INVITED PAPER

Acoustic Self-Awareness of Autonomous Systems in a World of Sounds

This article provides an overview of acoustic self-awareness signal processing techniques. The role of ego-noise representation and processing in autonomous systems is highlighted, and application case studies are presented.

By Alexander Schmidt, *Member IEEE*, Heinrich W. Löllmann, *Senior Member IEEE*, and Walter Kellermann, *Fellow IEEE*

ABSTRACT | Autonomous systems (ASs) operating in real-world environments are exposed to a plurality and diversity of sounds that carry a wealth of information for perception in cognitive dynamic systems. While the importance of the acoustic modality for humans as "ASs" is obvious, it is investigated to what extent current technical ASs operating in scenarios filled with airborne sound exploit their potential for supporting self-awareness. As a first step, the state of the art of relevant generic techniques for acoustic scene analysis (ASA) is reviewed, i.e., source localization and the various facets of signal enhancement, including spatial filtering, source separation, noise suppression, dereverberation, and echo cancellation. Then, a comprehensive overview of current techniques for ego-noise suppression, as a specific additional challenge for ASs, is presented. Not only generic methods for robust source localization and signal extraction but also specific models and estimation methods for ego-noise based on various learning techniques are discussed. Finally, active sensing is considered with its unique potential for ASA and, thus, for supporting self-awareness of ASs. Therefore, recent techniques for binaural listening exploiting head motion, for active localization and exploration, and for active signal enhancement are presented, with humanoïd robots as typical platforms. Underlining the multimodal nature of self-awareness, links to other modalities and nonacoustic reference information are pointed out where appropriate.

KEYWORDS | Acoustic scene analysis (ASA); active enhancement; active localization; autonomous systems (ASs); ego-noise; exploration; self-awareness

I. INTRODUCTION

Recent decades spawned striking examples of what is commonly referred to as autonomous systems (ASs), such as self-driving cars, robots operating in our daily environment or exploring unknown worlds in deep sea or outer space, unmanned aerial vehicles (UAVs) for logistics, air surveillance and combat, autonomous weapon systems, and many more. What they all have in common is the use of a large variety of sensor modalities for perceiving their environment. The challenge to develop efficient techniques for processing the according signals and intelligent concepts for exploiting the extracted information triggered an enormous amount of scientific contributions over the past half-century. While the importance of the acoustic modality for humans as "ASs" is obvious, relatively little attention was paid, however, so far to the acoustic modality of airborne sound[1] for ASs, and especially the specific challenges of ASs with its self-created noise resulting, e.g., from its own movements. This can be attributed to the fact that acoustic human–machine communication and acoustic scene analysis (ASA) only recently reached a stage of maturity which allows operation outside highly constrained acoustic environments, and the complexity of the acoustic scenario faced by ASs is often viewed as

Manuscript received May 31, 2019; revised October 31, 2019 and January 31, 2020; accepted February 21, 2020. Date of publication March 18, 2020; date of current version June 18, 2020. This work was supported in part by the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant 609465. *(Corresponding author: Alexander Schmidt.)*

The authors are with the Chair of Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91058 Erlangen, Germany (e-mail: alexander.as.schmidt@fau.de; heinrich.loellmann@fau.de; walter.kellermann@fau.de).

Digital Object Identifier 10.1109/JPROC.2020.2977372

[1] It should be noted that the large areas of underwater acoustics and structure-borne sound processing are excluded for the remainder of this article. Interested readers are referred to [1] and [2], respectively.

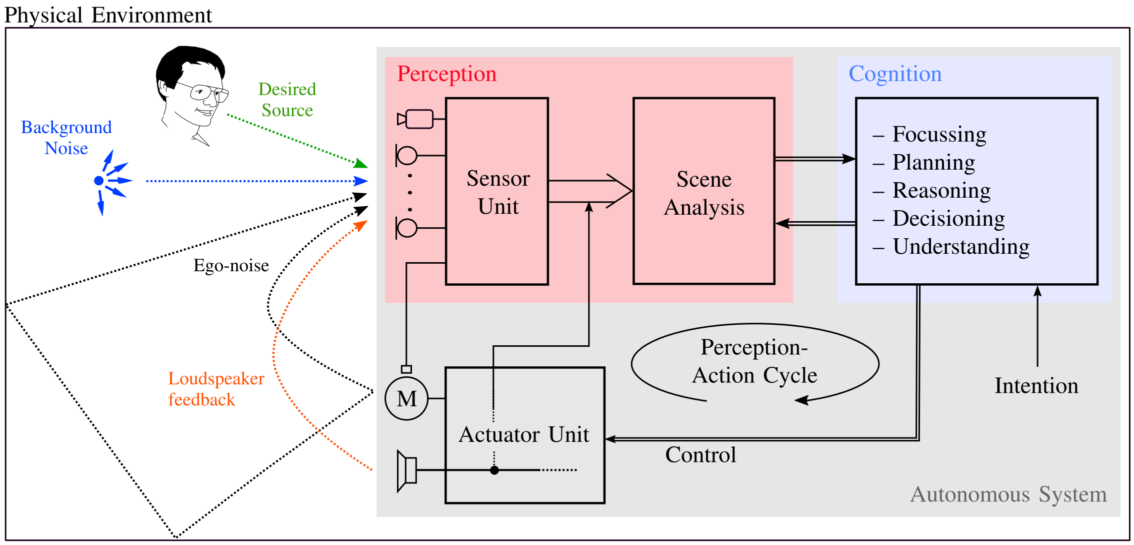

Fig. 1. ***Generic AS exposed to an acoustic scene comprising human speakers, noise sources, ego-noise, loudspeaker signals, and acoustic echoes and reverberation.***

lying beyond more pressing and still unsolved problems of simpler scenarios. While speech recognition in real-world environments progressed enormously over recent years [3], the wider domain of computational sound scene analysis [4], [5] is still in its infancy and current research is still restricted to limited scenarios and generally does not consider the spatial aspects of acoustic scenes [6]. For an AS, however, spatial information on the sound sources in its environment is crucial for proper understanding. Beyond the limitations of the technology in traditional acoustic application scenarios, the broad category of mobile ASs typically poses an additional major challenge by creating noise and interference itself, so-called ego-noise, as resulting, e.g., from the mechanical noise of a moving robot or a self-driving car, or the airflow around a drone. Stimulated by recent progress in acoustic signal processing and related learning algorithms, however, it can be safely expected that in the near future, ASs will be able to greatly augment their self-awareness by increasingly exploiting the acoustic domain with its plurality and diversity of sounds carrying a wealth of information, as a crucial modality for the perception part of a cognitive dynamic system. The importance of the acoustic modality should also be recognized in its complementary role in multimodal tasks, such as audiovisual source localization, where the visual modality is strongly supported in situations of bad lighting conditions, occlusions, or if the target is outside of the current field of view [7].

A generic AS is illustrated in Fig. 1, which highlights perception and cognition units as key elements of the perception–action cycle [8]. Perception comprises the various sensing modalities and the according signal processing-based scene analysis collaborating with the cognition unit. The illustrated sensor unit exhibits not only a camera and multiple microphones but also a proprioceptor providing information about the internal state of the AS, e.g., mechanical control information such as joint angle or motor rotation. Obviously, numerous other sensing modalities may be included, e.g., acceleration sensors or GPS units providing spatial information relative to an external coordinate system. As such, the sensor unit enables the AS to perceive on the one hand its own contribution to the acoustic scene, e.g., by emitting ego-noise, and on the other hand its ambient environment by, e.g., localizing surrounding acoustic sources and extracting their signals. Both tasks are crucial for an AS to achieve acoustic self-awareness.

The collected sensor data are then further processed and analyzed to extract the desired information about the environment (scene analysis). In the acoustic domain, it includes several subtasks such as source localization and tracking, signal extraction, and enhancement, but also higher-level tasks such as source detection and signal classification. Based on information extracted from the acoustic scene, the cognitive unit of the AS controls various actuators of the AS (actuator unit), e.g., motors to drive joints or propellers, but also loudspeakers for human–machine communication. The actuator unit therefore contributes to the acoustic world which the AS perceives. Therefore, in the acoustic domain, perception and action are tightly connected in the perception–action cycle.

Following the exemplary illustration in Fig. 1, the first part of the ensuing treatment aims at a comprehensive presentation of the relevant state of the art in acoustic signal processing and scene analysis as the basis of acoustic self-awareness for ASs, emphasizing localization, signal extraction, and enhancement (see Section II). The ability of an AS to respond to acoustic events has major potential in ASA, e.g., for improved localization and extraction of acoustic signals by motion and changing sensor

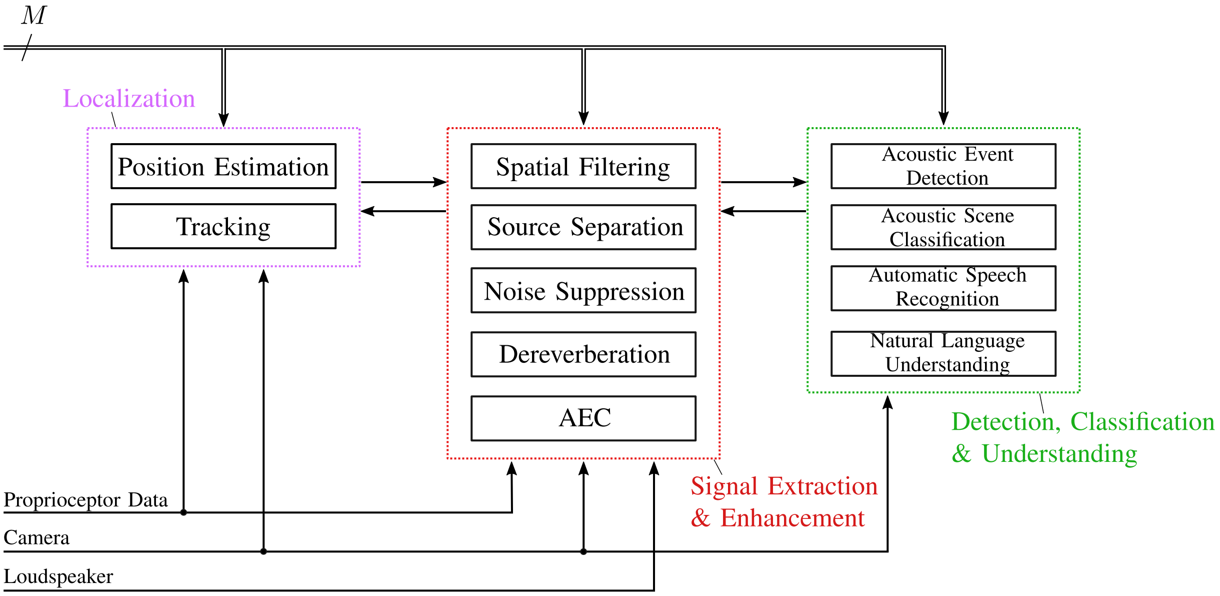

Fig. 2. ***Overview of methods and algorithms for ASA (see Section II).***

topologies. This, however, requires that the AS is aware of its current state and its impact on the acoustic environment. These aspects will be discussed in the second part of this article where we first focus on the specific challenge for an AS resulting from self-created noise (see Section III). We then consider motion and time-varying sensor topologies (see Section IV) and discuss resulting challenges and opportunities.

Note that counting on a continuing increase of processing capabilities, the scope for the discussed algorithms is not tailored to currently available hardware platforms for ASs, but is determined by the technical merits and potential of the methods. Accordingly, only some of the approaches presented in the following have been evaluated on one of the few commercially available hardware platforms, which are still characterized by limited processing capabilities due to space and power constraints. Besides this, such platforms are rarely fully programmable, which makes it difficult for the user to employ the AS for the required purposes. Consequently, self-designed customized hardware platforms prevail among experimental systems. Especially, computationally expensive algorithms are implemented and processed offline on external hardware.

II. ANALYZING A WORLD OF SOUNDS: BASIC CONCEPTS

This section provides an introduction of the signal model supporting the description of algorithms and an overview of fundamental acoustic signal processing techniques as used for ASA, notably for source localization and tracking, signal extraction and enhancement, scene classification, and event detection (see Fig. 2). Since numerous approaches for ASs in this area are inspired by more generic methods, we first present the general underlying problems and methods for addressing them, and then concentrate on approaches where generic techniques were extended for and adapted to ASs.

In the following, we distinguish between ASA and computational auditory scene analysis (CASA) [9], [10]: ASA does not refer to the functionality of the human auditory system. On the other hand, CASA is inspired by auditory scene analysis, as proposed in [9], and aims at translating important features of the human auditory system to computer systems, e.g., for extracting desired sources among other competing sources. The principles of CASA are often highly relevant also for ASA of ASs, especially for humanoïd robots [11].

A. Signal Model

The algorithms considered in the following do not typically operate on raw time-domain signals but rather use representations in a time–frequency domain, where the temporal dependence supports capturing both the nonstationarity of signals and the time-variance of systems. In the following, we use the discrete short-time Fourier transform (STFT) as a widely used generic time–frequency representation with uniform resolution in time and frequency [12]. We assume that the acoustic transfer functions (ATFs) are time-invariant within an STFT window such that, e.g., the characteristic shape or the statistics of a signal can be learned or estimated. Typically, time-invariance of the ATFs within an STFT window can be assured to a sufficient extent by a proper choice of the STFT window size. If this is not possible, e.g., if the AS is moving very fast, purely time domain-based signal processing algorithms should be considered as an alternative.

Let an AS be equipped with M microphones. We denote the discrete-time signal of channel m at time instant k by $x_k^{(m)}$, $m = 1, \dots, M$, and the corresponding discrete STFT domain representation by $\mathbf{X}^{(m)} \in \mathbb{C}^{F \times T}$, where F and T are the number of frequency bins and time frames, respectively. $X_{ft}^{(m)} \in \mathbb{C}$ denotes the ftth time–frequency bin of $\mathbf{X}^{(m)}$ with $f \in \{1, \dots, F\}$ and $t \in \{1, \dots, T\}$. For capturing all M channels at a given single time–frequency bin ft, we introduce $\mathbf{X}_{ft} = [X_{ft}^{(1)}, \dots, X_{ft}^{(M)}]^T \in \mathbb{C}^M$. Based on this, we define the spatial correlation matrices for time–frequency bin ft as $\mathbf{\Phi}_{\mathrm{X},ft} = \mathcal{E}\{\mathbf{X}_{ft}\mathbf{X}_{ft}^H\} \in \mathbb{C}^{M \times M}$ with $\mathcal{E}\{\cdot\}$ as the expectation operator and $\cdot^H$ denoting the Hermitian operator. For simplicity, signals will be referred to by their discrete STFT representation in the following.

