Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 87–112

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A survey on sound source localization in robotics: From binaural to array processing methods☆,☆☆

S. Argentieri[a,b,*], P. Danès[c,d], P. Souères[c,e]

[a] *Sorbonne Universités, UPMC Univ. Paris 06, UMR 7222, ISIR, F-75005 Paris, France*
[b] *CNRS, UMR 7222, ISIR, F-75005 Paris, France*
[c] *CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France*
[d] *Univ. de Toulouse, UPS, LAAS, F-31400 Toulouse, France*
[e] *Univ. de Toulouse, LAAS, F-31400 Toulouse, France*

Received 3 June 2014; received in revised form 13 February 2015; accepted 10 March 2015
Available online 20 March 2015

Abstract

This paper attempts to provide a state-of-the-art of sound source localization in robotics. Noticeably, this context raises original constraints—e.g. embeddability, real time, broadband environments, noise and reverberation—which are seldom simultaneously taken into account in acoustics or signal processing. A comprehensive review is proposed of recent robotics achievements, be they binaural or rooted in array processing techniques. The connections are highlighted with the underlying theory as well as with elements of physiology and neurology of human hearing.

Keywords: Robot audition; Source localization; Binaural audition; Array processing

1. Introduction

"Blindness separate us from things but deafness from people" said Helen Keller, a famous American author who was the first deafblind person to obtain a Bachelor in Arts, in 1904 (Kohlrausch et al., 2013). Indeed, hearing is a prominent sense for communication and socialization. In contrast to vision, our perception of sound is nearly omnidirectional and independent of the lighting conditions. Similarly, we are able to process sounds issued from a nearby room without any visual information on their origin. But human capabilities are not limited to sound *localization*. We can also *extract*, within a group of speakers talking simultaneously, the utterance emitted by the person we wish to focus on. Known as the term *Cocktail Party Effect* (Haykin and Chen, 2005), this separation capacity enables us to process efficiently and selectively the whole acoustic data coming from our daily environment. Sensitive to the slightest tone and level variations of an audio message, we have developed a faculty to *recognize* its origin (ringtone, voice of a colleague,

☆ This works has been conducted within the European FP7 TWO!EARS project under grant agreement n° 618075 and the BINAAHR project funded by ANR (France) and JST (Japan) under Contract n° ANR-09-BLAN- 0370-02.
☆☆ This paper has been recommended for acceptance by R.K. Moore.
* Corresponding author. Tel.: +33 144276355.
E-mail addresses: sylvain.argentieri@upmc.fr (S. Argentieri), patrick.danes@laas.fr (P. Danès), philippe.soueres@laas.fr (P. Souères).

http://dx.doi.org/10.1016/j.csl.2015.03.003
0885-2308/

etc.) and to *interpret* its contents. All these properties of localization, extraction, recognition and interpretation allow us to operate in dynamic environments, where it would be difficult to do without auditory information. All the above impressive human capabilities have stimulated developments in the area of *Robot Audition*. Likewise, the recent research topic of human–robot interaction (HRI) may have constituted an additional motivation to investigate this new field, with the aim to artificially reproduce the aforementioned localization, extraction, recognition and interpretation capabilities. Nevertheless, contrarily to computer vision, robot audition has been identified as a scientific topic of its own only since about 15 years (Nakadai et al., 2000). Since then, numerous works have been proposed by a growing community, with contributions ranging from sound source localization and separations in realistic reverberant conditions to speech or speaker recognition in the presence of noise. But as outlined in Argentieri et al. (2013), the robotics context raises several unexpected constraints, seldomly taken into account in signal processing or acoustics. Among them, one can mention:

Geometry constraint: Though the aim is to design an artificial auditory system endowed with performances inspired by human hearing, there is no need to restrict the study to a biomimetic sensor endowed with just two microphones. Indeed, bringing redundant information delivered by multiple transducers can improve the analysis and its robustness to noise. Straight connections thus appear with the field of array processing. Yet, the robotics context imposes an *embeddability* constraint. While array processing can consider large arrays of microphones—e.g. several meters long, robotics implies a trade-off between the size of the whole sensor and its performances, so that it can be mounted on a mobile platform, be it humanoid or not.

