Binaural Systems in Robotics

S. Argentieri, A. Portello, M. Bernard, P. Danès and B. Gas

1 Introduction

In the seventies, the word *robot* mainly termed a manipulator arm installed in a workshop. It was designed to perform repetitive and/or high-precision tasks, such as pick-and-place, assembly, welding or painting. Its environment was fully controlled with no human around, and its behavior was fully programmed in advance. Since then, robotics has dramatically evolved. Nowadays, a robot is endowed with advanced perception, decision and action capabilities. In essence, it is in interaction with its environment with humans and/or with other robots and is capable of autonomy and adaptation—full or shared. The spectrum of applications has kept broadening, and spans not only manufacturing and supply chain, but also exploration, health—such as in surgery, rehabilitation, assistance—and professional and personal-service robotics as, for instance, in mining, agriculture, transports, monitoring, rescue, guidance, cleaning, assistance, and games. Among the hot topics, one may cite robot deployment in uncontrolled and dynamic environments, *Human–robot interaction*, task-oriented behaviors and networked robotics devices in smart environments, or *ubiquitous robotics* [16]. For the last years, one could observe growing two-way connections between robotics and neurosciences, with the methods, models and experimental achievements of each discipline being a source of inspiration and strengthening for the other one [64].

Perception is a key requirement to robot autonomy, adaptation and self-awareness. Traditionally, a distinction is made between proprioception, that is, the ability for

S. Argentieri · M. Bernard · B. Gas (✉)
Inst. des Systèmes Intelligents et de Robotique,
Univ. Pierre et Marie Curie, National Centre for Scientific Research (CNRS), Paris, France
e-mail: bruno.gas@upmc.fr

A. Portello · P. Danès
Lab. d'Analyse et d'Architecture des Systèmes,
Univ. de Toulouse, Univ. Paul Sabatier, National Centre for Scientific Research (CNRS),
Toulouse, France

J. Blauert (ed.), *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-37762-4_9,

a robot to sense its own internal status, for instance, in terms of wheels angular positions/velocities, joint angles, odometry, gyroscope- and exteroception, which provides the robot with information on its environment. Among others, one can cite exteroceptive modalities, such as bumpers for emergency stop, ultrasound/infrared/laser scanning devices for range sensing, microphones, force sensors, tactile sensors and cameras—be it in the visible, infrared or multispectral range.

So far, the visual modality has undoubtedly received greatest interest. This is due to the richness of the information brought by images and to the high performance, low cost and embeddability of visual sensors. Vision has been used for decades in nearly all kinds of robotic tasks—from control loops and tracking routines to localization, map building, or scene modeling and interpretation. Numerous vision-based functions run nowadays in industry, for instance, non-destructive testing, scene/process monitoring, robot guidance, and other areas of application [5]. Besides, *computer vision* is a discipline by itself that does not take into account the specificities of robotics, such as real time constraints or changes in experimental conditions but, nevertheless, enriches the field.

Like vision, audition is a key sense in humans that plays a fundamental role in language and, consequently, in learning and communication. Quite surprisingly, *robot audition* was identified as a scientific topic of its own only since about the year 2000 [53], though related work existed before as part of bigger projects [12, 28, 30]. The reasons may be cultural as regards the importance of images in our society and also physiological—think of the predominance of vision in primates. More pragmatically, they are also certainly related to the difficult fulfillment of constraints like embeddability, high-performance acquisition, or real time processing. Consequently, while many theoretical results have long been developed in *acoustics* and *signal processing*, the literature on audition for robotics has remained scarce until recently. Fortunately, the timely topics of *cognitive robotics* and Human–robot interaction have promoted the field [25].

In its early days, robot audition benefited from developments in *computational auditory-scene analysis*, CASA [75]. Thereafter, Okuno et al. [63] identified the three main functions that any auditory robot should implement. These are

- Source localization, which may include a tracking system
- Separation of audio flows or source extraction
- Source recognition, which includes but is not limited to automatic speech recognition and can extend to scene interpretation

These functions encompass low-level issues as well as higher-level skills like emotion understanding or acoustic-scene meta-description.

In an attempt to provide a state of the art of binaural systems in robotics, this paper is structured as follows. The paradigms and constraints of robot audition are first summarized. Then the most prominent binaural approaches to canonical low- and high-level auditory functions are presented. After a review of robotics platforms, research projects and hard- and software related to the field, some novel *active* approaches developed by the authors are outlined, namely, combining binaural sensing and robot motion. A conclusion ends the chapter.

2 Paradigms and Constraints of Robot Audition

Two main paradigms for robot audition exist in the literature. On the one hand, *microphone arrays* have been used in a lot of applications. Various geometries have been selected, such as a line, a circle, a sphere, or the vertices of a cube. The redundancy in the sensed data is known to improve the acoustic-analysis performance and/or robustness [85]. Specific contributions have been concerned with

- The detection of the number of active sources, for example, through statistical identification [21]
- Source localization, for instance, through beamforming [59] or broadband beamspace MUSIC [2]
- Source extraction, for example, through geometrical source separation [84]
- Online assessment of uncontrolled dynamic environments
- Adaptation of speaker/speech recognition techniques to the robotics context [37]

On the other hand, *binaural approaches* have been developed.[1] These rely on a single pair of microphones that can be in free field, mounted inside an artificial pinna—not necessarily mimicking a human outer ear—and/or placed on a dummy head. From an engineering viewpoint, the possible use of cheap and efficient commercial stereo devices and drivers greatly eases the implementation. However, this simplification may imply an increased computational complexity.

Even though there is no fundamental need to restrict an acoustic sensor to only two microphones, for instance, when advanced data-acquisition and processing units are available, other arguments justify the binaural paradigm. First, robotics can be advocated as a privileged context to the investigation of some aspects of human perception. Indeed, as robots are endowed with locomotion and can incorporate multiple sensory modalities, they constitute a versatile experimental test bed to the validation/refutation of assumptions regarding the sensing structures as well as the processing and cognitive functions in humans. Conversely, these functions can be a source of inspiration for engineering approaches to perception. Last, there is an increasing demand for symbiotic interaction between humans and robots. This may imply the design of humanoid platforms, endowed with bioinspired perception and able to acceptably interact with humans in uncontrolled environments. Important constraints are, however raised by the robotics context, such as

Embeddability In the field of array processing, a large antenna involving a high number of microphones is often used. If such a sensor is to be embedded on a mobile robot, then a tradeoff must be handled between its size and, thus, aperture, and potential performances. Binaural sensors do not suffer from this geometrical constraint. Whatever the number of microphones is, the size and power consumption of the data acquisition and processing unit also constitute an important issue.

Real time Significantly distinct computation times are sometimes required by algorithms targeting the same goal. In robotics, low-level auditory functions such as

[1] Single-sensor approaches exist, such as [76, 83], but are rarely addressed in the literature.

binaural-cue calculation or source detection/localization must often run within a guaranteed short-time interval. Typically, up to 150 ms are demanded when their output is needed in reflex tasks such as exteroceptive control or people tracking. Specific processing units may have to be designed in order to guarantee real time behavior.

Environment Robotics environments are fundamentally dynamic, unpredictable and subject to noise, reverberation, or spurious sound sources. In order to ensure a guaranteed performance, adaptive or robust auditive functions must be designed.

Sources Most meaningful sound sources involved in robotics are broadband, with a spectrum spreading over the whole audible frequency bandwidth. This precludes the direct use of narrowband approaches developed elsewhere. Source non-stationarity is also an important issue. In addition, specific methods may be required depending on the source distance. Last, source motion can complexify the processing of the sensed signals, for example, when it breaks their assumed joint stationarity.

Robot Robot parts and robot motion generate so-called *self-noise*, or *ego-noise*, which may of course disturb the acoustic perception. Besides, in so-called *barge-in* situations, some sounds emitted intentionally by the robot may be interrupted, for example, by human utterance during a spoken dialog. Hence, they must be filtered-out online for not to damage the analysis of the scene.

Historically, most initial contributions to robot audition took place within the binaural paradigm. However, the results remained mixed when facing wideband non-stationary sources in noisy environments. Nevertheless, the last years have witnessed a renewal of interest for such approaches. A particular focus is put on *active* variations, which, thanks to the coupling of binaural sensing and robot motion, can overcome limitations of their passive counterparts. In computer vision, the coupling between visual perception and behavior has long been envisaged [1, 6]. The usefulness of active processes in hearing is discussed in [17, 53]. Importantly, the increased availability of cheap and accurate head-tracking solutions has given rise to related research amongst the hearing community with the potential of contributing to robot-listening strategies—see, for instance Ref. [11].