In general, the M microphones mounted to an AS capture a single or multiple desired ("target") source signals in the presence of interfering sound sources and additive noise. This is expressed in the STFT domain as follows:

$$\mathbf{X}_{ft} = \sum_{d=1}^{D} \mathbf{H}_{\mathrm{S},d,ft} S_{d,ft} + \sum_{j=1}^{J} \mathbf{H}_{\mathrm{U},j,ft} U_{j,ft} + \mathbf{N}_{\mathrm{U},ft} + \sum_{l=1}^{L} \mathbf{H}_{\mathrm{V},l,ft} V_{l,ft} + \mathbf{N}_{\mathrm{EN},ft}. \quad (1)$$

Vector $\mathbf{H}_{\mathrm{S},d,ft}$ represents the M possibly time-varying ATFs between the dth desired target source, which emits signal $S_{d,ft}$ and the microphone array. The often challenging task of detecting and classifying the desired target source(s) is addressed in Section II-D. Vector $\mathbf{H}_{\mathrm{U},j,ft}$ represents the M ATFs between the jth interfering source emitting signal $U_{j,ft}$ and the microphone array. ATFs $\mathbf{H}_{\mathrm{S},d,ft}$ and $\mathbf{H}_{\mathrm{U},j,ft}$ contain information about the directions of arrival (DOAs) of the different source signals. Methods to estimate DOAs are reviewed in Section II-B. Vector $\mathbf{N}_{\mathrm{U},ft}$ contains additive noise captured by the M microphones which is not caused by the AS, such as background noise or diffuse room reverberation. Methods to suppress this noise and the J interfering source signals are treated in Section II-C. Vector $\mathbf{N}_{\mathrm{EN},ft}$ represents the so-called ego-noise captured by the M microphones, which is distinctively characteristic relative to other acoustic signal processing scenarios as it is directly or indirectly caused by the AS itself (self-created noise). A comprehensive treatment of methods to suppress this ego-noise is presented in Section III. If the AS is equipped with L loudspeakers, e.g., for human–machine communication, the audio signals emitted by the loudspeakers $V_{l,ft}$ feedback into the microphone array. Vector $\mathbf{H}_{\mathrm{V},l,ft}$ contains the M ATFs between loudspeaker l and the microphone array. Since the loudspeaker signals are known to the system, suppression of acoustic echoes is most commonly treated as a supervised system identification problem, as discussed in Section II-C4. We note that the considered model assumes linear transfer functions between the audio sources and receivers. This assumption underlies most audio signal processing algorithms but might be violated if, e.g., low-cost loudspeakers with nonlinear transmission characteristics are involved.

B. Source Localization

In this section, we will first review basic concepts for sound source localization (SSL). A more detailed overview for SSL in general and specifically for robots can be found in [13]–[16], respectively.

Localization is viewed here as estimating the position of an active sound source relative to a sensor array and includes tracking techniques, i.e., techniques for exploiting previous observations and hypothesized models for this estimation. Typically, position estimation requires estimation of DOA and range, or triangulation techniques [17], [18] using multiple DOA estimates. For static scenarios, acoustic range estimation approaches such as [19] are so far much less common than DOA estimation techniques. Range estimation for mobile ASs will be addressed in Section IV. Very often, however, single DOA estimates are sufficient for the given task, e.g., when steering a beamformer (BF) [20] toward a target source. Therefore, DOA estimation approaches dominate the following review.

1) Position Estimation in a Free Sound Field: We first consider microphones as point-like sensors in an obstacle-free sound field (disregarding its mounting to scattering structures in the real world), which receive delayed versions of the original source signals, i.e., direct path components, possible room reflections, and uncorrelated noise components. For such scenarios, approaches for position estimation are considered that can be categorized as time difference of arrival (TDOA)-, steered response power (SRP)-, and subspace-based methods.

TDOA-based position estimation methods comprise a two-step strategy where first the relative time delays (i.e., TDOAs) are estimated and subsequently employed to estimate the position of the source. The performance of this approach depends heavily on the accuracy of the TDOA estimation, while the computation of the position is a purely geometrical, however, not straightforward problem [21]. The most widely used approach for TDOA estimation is the generalized cross-correlation (GCC) method [22], where the TDOA is extracted from a weighted version of the cross power spectral density (PSD) of two sensor signals. A popular weighting function is the phase transform (PHAT) [22], especially for speech [23]. In the context of ASs, GCC-PHAT was employed in, e.g., [24]. As an alternative, Valin *et al.* [25] proposed a weight function that depends on an estimate of the local signal-to-noise ratio (SNR) in each frequency bin. If multiple sound sources are active, identifying the correct source-specific TDOAs typically becomes difficult [26].

SRP methods can be interpreted as a generalization of the GCC approach to multiple microphone pairs by coherently adding up signals originating from a certain point in space and hence estimating the source's likelihood to be located at this position. For appropriate array geometries, the source's likelihood can incorporate both DOA- and range- estimates. For ASs, SRP-PHAT [27] is predominantly used (see [28] for robots and [29], [30] for drones). A microphone-dependent weight based on the estimated *a priori* SNR in each frequency bin was introduced in [31] and evaluated for a robot.

Subspace methods for DOA-estimation are based on the eigenanalysis of the spatial correlation matrix $\hat{\mathbf{\Phi}}_{\mathrm{X},ft}$ and a subsequent separation of signal and noise subspace. An in-depth overview of subspace methods can be found in [32]. As an example, we consider multiple signal classification (MUSIC), where the so-called MUSIC pseudospectrum is computed by projecting steering vectors pointing to different spatial directions onto the noise subspace, resulting in a minimum if a source is located in the hypothesized spatial direction. MUSIC is by far the most used subspace method for SSL in the field of ASs (see [33]–[40]). Since MUSIC explicitly assumes narrowband sources, several broadband extensions for audio signals have been proposed and evaluated regarding their real-time capability for ASs (see [36]). If the AS is exposed to directional noise sources U_j, eigenvectors of noise can mistakenly be identified as source eigenvectors. For this, generalized eigenvalue decomposition MUSIC (GEVD-MUSIC) was proposed in [33] and [35] where an estimate of the noise spatial correlation matrix is used to whiten the microphone signals. A more efficient extension of GEVD-MUSIC was proposed in [35]. Further relevant approaches for estimating the noise spatial correlation matrix in the AS scenario will be discussed in Section III.

2) Position Estimation With Embedded Microphone Arrays: As opposed to Section II-B1, we now consider sensors which are embedded into or mounted to scatterers, e.g., microphones embedded into the head of a humanoïd robot. This example provides also the link to the first use cases of DOA estimation, where the binaural hearing capability of humans was mimicked by humanoïd robots [11] using the long-known concepts of interaural level differences (ILDs) and interaural time differences (ITDs) [41] as predominant features to localize sources. ILDs and ITDs [or its frequency-domain counterpart interaural phase differences (IPDs)] are captured by the so-called head-related transfer function (HRTF), which additionally describes the filtering effect of the head, pinnae, and torso. For humanoïd robots, a variety of HRTF models have been proposed to map measured ILDs and ITDs to position estimates (see [11] and [42]–[44]). The proposed models are typically simple and hence allow implementation on real-time systems. More involved methods map extracted binaural cues to azimuths of multiple sound sources using Gaussian mixture models (GMMs) [45] or deep neural networks (DNNs) [46], [47]. To resolve the remaining limitations of binaural cues regarding the front-back ambiguity, the exploitation of motion will be discussed in Section IV. The almost spherical shape of some robot heads suggests to represent the received sound field by spherical harmonics (SHs), i.e., harmonic functions defined on the surface of a sphere, which allow efficient modeling of the scattering effect on localization [48]–[50]. Microphone array configurations on a robot's head have been optimized with respect to number and placement of sensors for, e.g., maximizing the amount of acquired spatial information [51] or extending the aliasing-free frequency range of the array [52], [53]. Beyond robot heads and their HRTFs, object-related transfer functions (ORTFs) [54] can be used to account for more general scatterers. ORTFs are typically measured or simulated and subsequently incorporated into the previously presented approaches.

3) Tracking: Tracking accounts for the estimation of source positions of moving sources over time based on previous observations and motion models. Traditional tracking systems [55] are based on Kalman filters (KFs) [56] and particle filters (PFs) [57], [58]. For robots, single-source tracking was investigated, e.g., in [59], including voice activity detection. In [60], a method for acoustic tracking of a single source or for drones is reported where the latter is especially challenging due to the very low SNR caused by massive ego-noise interference (see Section III). The fusion of video and audio information for tracking was exemplarily demonstrated in [61]–[63].

For tracking multiple targets, a data association stage is required assigning each position or DOA measurement to one of the hypothesized targets before they can be tracked individually, e.g., using KF- or PF-based approaches. However, those methods generally require the knowledge of the number of targets as input. In the context of robot audition, a real-time capable system was proposed in [31]. An alternative rigorous Bayesian framework for multi-target tracking is based on random finite sets, i.e., sets with random numbers of elements, where each element is represented by a random variable [64], therefore explicitly accounting for missing measurements, clutter, track initiation, and termination of a source, so that no data association step is needed. An approximate recursive, first-order moment-based solution to this systematic approach is given by the probability hypothesis density (PHD) filter [65]. Multispeaker tracking using PHD filters and TDOA measurements was presented in [66], demonstrating the robustness of the concept against reverberation. For robot audition, a bearing-only tracking approach employing PHD filters was investigated in [67].

C. Signal Extraction and Enhancement

In ASA, signal extraction aims at recovering the desired source signal in the presence of interfering sound sources and/or background noise and reverberation. Since ASs are usually equipped with multiple microphones to allow source localization, multichannel filtering techniques are preferably utilized. Signal enhancement is usually already a byproduct of the signal extraction step but is often be applied as additional postprocessing to the extracted signal, e.g., for further suppression of noise or reverberation.

This section briefly reviews the main concepts for signal extraction and enhancement as they are relevant for current and promising for next-generation ASs (see Fig. 2). We refer to, e.g., [68]–[70] for a more comprehensive treatment of multichannel acoustic signal enhancement methods and to, e.g., [32] and [71] for the general signal processing background.

The main approaches to multichannel signal extraction and enhancement can be categorized as either data-independent or data-dependent [20], [32]. Data-independent approaches perform a spatial filtering where, typically, signals arriving from an estimated or known DOA or position are emphasized, while signals arriving from other directions are suppressed, i.e., a "beam" is steered toward the source of interest ("beamforming"), without accounting for the statistics of the source signal. In contrast, data-dependent approaches include the statistics of the microphone signals when optimizing the signal extraction performance. In general, algorithms for data-independent multichannel signal enhancement are more robust and have lower computational complexity than data-dependent approaches (see [72] and [73]), whereas data-dependent approaches will outperform data-independent methods as long as the required signal statistics can be estimated reliably [32], [69].

1) Data-Independent Spatial Filtering: Data-independent spatial filtering is typically realized as a filter-and-sum BF [71]. The output signal of this linear filtering operation can be expressed in the STFT domain as

$$Y_{ft} = \sum_{m=1}^{M} W_f^{(m)*} X_{ft}^{(m)} = \mathbf{W}_f^H \mathbf{X}_{ft} \tag{2}$$

with generally complex-valued vector $\mathbf{W}_f$ containing time-invariant weights for each of the M sensor signals at frequency f. $\cdot^*$ denotes the conjugate complex operator. As input, we consider here

$$\begin{aligned}\mathbf{X}_{ft} &= \mathbf{H}_{\mathrm{S},ft} S_{ft} + \sum_{j=1}^{J} \mathbf{H}_{\mathrm{U},j,ft} U_{j,ft} + \mathbf{N}_{\mathrm{U},ft} \\ &= \mathbf{H}_{\mathrm{S},ft} S_{ft} + \mathbf{X}_{\mathrm{U},ft}\end{aligned} \tag{3}$$

where $\mathbf{X}_{\mathrm{U},ft}$ denotes all unwanted signal components. This signal model is a special case of (1) as acoustic feedback and ego-noise are neglected and only a single target source S_{ft} with ATFs $\mathbf{H}_{\mathrm{S},ft}$ is considered.

In its simplest form, weights $W_f^{(m)}$ in (2) only compensate TDOA of the desired signal components in each sensor signal prior to summation. This yields the popular delay-and-sum beamformer (DSB) which leads to minimum sensitivity to sensor noise or sensor mismatch at the cost of low directivity [32]. For audio signals, broadband BFs offering frequency-invariant beamwidth are attractive [74], [75] and typically involve noise-sensitive differential beamforming ("superdirectivity") at low frequencies [76], [77]. Robustness to microphone mismatch has been optimized in [78] and more general optimum tradeoffs reconciling spatial selectivity with beampattern constraints and noise robustness can be determined via convex optimization (see [75]). To allow ASs operating in dynamic environments to steer the main beam to arbitrary directions, polynomial BF designs can be used [79], [80].

The aforementioned BF designs implicitly assume free-field sound propagation around the sensor array and thus neglect scattering effects caused by the physical embedding of the microphones. In [81], the method of [75] is extended by incorporating HRTFs of a robot head to account for its scattering so that a significantly lower word error rate (WER) for automatic speech recognition (ASR) and increased speech quality could be reported. This design was also generalized to allow flexible beamsteering by polynomial beamforming [82]. Note that the robot-specific HRTF-based designs can also be applied to other ASs by considering ORTFs instead of HRTFs. BF design for spherical microphone arrays or arrays mounted to a rigid, approximately spherical scatterer can be conveniently carried out in the SH domain using the spherical Fourier transform [52], [83].

2) Data-Dependent Spatial Filtering: In contrast to data-independent BFs, data-dependent BFs exploit spectro-temporal as well as spatial information of the microphone signals to extract the desired source signal from a mixture including interferers and noise. The BF weights are obtained by solving a supervised optimum multichannel filtering problem relying either on estimates of the interference and noise, and/or target source statistics. A prominent example for this approach is given by the minimum variance distortionless response (MVDR) beamformer (see [32]). Its weights are obtained by minimizing the power of the unwanted signal components in the BF output signal

$$\begin{aligned}\mathcal{E}\{|\mathbf{W}_{ft}^H \mathbf{X}_{\mathrm{U},ft}|^2\} &= \mathcal{E}\{\mathbf{W}_{ft}^H \mathbf{X}_{\mathrm{U},ft} \mathbf{X}_{\mathrm{U},ft}^H \mathbf{W}_{ft}\} \\ &= \mathbf{W}_{ft}^H \mathbf{\Phi}_{\mathrm{U},ft} \mathbf{W}_{ft}\end{aligned} \tag{4}$$

subject to the constraint that the desired component in the output signal is equal to the target signal, i.e., $\mathbf{W}_{ft}^H \mathbf{H}_{\mathrm{S},ft} = 1$. If target signal S_{ft}, interfering signals $U_{j,ft}$, and noise $\mathbf{N}_{\mathrm{U},ft}$ are mutually uncorrelated and of zero mean, the resulting weights are given by

$$\mathbf{W}_{ft}^{(\mathrm{mvdr})} = \frac{\mathbf{\Phi}_{\mathrm{U},ft}^{-1} \mathbf{H}_{\mathrm{S},ft}}{\mathbf{H}_{\mathrm{S},ft}^H \mathbf{\Phi}_{\mathrm{U},ft}^{-1} \mathbf{H}_{\mathrm{S},ft}}. \tag{5}$$

As ATFs $\mathbf{H}_{\mathrm{S},ft}$ are very difficult to estimate in practice, the MVDR BF is often formulated in terms of the so-called relative transfer functions (RTFs) which are much easier to estimate than ATFs [84] and are often assumed as relative time delays for simplicity.

The need to estimate the spatial correlation matrix of the unwanted signal components $\mathbf{\Phi}_{\mathrm{U},ft}$ can be circumvented by the use of the minimum power distortionless response (MPDR) BF. By replacing $\mathbf{\Phi}_{\mathrm{U},ft}$ with $\mathbf{\Phi}_{\mathrm{X},ft}$ in (5), it minimizes the power of the BF output signal $\mathbf{W}_{ft}^{H}\mathbf{X}_{ft}$ subject to the distortionless constraint $\mathbf{W}_{ft}^{H}\mathbf{H}_{\mathrm{S},ft} = 1$ at the cost of a higher sensitivity to sensor mismatch relative to the MVDR BF (see [72]).

The constrained optimization of the MVDR and MPDR BF is converted to an unconstrained optimization problem using the generalized sidelobe canceler (GSC) in [85]. The GSC structure consists of a BF to fulfill the distortionless constraint, a blocking matrix to create so-called noise references, and a multichannel interference canceller, which uses these noise references as inputs and minimizes the residual interference and noise in the output of the BF to the extent to which it is correlated with the noise references. The GSC structure is especially suitable to adapt to nonstationary signals such as speech and time-varying DOAs and RTFs [86]–[88].