Real time constraint: Many existing methods to sound analysis rely on heavy computations. For instance, a processing time extending over several tens of seconds is admitted to perform the acoustic analysis of a passenger compartment. Contrarily, localization primitives involved in low-level reflex robotics functions—e.g. sensor-based control or auditive/visioauditive tracking—must be made available within a guaranteed short time interval. So the algorithms computational complexity is a fundamental concern. This may imply the design of dedicated devices or computational architectures.

Frequency constraint: Most sound signals valuable to robotics are *broadband*, i.e. spread over a wide bandwidth w.r.t. their central frequency. This is the case of voice signals, which show significant energy on the bandwidth [300–3300 Hz] used for telephony. Consequently, narrowband approaches developed elsewhere do not straightly apply in such broadband contexts. Noticeably, this may imply a higher computational demand.

Environmental constraint: Robotics environments are fundamentally dynamic and unpredictable. Contrarily to acoustically fully controlled areas, unexpected noise and reverberations are likely to occur, which depend on the room characteristics—dimensions, walls, type of the building materials, etc.—and may singularly deteriorate the analysis performance. The robot itself participates to these perturbations, because of its self-induced noise, e.g. from fans, motors, and other moving parts. A challenge is to endow embedded sound analysis systems with robustness and/or adaptivity capabilities, able to cope with barge-in situations where both the robot and a human are possibly both speaking together.

Generally, most of embedded auditory systems in robotics follow the following classical bottom-up framework: as a first step, the sensed signals are analyzed to estimate sound sources positions; next the locations are used to separate sound of interests from the sensed mixture in order to provide clean noise or speech signals; finally, sound/speaker/speech recognition systems are fed with these extracted signals. Of course, other alternatives have also been proposed (Otsuka et al., 2012), but this approach remains by far the most used framework in robot audition. Nevertheless, it exhibits the importance of sound localization in the overall analysis process. It has been indeed the most widely covered topic in the community, and a lot of efforts have been made to provide efficient sound localization algorithms suited to the robotics context. Moreover, independently of any high-level interpretation of the acoustical scene, having access to the low-level source localization information itself is mandatory for any applications related to HRI. Indeed, a natural intuitive HRI might heavily depends on how responsive a robot will be to acoustical information. Among them, source localization allows the robot to quickly react to an auditory stimulus by turning the head towards the source (turn-to reflex), or even to focus its other sensors in the direction of interest (e.g. by moving a camera field of view towards a speaker). Since, in our opinion, Robot Audition has reached an undeniable level of scientific maturity, we feel that the time has come to summarize and organize the main publications of the literature. This paper then attempts to review the most notable contributions specific to sound source localization. Another intent is to underline their connections with theoretical foundations of the field, including with basics of human physiology and neuroscience.

The paper is organized into two parts. First, binaural methods to sound source localization are reviewed in Section 2, from the perspective of performances and human operation. Next, array processing approaches are expounded in Section 3, with a focus on the specificities raised by the robotics constraints.

2. Binaural approaches to sound source localization in robotics

This section describes a first set of methods which try to mimic diverse aspects of the human auditory system, thus defining the topic of *binaural robot audition*. The common point to all the following works is the use of only two microphones, generally positioned inside a human-like pinna. There is an obvious practical interest to develop biomimetic auditory sensors containing a small number of microphones: the size is minimal and the embedded electronics is simplified. Significant advances in understanding the biological processes which enable the handling of acoustic data by humans have been obtained up to the 1980s (Middlebrooks and Green, 1991). They constitute the natural basis of binaural developments in robotics. Having this in mind, the successive steps of the sound localization process can be described by following the sound propagation, from the source to the binaural sensor:

- As a first step, a sound wave generated by an external source is modified by the presence of the robotic torso, head and pinnae prior to interact with the ears. The induced scattering and spectral changes must be modeled so as to precisely capture the time-space relationship linking the sound source to the binaural signals. From a engineering point of view, this relationship is captured by the so-called *Head Related Transfer Function* (HRTF), which will be studied in the first subsection.
- Next, human localization capabilities mainly rely on some acoustic features extracted by our ears and integrated by our brain. Those features have been extensively studied; among them, one can cite *Interaural Cues* for horizontal localization, or spectral notches for vertical localization (Middlebrooks and Green, 1991). A lot of them have also been used in robotics. These will be reviewed in the second subsection.
- Finally, on the basis of these features, sound localization itself is performed. Numerous approaches have been proposed so far, and the most prominent one in robotics are outlined in the third subsection.