3 Binaural Auditory Functions in Robotics

As aforementioned, a robot should be able to understand several sources at a time by using its own ears in a daily environment, that is, with permanent background noises, intermittent acoustical events, reverberation, and so on. This challenging environment may be dynamic, due to sound sources moving or changing their acoustic properties. Additionally, the robot ego-noise is part of the problem. Whatever the conditions, most embedded auditory systems in robotics aim at generating an acoustic map of the robot's surroundings in real time. This map can then be used as an input to higher cognitive/decisional layers of the robot's software architecture in order to gaze or

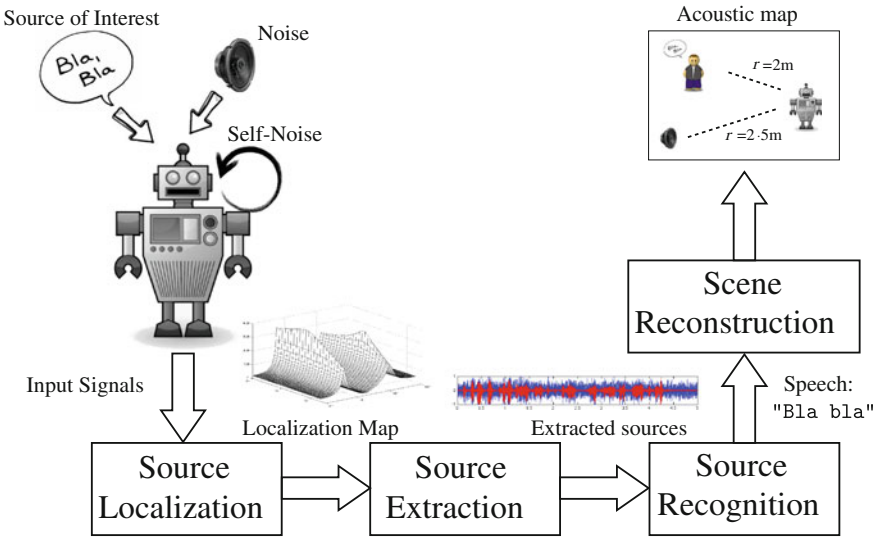

Fig. 1 Bottom-up working flow representation of classical robot audition systems

move towards the source of interest, answer to an interacting human partner, and other related issues. Such elementary behaviors can of course be enriched by considering low-level-reflex actions or multimodal approaches to artificial sound perception [80] or [81], this volume.

Acoustic maps of the environment are generally obtained along the bottom-up, signal-driven, workflow illustrated in Fig. 1. The successive computational steps are often split into two categories, namely *low-level* and *high-level* auditory functions, which are

Sound localization This stage is probably the most important low-level auditory function. Lots of efforts have been made to provide efficient binaural sound localization algorithms suited to robotics.

Extraction Once localized, each sound source can be separated so as to provide clean speech signals, noise signals, etc.

Ego-noise cancellation Its importance was acknowledged in the early days of active audition. Indeed, the noise of the motors enabling the robot motion may significantly degrade the robot's auditory perception.

Voice-activity detection, speaker recognition, speech recognition The need of these functions comes from the fact that most robotics systems are concerned with speech signals for interaction purposes.

High-level scene description Finally, all the above extracted information are gathered to accurately describe the acoustic scene.

Elements of this traditional bottom-up approach are reviewed in the following.

3.1 Sound Localization

In *binaural robot audition*, the azimuth and elevation of the multiple sources in the environment, and possibly their distance, are inferred from cues extracted from left and right signals. The literature reports the use of binaural cues, namely, *interaural time/phase difference*, ITD/IPD, *interaural level difference*, ILD, and monaural cues, that is, spectral information, and further characteristics, for instance, distance-related, [74]. In [94] various cue-extraction methods are reviewed. Whatever the localization policy, the problem is then to inverse the transformation that relates the spatial source locations to such cues. This requires a-priori knowledge about the propagation of an acoustic wave onto the robot's scatterers. This knowledge is generally captured in a simplified analytical model, or comes as experimental measurements of sound source properties, such as frequency contents, positions and induced auditory cues.

Considering source-azimuth estimation, the first model proposed in robotics was the *auditory epipolar geometry*, AEG [53]. It expresses the interaural phase difference, IPD, as a function of the source azimuth measured with the two microphones in the free field. As AEG does not take into account the shadowing effect of the head on sound propagation, some alternatives were proposed. Among them, one can cite the *revised auditory epipolar geometry*, RAEG [57], inspired by the Woodworth–Schlosberg formula [90]. These models are now commonly used in robotics, but are shown to be not so robust to changes in the environment. For instance, Nakadai et al. [56] showed that the simulated ITD obtained from RAEG is consistent with experimental measurements gathered in an anechoic room in the range of 500–800 Hz. Yet, if the comparison is made in a real robotics environment including reverberation and noises, then the basic models do not fit real-life data anymore.

Another analytical model, based on *scattering theory*, ST, was proposed in [55], considering the scattering induced by a perfectly-spherical head. In comparison with previous models, ST provides a more reliable closed-form model of the IPD as a function of the source azimuth. Of course, room acoustics is still not taken into account, so that the measured IPD remains heavily influenced by the environment. A similar approach was exploited in [27] and experimentally tested in [26] on a spherical plastic head.

In humans, source elevation is inferred from monaural cues taking the form of spectral notches. These are induced by the interference of the direct path of the wave to the concha and its reflections brought by the pinna. They are reproduced in robotics by artificial pinnae in charge of collecting the sound wave to the microphones. Importantly, both the placement of conchas on a robotics dummy head and the shape of pinnae can be optimized to maximize localization sensitivity to the source position. A solution to the first problem was proposed in [79]. The second one related to pinnae design has been more deeply investigated [41, 42, 76]. Yet, so far, both problems remain open. Numerical simulations may be required, which complicates the design. At large, the only emerging rule of thumb consists in designing an irregular or asymmetric pinna shape to provide elevation-dependent spectral cues.

Most of the previous approaches aim at getting closed-form equations of interaural or monaural cues. Such models could also involve the head-related transfer functions, HRTFs, of the robot that is, the transfer functions from the center of the head if it were absent and the two microphones. An HRTF captures all the effects exerted by the robot's body, head and pinnae on the incoming sound waves. In such a way it subsumes all the above head models. As closed-form HRTF expressions are very difficult to obtain for a generic mobile platform, a prior identification step is mandatory. It must be performed in an anechoic room, thus limiting its applicability to robotics. Nevertheless, in the case of well-identified environments, some HRTF-related applications to localization in robotics were developed. For instance, [50] estimated the position of a single talker in a known environment. This work was extended in [37] to simulate a moving talker. In [29] a learning approach is proposed for sound localization based on audio-motor maps that implicitly captures the HRTF. Self-organizing feature maps were also used in [60] to fuse binaural and visual information so as to estimate the 3-D position of a sound source. In the same vein, a multimodal approach to sound localization based on neural networks was proposed in [93], where vision is also combined with auditory cues to estimate the source azimuth. Further, a *gaussian-mixture model*, GMM, approach is proposed in [51], this volume, to evaluate the position of multiple sound sources. Learning approaches thus seem a promising generic tool to adapt an HRTF to various acoustic conditions.

3.2 Source Extraction

Once the sound sources have been localized in the robot's environment, the subsequent steps in low-level auditory analysis generally consist in extracting the sound sources of interest. Depending on the authors, source localization and extraction can be inverted or even gathered into a single function. Though binaural extraction of sources was addressed in the early days of robot audition, the number of approaches has remained quite small. One of the most famous solution is the *active direction-pass filter*, ADPF [57]. It works by collecting frequency sub-bands that are presumably linked to the same sound source in order to generate the extracted signal. This fusion of sub-bands is performed only if their associated binaural cues, that is, IPDs/ILDs, are spatially coherent according to a given head model, for instance, AEG or RAEG—see Sect. 3.1—taking into account the better spatial-discrimination ability for sources in front of the robot. This system has been proven to be effective to extract and recognize three simultaneous speakers [58].

One of the main advantages of ADPF is that it requires no a-priori knowledge of the signal contents. The same applies to other recent binaural separation techniques based on localization cues. Well-known approaches based on *independent-component analysis* [70] that can separate two sources at most from the output of a binaural sensor, are then generally overcome. For instance, Weiss et al. [89] proposed to combine binaural cues with higher-level information related to speaker identity

in order to separate more than two sources in the presence of reverberation. Another solution, already explored with ADPF, that outperforms beamforming approaches in reverberant scenes, is to derive time-frequency masks on the basis of binaural cues—see for instance [77]. In [23], multiple-speaker joint detection and localization was recasted as a probability-based spatial clustering of audio–visual observations into a common 3-D representation. Cooperative estimates of both the auditory activity and the 3-D position of each object were then deduced from a substantiated variation of the expectation-maximization algorithm. Experiments were conducted with the POPEYE platform—see Sect. 4.4—on single or multiple speakers in the presence of other audio sources.

Whatever the approach, it is still very difficult to extract multiple sources covering the same frequency range. In such cases, some extracted signals may mistakenly come from unrelated sources and may thus present missing or uncertain sections. In the downstream pattern-matching algorithms, for example, for speaker/speech recognition purposes, such problems can be handled within the *missing-feature theory*, MFT [18, 19, 45], which consists in tagging the missing sections by a null confidence weight, as will be shown further down.

3.3 Ego-Noise Cancellation

Two very restrictive solutions to *ego-noise cancellation* are generally proposed for the binaural case. On the one hand, loud enough sources can mask ego-noises during a movement, thus improving the signal-to-noise ratio, SNR, of the perceived signals. On the other hand, stop-and-listen approaches are sometimes used, so as to process sounds while the robot is at rest. Both approaches are unsatisfactory, and recent developments have tried to overcome these limitations.

Canceling the noise originating from the robot can be considered by source separation techniques. But existing studies mainly rely on microphone arrays, which makes them inappropriate for the binaural context. Additionally, as motors are generally placed in the vicinity of the microphones, a diffuse sound field should be used to model the noise, precluding the direct use of standard state-of-the-art approaches to source separation.

One of the first solutions to the specific ego-noise-cancellation problem in robotics was proposed in [54] on the SIG humanoid robot—see Sect. 4.2. It relies on two pairs of microphones, one of them being dedicated to the perception of the interior of the robot. With the help of these inner transducers, the method was able to classify spectral bands as being related either to ego-noise or to the source of interest, but could not suppress the noise from the perceived signals. Other approaches consist in predicting the noise to be emitted on the basis of the generated motion. For instance, Ito et al. [35] did this on a time-frame basis with a neural network. The most promising solution relies on noise patterns. In this vein, joint noise templates related to specific movements were stored offline into a large ego-noise database [32, 62], then were

identified online according to the robot motion to be performed, and subsequently subtracted from the perceived signals [31].