The MVDR BF is a special case of the linearly constrained minimum variance (LCMV) BF, which minimizes the output noise power subject to multiple linear constraints to account for multiple desired or interfering point sources [32]. Just as for MVDR BFs, time-varying blocking matrices can be estimated by blind source separation (BSS)-based methods allowing adaptation during a simultaneous activity of multiple sources [54], [89].

In addition to the MVDR BF and its variants, the multichannel Wiener filter (MWF) is another widely used approach for data-dependent signal extraction [32]. It minimizes the mean-squared error (MSE) between the BF output and a target signal and it can be shown that the MWF is equivalent to an MVDR BF according to (5) followed by a single-channel Wiener filter (WF)-based postfilter [90], [91]:

$$\mathbf{W}_{ft}^{(\mathrm{mwf})} = \underbrace{\mathbf{W}_{ft}^{(\mathrm{mvdr})}}_{\text{MVDR beamformer}} \cdot \underbrace{\frac{\Phi_{\mathrm{S},ft}}{\Phi_{\mathrm{S},ft} + \Phi_{\mathrm{U},ft}}}_{\text{Wiener postfilter}} \tag{6}$$

with $\Phi_{\mathrm{S},ft}$ and $\Phi_{\mathrm{U},ft}$ denoting the PSDs of the desired and undesired signal components at the MVDR BF output, respectively. Thus, compared to the MVDR, the MWF attains further noise suppression at the cost of target signal distortion. Tradeoffs between noise suppression and signal distortion can be formulated as tradeoffs between (5) and (6) and have been proposed in [73] and [92].

The approaches discussed so far require information about the DOA of the desired source, the microphone configuration and/or spatiotemporal statistics of the signals. In the following, BSS methods are discussed which separate signals emitted by point sources without such prior information. Therefore, multiple desired sources $S_{d,ft}$ can be separated from each other, or the separation of a single desired source S_{ft} from interferers $U_{j,ft}$ can be achieved. Representing a broad class of BSS algorithms, independent component analysis (ICA) is based on the assumption that the signals which are to be separated are mutually statistically independent and only requires knowledge about the number of sources (see [93]). Moreover, for most multichannel ICA-based algorithms, the number of sources should not exceed the number of microphones. The demixing weights for source separation are obtained by maximizing cost functions reflecting the statistical independence of the output signals [94]–[98]. ICA can be performed either in the time domain or in the frequency domain (see [99] and [100]). Frequency-domain approaches are computationally more attractive than time-domain approaches but, in contrast to time-domain approaches, require coping with the internal permutation and scaling problems [99], [101]. Comprehensive treatments of various ICA-based algorithms can be found, e.g., in [69], [93], [102], and [103].

The cost function of ICA-based BSS can also be formulated with an additional geometric constraint in the frequency domain such that, e.g., a spatial null is forced in a given direction [104], [105]. If the spatial null is directed toward a desired speaker, the unwanted signals can be extracted, which can then be used as noise reference in a GSC BF [54], [89], [105]. Favored by its relatively low computational complexity, the use of this geometrically constrained (GC) BSS scheme has been proposed for various mobile robots [25], [106], [107] (see Section IV-C). For the possibly rapidly changing scenarios faced by ASs, stepsize control for iterative optimization algorithms plays a crucial role and has been addressed in [108] and evaluated for the humanoïd robot ASIMO in [109].

As a second class of BSS algorithms, binary masking (see [110]–[112]) relies on the assumption that audio signals are sparse in the time–frequency plane so that, for any time–frequency bin tf, only one of the sources is dominant and other sources can be neglected (W-disjoint orthogonality) [111]. Each source is then represented only by those time–frequency bins where it is dominant, and its entire spectrogram is estimated from this mask [111], [112].

Various signal enhancement methods for mobile robots capitalize on the sparsity assumption: An ICA-based separation scheme for instantaneous mixtures followed by binary masking is introduced in [113]. In [114], a signal extraction scheme was proposed where the demixing weights are obtained by maximizing the sparsity of the BSS output signals. In [115], this scheme is extended by performing HRTF-based beamforming followed by BSS.

While the aforementioned spatial filtering techniques primarily aim at suppressing undesired sources or separating competing sources from each other, multichannel dereverberation aims at minimizing the distortion of the desired signals $S_{d,ft}$ due to acoustic reflections and reverberation as captured by ATFs $\mathbf{H}_{\mathrm{S},d,ft}$ (see [116]–[121]). If, in the multichannel case, the ATFs $\mathbf{H}_{\mathrm{S},d,ft}$ are exactly known, they can be perfectly equalized by multichannel inverse filtering according to the multiple-input–output inverse theorem (MINT) theorem [116]. Techniques for increasing robustness to errors in ATF estimation have been proposed, e.g., in [117] and [118]. Avoiding the need to explicitly estimate the ATFs, the weighted prediction error (WPE) method for speech dereverberation [119]–[121] is based on linear prediction inverting an autoregressive model for a desired speech signal and the acoustic channels. This widely used multichannel dereverberation concept has also been proposed for the SH domain [122]. A generic BSS-based approach for speech and audio dereverberation was presented in [123].

3) Single-Channel Signal Enhancement: Single-channel signal enhancement is applied either to a single microphone signal or to the output of a multichannel filtering scheme (postfiltering) [see (6)]. Commonly implemented in the STFT domain, the enhanced signal $\hat{S}_{ft}$ is obtained by multiplying the input signal X_{ft} with nonnegative real-valued weights W_{ft} [124]

$$\hat{S}_{ft} = X_{ft}\, W_{ft} \text{ with } 0 \leq W_{ft} \leq 1. \tag{7}$$

For signal enhancement, large weights should be assigned to undistorted time–frequency bins X_{ft} and low weights to distorted X_{ft}. A large variety of approaches to optimize the (time-varying) weights W_{ft} have been proposed (see [125] and [126]) including, e.g., the spectral subtraction filter [124] and the single-channel WF as contained in (5). Several spectral subtraction schemes have been proposed for robots (see [127]–[130]), which are discussed in Section III.

While the WF is a minimum mean-squared error (MMSE) estimator for the complex-valued spectral coefficients, optimal amplitude estimators, as desirable for (7), have been proposed in [131] and [132]. The use of an MMSE short-term spectral amplitude estimator for speech enhancement in mobile robots is proposed, e.g., in [133].

All schemes for computing weights W_{ft} in (7) require an estimate for the PSD of either the desired signal $\Phi_{\mathrm{S},ft}$ or of the unwanted signal components $\Phi_{\mathrm{U},ft}$. In scenarios where the desired signal is speech, noise estimates are usually obtained during speech pauses and then are also used during subsequent speech activity, assuming sufficient stationarity of the noise [134]–[136]. Obviously, these methods are not well suited for interfering speech sources or other nonstationary and unpredictable noise.

The discussed enhancement schemes based on spectral weighting according to (7) can also be used for suppressing late reverberation in speech signals. To this end, the late reverberation is considered as the unwanted noise $\Phi_{\mathrm{U},ft}$ in the postfilter [see (7)] [137], [138]. Estimation of $\Phi_{\mathrm{U},ft}$ then requires knowledge about the reverberation time T_{60}, which can be obtained by blind T_{60} estimators (see [139] and [140]). Joint noise suppression and dereverberation is achieved if $\Phi_{\mathrm{U},ft}$ represents both the PSD of background noise and the PSD of the late reverberant speech (see [138] and [141]).

4) Acoustic Echo Control: When ASs communicate via loudspeaker signals with their environment, e.g., service robots or humanoïd robots communicating with humans via voice, or other ASs emitting warning sounds, the echoes of the according signals feedback into the microphones and thus act as interference for the desired signals (see Fig. 1). Extensively studied for telecommunications, acoustic echo cancellation is a nontrivial supervised system identification problem [142], [143], and efficient schemes for human–machine interfaces using microphone arrays [144], [145] can be directly applied to ASs [146] for identifying the acoustic paths $\mathbf{H}_{\mathrm{V},l,ft}$ in (1). Postfiltering is common for suppressing residual echoes after echo cancellation [143], [147]. Other techniques renouncing on the loudspeaker signals as reference information have also been proposed for mobile robots, e.g., a scheme for acoustic echo control (AEC) and dereverberation using ICA [148]. Nonlinear behavior of loudspeakers and/or microphones calls for nonlinear echo path models [149]–[151] or according postfiltering [147], [152]. Since ASs are typically using single-channel audio output only, multichannel extensions for AEC, as investigated in [153]–[156], have not been studied for ASs yet.

D. Detection, Classification, and Understanding

The aforementioned algorithms for source extraction and enhancement, see Section II-C, presuppose that a target source was identified as such. If so, ASR and natural language understanding (NLU) can be employed to recognize the utterance and determine whether the semantic content is relevant for the intention of the AS or not. Compared to localization, signal extraction, and enhancement, methods for detection, classification, and understanding fall conceptually into a different category and despite their relevance for acoustic self-awareness, a more detailed description of generic concepts for these methods is out of the scope of this article. In the following, we therefore concentrate on describing their relation to acoustic self-awareness and point to relevant references for further reading.

In simple scenarios with a single acoustic source, and an AS that is, e.g., supposed to direct a camera to any sound source, the identification reduces to detecting a certain sound level and localizing the origin of the sound. In a more complex smart-home scenario, where the intention of a humanoïd robot may be to react upon certain utterances of a specific person, this involves: 1) detection and classification of all sound sources; 2) possibly authenticating the voice of a specific speaker; 3) recognizing multiple speech streams; and 4) understanding their content, before the action can be defined. Obviously, source detection and classification needs to precede a distinction between desired and undesired sources. The task of detecting occurrences of a specific type of sound(s) in an audio recording is referred to as acoustic event detection (AED), which is a rapidly growing research area in audio-related signal processing and machine learning [157]. Closely related to AED is acoustic scene classification (ASC), which classifies recordings along acoustic environments (such as street life, subway station, and living room) without classifying individual sound sources. An overview of state-of-the-art approaches for AED and ASC can be found in [4]. So far, most approaches in AED and ASC are based on single-channel recordings, sometimes also jointly processed with data from other modalities [158], e.g., cameras for audio-visual event detection. Features and acoustic models for describing sound events, and classification methods are largely adopted from the wealth of techniques offered by state-of-the-art ASR [4]. However, it should be noted that some recently proposed features were specifically designed for AED and ASC (see [159] and [160]). An overview of publicly available data sets for sound classification and detection, including a comparison of their structure, sizes, and annotation quality can be found in [161]. While the performance of single-channel approaches for AED and ASC rapidly degrades with the growing complexity of the scenario if the acoustic events are not separated in time or frequency, promising multichannel techniques for exploiting the spatial information are still in their infancy [4].

Despite the enormous recent progress of ASR and NLU supported by efficient deep learning techniques and large amounts of training data [162], AS-typical scenarios, like a smart home, with a combination of large distances between target sources and microphones, a potentially large variety and variability of previously unseen noise sources, interfering voices, and reverberant acoustic environments are still posing serious challenges for the state of the art in ASR [162], [163].

III. CONTRIBUTING TO A WORLD OF SOUNDS: EGO-NOISE

With its self-created noise or ego-noise, the AS has a self-inflicted impact on the acoustic scene which distinguishes ASA for ASs decisively from conventional ASA. To allow the AS to pursue its intentions even while ego-noise is emitted, robust strategies and algorithms are required to cope with ego-noise [11], [165], [166]. In the following, we first discuss the various origins and resulting properties of ego-noise. Then, in Section III-B, generic approaches for robust localization and source extraction under ego-noise are introduced. This discussion is continued in Section III-C by introducing ego-noise modeling and estimation approaches that explicitly incorporate knowledge about the AS and the properties of ego-noise. An overview of all presented methods for ego-noise estimation and modeling is given in Table 1.

A. Origins and Properties of Ego-Noise

Typically, an AS has various ego-noise sources producing signals that generally have specific temporal, spectral and spatial characteristics. In many cases, there is a primary ego-noise source, e.g., a motor. In addition, the interaction of an AS with the physical environment also causes ego-noise, e.g., the noise of the footsteps of a humanoïd robot or the tire noise of a self-driving vehicle. A common property of ego-noise—independent of the application—is that it is usually louder than a signal of interest since the ego-noise sources are typically located in the immediate proximity of the microphones.

Ego-noise was first investigated in the context of robot audition for humanoïd robots [11], [165], [166], where ego-noise is primarily caused when the robot is moving and rotating joints as well as the moving parts of its body cause significant noise. The main motivation to develop robust algorithms for ASA in the presence of ego-noise was the quest for an intuitive human–robot interaction by overcoming the limitations of the so-called stop-perceive-act principle [11], where the humanoïd needs to stall its activity while sensing acoustic signals. A typical ego-noise spectrogram recorded by a humanoïd robot is shown in Fig. 3(a). Ego-noise is highly nonstationary as the robot moves with varying speeds and accelerations. Furthermore, ego-noise cannot be modeled as a single static interfering point source as, e.g., the noise originating from each joint is radiated to the environment as structure-borne sound (see Fig. 4). On the other hand, ego-noise often exhibits characteristic spectral structure, e.g., harmonic components, and distinctive radiation characteristics as the different ego-noise sources, e.g. joints, are distributed over the robot's body (see Fig. 4). Both spectral and spatial characteristics of ego-noise can be used advantageously for its modeling (see Section III-C).

Ego-noise of drones mainly consists of multiple narrowband harmonic components caused, similar to humanoïds, by motors and broadband noise due to the airflow by the propellers and wind [167]. Fig. 3(b) shows an ego-noise spectrogram of a single motor and propeller rotating with constant speed. The visible harmonics would exhibit highly nonstationary transient characteristics if the rotation speed varied. If four simultaneously active motors and propellers are considered, the harmonic structure of the propeller noise becomes less pronounced [see Fig. 3(c)], which renders ego-noise suppression for drones to a highly challenging task. On the other hand, however, the relative position between motors and microphones is fixed, which has been proven to be beneficial for robust SSL and source extraction [60]. Significant effort was spent reducing the harmful impact of ego-noise of drones by construction, e.g., optimizing a drone's propulsion system [168], [169] or by the optimum placement of the microphones on a drone. For this, Ishiki and Kumon [170] propose to simulate the intensity of ego-noise at different locations using a kinematic model of the drone.