In all the following, the left and right microphone signals will be referred to as $l(t)$ and $r(t)$ respectively, with t the time (in s). Their frequency counterparts, obtained through a Fourier transform, will be denoted by $L(f)$ and $R(f)$ respectively, with f the frequency (in Hz). Sound source position is expressed in terms of horizontal azimuth angle θ, elevation angle φ in the median plane (in radian), and distance r, all of them being expressed w.r.t. an origin located at the robot's head center. In the remaining of the paper, the position $(\theta, \varphi) = (\pi/2, 0)$ corresponds to a sound source in front of the head (i.e. at boresight, see Fig. 2).

2.1. The head related transfer function

2.1.1. Definition

The HRTF captures the relationships between the signal $s(t)$ originating from a sound source and captured at a certain arbitrary reference position in space and the two signals perceived by the two ears. These relationships can be written in the form

$$\begin{cases} L(f) = H_L(r_s, \theta_s, \varphi_s, f)S(f), \\ R(f) = H_R(r_s, \theta_s, \varphi_s, f)S(f), \end{cases} \tag{1}$$

where $H_L(\cdot)$ and $H_R(\cdot)$ represent the left and right HRTFs respectively, $(r_s, \theta_s, \varphi_s)$ is the actual sound source position w.r.t. the chosen reference position, and $S(f)$ the Fourier transform of $s(t)$. Importantly, the HRTFs account for all the modifications brought by the body of the listener, including torso, head and pinnas effects, to the incident sound wave. So it varies significantly from a human listener to another as its reflects his/her own morphology. The same applies in robotics, where all the possible acoustic scatters impact on the sensed signals, and is thus captured by the corresponding HRTFs. But it is fundamental to understand that the HRTF strictly corresponds to propagation in free field and does not include room reflections or reverberations. Consequently, the HRTF can be obtained in two ways. The first solution is to accurately model the body and head effects. If the robot shapes are simple, then basic acoustic equations can be sufficient

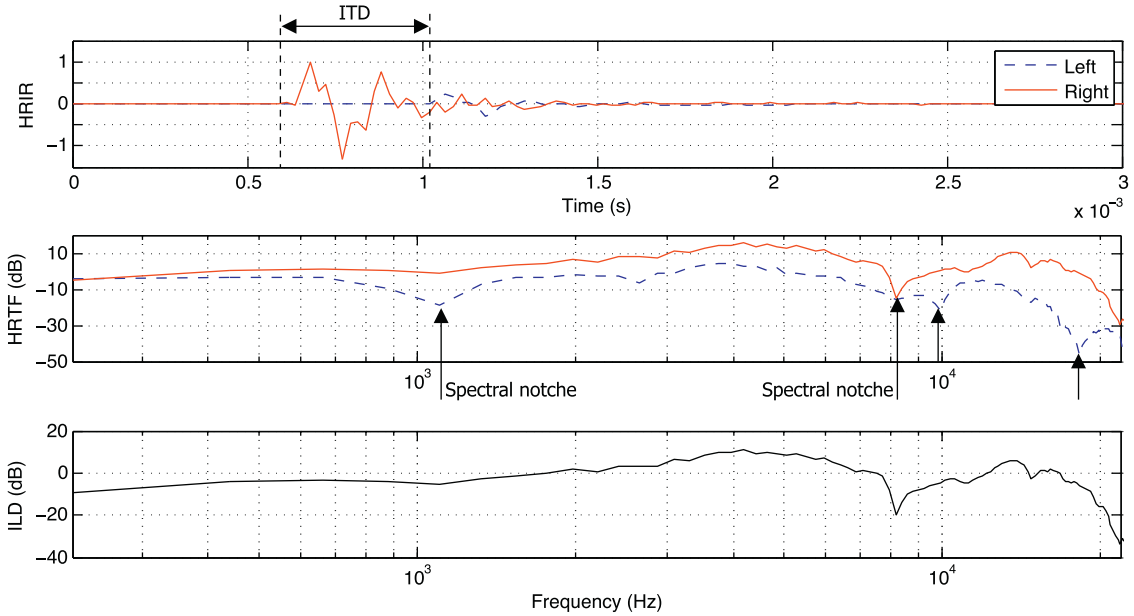


Fig. 1. HRIR and HRTF data for a subject of the CIPIC database (Algazi et al., 2001). Interaural time difference (ITD), interaural level difference (ILD) and monaural (spectral notches) cues are also reported (see Section 2.2.1 for their definitions).