3.4 Voice-Activity Detection

When considering Human–robot interaction applications, the perceived signals are mainly composed of speech information, non-informative signals, and various types of noise. Efficient speech extraction is necessary to decrease the error rate in high-level auditory analysis. It can be performed by detecting speech segments in the sensed signals, prior to localizing the corresponding sources of interest and spatially filtering them out of the noise. *Voice-activity detection*, VAD, algorithms have often been used, which generally classify signal snippets as either *noise-and-speech* or *only-noise*. Again, multiple solutions to VAD have been proposed in the literature, such as energy feature, zero-cross rate [71] or higher-order statistics [61]. However, very few of them are specifically dedicated to binaural audition and/or suited to robotics. Energy feature can hardly cope with individual differences and dynamic volume changes. Zero-cross rate is better in this respect, but is more sensitive to noise. Statistics show good behavior but their performance decrease in an acoustic environment which shows significant differences with the one used to learn the statistics. As a solution suited to robotics, Kim et al. [37] proposed an enhanced speech detection method that can be used to separate and recognize speech in a noisy home environment. Nevertheless, the detected utterances should take place in front of the robot. Another approach is outlined in [20], inspired by wireless sensor network applications in the context of hearing aids. Therein, a basic energy-based VAD was combined with a cross-correlation based VAD to detect speech in the two signals. But again, the speaker should be uttering in front of the system. Besides, a Bayesian network based integration of audio–visual data for VAD was proposed in [91] as the first layer of an automatic speech recognition system.

3.5 Speaker and Speech Recognition

In robotics, speaker and speech recognition are probably the key high-level auditory functions required to perform natural Human–robot interaction. Traditional *automatic speech recognition*, ASR, algorithms are known to be very sensitive to the speech signal-acquisition conditions, such as quality of the microphones, distance to the speaker, environmental noise, and so on. Therefore, considering this problem at large, the robotics context requires a trade-off between large-vocabulary and multiple-speaker applications in the well-known framework of the *cocktail-party problem*.

Environmental noise is probably the most prominent challenge to be faced by ASR systems in robotics. This is probably the reason why most recent studies have focused

on noise removal from the speech signals, that is, *speech enhancement*. The aim here is to retrieve ideal acquisition conditions, generally by applying a set of spatial filters which enable the attenuation of noisy sources or echoes in the perceived signal in order to use traditional ASR systems. In this topic, the aforementioned missing features approaches are of particular interest. They are able to cope with additive, possibly non-stationary noise sources [72, 78] by discarding specific regions in the speech spectrogram with low SNRs. Likewise, one can mention *missing-feature compensation* techniques, which are able to accurately estimate the omitted regions of these incomplete spectrograms. These approaches were used in [82] in order to separate two speakers uttering simultaneously from the front of a humanoid robot endowed with a binaural sensor, on the basis of an *independent-component analysis* based source-separation technique. This allows for improving the speech-recognition rates by 15 % with respect to a state-of-the-art-based *hidden Markov model*, HMM, recognition system. Recent developments in the missing-features framework were concerned with adaptive recognition systems, that is, with MFT–ASR approaches which allow to change the weight of those spectrogram sections that are considered damaged [73]. In this domain, Ince et al. [33] exploits an MFT–HMM approach to cancel ego-noise by applying a time-frequency mask. This makes it possible to decrease the contribution of unreliable parts of distorted speech in signals extracted from a microphone array. Such an approach can also be applied in binaural systems.

Compared to speech recognition, speaker recognition has rarely been addressed in robotics. Like speech recognition, it is usually analyzed for the monaural case under similar recording conditions. Initially, the subject was already addressed by means of microphones arrays [44]. For applications to robotics, one can refer to [36], and to the recent preliminary study [95]. The latter paper shows that in a reverberant and noisy environment, the success rate of binaural speaker recognition is much higher than with monaural approaches.

3.6 High-Level Scene Description

Several studies have been conducted in order to endow a user with auditory awareness about a complex auditory scene. A basic, though incomplete, solution consists in applying 3-D acoustic spatialization techniques to an extracted and labeled source in such a way that the users sense better where the corresponding message comes from. Intuitive tools exist in our daily lives to obtain awareness on a visual scene, such as overviewing, zooming, scrutinizing, (re)playing at various places, browsing, indexing, etc. In the same vein, a 3-D auditory scene visualizer according to the mantra *"overview first, zoom and filter—and then detail on demand"* was developed on top of face-tracking and auditory functions [40] and integrated into the $\mathcal{HARK}$ robot-audition toolbox—Sect. 4.2. This system has been improved to get better immersive feeling. Though a microphone array is assumed, the underlying concepts extend to binaural techniques.

4 Platforms and Research Projects

This section reviews some notable auditory robots and/or research projects. Associated hard- and software dedicated to robot audition are also mentioned as far as available.

4.1 Cog

The upper torso humanoid Cog, from MIT, is probably the first platform endowed with audition[2] It was targeted towards the scientific goal of understanding human cognition and the engineering goal of building a general purpose flexible and dextrous robot. Interestingly, the authors pointed out in their manifesto [12] that the first goal implied the study of four essential aspects of human intelligence that, by themselves, involve manufacturing a human-like platform. These topics, discarded in conventional approaches to *artificial intelligence*, were

- *Development* Considering the framework by which humans successfully acquire increasingly more complex skills and competencies
- *Social interaction* Enabling humans to exploit other humans for assistance, teaching, and knowledge
- *Physical embodiment and interaction* Humans use the world itself as a tool for organizing and manipulating knowledge
- *Integration* Humans maximize the efficacy and accuracy of complementary sensory and motor systems

Cog was endowed with an auditory system, comprising two omni-directional microphones mounted on its head, and crude pinnae around them to facilitate localization. The sound acquisition and processing units were standard commercial solutions. Companion development platforms were built, similar in mechanical design to Cog's head with identical computational systems. One of them, oriented towards the investigation of the relationships between vision and audition, complemented the binaural auditory system with one single color camera mounted at the midline of the head. Visual information was used to train a neural network for auditory localization [34].

4.2 SIG/SIG2 and HARK

The SIG Project [54] was initiated by Kitano Symbiotic Systems, ERATO and JST Corp., Tokyo. The pioneering program has then been pursued further in collaboration of Kyoto University and the Honda Research Institute.[3] In an effort to understand

[2] http://www.ai.mit.edu/projects/humanoid-robotics-group/cog/

[3] http://winnie.kuis.kyoto-u.ac.jp/SIG/

high-level perceptual functions and their multi-modal integration towards intelligent behavior, an unprecedented focus was put on *computational auditory-scene analysis*, CASA, in robotics. The authors promoted the coupling of audition with behaviors also known as *active audition* for CASA, so as to dynamically focus on specific sources for gathering further multimodal information through active motor control and related means. This approach paved the road to many developments, the first one being ego-noise cancellation.

SIG is an upper-torso humanoid. It has a plastic cover designed to acoustically separate its interior from the external world. It is fitted with a pair of CCD cameras for stereo vision and two pairs of microphones—one in the left and right ears for sound-source localization and the other one inside the cover, mainly for canceling self-motor noise in motion. A second prototype, named SIG2, was designed to solve some problems in SIG, such as the loud self-noise originating from motors, an annoying sound reflection by the body, sound resonance and leakage inside the cover, and the lack of pinnae. This implied changes in the material and actuators, as well as the design of human-shaped ears. Many striking achievements were obtained on SIG/SIG2, such as multiple sound-source localization and tracking from binaural signals while in motion, multiple-speaker tracking by audio–visual integration, human–robot interaction through recognition of simultaneous speech sources.

Subsequently and importantly, array-processing techniques for source localization and source separation were designed and implemented on SIG/SIG2. This gave rise to the *open-source robot-audition toolbox*, $\mathcal{HARK}$,[4] that gathers a comprehensive set of functions enabling computational auditory-scene analysis with any robot, any microphone configuration and various hardware. Within the recent revival of active binaural audition, $\mathcal{HARK}$ has been complemented with a package for binaural processing.

4.3 iCub

An open-source platform, comprizing hardware and software, well suited to robot audition is the iCub humanoid robot. iCub has been developed since 2004 within the RobotCub project,[5] and disseminated into more than twenty laboratories. Sized as a 3.5 year-old child, it is endowed with many degrees of freedom and human-like sensory capabilities, including binaural audition. It has also been designed towards research in embodied cognition, including study of cognition from a developmental perspective in order to understand natural and artificial cognitive systems.

Two electret microphones are placed on the surface of its 5-DOF head and plastic reflectors simulate pinnae. The shape of these ears has been kept simple, so as to ease their modeling and production while preserving the most prominent acoustic characteristics of the human ear. To better manage the frequencies of the resonances and notches to be used for vertical localization, a spiral geometry was selected [29].

[4] HRI-JP audition for robots with Kyoto University, http://winnie.kuis.kyoto-u.ac.jp/HARK/. In Ariel's Song, The Tempest, from Shakespeare, *hark* is an ancient english word for *listen*.

[5] http://www.icub.org/projects.php

Small asymmetries between right and left pinnae, namely, a rotation of 18°, enable to tell the notches due to the source contents from these due to its spatial location just by comparing the binaural spectral patterns. A supervised learning phase matches the (ITD, ILD, notches)-tuples extracted from binaural snippets with sound-source positions inside audio-motor maps. These maps are then used online to drive the robot by sound. The maps are seamlessly updated using vision to compensate for changes in the HRTFs as imposed by ears and/or environment. Experiments show that the robot can keep the source within sight by sound-based gaze control, with worst-case errors of pan and tilt below 6°.