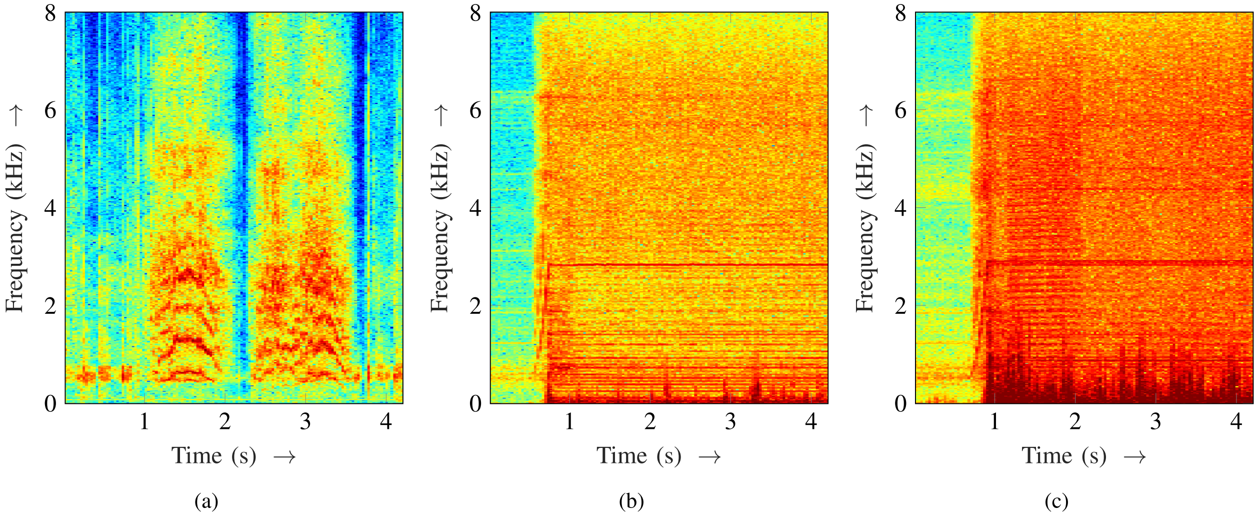

Fig. 3. ***Spectrograms of exemplary ego-noise recordings (logarithm of magnitude). (a) Ego-noise of a right arm waving movement of the humanoïd robot NAO, SoftBank Robotics. The movement incorporates all six joints in the arm. (b) Motor and propeller ego-noise of an quadrotor UAV MK-Quadro, MikroKopter. The ego-noise originates from a single motor rotating with constant speed. (c) Ego-noise of all four motors and propellers for the UAV from (b). Recordings (b) and (c) taken with permission from [164].***

Since the primary ego-noise source is the AS itself, knowledge about the current internal state of the AS can provide important information about the emitted ego-noise. Inspired by this, various approaches investigated the question of how this knowledge can be beneficially used in addition to the recorded audio signals. In the following, we refer to information about the internal state of an AS as nonacoustic reference information (NARI) aside from acoustic reference signals like loudspeaker signals. Various kinds of NARI have been employed in the literature, e.g., ratios between pulsewidth and period length in pulsewidth modulation (PWM) for electric motors [171] or piezoelectric sensors measuring the vibration of a drone [172]. Most prominently, motor data are used, i.e., angle information collected by proprioceptors of joints or rotation speed of motors, therefore immediately describing the primary source of ego-noise. Integrating motor data to the audio processing pipeline is generally referred to as audio-motor integration [173] and corresponding methods will be discussed in Section III-C.

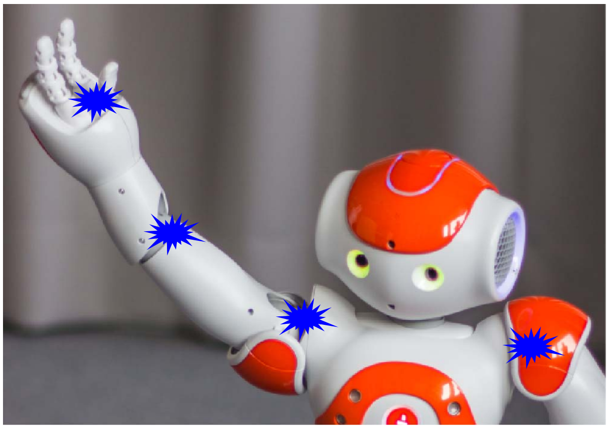

Fig. 4. ***Humanoïd robot NAO waving its arm. During the movement, different joints are activated emitting ego-noise (hypothesized source positions shown in blue). This motivates modeling ego-noise comprising multiple spatially distributed sources.***

B. Generic Methods for Robust SSL and Source Extraction

Since ego-noise heavily affects the microphone signals, the question of how to achieve robust performance for subsequent signal processing algorithms has gained considerable attention in the last two decades. In this section, we discuss generic approaches both for localization and signal enhancement, i.e., methods which do not or only very weakly rely on the characteristics or an explicit model of ego-noise.

1) Ego-Noise Suppression Using Reference Signals: As one of the first approaches for ego-noise reduction, the SIG humanoïd robot [11] was equipped with additional microphones mounted inside the robot's housing near the motors in order to record an ego-noise reference signal. The internal microphones were interpreted as additional auditory perception channels of the robot and were subsequently used as reference signals for adaptive filtering-based ego-noise cancellation. Internal microphones were also employed in [174] for improving the performance of a human–robot dialogue system, which is exposed to ego-noise as well as external environmental noise sources. The authors propose a frequency-domain semi-blind source separation algorithm to estimate the noise signals and obtain an enhanced desired signal by applying an MWF. The idea of reference microphones was also adopted for drones, where separate microphones were mounted next to the propellers, e.g., in [175], which applies the conventional least mean square (LMS) adaptive filtering for noise cancellation. In general, the noise reduction performance of such reference signal-based methods depends heavily on how accurately the noise in the sensor signal can be estimated from the reference signal. Especially in the context of drones, this is a crucial problem since the microphones are typically not shielded from other sounds [175].

2) Spatial Filtering-Based Signal Extraction for Drones: Other methods address the extraction of desired signals from drone ego-noise by spatial filtering-based techniques. In [176], the performance of ICA-based BSS and several beamforming techniques was compared for extracting a target source at a known position from a microphone signal recorded by a drone hovering with varying motor speeds and distances to the target source. For MVDR beamforming, the required spatial correlation matrices of ego-noise were estimated either incrementally (see below) or during noise-only periods and then kept constant. In addition, an MWF was evaluated for source extraction, assuming that ego-noise and the target signal are sparse in the STFT domain, i.e., each time–frequency bin is dominated by one of the two sources. Based on this sparseness assumption, a probabilistic time–frequency mask can be computed to estimate the spatial correlation matrix of the desired source signal [177], [178]. In [167], it is argued that the sparseness assumption is valid for the harmonic components of drone ego-noise, which for broadband propeller noise holds as an approximation at best [see Fig. 3(c)]. The evaluation in terms of SNR improvement shows best results for BSS- and MWF-based techniques, especially for high SNR values, i.e., if ego-noise is less dominant compared to the target source.[2] In [179], a similar comparison is presented; however, it assumed that the position of the target source is not known and its location is estimated by visual information captured by a camera mounted to the drone. In [180], a fixed BF with subsequent postfilter is proposed to enhance a desired source signal. For this, the drone is equipped with two microphone arrays, installed both on a metallic beam attached to the UAV. Ego-noise is assumed to be comprised of several coherent noise sources and its PSD is estimated in the beamspace [181] by directing a set of nonadaptive BFs to different directions in space. The BF output PSDs are subsequently used to estimate the required ego-noise PSD which is employed for a WF-based postfiltering. For simulated data, it is shown that the proposed method significantly reduces ego-noise, while also robustly suppressing an additional interfering speaker.

3) Spatial Filtering-Based SSL for Drones: Robust SSL under ego-noise is usually addressed by GEVD-MUSIC, i.e., whitening the microphone array's spatial correlation matrix in a preprocessing step (see Section II-B). For obtaining reliable estimates for the spatial correlation matrix of the emitted ego-noise, Okutani *et al.* [37] used a time-averaging method for drones (called incremental GEVD-MUSIC, iGEVD-MUSIC), taking the dynamically changing characteristics of ego-noise into account. In [38], this idea is extended for GSVD-MUSIC (iGSVD-MUSIC). The duration of the averaging window is adapted to the noise dynamics, i.e., a large window can be chosen if the drone is hovering, while the windows need to be short if the drone is flying with varying speeds and accelerations. As an intrinsic problem of these methods, performance degrades drastically if, within the averaging window, other sources become active and bias the estimation. In [182], two different UAV microphone array designs are presented and MUSIC and iGSVD-MUSIC are evaluated and compared for an outdoor SSL scenario. SSL success rates, defined as the ratio between the number of successful and total number of localization experiments, of almost 100% are reported even for low SNRs. It is proposed to adapt the algorithms specifically to the scenario and emphasized that the significant computational costs are a major challenge for real-time applicability. In [183], an alternative SSL method was presented localizing a single source. A set of MWF-based spatial filters, constructed similar to the method presented in [176] (see Section III-B2), are steered to different candidate DOAs. If pointing toward a speech target source, the filter output is argued to be non-Gaussian which was detected using a Kurtosis-based criterion. The approach shows superior results compared to different competing methods including iGEVD-MUSIC.

C. Ego-Noise Modeling and Estimation Methods

The use of classical denoising techniques such as spectral subtraction or single- and multichannel Wiener filtering (see Section II-C) requires ego-noise estimates, e.g., in the form of PSDs or spatial correlation matrices. Beyond the discussion of generic methods in Section III-B, we now consider approaches that use signal models to explicitly take into account the spectral and spatial properties and characteristics of ego-noise.

1) Generic Dictionary Approaches: The basic idea of a dictionary representation is to approximate an ego-noise signal $\mathbf{N}_{\mathrm{EN},ft}$ in the STFT domain by a linear combination of prototype signals, called atoms, which are specifically designed to match the characteristics of ego-noise. The atoms are collected in a dictionary, and for each time frame, a linear combination of those atoms has to be found which optimally fits the current ego-noise signal with respect to the chosen criterion. For ego-noise modeling, a dictionary is typically learned from recordings which contain ego-noise only as a data set for training. Subsequently, the trained dictionary can be employed to estimate the ego-noise for an STFT frame which contains ego-noise and other signals, e.g., speech. Typically, these approaches are semi-supervised [184] and entirely based on information from the audio modality.

[2] Audio demos for methods described in [176] could be found on //www.eecs.qmul.ac.uk/~andrea/auditory-mav.html at the time of writing.

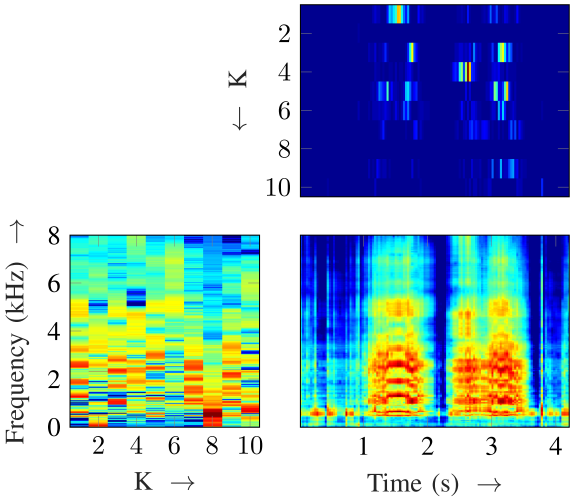

Fig. 5. ***Illustration of the dictionary method: dictionary with K = 10 atoms (bottom left), activation matrix (top right), and estimated spectrogram (bottom right).***

In the following, we consider single- and multichannel nonnegative matrix factorization (MNMF), denoted as SC-nonnegative matrix factorization (NMF) [69], [185], [186] and MC-NMF [187], [188], respectively, and phase-optimized KSVD (PO-KSVD) [189] as exemplary dictionary learning methods.

The objective of NMF is to model PSDs of ego-noise, i.e., spatial correlation matrices for MC-NMF, which degenerate to scalar PSD values for SC-NMF. SC-NMF aims at approximating the squared magnitude of the mth microphone channel $\tilde{\mathbf{N}}_{\text{EN}} = |\mathbf{N}_{\text{EN}}^{(m)}|^2 \in \mathbb{R}_+^{F\times T}$ [190] by a product of the nonnegative dictionary $\mathbf{D} \in \mathbb{R}_+^{F\times K}$ and the activation matrix $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_T] \in \mathbb{R}_+^{K\times T}$

$$\tilde{\mathbf{N}}_{\text{EN}} \approx \mathbf{DH} = [\mathbf{Dh}_1, \dots, \mathbf{Dh}_T]. \tag{8}$$

The factorization is achieved by minimizing a cost function which measures the similarity between $\tilde{\mathbf{X}}$ and $\mathbf{DH}$ with respect to the model parameters. A common choice as cost function for audio applications is the so-called Itakura-Saito (IS) divergence since it depends only on the power ratios between the true and approximated signal [191]. $\mathbf{D}$ and $\mathbf{H}$ are typically obtained using iterative update rules that can be derived using, e.g., majorization–minimization algorithms [192] or heuristic approaches [186]. Fig. 5 illustrates the fundamental idea of SC-NMF by estimating the PSD of an exemplary ego-noise signal. Semi-supervised approaches for SC-NMF for noise reduction, in general, were initially addressed in [184] and specifically applied for the reduction of ego-noise, e.g., in [193].

For MC-NMF, (8) is adopted as source model and extended by an additional spatial model. Following this idea, MC-NMF can be applied to the estimation of spatial correlation matrices of ego-noise assuming again that ego-noise originates from various mechanical parts of a robot or drone and thus is reasonably well modeled by a set of time-variant, spectrally structured and spatially distributed sources. Considering a dictionary with K atoms $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_K]$, the spatial correlation matrix of ego-noise for frequency bin f and time frame t is given by

$$\hat{\mathbf{\Phi}}_{\text{EN},ft} = \sum_{k=1}^{K} \mathbf{\Phi}_{fk} d_{fk} h_{kt} \tag{9}$$

where d_{fk} is the fth entry of atom k, $\mathbf{d}_k$, modeling the squared STFT magnitudes of the ego-noise sources. $\mathbf{\Phi}_{fk} \in \mathbb{C}^{M\times M}$ is an estimated correlation matrix modeling the spatial characteristics of d_{fk}. Thus, while the atoms in $\mathbf{D}$ describe the spectral properties of each ego-noise source, matrices $\mathbf{\Phi}_{fk}$ assigned to each atom add the spatial characteristics to the model. The contribution of each atom and spatial correlation matrix to the overall $\hat{\mathbf{\Phi}}_{\text{EN},ft}$ is determined by activation h_{kt} which allows a flexible and time-varying modeling of $\hat{\mathbf{\Phi}}_{\text{EN},ft}$. Depending on the assumptions for $\mathbf{\Phi}_{fk}$, different algorithms to learn a dictionary and the associated spatial correlation matrices can be derived, varying widely in complexity and modeling accuracy. For example, $\mathbf{\Phi}_{fk}$ can be assumed to be a rank-1 matrix [194] or alternatively full-rank [188]. MC-NMF-based ego-noise modeling has successfully been used for ego-noise reduction for a humanoïd robot in [195], following a semi-supervised approach (see Section III-C1) and an MWF. In [196] and [197], a rank-1 MC-NMF model is applied to blindly estimate a demixing filter to extract speech from the ego-noise of a hose-shaped rescue robot. This approach, however, assumes oracle knowledge to solve the permutation problem after demixing, i.e., to identify the channel which contains the desired signal component. A similar approach for ego-noise suppression was presented in [198] and is shown to require significantly less computational effort compared to a robust PCA-based method at the cost of reduced noise suppression. This illustrates the typical tradeoff between computational effort and achievable performance in the design of algorithms for ASs.

While SC-NMF captures only the spectral characteristics of ego-noise, multichannel approaches are able to capture also the characteristic spatial structure of ego-noise. A corresponding approach explicitly addressing ego-noise reduction was presented in [189]. The assumed signal model interprets each atom as a contribution of a set of sound sources that are spatially distributed over the body of the robot (see Fig. 4). Per frequency bin, the M microphone channels are concatenated, giving a signal matrix $\tilde{\mathbf{N}}_{\text{EN}} \in \mathbb{C}^{MF\times T}$ and a dictionary $\mathbf{D} \in \mathbb{C}^{MF\times K}$. In [189], a time-varying phase matrix $\mathbf{\Phi}_t \in \mathbb{C}^{F\times K}$ is introduced that allows to adjust the phase of the atoms, e.g., in order to compensate for TDOAs for the noise components of the various hypothesized ego-noise sources. This phase-corrected dictionary is denoted by $\mathbf{D}\{\mathbf{\Phi}_t\}$, where the curly brackets indicate that the ftth element of $\boldsymbol{\Phi}_t$ is multiplied with the M bins associated with the frequency index f in atom t. Analogously to (8), $\tilde{\mathbf{N}}_{\text{EN}}$ is then modeled by

$$\tilde{\mathbf{N}}_{\text{EN}} \approx [\mathbf{D}\{\boldsymbol{\Phi}_1\}\mathbf{h}_1, \ldots, \mathbf{D}\{\boldsymbol{\Phi}_T\}\mathbf{h}_T] \tag{10}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_T] \in \mathbb{C}^{K \times T}$ is the activation matrix. While in (8) the number of activated atoms per time frame is not bounded, Deleforge and Kellermann [189] constrained (10) such that only a fixed number of atoms can be chosen per time frame. This constraint is based on the assumption that the gain and the activation of each source is sparse along the time axis. The dictionary is learned using the PO-KSVD algorithm which is based on [199], including an additional phase optimization step. Subsequently, the best matching atoms from the dictionary can be chosen using a sparse coding algorithm [189].

