D. Yogatama, B. Yuan, J. Zhan, Z. Zhu, Deep speech 2 : End-to-end speech recognition in english and mandarin, in: Proceedings of the International Conference on Machine Learning, vol. 48 (2016) 173–182.

Caleb Rascon is a researcher in the Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas (IIMAS) of the Universidad Nacional Autónoma de México (UNAM) and member of the service robotics group Golem. He received his bachelor degree in Electronic Systems Engineering in ITESM, and his Ph.D. in the University of Manchester. He was awarded as an Innovator under 35 in the Mexico 2014 edition of the MIT Technology Review, and has authored several papers and conducted keynotes on the topic of Robot Audition. His other research interest are Control Engineering, Machine Learning, and Service Robots.

Ivan Meza is a research assistant at IIMAS, UNAM working at the Department of Computer Science on the Golem group. He received his bachelor degree in Computer Engineering in UNAM, and his Ph.D. in the School of Informatics of the University of Edinburgh. He has authored several papers on the topics of Human–Robot Interaction, Computational Linguistics, and Machine Learning, as well organized summer internships and conference workshops on the matter. His other research interest are Deep learning, Natural Language Processing, Dialogue Systems and Service Robots.