4.4 POP and HUMAVIPS, and Their Associated Platform/Datasets

A recent milestone in robot-audition research is undoubtedly the *perception-on-purpose*, POP, project.[6] This European scientific collaboration in 2006–2008 was oriented towards the understanding and modeling of the interactions between an agent and its physical environment from the biological and computational points of view, concentrating on the perceptual modalities of vision and audition. Aside from a fundamental investigation of cognitive mechanisms of attention on the basis of measures of brain physiology brought about by *functional magnetic-resonance imaging*, fMRI, and *electro/magneto-encephalography*, EEG/MEG, a sound mathematical framework was targeted, enabling a robot to feature purposeful visio–auditive perception by stabilizing bottom-up perception through top-down cognition. A specific focus was put on crossmodal integration of vision and audition along space and time, the design and development of methods and algorithms to coordinate motor activities and sensor observations, the design and thorough evaluation of testable computational models and on the provision of an experimental testbed.

The following achievements can be mentioned. A two-microphone binocular robotic platform, POPEYE, was built [15]. This highly repeatable system can undergo high velocities and accelerations along 4-DOFs, namely, pan, tilt, and the two camera-independent pan angles. POPEYE allows the use of a dummy head for binaural audition but is not fully bio-mimetic since the binaural axis is higher than the stereo-vision axis. Novel algorithms for real-time robust localisation of sound sources in multisource environments were proposed, based on a fusion of interaural time difference and pitch cues, using source-fragment methods inspired by glimpsing models of speech perception. Active-listening behaviors were defined that can use planned movement to aid auditory perception, namely, head rotation in order to maintain a tracked source in the auditory fovea, judgement of distance by triangulation, and others. As mentioned in Sect. 3.2, an original approach to detection and localization of multiple speakers from audio–visual observations was also developed and experimented on POPEYE [23].

[6] http://perception.inrialpes.fr/POP/

The dataset CAVA, which stands for *computational audio–visual analysis of binaural–binocular recordings* [3] was made freely available for non-profit applications. It was recorded from sensors mounted on a person's head in a large variety of audio–visual scenarios, such as multiple speakers participating in an informal meeting, static/dynamic speakers, presence of acoustic noise, and occluded or turning speakers.

The subsequent HUMAVIPS project, *humanoids with auditory and visual abilities in populated spaces*, runs from 2010 to 2013 and concerns multimodal perception within principled models of Human–robot interaction and humanoid behavior.[7] In this project coordinated audio–visual, motor and communication abilities are targeted, enabling a robot to explore a populated space, localize people therein, assess their status and intentions, and then decide to interact with one or two of them by synthesizing an appropriate behavior and engaging a dialog. Such cocktail-party and other social skills are being implemented on an open-source-software platform and experienced on a fully-programmable humanoid robot.

Open-source datasets have also been disseminated in this framework. Two of them have been recorded with the aforementioned POPEYE system. To investigate audio–motor contingencies from a computational point of view and to experiment with new auditory models and techniques for computational auditory-scene analysis, the CAMIL dataset, *computational audio–motor integration through learning*, provides recordings of various motionless sources, like random spectrum sounds, white noise, speech, music, from a still or moving dummy head equipped with a binaural pair of microphones.[8] Over 100 h of recordings have been elaborated, each of them being annotated with the ground-truth pan-and-tilt motor angles undergone by the robot.

Likewise, to benchmark Human–robot interaction algorithms, the RAVEL corpora, *robots with auditory and visual abilities*, provides synchronized binaural auditory and binocular visual recordings by means of a robocentric stable acquisition device in realistic natural indoor environments.[9] It gathers high-quality audio–visual sequences from two microphone pairs and one camera pair in various kinds of scenarios concerning human-solo- action recognition, identification of gestures addressed to the robot, and human–human as well as Human–robot interaction. The scenes may be affected by several kinds of audio and visual interferences and artifacts. To ease the statement of ground truth, the absolute position and utterances of actors in the scene are also recorded by external cross-calibrated and synchronized devices, namely, a commercial 3-D tracking system and four distributed headset microphones.

[7] http://humavips.inrialpes.fr/

[8] http://perception.inrialpes.fr/~Deleforge/CAMIL_Dataset/index.html

[9] http://ravel.humavips.eu

4.5 BINAAHR

BINAAHR, *binaural active audition for humanoid robots*, was established as a french–japanese collaboration focused on two accepted concepts of active (binaural) robot audition.[10] On the one hand, a low-level auditory function is said to be active if it combines, and is improved by, the perception and the motor commands of the sensor. On the other hand, a high-level interaction is active if it is bidirectional and involves the robot and multiple parties. The project has contributed to the design of artificial binaural systems, active binaural localization (Sect. 5.2), binaural separation of more than two sources, ego-noise cancellation, binaural speaker recognition, audio–visual speech recognition and other significant issues.

A separate line of research in BINAAHR champions robotics, because of its locomotion and multimodal-sensing capabilities, as a privileged context to investigate *psychology-of-perception* theories of *active human perception*, that is, theories that hypothesize an interweave of human perception and action. In this context, some innovative developments have been tightly connected with the sensorimotor contingency theory [65, 66]. These developments enable the analysis of the sensorimotor flow of a naive agent, be it endowed with hearing only or with both vision and audition, in order to characterize the dimension of the rigid group of the space underlying its input–output relationship, that is, the dimension of its physical space—see Sect. 5.1. Experimental issues concern unitary testing of low-level functions on binaural prototypes as well as the integration of functions on a HRP-2 humanoid robot.

4.6 Further Hardware

Further specific hardware suited to the needs of robot audition has been developed with the aim of pushing forward the integration of auditory functions on embeddable autonomous sensors. Corresponding achievements have mainly been oriented towards array processing, partly because off-the-shelf multichannel data-acquisition devices are often unsatisfactory because of limited embeddability, and high cost due to a too high genericity and other reasons. Although suitable commercial stereo devices can be used for binaural acquisition, array processing-oriented hardware may still constitute an inspiration, for example, for computational issues.

The active direction-pass filter (Sect. 3.2, [57]) was integrated in a dedicated reconfigurable processor and could separate a mixture of sounds in real time with good accuracy [43]. A more generic low-cost low-consumption sound card was developed in order to fit the requirements of the ManyEars project, which features an array based source localization, tracking and separation system [52]. This board, named 8-Sounds-USB, performs an eight-channel synchronous audio capture and embeds computational power. Likewise, the *embedded audition for robotics*, EAR, sensor,

[10] http://projects.laas.fr/BINAAHR

based on a fully programmable eight-channel data-acquisition board and a powerful FPGA, has recently been revisited within a *system-on-a-programmable-chip* approach [47], namely, a C/C++ compatible soft-processor has been implemented on the FPGA, together with dedicated hardwired modules such as co-processing units, memory controllers, communication and data acquisition modules. A release suited to binaural audition is under development. Several complex and intensive operations will be hardwired. The device will process data sensed by MEMS microphones and be compatible with standard audio interfaces.

4.7 Conclusion

While robot audition is a fairly recent field of research, various solutions have been proposed to cope with the constraints of robotics. In the above, most low- or high-level functions have been reviewed independently, yet many contributions have been considering them jointly—see for instance Ref. [23] in Sect. 3.2 for joint source localization and separation. The same holds for higher-level auditory functions. For instance, it was shown in [92] that the design of a recognition system should take into account a trade-off between the recognition rates and the sensitivity to speaker locations. Last, the order of the successive computation steps involved in a complex auditory task can differ from one author to another. To conclude, no universal strategy is available at this time and the optimal architecture for a CASA system in robotics is still a matter of debate. Some related activities were conducted in Europe, for example, in the context of the research cluster CoTeSys, *cognition for technical systems*. For instance, a multi-modal robotics head, Mask-Bot, was built [67] to feature face animation, speech communication and basic sound localization. A generic comprehensive model of binaural listening that could also be of high interest for robotics is proposed in [10], this volume.

5 Active Approaches to Perception: Applications to Binaural Audition in Robotics

Theoretical approaches to perception are many, and some of them show significant divergences. For instance, Marr's celebrated computational approach to visual perception [48], which prevailed in the development of *artificial intelligence*, proposes a viewpoint of passive perception where the representation is predominant and behavioral aspects are overlooked. Nowadays, it is still a debate whether this conception should be traded off for another theory, namely, one that hypothesizes perception and action to be interweaved. The latter viewpoint is usually related to Gibson's theory, which puts forward the active and exploratory nature of perception [24]. Such considerations also apply to robotics, which has for long considered perception as a bottom-up process in the sense that action are results of sensory analysis.

This historical viewpoint on perception is currently being questioned experimentally, all the more since the exploratory abilities of robotics platforms can be exploited to improve analysis and understanding of the environment. In this context, the current section gathers two original contributions of the authors with regard to binaural auditive perception. Both entail an active behavior of the robotic platform but along distinct approaches.

In a first subsection, an active strategy for *auditory-space learning* is proposed together with its application to sound source localization. It relies on a general theoretical approach to perception, grounded in sensorimotor theory. In a second subsection, a stochastic-filtering strategy for *active binaural sound localization* is introduced, where it is shown how the motor commands of a moving binaural sensor can be fused with the auditive perception to actively localize a still or moving intermittent sound source in the presence of false measurements.

5.1 Active Hearing for Auditory-Space Learning and Sound-Source Localization

Action in robotics is usually viewed as a high-level process and is mostly used to address problems that cannot be solved by passive strategies alone, such as front-back disambiguation or distance perception. The method proposed here investigates an alternative paradigm, where the action is envisaged at the same level as perception. In this framework, action and perception interact so as to build an internal representation of the auditory space. As a first step, an active hearing process is used during the learning of an auditory-motor map. Next, this map is used for a-priori passive sound localization. In what follows the approach is introduced for azimuthal localization by considering a mobile listener endowed with a binaural auditory system and perceiving a single stationary sound source of random azimuth in a $\pm 90°$ range.