2) Motor Information-Guided Methods: Besides methods purely based on audio signals, other ego-noise modeling approaches exploit available motor information given by, e.g., (software) motor commands or motor data such as engine rotation frequency, joints' angle, or angular velocities collected by proprioceptors.

The benefit of motor data compared to motor commands is that the emitted ego-noise is directly related to the measured physical state of the robot which may differ from the reference physical states defined by the motor command. However, a complete analytical modeling of the dependence between motor data and emitted ego-noise is usually infeasible due to the complex mechanical dependencies and interactions between structure- and airborne sounds. Nevertheless, some prior knowledge of the structure and shape of ego-noise spectrograms can be straightforwardly inferred from motor data. In some cases, ego-noise exhibits pronounced harmonic, deterministic spectral components caused by rotating engines, appearing at half of the rotation frequency of the motors. In [130], this knowledge is used to perform an order analysis (OA) [203] of the signal, so that the position of the harmonics becomes independent of the rotation frequency, which allows for simple subtraction-based denoising.

Other approaches model the complex dependencies between motor data and emitted ego-noise entirely by learning-based strategies, which, similar to dictionary methods, involve a prior training of the model that in turn can subsequently be used for ego-noise estimation in mixtures of multiple sounds. In [127], a feedforward neural network with two hidden layers containing thirty nodes each is trained to predict the PSD of ego-noise caused by the Aibo robot. The neural network is fed with angular velocities of Aibo's joints collected by proprioceptors. The obtained estimate was subsequently used for spectral subtraction resulting in a significant improvement of speech recognition rates, especially for low SNRs. In [128], an approach is considered where the characteristic spectral shape of the ego-noise for each movement was saved as PSD templates in a database and associated with the motor command which triggers the movement. During application, an additional preprocessing stage is required to ensure that the templates are temporally aligned to the recorded signal. An alternative template-based method was proposed in [129] and [200], where the choice of spectral templates is, however, not motor command- but motor data-guided. For a new motor data sample, the nearest neighbor in the motor data-space is searched and the associated ego-noise template is used for spectral subtraction. The idea to associate motor data with ego-noise templates was also followed in [201], where, however, nonlinear classifiers in the motor data-space are used to link a new motor data sample to a set of atoms from a previously trained dictionary-based ego-noise model. Therefore, the classifiers replace the costly search for atoms in the dictionary.

In [204], prior knowledge of the harmonic structure is incorporated into the learning-based ego-noise modeling. It is proposed to decompose the observed ego-noise spectrogram into a part which captures the harmonic structure and a residual part. Each component is modeled by an NMF-based dictionary to yield

$$\tilde{\mathbf{N}}_{\text{EN}} = \tilde{\mathbf{N}}_{\text{EN}}^{(\text{h})} + \tilde{\mathbf{N}}_{\text{EN}}^{(\text{r})} = \mathbf{D}^{(\text{h})}\mathbf{H}^{(\text{h})} + \mathbf{D}^{(\text{r})}\mathbf{H}^{(\text{r})} \tag{11}$$

where $\mathbf{D}^{(\text{h})}$ is completely motor data-driven and only $\mathbf{D}^{(\text{r})}$ requires prior training. This approach is evaluated for speech enhancement and outperforms an audio only-based method in scenarios which are insufficiently captured in the beforehand training.

3) NARI-Based Estimation of Spatial Correlation Matrices: Similar to the methods described in Section III-C2, spatial correlation matrices of ego-noise can be estimated based on NARI describing the physical state of the AS. For this, several approaches have been proposed in the context of estimating $\boldsymbol{\Phi}_{\text{EN},ft}$, which is subsequently employed for GEVD-MUSIC. In [171], NARI is composed of pitch-, roll-, and yaw-angle, as well as the vertical and horizontal speed of an UAV. In addition, the ratio between pulsewidth and period length of the propulsion system's PWM is considered. The resulting parameter reference vector subsequently serves as input to a Gaussian process (GP) regression framework [205] to estimate the vectorized spatial correlation matrix of ego-noise. Localization results show an improvement compared to conventional MUSIC, especially for high SNR values.

Inspired by the search for ego-noise PSD templates in a motor data-space, Ince *et al.* [202] extended this idea to the estimation of spatial correlation matrices by the outer product of identified ego-noise templates. The presented approach showed a significant reduction of the mean localization error compared to conventional MUSIC for both sinusoidal and white noise target signals.

Table 1 Overview of Different Ego-Noise Modeling and Estimation Methods

Task	Type of AS	Ref.	Method & Objective	Sensors	
				# Mic	ref. parameter
Speech Enhancement	humanoïd robot	[11]	reference microphones, adaptive noise cancelling	2	-
	robot	[174]	reference microphones, semi-blind source separation	3	-
	UAV	[176]	spatial filter based on inst. DoA estimation	8	-
	humanoïd robot	[193]	dictionary, WF-type spectral enhancement	1	-
	humanoïd robot	[189]	phase-corrected dictionary, spectral subtraction	4	-
	UAV	[130]	order analysis-based spectral subtraction	1	motor data
	robot dog	[127]	neural network-based noise estim., spectral subtraction	1	motor data
	humanoïd robot	[129], [200]	template-based noise estim., spectral subtraction	1	motor data
	humanoïd robot	[128]	template-based noise estim., spectral subtraction	1	motor commands
	humanoïd robot	[201]	motor data-guided dictionary, spectral subtraction	4	motor data
	hose-shaped robot	[196], [197]	MNMF, adaptive noise cancelling	8	-
	humanoïd robot	[195]	MNMF, MWF-type spectral enhancement	4	-
Localization	UAV	[37]	incremental estim. of spatial cov. matrix, iGEVD-MUSIC	8	-
	UAV	[38]	incremental estim. of spatial cov. matrix, iGSVD-MUSIC	16	-
	UAV	[183]	comparing output statistics of spatial filters	8	-
	UAV	[171]	GP-based spatial covariance estim., GEVD-MUSIC	8	PWM, kinematic values
	humanoïd robot	[202]	template-based noise estim., GEVD-MUSIC	1	motor data

IV. EXPLORING A WORLD OF SOUNDS: ACTIVE SENSING

While in the previous sections the surrounding acoustic scene was already allowed to be dynamical, the capability of the AS to move has only been viewed as a source of ego-noise. In the following, we will consider the AS's ability to move actively in space. Then, motion implies both: 1) the capability to change and adapt its own pose and internal sensor topology and 2) changing its position in a room. As main promises, the capability to move allows an enhanced localization and source extraction performance by overcoming limitations linked to static sensor arrangements such as front-back disambiguation or the problem of range estimation in localization. Besides, it allows the AS to actively explore its environment and gain knowledge about the surrounding acoustic scene. The idea to combine audition with behavior and motion, referred to as active audition, was initially proposed by Nakadai *et al.* [11].

Inspired by humans and mammals, we first review methods which utilize the AS's capability to move explicitly for SSL. Subsequently, we address active localization and exploration, i.e., using motion on purpose to enhance localization and to explore the environment. This involves several subproblems such as fusing localization measurements over space and time, self-localization, and evaluation of different motion strategies. Finally, we present approaches that demonstrate that motion can be beneficial for speech enhancement, referred to as active enhancement [11]. Fig. 6 summarizes the different tasks for active sensing with an AS.

A. Motion-Based Auditorimotor Maps

Approaches presented in this section fall into the category of binaural listening (see Section II-B2) and require head rotation for localization. If we assume that an AS is localizing an emitting acoustic source in a room, localization can be interpreted as an association of auditory features, like ILD or IPD, to particular points in space. These mappings are referred to as auditorimotor maps [206] or audio-motor mapping [207], since the perceived audio features depend in general on the motor state of the AS, describing, e.g., position, orientation, and velocity in the room relative to the emitting source. It is generally assumed that these mappings can be learned. An example of such a learning method is based on the sensorimotor theory of perception [208], [209], suggesting that experiencing the sensory consequences of motor actions is necessary for learning and that the auditorimotor maps can be represented by low-dimensional manifolds of according dimensions [206], [207], [209], [210]. Following this idea, several approaches have been implemented for binaural localization for humanoïd robots using specific binaural cues as audio-motor features. In [206], Laplacian eigenmaps are used as manifold learning technique and constructed by an unsupervised approach using so-called auditory-evoked orientation behavior: as soon as an auditory event is perceived, the head is rotated until the first zero-crossing of the ILD is detected. The performed head rotation is associated with the initially perceived audio-motor feature. Once a map is built, the nearest neighbor in the map is identified for any new audio-motor feature and the associated head rotation is performed.

In [210], a modified version of self-organizing maps (SOMs) is used to represent the manifold. SOMs are a specific type of neural network, which should represent the similarity between the audio features, i.e., similar audio vectors are supposed to activate similar regions of the map. A map-architecture is suggested in [210], where the center of the map is activated if an audio feature originating from a source facing the robot is the input. Based on this, a reinforcement learning-inspired method for localization is proposed with the objective to turn the robot's head until the center of the map is activated. Movements that increase

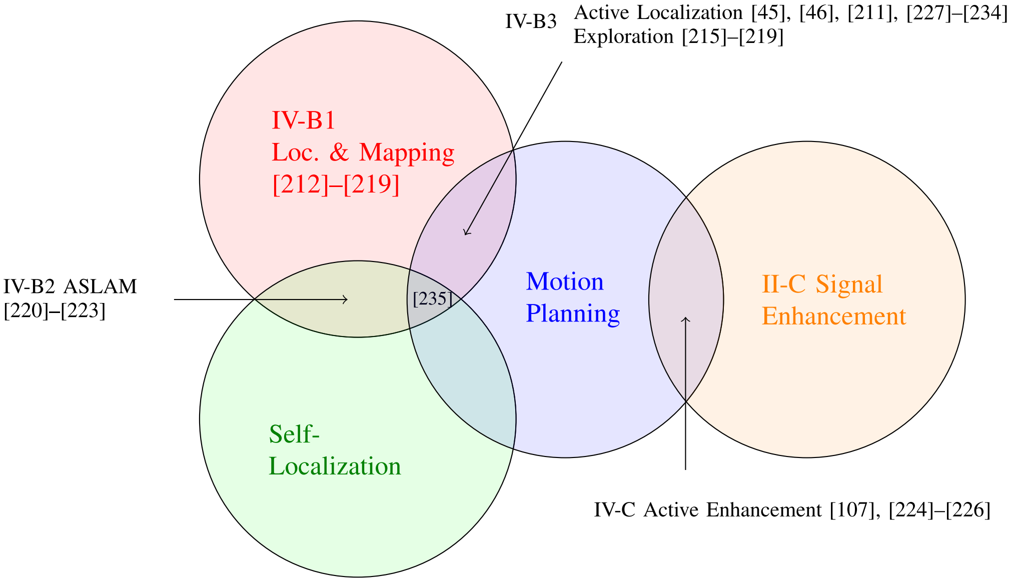

Fig. 6. ***Overview of different tasks for active acoustic sensing by a mobile AS.***

the distance to the center are penalized, while movements toward the center are rewarded. A similar approach addresses the problem of finding a low-dimensional representation with a regression-based method [207].

In [173], dynamic binaural cues are proposed, which explicitly require a head rotation for computation if the acoustic scene is assumed to be static. For a head rotation with constant speed, the dynamic cue is assumed to be specific for a certain source position. The auditorimotor map is learned in a supervised learning-based SSL framework [211].

B. Active Localization and Exploration

While the methods in Section IV-A were limited on binaural localization, we now turn to approaches for microphone array-equipped robots which use motion beyond head rotation for localization and exploration. After discussing how to fuse localization measurements over space and time, we review methods for self-localization, which is a necessary requirement for the subsequently presented motion strategies. An overview of the different tasks and their relations is given in Fig. 6.

1) Localization and Mapping: In the following, we discuss how localization measurements collected over time and space can be integrated into a single localization estimate. Moreover, we consider approaches which focus on integrating localization estimates into a map, representing the acoustic scene. At this point, it is not specified how the movements are planned—they could be, e.g., predefined or controlled by a user. Dedicated approaches for motion planning are presented in Section IV-B3.

If the source is assumed to be static, DOA estimates collected at different positions can be used for triangulation to estimate the distance to the source and has been proposed for robots in [212] and [213]. Estimating the orientation and distance of an acoustic source over time by a moving robot has been successfully implemented in real time using nonlinear extensions of a KF [214]. This approach was extended in [236], accounting explicitly for false measurements and intermittent source activity.

While the latter approaches localize and track a specific, however possibly time-varying, number of sources while moving, other methods aim at mapping the entire acoustic scene. Occupancy grids are one of the most successful frameworks for environment modeling in robotics [237] and have been introduced for mapping of acoustic scenes, then called auditory evidence grids (AEGs), in [215]. AEGs have been successfully implemented on mobile robots [215]–[217]. In this concept, a room is represented by a 2-D or 3-D map m consisting of an evenly spaced grid, where each cell m_i is associated with a probability that any sound source is located in this cell. The goal of AEG is to compute the posterior probability for the map m at time step t given all localization measurements $z_{1:t}$ and positions/poses of the AS $s_{1:t}$ collected so far

$$p(m|z_{1:t}, s_{1:t}).$$

Similar to the original occupancy grid idea, the AEG framework is based on the assumption that the occupancy in each grid cell is independent of all neighboring cells, which allows to formulate iterative Bayesian updates that are based on an inverse sensor model

$$p(m_i|z_t, s_t)$$

describing the probability that cell m_i is occupied given the current localization measurement and pose of the AS.

In [215]–[217], a rescaled version of the GCC-PHAT response is considered as pseudo-likelihood for $p(m_i|z_t, s_t)$, where the pose information of the AS s_t is measured by laser light sensors mounted on top of the AS. By moving into a room, scanning the environment from different positions, and updating the map, the AS can successively create an acoustic image of its surroundings. A similar approach was presented in [218], in which a PF was used to localize the AS within the map and a laser detection and ranging (LIDAR) sensor was employed to estimate the distances to hypothesized sound sources. Thus, the problem of poor range estimation from audio localization techniques could be overcome to some extent. The fundamental assumption of AEG to treat the occupancy in each cell independently of its neighboring cells can result in inconsistent maps [232], [238]. To solve this problem, forward sensor models have been proposed [238], specifying a probability distribution for localization measurements z_t given a map m. An according approach for SSL by optimal control of a robot will be presented in Section IV-B3.

An alternative acoustic mapping approach was proposed in [219], where radiated sound intensities are estimated instead of occupancy probabilities. Similar to [218], cells in which possible sound sources are located are identified by matching DOA measurements with objects in the scene recognized by an onboard LIDAR sensor. The corresponding cells are filled by the measured sound intensity. The presented grid-based approaches for estimating maps of the acoustic environment have all been evaluated on robots and are real-time capable. To the best of the authors' knowledge, there exists no study comparing the different methods among each other.