A Sensorimotor Definition of Source Localization

The present method is grounded in the sensorimotor theory [65, 68], claiming that the brain is initially a naive agent that interacts with the world via an unknown set of afferent and efferent connexions, with no a-priori knowledge about its own motor capacities or the space it is immersed in. The agent therefore extracts this knowledge by analyzing the consequences of its own movements on its sensory perceptions, building a sensorimotor representation of its embodying space. Generally speaking, consider that all the environments, motor states and sensations that an agent can experiment are depicted as the respective manifolds, $\mathcal{E}$, $\mathcal{M}$ and $\mathcal{S}$ [66].

A sensory state, $s \in \mathcal{S}$, is given as a function of the current motor and environment states, $m \in \mathcal{M}$ and $e \in \mathcal{E}$, through a sensorimotor law, Φ, so that $s = \Phi(m, e)$. Here e models the scene acoustics and spatial and spectral properties of the sound source, m

models the agent's body configuration, whereas Φ represents the body-environment interactions and neural processing that gives rise to the sensation, s. Moreover the sensory space, $\mathcal{S}$, lies on a low-dimensional manifold whose topology is similar to the embodying space and, consequently, the learning of spatial perception becomes the learning of such a manifold. Such a process has been applied to auditory-space learning using non-linear dimensionality-reduction techniques [4, 22]. Nevertheless the knowledge of this auditory space is not sufficient for sound localization—an association of a percept in this space and a spatial location have still to be done.

Classical localization methods express a source location in terms of angle or range in an Euclidean physical space. As the sensorimotor approach directly links perception and action in an internal representation of space, a spatial position is here directly expressed as a motor state and as such does not implies any notion of space [68]. Given a motor space, $\mathcal{M}$, and an environment state, $e \in \mathcal{E}$, the source localization problem can thus be defined as the estimation of the motor state, $\tilde{m}$, as follows:

$$\tilde{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} |\Phi(m, e) - \Phi(m_0, e_0)| \,, \tag{1}$$

where $|.|$ denotes a distance metric and $\Phi(m_0, e_0)$ represents a reference sensory state that has to be approximated. In the case of sound-source localization, $\Phi(m_0, e_0)$ corresponds to a source localized in front of the listener with the head in the rest position, which is the most obvious case of azimuthal localization.

Evoked Behavior for Active Hearing

An evoked or reflex behavior is a simple *hard-wired* behavior allowing a naive agent to react to a stimulus. Considering active hearing, a simple behavior enabling head-to-source orientation can be implemented from ILD cues in a simple way as follows. Once a sound is perceived, the agent orients its head toward the loudest side while the ILD is non-zero. Once the behavior is completed, that is, when the ILD reaches 0, the head of the agent has arrived at an orientation facing the sound source. Because the source azimuth is in a range of $\pm 90°$ only, the agent is not exposed to front-back ambiguity—although front-back disambiguation has also been addressed in the literature, for instance, Ref. [9]. This active hearing process allows an *a-posteriori* localization, $\tilde{m}$ being given after motion as the difference between the initial and final motor states. Moreover, the agent's final configuration with the source in front of the head corresponds to the reference sensory state, $\Phi(m_0, e_0)$, as introduced in (1). This reference state, initially unknown, is approximated through successive executions of the behavior. The orientation behavior has been successfully demonstrated on a robotic platform and been extended to phonotaxis, that is, allowing for a reactive approach of the robot towards the source [8].

Autonomous Online Learning of Sound Localization

The evoked behavior that links the initial sensory state in $\mathcal{S}$ to the final motor state in $\mathcal{M}$, provides the sensorimotor association required for an a-priori passive localization. Figure 2 shows an auditory space representation after the learning of high dimensional ILD cues from 1000 auditory stimuli. Each point, corresponding to a different sensory state, is associated with its localization estimation, $\tilde{m}$, computed after the orientation behavior.

After learning of such an association, it becomes possible to localize new percepts based on neighborhood relationships. Suppose a new stimulus corresponding to a sensory state, $s \in \mathcal{S}$, perceived by the agent. s is firstly projected in the sensory-space representation and, if this projection has close neighbors—s_1 in Fig. 2—its corresponding motor state, $\tilde{m}$, is interpolated from the neighborhood, giving a passive localization estimation. If the projection is outlying in an area with no neighbors —s_2 in Fig. 2—this sensory state is not yet represented and $\tilde{m}$ can not be estimated passively. In this case the orienting behavior is executed, giving an active estimation of $\tilde{m}$.

Instead of learning an auditory space representation from a database, an iterative process can be used by mixing the dimensionality reduction with the evoked behavior. This allows therefore the representation to be learned online, experience after experience. Thus each new percept which projection is outlying in the representation is learned (s_2 in Fig. 2): its related sensory state is added in an updated representation and is associated with its related active estimation of $\tilde{m}$. Reference [9] provide more details on this learning algorithm and the auditory system used. The authors propose a simulated experiment where a mean localization error of about 1° is reached after an online learning of 200 iterations.

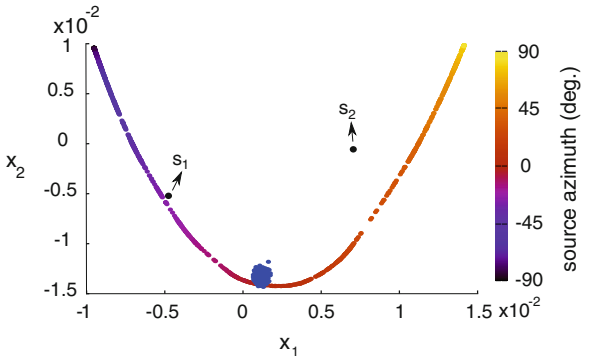

Fig. 2 Two dimensional manifold of auditory space learned from a set of 1000 sensory states. *Parabolic curve* obtained *before* the orienting behavior—each state corresponding to a sound source of random azimuth in the range of $\pm 90°$. *Point cluster close to* $x_1 = 0$: Projections on the manifold of sensory states obtained *after* the orienting behavior, approximating the reference state, $\Phi(m_0, e_0)$. New percepts, such as s_1 and s_2, can be localized on the manifold as well—see text for details

Discussion

The above method has, to be sure, been illustrated for a very simple case, namely, a single stationary sound source in the azimuthal plane. Yet, it seems to be basically suitable for hearing systems in autonomous robotics. In fact, the dimensionality reduction used for the computation of the auditory-space representation allows for unsupervised learning of scene non-linearities, such as reverberation or HRTF filtering. Also, this method requires almost no a-priori knowledge of either the agent or the environment. It mainly depends on the knowledge of the auditory-space representation dimension, typically 2-D or 3-D, and on a dimension-reduction technique robust enough to estimate the non-linear embedding of complex environments in an efficient hard-wired evoked behavior. Active hearing, binaural processing, representation learning and online estimation have the potential to be integrated into a single model that could be applied to more complex problems, thus opening new perspectives for sensorimotor approach to binaural robot audition.

5.2 A Stochastic-Filtering Strategy for Active Binaural Sound Localization

While the above approach aims at estimating the source position from a binaural sensor with no assumption regarding the environment, this chapter will now be concluded with a strochastic-filtering approach to binaural sound localization from a moving platform. As to this field, reference is due to cite [14, 46], where tracking algorithms based on the particle filtering framework are exploited to detect utterer changes and infer speaker location, respectively. The work presented in this section shows how binaural perception and motor commands of the sensor can be fused to localize an intermittent source in the presence of false measurements.

In the context of binaural audition, sound-source localization relies prominently on *time-delay estimation*, TDE, that is, on an estimation of the arrival-time differences of the sound signals at the two acoustic sensors. The topic of TDE has been widely addressed. In robotics, the most common approach is undoubtedly *generalized cross-correlation*, GCC [38], which consists in cross-correlating truncated and filtered versions of the raw sensed signals and picking the `argmax` of the resulting function. However, given a state vector, X, that is, a vector fully characterizing the sensor-to-source relative position, the time delay comes as a nonlinear and noninvertible function, h, of X. Without any additional information it is not possible to recover the complete state vector from just a time-delay estimate. For instance, consider for simplification, that source and microphones lie on a common plane parallel to the ground, and let the Cartesian coordinates vector, $X = (x, y)^T$, represent the source position in a frame, (R, e_X, e_Y), rigidly linked to the microphone pair, $\{R_l, R_r\}$, with $R_lR = \frac{R_lR_r}{2}$ and $e_Y = \frac{RR_l}{|RR_l|}$. It can be shown that given a time delay, τ, all the pairs, (x, y), satisfying $h(x, y) = \tau$, describe a branch of hyperbola, referred in the literature as cone of confusion. In other words, given a time delay, one cannot

locate the true sound source on the associated hyperbola branch. However, with the microphones being mounted on a mobile robot, its motor commands, for instance, translational and rotational velocities, can be fused with audio perception to infer sound localization. Similarly, when the source is moving, prior knowledge about its dynamics can be used. One way to tackle this problem is to use a *Bayesian filtering framework*. In this context and in the presence of relative motion, X is now considered as a discrete hidden-Markov process, characterized by a dynamic equation of the form

$$X_{[k+1]} = f(X_{[k]}, u_{1[k]}, u_{2[k]}) + W_{[k]}. \tag{2}$$

Therein, $X_{[k]}$ is a random vector describing the process X at time step k. $u_{1[k]}$ is a deterministic vector gathering information about the robot's motor commands. $u_{2[k]}$ is a vector composed of the source velocities. $u_{2[k]}$ can be deterministic or random, depending on whether the source motion is fully described beforehand or not. $W_{[k]}$ is an additive random noise accounting for uncertainty in the relative motion. At each time, k, the time delay measurement, hereafter referred as $Z_{[k]}$, is a memoryless function of the state vector, according to

$$Z_{[k]} = h(X_{[k]}, u_{1[k]}, u_{2[k]}) + V_{[k]}, \tag{3}$$

with $V_{[k]}$ being an additive noise representing the TDE error. Given an initial probability-density function, pdf, of $p(x_{[0]})$ and a sequence of measurements, $z_{[1:k]} \triangleq z_{[1]}, \ldots, z_{[k]}$, considered as samples of $Z_{[1:k]} \triangleq Z_{[1]}, \ldots, Z_{[k]}$, the optimal Bayesian filter consists in the recursive computation of the posterior state probability-density function, that is, $p(x_{[k]}|Z_{[1:k]} = z_{[1:k]})$. When f, g are nonlinear functions and/or W, V are non-gaussian processes, the optimal filter has no closed-form expression. Approximate solutions are thus needed, such as the *extended/unscented Kalman filter*, EKF/UKF, particle filters, PF, or grid-based methods. Whatever the chosen strategy is, certain issues have to be dealt with, such as

Modeling The state space model must be defined in such a way that the state vector gathers a minimal set of parameters. For a still source—or a moving source with known velocities with respect to the world—the state vector can be made up with, for example, its Cartesian coordinates in the sensor frame. If the source is moving at unknown speed, an autonomous equation describing the structure of its motion in the world frame must be introduced, whose initial condition and parameters complete the vector X to be estimated. For the localization of human utterers, typically used models are Langevin processes or random walks [13, 86]. However that may be, the mathematical transcription of the prior knowledge of the source dynamics is of crucial importance.

Consistency and tuning Generally, the statistics of the dynamic and measurement noise processes are unknown, so they must be hypothesized. Setting a too-high covariance leads to too pessimistic conclusions, while setting them too low may result in an overconfident filter. In general, the noise statistics are set in an ad-hoc manner. However, inconsistency, that is, overconfidence or underconfidence, can

arise independently of the noise statistics. Indeed, the approximation of the true-state posterior density function propagated recursively by the filter can become inaccurate when the nonlinearities are not smooth enough and/or the filtering technique is not suited to the model and its parameters. For instance, the runs of basic particle filtering strategies conducted by the authors on simulated and experimental measurements showed that these estimation strategies are not suited to active binaural localization.

Initialization When no prior knowledge about the source location is available, one usually set the initial prior $p(x_{[0]})$ as a *flat prior*, that is, a probability distribution with zero-mean and infinite covariance matrix. However, due to the non-linearities involved in the considered source localization problem, the propagation of widely spread distributions often leads to overconfident conclusions. As a solution, the posterior state pdf can be approximated by a *Gaussian mixture*, GM, (GM) whose hypotheses are recursively propagated using a bank of non-interactive filters [7].

Time-delay extraction At each time, k, the measurement, $Z_{[k]}$, is obtained from a TDE algorithm using audio data collected over a finite time window. Generally, this time window is of short duration for distinct reasons. First, TDE algorithms rely on the hypothesis that the two windowed signals can be regarded as sample sequences of individually and jointly *wide-sense stationary*, WSS, processes. Individual stationarity implies that the source signal is itself WSS, while speech, for instance, cannot be considered as WSS unless the time window is sufficiently short. Joint stationarity implies that the time delay must be approximately constant over the time window. This is of crucial importance when source and sensor move. Classical TDE algorithms do not provide reliable and meaningful estimates if this hypothesis is not satisfied—unless the time-delay variations are specifically taken into account and compensated in the algorithm, like in [39]. Finally, in an embedded application, a cross-correlation cannot have an unreasonable length, due to finite time and space resources.

Because of the environment noise, the non-stationarity of the source and the short duration of TDE windows, the TDE statistics may change significantly over time, namely, if at a considered instant, the SNR and time-bandwidth product, TBP, of the signals are sufficiently high, then the TDE algorithm outputs an accurate estimate of the genuine time delay. If not, the estimate might be unreliable, that is, drawn from a process with large variance or, in the worst case, uncorrelated with the state [87, 88]. Such spurious measurements must be taken into account in order to prevent filter inconsistency/instability. They can be handled in a *hard manner* with an external decision rule that selects, according to some criteria, such as estimated SNR, the measurements that are to be incorporated into the filter—or in a probabilistic way, for example, by probabilistic data association.

Taking all these considerations into account, a filtering strategy was proposed in [69]. It relies on a multiple hypothesis UKF with probabilistic data association and a *source-activity detection*, SAD, system based on *generalized likelihood-ratio test*, GLRT. The results from an experiment conducted in an acoustically prepared room are shown in Fig. 3. Each subfigure represents a time snapshot, the left figure

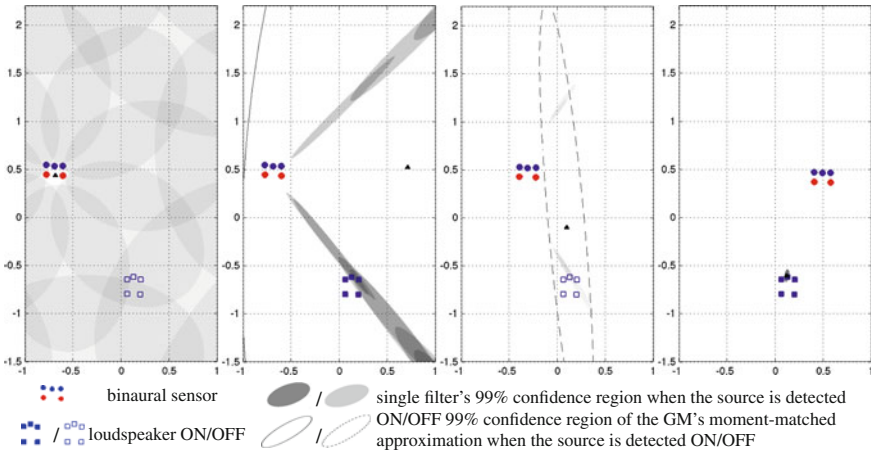

Fig. 3 Localization results from an experiment conducted in an acoustically prepared room. Each figure represents a time snapshot. *Far left* initial time. *Far right* final time. The results show that the motion of the sensor allowed to disambiguate front and back and to provide information regarding source distance. For more explanations and details see [69]

corresponding to initial time, and the right figure corresponding to final time. At the beginning, the filter is initialized with 24 hypotheses so that the union of the 99 % probability ellipsoids defined from the 24 modes of the initial GM-prior-pdf covers a 4 m-radius circular region around the sensor. In the second snapshot, a part of the hypothesis is spread along the source-to-sensor direction, while another part is spread along the symmetric direction with respect to the $(R_l R_r)$-axis. This behavior depicts that so far there is a large uncertainty on the distance to the source and there is a front-back ambiguity. This originates from the aforementioned time-delay characteristics. In the third snapshot, the loudspeaker is switched off, and the transition from on to off has been detected by the filter. In the fourth snapshot, the loudspeaker is emitting again, and the transition from off to on has been detected correctly. Note that the state pdf is now very sharp around the true-source location. In other words, thanks to motion, front and back have been disambiguated and the distance uncertainty has significantly decreased. The experimental results show that the standard deviation of the errors at the end of the listener's motion is about $\pm 2^\circ$ in terms of azimuth and ± 5 cm in terms of range. However, the good performance of the system is partially due to the favorable experimental conditions, namely, the sensor speeds were precisely known, and the acoustic environment was particularly clean—that is, with only little reverberation and noise. Experiments should be performed in a more realistic environment such as an office, with possibly non-negligible background noise. This is subject to future work.

6 Conclusion

This chapter has discussed binaural systems in robotics along several dimensions. After the statement of key constraints raised by this context, the canonical binaural functions underlying the analysis of any auditory scene were detailed. Prominent platforms and research projects were then reviewed. Finally, two recent approaches to binaural audition developed by the authors were presented. These are termed to be *active* because they consider the coupling of binaural sensing and robot motion.

As mentioned before, binaural audition is an attractive paradigm, regarding engineering issues, cross-fertilization between robotics and neurosciences, as well as Human–robot interaction. Though the field is studied by only very few laboratories as compared to binocular vision, it has now reached a certain level of maturity. But there remains ample space for methodological and technological contributions, particularly, to the end of better coping with uncontrolled and dynamic environments. This could then allow to better understand how to use binaural audition as a mechanism for acoustic-scene analysis in general.

To conclude, some broader research areas connected to binaural audition have been mentioned that hopefully bring new researchers to the field. First, the coupling of bottom-up and top-down processes in active audition and, to a larger extent, to active perception, deserves attention. As shown above, two distinct viewpoints have been developed towards purposeful auditory perception. One addresses the definition of top-down feedback from symbolic levels, while the other approaches—including those developed by the current authors—have addressed this topic right at the sensorimotor level. In the authors' opinion, these two lines of research are to be reinforced and must join each other in order to define a comprehensive computational architecture for active analysis of auditory scenes. Some subtopics should finally be mentioned, such as, the definition of binaural audition based control/estimation strategies that explicitly include an exploration goal to collect information about the source location, their interlinking with decision processes, the assimilation of available data over space and time, the generality of such an approach and its ability to tackle multimodality, adaptive approaches to binaural cues extraction and exploitation where auditory cues and algorithms are dynamically changed according to the context, just to name a few. A second fruitful broad research area is the involvement of binaural audition in ubiquitous robotics. The idea here is to outfit rooms with embedded sensors, such as microphone arrays, cameras and/or RFID antennas. These areas would be shared by humans and robots interacting with each other, using binaural audition, vision and maybe, further available modalities. The mobility of the robots enables the possibility of combining their motor commands with binaural perception, not only to improve a *local* binaural function, but also to dynamically reconfigure the global network constituted by the microphone arrays and the binaural heads—possibly including prior knowledge. Some contributions have already been developed in this field on the basis of dynamically reconfigurable microphone arrays [49]. They could constitute a valuable source of inspiration for enhancing binaural robot audition.