2) Self-Localization and Mapping: The aforementioned methods for fusing localization measurements are based on the assumption that the AS has knowledge about its own position in space, either by being tracked by an external system or alternatively using own sensor data, e.g., LIDAR or radar sensors. Employing motor controls or inertial sensor data is reported to be subject to errors due to physical and mechanical limitations (see [239], [240])

Alternative approaches aim at creating maps of the environment, while simultaneously estimating the location of the AS within this map [simultaneous localization and mapping (SLAM)], which can be interpreted as anchoring the AS's position and orientation such that both align best to the localization estimates in the map. The SLAM problem has gained significant attention in the robotics community [241], [242], where vision is predominantly used to obtain instantaneous localization estimates. In contrast, using the acoustic modality for SLAM is still in its early stages of the development. A multipath propagation-based concept was proposed in [220], referred to as EchoSLAM where the position of the AS was inferred with the help of echoes perceived at multiple viewpoints in a room. Interestingly, only one microphone is required in this approach.

Most other approaches extend concepts from visual SLAM by TDOA-based estimation using microphone arrays [221], [222]. However, the prerequisites for vision-based SLAM are often fundamentally conflicting with the properties of acoustic sources. For example, factored solution to SLAM (FastSLAM) [243] requires landmarks in the scene assisting the AS to locate itself within the map. However, acoustic sources are not appropriate for this, since they are not continuously active and a permanent emission of sound stimuli is usually highly undesirable.

To address this challenge, acoustic SLAM (aSLAM) was proposed [223]. To avoid the need for permanently active acoustic landmarks in the scenery, Evers and Naylor [223] introduce PHD filters for SLAM, explicitly modeling multiple, intermittent sources subject to erroneous and missing DOA estimates. As a consequence, the approach is reported to be robust against reverberation and noise. The locations of the acoustic sources are inferred from 2-D DOA measurements, making use of a novel probabilistic triangulation method benefiting from the mobility of the ASs.

3) Motion Strategies for Active Localization and Exploration: In this section, we present concepts that use the capability of an AS to move actively to improve localization, referred to as active localization, and to explore the acoustic environment. Both active localization and exploration have in common that they are feedback-controlled, i.e., based on a current localization measurement or acoustic map, the behavior of the AS is adapted.

Approaches presented in Section IV-B1 for building an acoustic map do not fall into this category, since there the movement of the AS is either predefined or user-controlled, i.e., a specific feedback loop is missing. However, Martinson [215], [216] stipulated a subsequent exploration stage to refine the previously established acoustic map. For this, the AS defines waypoints around a hypothesized source in the acoustic map, which are subsequently approached according to the nearest-neighbor-principle. The AS stops at each waypoint, collects samples, and finally refines the position of the investigated source within the map.

A variety of active localization strategies are human-inspired and heuristically adopted for robots, mainly to solve the front-back-ambiguity problem in the context of binaural localization. Motion is then typically limited to horizontal head movements, where the head turns randomly within a predefined range [45], [46], [227]. In [228], also tipping movements of the head are considered. In other approaches, the head is turned by a predefinded range toward the hypothesized source [210], [227], which implies that these approaches are feedback-based, i.e., a detected front-back-ambiguity triggers the head movement, followed by a subsequent localization and possibly further head movement.

While these head movement strategies were heuristically motivated, an analytic strategy to design optimal movements for localization is presented in [244] and [245] for localization in the SH domain. If the acoustic source is assumed to be static, sampling the sound field by a moving array can be thought of as producing additional virtual microphones at each of the array positions in the different time frames. The quality of this virtual microphone array can be measured by the effective rank [246] of the so-called measurement matrix, which can be shown to be directly related to DOA estimation performance. In [245], the effective rank for varying head movements and microphone geometries was evaluated, confirming its effectiveness as a measure to design movements.

An information-theoretic approach for the design of head movements for active binaural localization has been presented in [229] and [230]: first, for timestep t, a posterior probability describing the head-to-source relation p_t given all previously collected measurements $z_{1:t}$ is derived, $p(p_t|z_{1:t})$. Based on this, an information-theoretic one-step-look-ahead control scheme is proposed, determining the head movement which maximizes the posterior probability $p(p_{t+1}|z_{1:t+1})$ computed by a gradient ascent method. The evaluation demonstrates that limitations linked to static sensor arrangements such as front-back ambiguities or missing range information can be overcome. Using a similar approach, in [231], the movement of external movable reflectors mounted next to the head microphones of a robot, mimicking the pinnae of animals, is determined.

An extension to the previously described information-theoretic approach has been used in [232], where arrays with four microphones are considered and the AS is able to change its position in a room. Here, AEGs from Section IV-B1 are extended by learning forward sensor models expressed by

$$p(z_t|m_i, s_t). \tag{12}$$

This method avoids inconsistencies in the map and allows to compute an entropy over all grid cells bearing uncertainty information about the position of the acoustic source. Over a finite time horizon, the expected map entropy is accumulated assuming that the entropy at each future pose does not depend on the trajectory used to reach that pose. This simplifying assumption is required for the employed dynamic programming method to identify those poses which minimize the accumulated entropy. In [233], Gaussian mixtures instead of AEGs are chosen to represent the belief over the acoustic scene based on current measurements, and Monte Carlo tree search (MCTS) is proposed to find optimal sequences of poses which minimize the expected entropy of the estimated source location. A similar approach was chosen in [234] for binaural active localization. However, here, the optimization is reformulated, allowing a tradeoff between two conflicting goals, namely, minimizing the localization uncertainty of a source and reaching a specific final target along the shortest path.

An exploration scheme that also considers self-localization within the acoustic map has been presented in [235]. Within an assumedly static acoustic scene, the state of the AS and the sources is tracked by an unscented Kalman filter (UKF). Each hypothesized source applies a potential field (PoF) on the AS, being composed of an attractive PoF, whose strength is determined by the uncertainty of the source's state, and a repulsive PoF, which on the one hand should avoid a collision of the source with the AS and on the other hand rewards circular trajectories around the source. For exploration, the AS is guided by following the steepest descent along the gradient of the superposition of the PoFs of all sources. By simulation, it was shown that the proposed method clearly outperforms other recent approaches regarding acoustic scene mapping performance, while at the same time being efficient enough to allow a real-time implementation on a robot.

C. Active Signal Enhancement

Complementary to active localization and exploration, the ability of an AS to move can also lead to an improved signal enhancement performance. Intuitively, a human would direct its head toward an acoustic source or approach it along a direct path in order to better understand the signal of interest. Indeed, both strategies were implemented and tested on mobile robot platforms, see [224] and [225], respectively. A more sophisticated path-planning strategy has been presented in [226]: If an acoustic map is given, i.e., the position and sound intensity of the desired and interfering sources are known, the expected signal-to-interference (SIR) ratio at each point in the room can be estimated, assuming that the signal amplitudes are proportional to the inverse of the source-AS distance. Then, the AS approaches the neighboring cell with the highest SIR, updates the map, and continues its approach toward the desired source. This method was implemented on a robot and was shown to be especially beneficial compared to directly approaching the source if an interferer is located along the direct path between AS and the desired source.

Exploiting additional degrees of freedom, Barfuss and Kellermann [107] and Tourbabin *et al.* [244] placed microphones not only on the head but also on the movable limbs of a humanoïd robot, thus allowing changes of the aperture of the microphone array, e.g., by letting the robot stretch out its arms or pull them back in. In [107], this adaptive microphone array is employed to control the spatial characteristics of a GC-BSS-based blocking matrix [247], used to estimate a noise reference, and thus to enhance spatial filtering performance.

V. SUMMARY AND OUTLOOK

Aiming at exploiting the acoustic domain for supporting self-awareness of ASs, this article attempts to provide a structured and comprehensive survey on relevant signal processing techniques for perception tasks in the perception–action cycle of ASs as cognitive dynamic systems. With multiple microphones as sensors, many well-established techniques for multichannel acoustic signal processing from other application domains can be adopted for ASA by ASs. Along with brief generic descriptions of relevant state-of-the-art methods, their adaptation to the goals and constraints of ASs is discussed and known realizations are described. The resulting overview covers the areas of source position estimation and tracking, as well as the various facets of signal enhancement, including spatial filtering, source separation, noise suppression, dereverberation, and echo cancellation, as far as they are deployed with ASs.

Beyond the challenges that ASs share with other acoustic application domains, ASs often need to cope with high levels of ego-noise and also should exploit their ability to move and explore for active sensing. It is demonstrated how ego-noise suppression can benefit from prior knowledge on its origins, and powerful and efficient techniques are presented which merge according models—often benefiting from NARI—with advanced learning techniques. Most of the proposed ego-noise suppression algorithms concentrate on humanoïd robots and drones, but even within these domains, concepts vary widely, and published results for the respective implementations are hardly comparable. Similar observations hold for the state of the art in active sensing: While the benefit of robot head movement for disambiguating binaural localization information has been recognized early on, other concepts for combining localization or self-localization with mapping in the acoustic domain are not yet developed beyond a prototypical stage. Unsurprisingly, active signal enhancement, either requiring location information as a precondition or estimating it simultaneously, also has not matured beyond this stage.

In summary, the state of the art suggests that, for supporting self-awareness of ASs, the acoustic modality is still much less exploited than humans use it. While robotic vision is a well-established and broad technical area, robot audition has not reached the same level of maturity yet. This may largely be attributed to the complexity of real-world acoustic scenes that ASs will typically face, where, e.g., unlike in the visual domain, focusing on a specific target already requires sophisticated spatial filtering and not just narrowing a viewing angle. Moreover, as a typical challenge for ASs in real-world acoustic scenarios, several signal processing tasks have to be solved simultaneously and in real time, e.g., algorithms for ego-noise suppression, echo cancellation, and BSS need to collaborate efficiently and demand a holistic algorithm design beyond the currently known single-task solutions. However, with recent progress in acoustic signal processing, ASA, and automatic recognition of acoustic events and speech on the one hand, and the imminent demand from application areas such as autonomous cars on the other hand, it can be expected that the acoustic dimension of self-awareness of ASs will assume a significantly more important role in the near future. Therefore, sharing a fast and synchronized communication infrastructure with high throughput, and adequate energy-efficient computing resources will play a crucial role for exploiting synergies with other modalities, not just with the visual modality, but also with all other sensors of the AS. Then, based on the techniques presented earlier, the wealth of acoustic information can be collectively extracted from complex acoustic scenes by next-generation ASs. ■

REFERENCES

[1] D. A. Abraham, *Underwater Acoustic Signal Processing*, 1st ed. Cham, Switzerland: Springer, 2019.

[2] L. Cremer, M. Heckl, and B. Petersson, *Structure-Borne Sound*, 3rd ed. Berlin, Germany: Springer-Verlag, 2005.

[3] W. Xiong *et al.*, "Toward human parity in conversational speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2410–2423, Dec. 2017.

[4] T. Virtanen, M. Plumbley, and D. Ellis, Eds., *Computational Analysis of Sound Scenes and Events*, 1st ed. Cham, Switzerland: Springer-Verlag, 2018.

[5] A. Mesaros *et al.*, "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 6, pp. 992–1006, Jun. 2019.

[6] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 776–780.

[7] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proc. IEEE*, vol. 103, no. 9, pp. 1635–1653, Sep. 2015.

[8] S. Haykin, *Cognitive Dynamic Systems: Perception-Action Cycle, Radar and Radio*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[9] A. S. Bregman, *Auditory Scene Analysis*, 1st ed. Cambridge, MA, USA: MIT Press, 1990.

[10] M. Cooke, G. J. Brown, M. Crawford, and P. Green, "Computational auditory scene analysis: Listening to several things at once," *Endeavour*, vol. 17, no. 4, pp. 186–190, Jan. 1993.

[11] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. Conf. Artif. Intell.* Austin, TX, USA: Association for the Advancement of Artificial Intelligence, 2000, pp. 832–839.

[12] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, 1st ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1976.

[13] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *Proc. Hands-Free Speech Commun. Microphone Arrays (HSCMA)*, Trento, Italy, May 2008, pp. 69–72.

[14] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Adv. Signal Process.*, vol. 2006, no. 1, pp. 1–19, Dec. 2006.

[15] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robot. Autom. Syst.*, vol. 96, pp. 184–210, Oct. 2017.

[16] S. Argentieri, P. Danès, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Comput. Speech Lang.*, vol. 34, no. 1, pp. 87–112, Nov. 2015.

[17] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Process.*, vol. 107, pp. 54–67, Feb. 2015.

[18] Q. Nguyen and J. Choi, "Selection of the closest sound source for robot auditory attention in multi-source scenarios," *J. Intell. Robot. Syst.*, vol. 83, no. 2, pp. 239–251, Aug. 2016.

[19] A. Brendel and W. Kellermann, "Distributed source localization in acoustic sensor networks using the coherent-to-diffuse power ratio," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 61–75, Apr. 2019.

[20] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. ASSP-5, no. 2, pp. 4–24, Apr. 1988.

[21] F. Gustafsson and F. Gunnarsson, "Positioning using time-difference of arrival measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, Apr. 2003, pp. 553–556.

[22] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[23] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 288–292, May 1997.

[24] D. Bechler, M. S. Schlosser, and K. Kroschel, "System for robust 3D speaker tracking using microphone array measurements," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sendai, Japan, Sep. 2004, pp. 2117–2122.

[25] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization

using a microphone array on a mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, Oct. 2003, pp. 1228–1233.

[26] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 8, pp. 1479–1489, Nov. 2008.

[27] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays* (Digital Signal Processing), M. Brandstein and D. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001.

[28] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, St. Louis, MO, USA, Oct. 2009, pp. 2033–2038.

[29] M. Basiri, F. Schill, P. U. Lima, and D. Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Vilamoura, Portugal, Oct. 2012, pp. 4737–4742.

[30] M. Basiri, F. Schill, P. Lima, and D. Floreano, "On-board relative bearing estimation for teams of drones using sound," *IEEE Robot. Autom. Lett.*, vol. 1, no. 2, pp. 820–827, Jul. 2016.

[31] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst.*, vol. 55, no. 3, pp. 216–228, Mar. 2007.

[32] H. L. V. Trees, *Optimum Array Processing* (Detecion, Estimation and Modulation Theory), no. 4. New York, NY, USA: Wiley, 2002.

[33] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2009, pp. 664–669.

[34] K. Nakadai, G. Ince, K. Nakamura, and H. Nakajima, "Robot audition for dynamic environments," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Hong Kong, Aug. 2012, pp. 125–130.

[35] K. Nakamura, K. Nakadai, and H. G. Okuno, "A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition," *Adv. Robot.*, vol. 27, no. 12, pp. 933–945, Aug. 2013.

[36] S. Argentieri and P. Danes, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, San Diego, CA, USA, Oct. 2007, pp. 2009–2014.

[37] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Vilamoura, Portugal, Oct. 2012, pp. 3288–3293.

[38] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and K. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Chicago, IL, USA, Sep. 2014, pp. 1902–1907.

[39] K. Hoshiba *et al.*, "Design of UAV-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, p. 2535, Nov. 2017.

[40] C. T. Ishi, J. Even, and N. Hagita, "Using multiple microphone arrays and reflections for 3D localization of sound sources," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Tokyo, Japan, Nov. 2013, pp. 3937–3942.

[41] J. Strutt, "On our perception of sound direction," *Phil. Mag.*, vol. 13, no. 1, pp. 214–232, 1907.

[42] K. Nakadai, H. G. Okuno, and H. Kitano, "Epipolar geometry based sound localization and extraction for humanoid audition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Maui, HI, USA, Oct. 2001, pp. 1395–1401.

[43] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition," in *Proc. Eur. Conf. Speech Commun. Technol. (INTERSPEECH-Eurospeech)*, Denver, CO, USA: International Speech and Communication Association, 2002, p. 4.

[44] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, Oct. 2003, pp. 1147–1152.

[45] T. May, N. Ma, and G. J. Brown, "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 2679–2683.

[46] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[47] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Tokyo, Japan, Nov. 2013, pp. 2927–2932.