Acknowledgments This work was conducted within the project *binaural active audition for humanoid robots*, BINAAHR, funded under contract # ANR-09-BLAN-0370-02 by ANR, France, and JST, Japan. The authors would like to thank two anonymous reviewers for valuable suggestions.

References

1. J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *Intl. J. Computer Vision*, 1:333–356, 1988.
2. S. Argentieri and P. Danès. Broadband variations of the MUSIC high-resolution method for sound source localization in robotics. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2007, pages 2009–2014, 2007.
3. E. Arnaud, H. Christensen, Y.-C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant, F. Forbes, and R. Horaud. The CAVA corpus: Synchronised stereoscopic and binaural datasets with head movements. In *ACM/IEEE Intl. Conf. Multimodal, Interfaces*, ICMI'08, 2008.
4. M. Aytekin, C. Moss, and J. Simon. A sensorimotor approach to sound localization. *Neural Computation*, 20:603–635, 2008.
5. P. Azad, T. Gockel, R. Dillmann. *Computer Vision: Principles and Practice*. Elektor, Electronics, 2008.
6. R. Bajcsy. Active perception. *Proc. of the IEEE*, 76:966–1005, 1988.
7. Y. Bar-Shalom and X. Li. *Estimation and Tracking: Principles, Techniques and Software*. Artech House, 1993.
8. M. Bernard, S. N'Guyen, P. Pirim, B. Gas, and J.-A. Meyer. Phonotaxis behavior in the artificial rat Psikharpax. In *Intl. Symp. Robotics and Intelligent Sensors*, IRIS'2010, pages 118–122, Nagoya, Japan, 2010.
9. M. Bernard, P. Pirim, A. de Cheveigné, and B. Gas. Sensorimotor learning of sound localization from an auditory evoked behavior. In IEEE *Intl. Conf. Robotics and Automation*, ICRA'2012, pages 91–96, St. Paul, MN, 2012.
10. J. Blauert, D. Kolossa, K. Obermayer, and K. Adiloglu. Further challenges and the road ahead. In J. Blauert, editor, *The technology of binaural listening*, chapter 18. Springer, Berlin-Heidelberg-New York NY, 2013.
11. W. Brimijoin, D. Mc Shefferty, and M. Akeroyd. Undirected head movements of listeners with asymmetrical hearing impairment during a speech-in-noise task.*Hearing Research*, 283:162–8, 2012.
12. R. Brooks, C. Breazeal, N. Marjanović, B. Scassellati, and M. Williamson. The Cog project: Building a humanoid robot. In C. Nehaniv, editor, *Computations for Metaphors, Analogy, and Agents*, volume 1562 of LNCS, pages 52–87. Springer, 1999.
13. Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. *Proc. of the IEEE*, 920:485–494, 2004.
14. H. Christensen and J. Barker. Using location cues to track speaker changes from mobile binaural microphones. In *Interspeech 2009*, Brighton, UK, 2009.
15. H. Christensen, J. Barker, Y.-C. Lu, J. Xavier, R. Caseiro, and H. Arafajo. POPeye: Real-time binaural sound-source localisation on an audio-visual robot head. In *Conf. Natural Computing and Intelligent Robotics*, 2009.
16. Computing Community Consortium. *A roadmap for US robotics. From Internet to Robotics*, 2009. http://www.us-robotics.us/reports/CCC%20Report.pdf.
17. M. Cooke, Y. Lu, Y. Lu, and R. Horaud. Active hearing, active speaking. In *Intl. Symp*. Auditory and Audiological Res., 2007.
18. M. Cooke, A. Morris, and P. Green. Recognizing occluded speech. In *Proceedings of the ESCA Tutorial and Res.arch Worksh. Auditory Basis of Speech Perception*, pages 297–300, Keele University, United Kingdom, 1996.

19. M. Cooke, A. Morris, and P. Green. Missing data techniques for robust speech recognition. In *Intl. Conf. Acoustics, Speech, and Signal Processing*, ICASSP'1997, pages 863–866, Munich, Germany, 1997.
20. B. Cornelis, M. Moonen, and J. Wouters. Binaural voice activity detection for MWF-based noise reduction in binaural hearing aids. In *European Signal Processing Conf.*, EUSIPCO'2011, pages Barcelona, Spain, 2011.
21. P. Danès and J. Bonnal. Information-theoretic detection of broadband sources in a coherent beamspace MUSIC scheme. In *IEEE/RSJ Intl. Conf. Intell. Robots and Systems*, IROS'2010, pages 1976–1981, Taipei, Taiwan, 2010.
22. A. Deleforge and R. Horaud. Learning the direction of a sound source using head motions and spectral features. Technical Report 7529, INRIA, 2011.
23. A. Deleforge and R. Horaud. The Cocktail-Party robot: Sound source separation and localisation with an active binaural head. In *IEEE/ACM Intl. Conf. Human Robot Interaction*, HRI'2012, Boston, MA, 2012.
24. J. Gibson. *The Ecological Approach to Visual Perception*. Erlbaum, 1982.
25. M. Giuliani, C. Lenz, T. Müller, M. Rickert, and A. Knoll. Design principles for safety in human-robot interaction. *Intl. J. Social Robotics*, 2:253–274, 2010.
26. A. Handzel, S. Andersson, M. Gebremichael, and P. Krishnaprasad. A biomimetic apparatus for sound-source localization. In *IEEE Conf. Decision and Control*, CDC'2003, volume 6, pages 5879–5884, Maui, HI, 2003.
27. A. Handzel and P. Krishnaprasad. Biomimetic sound-source localization. *IEEE Sensors J.*, 2:607–616, 2002.
28. S. Hashimoto, S. Narita, H. Kasahara, A. Takanishi, S. Sugano, K. Shirai, T. Kobayashi, H. Takanobu, T. Kurata, K. Fujiwara, T. Matsuno, T. Kawasaki, K. Hoashi. Humanoid robot-development of an information assistant robot, Hadaly. In *IEEE Intl. Worksh. Robot and Human, Communication*, RO-MAN'1997, pages 106–111, 1997.
29. J. Hörnstein, M. Lopes, J. Santos-victor, and F. Lacerda. Sound localization for humanoid robots - building audio-motor maps based on the HRTF. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2006, pages 1170–1176, Beijing, China, 2006.
30. J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie. A model-based sound localization system and its application to robot navigation. *Robotics and Autonomous Syst.*, 270:199–209, 1999.
31. G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura. Ego noise suppression of a robot using template subtraction. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2009, pages 199–204, Saint Louis, MO, 2009.
32. G. Ince, K. Nakadai, T. Rodemann, J. Imura, K. Nakamura, and H. Nakajima. Incremental learning for ego noise estimation of a robot. In*IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2011, pages 131–136, San Francisco, CA, 2011.
33. G. Ince, K. Nakadai, T. Rodemann, H. Tsujino, and J. Imura. Multi-talker speech recognition under ego-motion noise using missing feature theory. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2010, pages 982–987, Taipei, Taiwan, 2010.
34. R. Irie. Multimodal sensory integration for localization in a humanoid robot. In *IJCAI Worksh. Computational Auditory Scene Analysis*, pages 54–58, Nagoya, Aichi, Japan, 1997.
35. A. Ito, T. Kanayama, M. Suzuki, and S. Makino. Internal noise suppression for speech recognition by small robots. In *Interspeech*'2005, pages 2685–2688, Lisbon, Portugal, 2005.
36. M. Ji, S. Kim, H. Kim, K. Kwak, and Y. Cho. Reliable speaker identification using multiple microphones in ubiquitous robot companion environment. In *IEEE Intl. Conf. Robot & Human Interactive Communication*, RO-MAN'2007, pages 673–677, Jeju Island, Korea, 2007.
37. H.-D. Kim, J. Kim, K. Komatani, T. Ogata, and H. Okuno. Target speech detection and separation for humanoid robots in sparse dialogue with noisy home environments. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2008, pages 1705–1711, Nice, France, 2008.
38. C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoustics, Speech and, Signal Processing*, 24:320–327, 1976.

39. C. Knapp and G. Carter. Time delay estimation in the presence of relative motion. In *IEEE Intl. Conf. Acoustics, Speech, and Signal Processing*, ICASSP'1977, pages 280–283, Storrs, CT, 1977.
40. Y. Kubota, M. Yoshida, K. Komatani, T. Ogata, and H. Okuno. Design and implementation of a 3D auditory scene visualizer: Towards auditory awareness with face tracking. In *IEEE Intl. Symp. Multimedia*, ISM'2008, pages 468–476, Berkeley, CA, 2008.
41. M. Kumon and Y. Noda. Active soft pinnae for robots. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2011, pages 112–117, San Francisco, CA, 2011.
42. M. Kumon, R. Shimoda, and Z. Iwai. Audio servo for robotic systems with pinnae. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2005, pages 885–890, Edmonton, Canada, 2005.
43. S. Kurotaki, N. Suzuki, K. Nakadai, H. Okuno, and H. Amano. Implementation of active direction-pass filter on dynamically reconfigurable processor. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2005, pages 3175–3180, Edmonton, Canada, 2005.
44. Q. Lin, E. E. Jan, and J. Flanagan. Microphone arrays and speaker identification. *IEEE Trans. Speech and Audio Processing*, 2:622–629, 1994.
45. R. Lippmann and B. A. Carlson. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In *Eurospeech*'1997, pages 863–866, Rhodos, Greece, 1997.
46. Y.-C. Lu and M. Cooke. Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners. *Speech Comm.*, 53:622–642, 2011.
47. V. Lunati, J. Manhès, and P. Danès. A versatile system-on-a-programmable-chip for array processing and binaural robot audition. In*IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2012, pages 998–1003, Vilamoura, Portugal, 2012.
48. D. Marr. Vision: A Computational Investigation into the Human Representation and Processing of *Visual Information*. Feeeman, W.H., 1982.
49. E. Martinson and B. Fransen. Dynamically reconfigurable microphone arrays. In *IEEE Intl. Conf. Robotics and Automation*, ICRA'2011, pages 5636–5641, Shangai, China, 2011.
50. Y. Matsusaka, T. Tojo, S. Kubota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi. Multi-person conversation via multi-modal interface - a robot who communicate with multi-user -. In *Eurospeech*'1999, pages 1723–1726, Budapest, Hungary, 1999.
51. T. May, S. van de Par, and A. Kohlrausch. Binaural localization and detection of speakers in complex acoustic scenes. In J. Blauert, editor, *The Technology of Binaural Listening*, chapter 15. Springer, Berlin-Heidelberg-New York NY, 2013.
52. F. Michaud, C. Côté, D. Létourneau, Y. Brosseau, J.-M. Valin, E. Beaudry, C. Raïevsky, A. Ponchon, P. Moisan, P. Lepage, Y. Morin, F. Gagnon, P. Giguère, M.-A. Roux, S. Caron, P. Frenette, and F. Kabanza. Spartacus attending the 2005 AAAI conference.*Autonomous Robots*, 22:369–383, 2007.
53. K. Nakadai, T. Lourens, H. Okuno, and H. Kitano. Active audition for humanoids. In *Nat. Conf. Artificial Intelligence*, AAAI-2000, pages 832–839, Austin, TX, 2000.
54. K. Nakadai, T. Matsui, H. Okuno, and H. Kitano. Active audition system and humanoid exterior design. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2000, pages 1453–1461, Takamatsu, Japan, 2000.
55. K. Nakadai, D. Matsuura, H. Okuno, and H. Kitano. Applying scattering theory to robot audition system: Robust sound source localization and extraction. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2003, pages 1147–1152, Las Vegas, NV, 2003.
56. K. Nakadai, H. Okuno, and H. Kitano. Epipolar geometry based sound localization and extraction for humanoid audition. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2001, volume 3, pages 1395–1401, Maui, HI, 2001.
57. K. Nakadai, H. Okuno, and H. Kitano. Auditory fovea based speech separation and its application to dialog system. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, IROS'2002, volume 2, pages 1320–1325, Lausanne, Switzerland, 2002.
58. K. Nakadai, H. Okuno, and H. Kitano. Robot recognizes three simultaneous speech by active audition. In *IEEE Intl. Conf. Robotics and Automation*, ICRA'2003, volume 1, pages 398–405, Taipei, Taiwan, 2003.

59. H. Nakajima, K. Kikuchi, T. Daigo, Y. Kaneda, K. Nakadai, and Y. Hasegawa. Real-time sound source orientation estimation using a 96 channel microphone array. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2009, pages 676–683, Saint Louis, MO, 2009.
60. H. Nakashima and T. Mukai. 3D sound source localization system based on learning of binaural hearing. In *IEEE Intl. Conf. Systems*, Man and Cybernetics, SMC'2005, pages 3534–3539, Nagoya, Japan, 2005.
61. E. Nemer, R. Goubran, and S. Mahmoud. Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Trans. Speech and Audio Processing*, 9:217–231, 2001.
62. Y. Nishimura, M. Nakano, K. Nakadai, H. Tsujino, and M. Ishizuka. Speech recognition for a robot under its motor noises by selective application of missing feature theory and MLLR. In *ISCA Tutorial and Research Worksh. Statistical and Perceptual Audition*, Pittsburgh, PA, 2006.
63. H. Okuno, T. Ogata, K. Komatani, and K. Nakadai. Computational auditory scene analysis and its application to robot audition. In *IEEE Intl. Conf. Informatics Res. for Development of Knowledge Society Infrastructure*, ICKS'2004, pages 73–80, 2004.
64. J. O'Regan. How to build a robot that is conscious and feels. *Minds and Machines*, pages 117–136, 2012.
65. J. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24:939–1031, 2001.
66. D. Philipona and J. K. O'Regan. Is there something out there? inferring space from sensorimotor dependencies. *Neural Computation*, 15:2029–2049, 2001.
67. B. Pierce, T. Kuratate, A. Maejima, S. Morishima, Y. Matsusaka, M. Durkovic, K. Diepold, and G. Cheng. Development of an integrated multi-modal communication robotic face. In *IEEE Worksh. Advanced Robotics and its Social Impacts*, RSO'2012, pages 101–102, Munich, Germany, 2012.
68. H. Poincaré. L'espace et la géométrie. *Revue de Métaphysique et de Morale*, pages 631–646, 1895.
69. A. Portello, P. Danès, and S. Argentieri. Active binaural localization of intermittent moving sources in the presence of false meaurements. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2012, pages 3294–3299, Vilamoura, Portugal, 2012.
70. R. Prasad, H. Saruwatari, and K. Shikano. Enhancement of speech signals separated from their convolutive mixture by FDICA algorithm. *Digital Signal Processing*, 19:127–133, 2009.
71. L. Rabiner and M. Sambur. An algorithm for determining the endpoints of isolated utterances. *The Bell System Techn. J.*, 54:297–315, 1975.
72. B. Raj, R. Singh, and R. Stern. Inference of missing spectrographic features for robust speech recognition. In *Intl. Conf. Spoken Language Processing*, Sydney, Australia, 1998.
73. B. Raj and R. M. Stern. Missing-feature approaches in speech recognition. *IEEE Signal Processing Mag.*, 22:101–116, 2005.
74. T. Rodemann. A study on distance estimation in binaural sound localization. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2010, pages 425–430, Taipei, Taiwan, 2010.
75. D. Rosenthal and H. Okuno, editors. *Computational Auditory Scene Analysis*. Lawrence Erlbaum Associates, 1997.
76. A. Saxena and A. Ng. Learning sound location from a single microphone. In *IEEE Intl. Conf. Robotics and Automation*, ICRA'2009, pages 1737–1742, Kobe, Japan, 2009.
77. S. Schulz and T. Herfet. Humanoid separation of speech sources in reverberant environments. In *Intl. Symp. Communications*, Control and Signal Processing, ISCCSP'2008, pages 377–382, Brownsville, TX, 2008.
78. M. L. Seltzer, B. Raj, and R. Stern. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Comm.*, 43:379–393, 2004.
79. A. Skaf and P. Danès. Optimal positioning of a binaural sensor on a humanoid head for sound source localization. In *IEEE Intl. Conf. Humanoid Robots*, Humanoids'2011, pages 165–170, Bled, Slovenia, 2011.
80. D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten. A study of lip movements during spontaneous dialog and its application to voice activity detection. *J. Acoust. Soc. Am.*, 125:1184–1196, 2009.

81. M. Stamm and M. Altinsoy. Employing binaural-proprioceptive interaction in human machine interfaces. In J. Blauert, editor, *The technology of binaural listening*, chapter 17. Springer, Berlin-Heidelberg-New York NY, 2013.
82. R. Takeda, S. Yamamoto, K. Komatani, T. Ogata, and H. Okuno. Missing-feature based speech recognition for two simultaneous speech signals separated by ICA with a pair of humanoid ears. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2006, pages 878–885, Beijing, China, 2006.
83. K. Tanaka, M. Abe, and S. Ando. A novel mechanical cochlea "fishbone" with dual sensor/actuator characteristics. *IEEE/ASME Trans. Mechatronics*, 3:98–105, 1998.
84. J. Valin, J. Rouat, and F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2004, pages 2123–2128, Sendai, Japan, 2004.
85. H. Van Trees. *Optimum Array Processing (Detection, Estimation, and Modulation Theory, Part IV)*. Wiley-Interscience, 2002.
86. D. Ward, E. Lehmann, and R. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Trans. Speech and Audio Processing*, 11:826–836, 2003.
87. E. Weinstein and A. Weiss. Fundamental limitations in passive time delay estimation - Part II: Wideband systems. *IEEE Trans. Acoustics, Speech and Signal Processing*, pages 1064–1078, 1984.
88. A. Weiss and E. Weinstein. Fundamental limitations in passive time delay estimation - Part I: Narrowband systems. *IEEE Trans. Acoustics, Speech and Signal Processing*, pages 472–486, 1983.
89. R. Weiss, M. Mandel, and D. Ellis. Combining localization cues and source model constraints for binaural source separation. *Speech Comm.*, 53:606–621, 2011.
90. R. Woodworth and H. Schlosberg. *Experimental Psychology*. Holt, Rinehart and Winston, 3rd edition, 1971.
91. T. Yoshida and K. Nakadai. Two-layered audio-visual speech recognition for robots in noisy environments. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2010, pages 988–993, 2010.
92. K. Youssef, S. Argentieri, and J. Zarader. From monaural to binaural speaker recognition for humanoid robots. In *IEEE/RAS Intl. Conf. Humanoid Robots*, Humanoids'2010, pages 580–586, Nashville, TN, 2010.
93. K. Youssef, S. Argentieri, and J.-L. Zarader. A binaural sound source localization method using auditive cues and vision. In *IEEE Intl. Conf. Acoustics, Speech and Signal Processing*, ICASSP'2012, pages 217–220, Kyoto, Japan, 2012.
94. K. Youssef, S. Argentieri, and J.-L. Zarader. Towards a systematic study of binaural cues. In *IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, IROS'2012, pages 1004–1009, Vilamoura, Portugal, 2012.
95. K. Youssef, B. Breteau, S. Argentieri, J.-L. Zarader, and Z. Wang. Approaches for automatic speaker recognition in a binaural humanoid context. In *Eur. Symp. Artificial Neural Networks*, Computational Intelligence and Machine Learning, ESANN'2011, pages 411–416, Bruges, Belgium, 2011.