[48] D. Jarrett, A. Habets, and P. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*. Aalborg, Denmark: European Association for Signal Processing, Aug. 2010, pp. 442–446.

[49] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. 14th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Antibes, France, Sep. 2014, pp. 258–262.

[50] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 178–192, Jan. 2017.

[51] V. Tourbabin and B. Rafaely, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1803–1814, Dec. 2014.

[52] V. Tourbabin and B. Rafaely, "Optimal design of microphone array for humanoid-robot audition," in *Proc. Israeli Conf. Robot. (ICR)*. Herzliya, Israel: Israeli Robotics Association, Mar. 2016.

[53] *Seventh Framework Programme 'Embodied Audition for RobotS' (EARS)*. Accessed: Sep. 25, 2018. [Online]. Available: https://robot-ears.eu/

[54] S. Markovich-Golan, W. Kellermann, and S. Gannot, "Spatial filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Hoboken, NJ, USA: Wiley, 2018.

[55] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Boston, MA, USA: Artech House, 1999.

[56] D. E. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone-array measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Munich, Germany, Apr. 1997, pp. 371–374.

[57] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[58] E. A. Lehmann and R. C. Williamson, "Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP J. Adv. Signal Process.*, vol. 2006, no. 1, pp. 17–21, Jun. 2006.

[59] I. Marković and I. Petrović, "Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering," *Robot. Auton. Syst.*, vol. 58, no. 11, pp. 1185–1196, Nov. 2010.

[60] L. Wang, R. Sanchez-Matilla, and A. Cavallaro, "Tracking a moving sound source from a multi-rotor drone," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Madrid, Spain, Oct. 2018, pp. 2511–2516.

[61] H. Asoh *et al.*, "An application of a particle filter to Bayesian multiple sound source tracking with audio and video information fusion," in *Proc. Int. Conf. Inform. Fusion*, 2004, pp. 805–812.

[62] H. Stiefelhagen *et al.*, "Enabling multimodal humanoid robot interaction for the Karlsruhe humanoid robot," *IEEE Trans. Robot.*, vol. 23, no. 5, pp. 840–851, Oct. 2007.

[63] S. Spors, R. Rabenstein, and N. Strobel, "Joint audio-video object tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Thessaloniki, Greece, Oct. 2001, pp. 393–396.

[64] R. P. S. Mahler, *Statistical Multisource-Multitarget Information Fusion*, 1st ed. Norwood, MA, USA: Artech House, 2007.

[65] R. P. S. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.

[66] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, Sep. 2006.

[67] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearing-only acoustic tracking of moving speakers for robot audition," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Singapore, Jul. 2015, pp. 1206–1210.

[68] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[69] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*, 1st ed. Hoboken, NJ, USA: Wiley, 2018.

[70] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech Processing in Modern Communication* (Springer Topics in Signal Processing), vol. 3, 1st ed. Berlin, Germany: Springer-Verlag, 2010.

[71] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Upper Saddle River, NJ, USA: Prentice-Hall, Feb. 1993.

[72] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.

[73] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, nos. 7–8, pp. 636–656, Jul. 2007.

[74] M. M. Goodwin and G. W. Elko, "Constant beamwidth beamforming," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Minneapolis, MN, USA, Apr. 1993, pp. 169–172.

[75] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 77–80.

[76] G. W. Elko and A.-T. N. Pong, "A simple adaptive first-order differential microphone," in *Proc. Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 1995, pp. 169–172.

[77] G. W. Elko and J. Meyer, "Second-order differential adaptive microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 73–76.

[78] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 2, pp. 617–631, Feb. 2007.

[79] M. Kajala and M. Hamalainen, "Filter-and-sum beamformer with adjustable filter characteristics,"

in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Salt Lake City, UT, USA, May 2001, pp. 2917–2920.

[80] E. Mabande and W. Kellermann, "Design of robust polynomial beamformers as a convex optimization problem," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement (IWAENC)*, Tel Aviv, Israel, Sep. 2010, pp. 1–4.

[81] H. Barfuss, C. Huemmer, G. Lamani, A. Schwarz, and W. Kellermann, "HRTF-based robust least-squares frequency-invariant beamforming," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[82] H. Barfuss, M. Bachmann, M. Buerger, M. Schneider, and W. Kellermann, "Design of robust two-dimensional polynomial beamformers as a convex optimization problem with application to robot audition," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 106–110.

[83] B. Rafaely, "Phase-mode versus delay-and-sum spherical microphone array processing," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 713–716, Oct. 2005.

[84] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[85] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[86] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[87] W. Herbordt, S. Nakamura, and W. Kellermann, "Joint optimization of LCMV beamforming and acoustic echo cancellation for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Philadelphia, PA, USA, Mar. 2005, pp. 77–80.

[88] Y. Zheng, K. Reindl, and W. Kellermann, "Analysis of dual-channel ICA-based blocking matrix for improved noise estimation," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, pp. 1–24, Mar. 2014.

[89] K. Reindl, S. Meier, H. Barfuss, and W. Kellermann, "Minimum mutual information-based linearly constrained broadband signal extraction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 6, pp. 1096–1108, Jun. 2014.

[90] L. Brooks and I. Reed, "Equivalence of the likelihood ratio processor, the maximum signal-to-noise ratio filter, and the Wiener filter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-8, no. 5, pp. 690–692, Sep. 1972.

[91] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Berlin, Germany: Springer, 2001, pp. 39–60.

[92] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Process.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[93] S. Makino, T. W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. Dordrecht, The Netherlands: Springer, 2007.

[94] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[95] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New Paltz, NY, USA: Wiley, 2001.

[96] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Boston, MA, USA: Springer, 2004, pp. 255–293.

[97] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 70–79, Jan. 2007.

[98] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2011, pp. 189–192.

[99] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[100] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

[101] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[102] M. Pederson, J. Larsen, U. Kjems, and L. Parra, "Convolutive blind source separation methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer, 2008, ch. 2, pp. 13–61.

[103] S. Makino, Ed., *Audio Source Separation* (Signals and Communication Technology), 1st ed. Cham, Switzerland: Springer, 2018.

[104] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 352–362, Sep. 2002.

[105] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *Proc. 3rd IEEE Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Aruba, The Netherlands, Dec. 2009, pp. 253–256.

[106] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno, "Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, New Orleans, LA, USA, Apr. 2004, pp. 1517–1523.

[107] H. Barfuss and W. Kellermann, "An adaptive microphone array topology for target signal extraction with humanoid robots," in *Proc. 14th Int. Workshop Acoust. Signal Enhancement (IWAENC)*, Antibes, France, Sep. 2014, pp. 16–20.

[108] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Sound source separation of moving speakers for robot audition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3685–3688.

[109] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Blind source separation with parameter-free adaptive step-size method for robot audition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1476–1485, Aug. 2010.

[110] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, vol. 5, Istanbul, Turkey, Jun. 2000, pp. 2985–2988.

[111] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[112] A. Aissa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani, and Y. Grenier, "Underdetermined blind separation of non-disjoint sources in the time-frequency domain," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 897–907, Mar. 2007.

[113] H. Saruwatari *et al.*, "Two-stage blind source separation based on ICA and binary masking for real-time robot audition system," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Edmonton, AB, Canada, Aug. 2005, pp. 2303–2308.

[114] M. Maazaoui, Y. Grenier, and K. Abed-Meraim, "Frequency domain blind source separation for robot audition using a parameterized sparsity criterion," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Barcelona, Spain: European Association for Signal Processing, Aug. 2011, pp. 1869–1873.

[115] M. Maazaoui, K. Abed-Meraim, and Y. Grenier, "Adaptive blind source separation with HRTFs beamforming preprocessing," in *Proc. IEEE 7th Sensor Array Multichannel Signal Process. Workshop (SAM)*, Innsbruck, Austria, Jun. 2012, pp. 269–272.

[116] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[117] W. Zhang, E. Habets, and P. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Tel Aviv, Israel, Sep. 2010, pp. 1–5.

[118] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[119] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[120] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[121] A. Jukic, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 1509–1520, Sep. 2015.

[122] A. H. Moore and P. A. Naylor, "Linear prediction based dereverberation for spherical microphone arrays," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement (IWAENC)*, Xi'an, China, Sep. 2016, pp. 1–5.

[123] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*, P. A. Naylor and N. Gaubitch, Eds. London, U.K.: Springer, 2010, ch. 2, pp. 311–385.

[124] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[125] J. Benesty, M. Sondhi, and Y. Huang, Eds., *Handbook of Speech Processing*. Berlin, Germany: Springer, 2008.

[126] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer, 2009.

[127] A. Ito, T. Kanayama, M. Suzuki, and S. Makino, "Internal noise suppression for speech recognition by small robots," in *Proc. Eur. Conf. Speech Commun. Technol. (INTERSPEECH-Eurospeech)*, Lisbon, Portugal: International Speech and Communication Association, 2005, pp. 2685–2688.

[128] Y. Nishimura, M. Ishizuka, K. Nakadai, M. Nakano, and H. Tsujino, "Speech recognition for a humanoid with motor noise utilizing missing feature theory," in *Proc. 6th IEEE-RAS Int. Conf. Hum. Robots*, Cancun, Mexico, Dec. 2006, pp. 26–33.

[129] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura, "Ego-noise suppression of a robot using template subtraction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, St. Louis, MO, USA, Oct. 2009, pp. 199–204.

[130] P. Marmaroli, X. Falourd, and H. Lissek, "A UAV

motor denoising technique to improve localization of surrounding noisy aircrafts: Proof of concept for anti-collision systems," in *Proc. Acoust.* Nantes, France: Société Française d'Acoustique, Apr. 2012, pp. 2938–2942.

[131] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[132] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[133] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sendai, Japan, vol. 3, Sep. 2004, pp. 2123–2128.

[134] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[135] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[136] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[137] K. Lebart, J.-M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[138] E. Habets, S. Gannot, and I. Cohen, "Dereverberation and residual echo suppression in noisy environments," in *Speech Audio Process. Adverse Environments*, E. Hänsler and G. Schmidt, Eds. Berlin, Germany: Springer, 2008, ch. 6, pp. 185–227.

[139] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 161–165.

[140] H. W. Löllmann, A. Brendel, and W. Kellermann, "Efficient ML-estimator for blind reverberation time estimation," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 1–5.

[141] H. W. Löllmann and P. Vary, "A blind speech enhancement algorithm for the suppression of late reverberation and noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3989–3992.

[142] M. M. Sondhi, "An adaptive echo canceller," *Bell Syst. Tech. J.*, vol. 46, no. 3, pp. 497–511, Mar. 1967.

[143] E. Hänsler and G. Schmidt, Eds., *Acoustic Echo and Noise Control*, 1st ed. Hoboken, NJ, USA: Wiley, 2004.

[144] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. Munich, Germany, vol. 1, Apr. 1997, pp. 219–222.

[145] M. Brandstein and D. Ward, Eds., *Microphone Arrays*. Berlin, Germany: Springer, 2001.

[146] J. Beh, T. Lee, I. Lee, H. Kim, S. Ahn, and H. Ko, "Combining acoustic echo cancellation and adaptive beamforming for achieving robust speech interface in mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nice, France, Sep. 2008, pp. 1693–1698.

[147] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral feature-based nonlinear residual echo suppression," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[148] R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, "ICA-based efficient blind dereverberation and echo cancellation method for barge-in-able robot audition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3677–3680.

[149] M. Zeller, L. A. Azpicueta-Ruiz, J. Arenas-Garcia, and W. Kellermann, "Adaptive volterra filters with evolutionary quadratic kernels using a combination scheme for memory control," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1449–1464, Apr. 2011.

[150] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 2065–2079, Sep. 2012.

[151] C. Hofmann and W. Kellermann, "Recent advances on LIP nonlinear filters and their applications: Efficient solutions and significance-aware filtering," in *Adaptive Learning Methods for Nonlinear System Modeling*, D. Comminiello and J. Principe, Eds. Oxford, U.K.: Butterworth-Heinemann, 2018, pp. 71–102.

[152] O. Hoshuyama and A. Sugiyama, "An acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toulouse, France, vol. 5, May 2006, pp. 269–272.

[153] M. Sondhi, D. Morgan, and J. Hall, "Stereophonic acoustic echo cancellation—An overview of the fundamental problem," *IEEE Signal Process. Lett.*, vol. 2, no. 8, pp. 148–151, Aug. 1995.

[154] H. Buchner, J. Benesty, and W. Kellermann, "Multichannel frequency-domain adaptive filtering with application to multichannel acoustic echo cancellation," in *Adaptive Signal Processing: Applications to Real-World Problems* (Signals and Communication Technology), J. Benesty and Y. Huang, Eds. Berlin, Germany: Springer, 2003, pp. 95–128.

[155] H. Buchner, J. Benesty, and W. Kellermann, "Generalized multichannel frequency-domain adaptive filtering: Efficient realization and application to hands-free speech communication," *Signal Process.*, vol. 85, no. 3, pp. 549–570, Mar. 2005.

[156] Y. Huang, J. Benesty, and J. Chen, "Identification of acoustic MIMO systems: Challenges and opportunities," *Signal Process.*, vol. 86, no. 6, pp. 1278–1295, Jun. 2006.

[157] M. D. Plumbley, C. Kroos, J. P. Bello, G. Richard, D. P. Ellis, and A. Mesaros, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*. Tampere, Finland: Tampere Univ. of Technology, Laboratory of Signal Processing, 2018.

[158] S. Essid *et al.*, "Multiview approaches to event detection and scene analysis," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. Plumbley, and D. Ellis, Eds., 1st ed. Cham, Switzerland: Springer-Verlag, 2018.

[159] S. Chu, S. Narayanan, and C.-J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[160] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6445–6449.

[161] A. Mesaros, T. Heittola, and D. Ellis, "Datasets and evaluation," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. Plumbley, and D. Ellis, Eds., 1st ed. Cham, Switzerland: Springer-Verlag, 2018.

[162] S. Watanabe, M. Delcroix, F. Metze, and J. Hershey, Eds., *New Era for Robust Speech Recognition: Exploiting Deep Learning*, 1st ed. Basel, Switzerland: Springer, 2017.

[163] M. Woelfel and J. McDonough, *Distant Speech Recognition*, 1st ed. Hoboken, NJ, USA: Wiley, 2009.

[164] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: Dataset and methods for UAV-embedded sound source localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Madrid, Spain, Oct. 2018, pp. 1–8.

[165] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 5610–5614.

[166] K. Nakadai, T. Matsui, H. G. Okuno, and H. Kitano, "Active audition system and humanoid exterior design," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Takamatsu, Japan, Oct. 2000, pp. 1453–1461.

[167] L. Wang and A. Cavallaro, "Acoustic sensing from a multi-rotor drone," *IEEE Sensors J.*, vol. 18, no. 11, pp. 4570–4582, Jun. 2018.

[168] G. Sinibaldi and L. Marino, "Experimental analysis on the noise of propellers for small UAV," *Appl. Acoust.*, vol. 74, no. 1, pp. 79–88, Jan. 2013.

[169] L. Marino, "Experimental analysis of UAV propeller noise," in *Proc. AIAA/CEAS Aeroacoustics Conf.* Reston, VI, USA: American Institute of Aeronautics and Astronautics, Jun. 2010, p. 3854.

[170] T. Ishiki and M. Kumon, "Design model of microphone arrays for multirotor helicopters," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Hamburg, Germany, Sep. 2015, pp. 6143–6148.

[171] K. Furukawa *et al.*, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Tokyo, Japan, Nov. 2013, pp. 3943–3948.

[172] R. Fernandes, E. Santos, A. Ramos, and J. Apolinario, "A first approach to signal enhancement for quadcopters using piezoelectric sensors," in *Proc. Int. Conf. Transformative Sci. Eng., Bus. Social Innov.*, Dallas, TX, USA, 2015, pp. 536–541.

[173] A. Deleforge, A. Schmidt, and W. Kellermann, "Audio-motor integration for robot audition," in *Multimodal Behavior Analysis in the Wild* (Computer Vision and Pattern Recognition), X. Alameda-Pineda, E. Ricci, and N. Sebe, Eds. London, U.K.: Academic, Nov. 2018, pp. 27–52.

[174] J. Even, H. Sawada, H. Saruwatari, K. Shikano, and T. Takatani, "Semi-blind suppression of internal noise for hands-free robot spoken dialog system," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, St. Louis, MO, USA, Oct. 2009, pp. 658–663.

[175] S. Yoon, S. Park, Y. Eom, and S. Yoo, "Advanced sound capturing method with adaptive noise reduction system for broadcasting multicopters," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Las Vegas, NV, USA, Jan. 2015, pp. 26–29.

[176] L. Wang and A. Cavallaro, "Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles," *IEEE Sensors J.*, vol. 17, no. 8, pp. 2447–2455, Apr. 2017.

[177] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 31–42, Mar. 2015.

[178] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 2182–2196, Dec. 2014.

[179] R. Sanchez-Matilla, L. Wang, and A. Cavallaro, "Multi-modal localization and enhancement of multiple sound sources from a micro aerial vehicle," in *Proc. ACM Multimedia Conf.*, Mountain View, CA, USA, 2017, pp. 1591–1599.

[180] Y. Hioka, M. Kingan, G. Schmid, and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement (IWAENC)*, Xi'an, China, Sep. 2016, pp. 1–5.

[181] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, "Underdetermined sound source separation using power spectrum density

estimated by combination of directivity gain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 6, pp. 1240–1250, Jun. 2013.

[182] K. Hoshiba *et al.*, "Design of UAV-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, p. 2535, Nov. 2017.

[183] L. Wang and A. Cavallaro, "Time-frequency processing for sound source localization from a micro aerial vehicle," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 496–500.

[184] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Proc. IEEE Workshop Mach. Learn. Signal Process. (IWAENC)*, Thessaloniki, Greece, Aug. 2007, pp. 431–436.

[185] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[186] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, vol. 13, 2001, pp. 556–562.

[187] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[188] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 971–982, May 2013.

[189] A. Deleforge and W. Kellermann, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 355–359.

[190] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2003, pp. 177–180.

[191] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[192] C. Févotte and J. Idier, "Algorithms for non-negative matrix factorization with the β-divergence," *Neural Comput.*, vol. 23, no. 9, pp. 2421–2456, 2011.

[193] T. Tezuka, T. Yoshida, and K. Nakadai, "Ego-motion noise suppression for robots based on semi-blind infinite non-negative matrix factorization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Florence, Itlay, May 2014, pp. 6293–6298.

[194] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

[195] T. Haubner, A. Schmidt, and W. Kellermann, "Multichannel nonnegative matrix factorization for ego-noise suppression," in *Proc. ITG Fachtagung Sprachkommunikation*. Oldenburg, Germany: VDE-Verlag, Oct. 2008, pp. 136–140.

[196] M. Takakusaki, D. Kitamura, N. Ono, T. Yamada, S. Makino, and H. Saruwatari, "Ego-noise reduction for a hose-shaped rescue robot using determined rank-1 multichannel nonnegative matrix factorization," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, Sep. 2016, pp. 1–4.

[197] N. Mae, D. Kitamura, M. Ishimura, T. Yamada, and S. Makino, "Ego noise reduction for hose-shaped rescue robot combining independent low-rank matrix analysis and noise cancellation," in *Proc Asia–Pacific Signal Inform. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Jeju, South Korea, Dec. 2016, pp. 1–6.

[198] Y. Bando *et al.*, "Low-latency and high-quality two-stage human-voice-enhancement system for a hose-shaped rescue robot," *J. Robot. Mechtron.*, vol. 27, no. 1, pp. 198–212, Feb. 2017.

[199] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Oct. 2006.

[200] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J.-I. Imura, "A hybrid framework for ego noise cancellation of a robot," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Anchorage, AK, USA, May 2010, pp. 3623–3628.

[201] A. Schmidt, A. Deleforge, and W. Kellermann, "Ego-noise reduction using a motor data-guided multichannel dictionary," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Daejon, South Korea, Oct. 2016, pp. 1281–1286.

[202] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai, "Assessment of general applicability of ego noise estimation," in *Proc. IEEE Int. Conf. Robot. Autom.*, San Francisco, CA, USA, May 2011, pp. 3517–3522.

[203] S. Gade, S. Herlufsen, H. Konstantin-Hansen, and N. Wismer, "Order tracking analysis: Technical review," Brüel & Kjær, Nærum, Denmark, Tech. Rep. 2, 1995.

[204] A. Schmidt and W. Kellermann, "Informed ego-noise suppression using motor data-driven dictionaries," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 116–120.

[205] C. E. Rasmussen and C. K. I. Williams, Eds., *Gaussian Processes for Machine Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 2006.

[206] M. Bernard, P. Pirim, A. de Cheveigne, and B. Gas, "Sensorimotor learning of sound localization from an auditory evoked behavior," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, St. Paul, MN, USA, May 2012, pp. 91–96.

[207] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots—Building audio-motor maps based on the HRTF," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Beijing, China, Oct. 2006, pp. 1170–1176.

[208] H. Poincaré, *The Foundations of Science: Science and Hypothesis, The Value of Science, Science and Method*, 1st ed. New York, NY, USA: The Science Press, 1929.

[209] M. Aytekin, C. Moss, and J. Simon, "A sensorimotor approach to sound localization," *Neural Comput.*, vol. 20, no. 3, pp. 603–635, Mar. 2008.

[210] E. Berglund and J. Sitte, "Sound source localisation through active audition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. Beijing, China, Aug. 2005, pp. 653–658.

[211] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *Int. J. Neural Syst.*, vol. 25, no. 1, pp. 1–18, Feb. 2015.

[212] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Beijing, China, Oct. 2006, pp. 380–385.

[213] L. Kneip and C. Baumann, "Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 3108–3119, Nov. 2008.

[214] A. Portello, P. Danès, and S. Argentieri, "Acoustic models and Kalman filtering strategies for active binaural sound localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Deajeon, South Korea, Sep. 2011, pp. 137–142.

[215] E. Martinson and A. Schultz, "Auditory evidence grids," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Beijing, China, Oct. 2006, pp. 1139–1144.

[216] E. Martinson and A. Schultz, "Discovery of sound sources by an autonomous mobile robot," *Auton. Robots*, vol. 27, no. 3, p. 221, Jun. 2009.

[217] B. P. DeJong, "Auditory occupancy grids with a mobile robot," *J. Autom. Mobile Robot. Intell. Syst.*, vol. 6, no. 3, pp. 3–12, 2012.

[218] N. Kallakuri, J. Even, Y. Morales, C. Ishi, and N. Hagita, "Probabilistic approach for building auditory maps with a mobile microphone array," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Karlsruhe, Germany, May 2013, pp. 2270–2275.

[219] J. Even, N. Kallakuri, Y. Morales, C. Ishi, and N. Hagita, "Creation of radiated sound intensity maps using multi-modal measurements onboard an autonomous mobile platform," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Tokyo, Japan, Nov. 2013, pp. 3433–3438.

[220] M. Kreković, I. Dokmanić, and M. Vetterli, "EchoSLAM: Simultaneous localization and mapping with acoustic echoes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 11–15.

[221] J. Hu, C.-Y. Chan, D.-K. Wang, and C.-C. Wang, "Simultaneous localization of mobile robot and multiple sound sources using microphone array," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Stockholm, Sweden, May 2009, pp. 29–34.

[222] S. Ogiso, T. Kawagishi, K. Mizutani, N. Wakatsuki, and K. Zempo, "Self-localization method for mobile robot using acoustic beacons," *J. Robomech*, vol. 2, no. 1, pp. 1–12, Sep. 2015.

[223] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018.

[224] K. Nakadai, H. G. Okuno, and H. Kitano, "Robot recognizes three simultaneous speech by active audition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, vol. 1, Taipei, China, Sep. 2003, pp. 398–405.

[225] K. Song, Q. Liu, and Q. Wang, "Olfaction and hearing based mobile robot navigation for odor/sound source search," *Sensors*, vol. 11, no. 2, pp. 2129–2154, Feb. 2011.

[226] N. B. Thomsen, "Signal-to-interference ratio driven positioning strategy for a social robot in spoken interactions," in *Speech Processing for Social Robots to Improve Interaction With Humans*, 1st ed. Aalborg, Denmark: Aalborg Univ. Press, 2017.

[227] N. Ma, T. May, H. Wierstorf, and G. Brown, "A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 2699–2703.

[228] H.-D. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Binaural active audition for humanoid robots to localise speech over entire azimuth range," *Appl. Bionics Biomech.*, vol. 6, nos. 3–4, pp. 355–367, 2009.

[229] G. Bustamante, A. Portello, and P. Danès, "A three-stage framework to active source localization from a binaural head," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 5620–5624.

[230] G. Bustamante, P. Danès, T. Forgue, and A. Podlubne, "Towards information-based feedback control for binaural active localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6325–6329.

[231] W. Odo, D. Kimoto, M. Kumon, and T. Furukawa, "Active sound source localization by pinnae with recursive Bayesian estimation," *J. Robot. Mechtron.*, vol. 29, no. 1, pp. 49–58, Feb. 2017.

[232] E. Vincent, A. Sini, and F. Charpillet, "Audio source localization by optimal control of a mobile robot," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 5630–5634.

[233] Q. V. Nguyen, F. Colas, E. Vincent, and F. Charpillet, "Long-term robot motion planning for active source localization with Monte Carlo tree search," in *Proc. IEEE Hands-Free Speech Commun. Microphone Arrays (HSCMA)*, San Francisco, CA, USA, Mar. 2017, pp. 61–65.

[234] C. Schymura, J. Grajales, and D. Kolossa, "Monte Carlo exploration for active binaural localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New Orleans, MS, USA, Mar. 2017, pp. 491–495.

[235] C. Schymura and D. Kolossa, "Potential-field-based active exploration for acoustic simultaneous localization and mapping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 76–80.

[236] A. Portello, P. Danès, and S. Argentieri, "Active binaural localization of intermittent moving sources in the presence of false measurements," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Vilamoura, Portugal, Oct. 2012, pp. 3294–3299.

[237] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, Jun. 1989.

[238] S. Thrun, "Learning occupancy grid maps with forward sensor models," *Autom. Robots*, vol. 15, no. 2, pp. 111–127, Sep. 2003.

[239] B. Barshan and H. F. Durrant-Whyte, "Inertial navigation systems for mobile robots," *IEEE Trans. Robot. Autom.*, vol. 11, no. 3, pp. 328–342, Jun. 1995.

[240] L. George and A. Mazel, "Humanoid robot indoor navigation based on 2D bar codes: Application to the NAO robot," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots (Humanoids)*, Atlanta, GA, USA, Oct. 2013, pp. 329–335.

[241] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–110, Jun. 2006.

[242] A. Nuchter, *3D Robotic Mapping: The Simultaneous Localization and Mapping Problem With Six Degrees of Freedom*, 1st ed. Berlin, Germany: Springer-Verlag, 2009.

[243] M. Montemerlo and S. Thrun, *FastSLAM: A Scalable Method for the Simultaneous Localization and Mapping Problem in Robotics* (Springer Tracts in Advanced Robotics), no. 27, 1st ed. Berlin, Germany: Springer-Verlag, 2007.

[244] V. Tourbabin, H. Barfuss, B. Rafaely, and W. Kellermann, "Enhanced robot audition by dynamic acoustic sensing in moving humanoids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 5625–5629.

[245] V. Tourbabin and B. Rafaely, "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 11, pp. 2046–2058, Nov. 2015.

[246] O. Roy and M. Vetterli, "The effective rank: A measure of effective dimensionality," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Poznań, Poland, Sep. 2007, pp. 606–610.

[247] K. Reindl *et al.*, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Comput. Speech Lang.*, vol. 27, no. 3, pp. 726–745, May 2013.

ABOUT THE AUTHORS

Alexander Schmidt (Member, IEEE) received the M.Sc. degree in electrical engineering from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, in 2015. He is currently a Research Assistant at the Chair of Multimedia Communications and Signal Processing, FAU, working toward a Ph.D. in the area of multichannel signal enhancement for robot audition.

He was with the EU FP7 Project, Embodied Audition for Robots (EARS). His special interest includes (sparse) dictionary learning for signal representation combined with physical-mechanical models.

Heinrich W. Löllmann (Senior Member, IEEE) received the Dipl.Ing. (univ.) degree in electrical engineering and the Dr.Ing. degree from RWTH Aachen University, Aachen, Germany, in 2001 and 2011, respectively.

From 2001 to 2012, he worked as a Scientific Co-Worker with the Institute of Communication Systems and Data Processing, RWTH Aachen University. He was a Research Manager of the EU FP7 Project, Embodied Audition for Robots (EARS). Since 2012, he has been a Senior Researcher with the Chair of Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany. He has authored one book chapter and more than 40 refereed articles. His research focuses on speech and audio signal processing, including filter-bank design, speech dereverberation and noise reduction, estimation of room acoustical parameters, and algorithms for robot audition.

Dr. Löllmann is currently a member of the Technical Committee on Audio and Acoustic Signal Processing of the IEEE Signal Processing Society.

Walter Kellermann (Fellow, IEEE) received the Dipl.Ing. (univ.) degree in electrical engineering from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany, in 1983, and the Dr.Ing. degree from Technical University Darmstadt, Darmstadt, Germany, in 1988.

From 1989 to 1990, he was a Postdoctoral Member of Technical Staff at AT&T Bell Laboratories, Murray Hill, NJ, USA. In 1990, he joined Philips Kommunikations Industrie, Nuremberg, Germany, to work on hands-free communication in cars. From 1993 to 1999, he was a Professor with Fachhochschule Regensburg, Regensburg, Germany, where he also became the Director of the Institute of Applied Research in 1997. In 1999, he co-founded DSP Solutions, Regensburg, Germany, a consulting firm in digital signal processing. He joined FAU as a Professor and the Head of the Audio Research Laboratory. In 2016, he was a Visiting Fellow with Australian National University, Canberra, ACT, Australia. Since 1999, he has been a Professor of communications with FAU. He has authored or coauthored 21 book chapters, and more than 300 refereed articles in journals and conference proceedings. He has authored or coauthored more than 70 patents. His current research interests include speech signal processing, array signal processing, and adaptive and learning algorithms and its applications to acoustic human–machine interfaces.

Dr. Kellermann was a member of the IEEE James L. Flanagan Award Committee from 2011 to 2014 and the SPS Board of Governors from 2013 to 2015. He is currently a member of the SPS Nominations and Appointments Committee. He received the Julius von Haast Fellowship by the Royal Society of New Zealand in 2012 and the Group Technical Achievement Award of the European Association for Signal Processing (EURASIP) in 2015. He was a co-recipient of ten best paper awards. He was the General Chair of seven mostly IEEE-sponsored workshops and conferences. He was the Chair of the IEEE SPS Technical Committee for Audio and Acoustic Signal Processing from 2008 to 2010. He served as an Associate Editor and a Guest Editor for various journals, including the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 2000 to 2004 and the *IEEE Signal Processing Magazine* in 2015. He also serves as an Associate Editor for the *EURASIP Journal on Applied Signal Processing*. He served as a Distinguished Lecturer for the IEEE Signal Processing Society (SPS) from 2007 to 2008. He was the Vice President of Technical Directions of the IEEE Signal Processing Society from 2016 to 2018.