to account for the acoustic effect on the signals. In the case of a more realistic robot, with complex body, shoulders, nose, pinnae, etc., an acoustic simulation software might be necessary to solve the problem through finite-element methods (Otani and Ise, 2006). The second solution consists in identifying the HRTF through real measurements, which must be performed in an anechoic room. This solution may not be so practical for every robotic platform, for it requires specific hardware/software. Consequently, HRTF identification is sometimes performed in the room where the robot will be located, resulting in measurements mixing together the head effect with the room acoustic. Additionally, various HRTF databases are proposed in the literature. One can cite, among others, the celebrated CIPIC database (Algazi et al., 2001), published by the CIPIC Interface Laboratory from the University of California Davis, and accessible from the website http://interface.cipic.ucdavis.edu/, or the recent HRTF database proposed by the Deutsche Telekom Laboratories (TU-Berlin) (Wierstorf et al., 2011), located at https://dev.qu.tu-berlin.de/projects/measurements/wiki. Note that those databases generally include HRTF measurements for human subjects, but also for some Head And Torso Simulators (HATS), like the KEMAR by G.R.A.S.. To the authors knowledge, no robotic HRTF databases have been proposed so far, but the $\mathcal{HARK}$ software framework (see Section 4) proposes a systemic approach for measuring them.

Typical HRTFs extracted from the CIPIC database is shown in Fig. 1, for the azimuth $\theta = 35^\circ$ and elevation $\phi = 20^\circ$.

Its time counterpart head related impulse response (HRIR) is also represented. This figure highlights the shadowing effect of the head: depending on the source position, the left and right signals differ in terms of time of arrival (cf. the delay between the left and right HRIR), but also in terms of spectral content (cf. the amplitude difference between the two HRTFs and the spectral notches positions). These last cues will often serve as the basis to infer localization (see Section 2.2). Readers interested in a more complete tutorial on HRTF can refer to Cheng and Wakefield (2001), where experimental and theoretical data are compared.

2.1.2. HRTF models in robotics

As already outlined, the complex structure of most robotic platforms prevents the access (through simulations or identifications) to the exact robot HRTFs at the two ears. Consequently, some simple head models have been proposed by the robot audition community, with the aim to capture the robot head effect on the binaural signals up to some extent. The three most classical models are depicted on Fig. 2. They consist in considering the left and right microphones in the free field—Auditory Epipolar Geometry (AEG) (Nakadai et al., 2000), placed at the surface of a disk—Revised Epipolar Geometry (RAEG) (Nakadai et al., 2001), or at the surface of a sphere—Scattering Theory (ST) (Nakadai

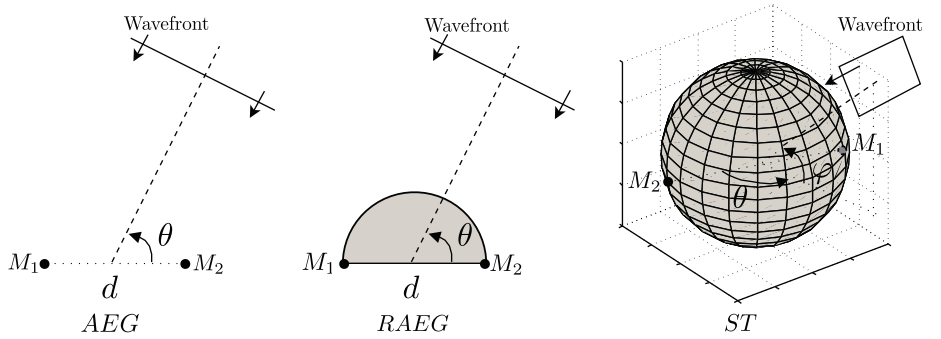


Fig. 2. The three classical head models: auditory epipolar geometry (AEG, left), revised auditory epipolar geometry (RAEG, middle), and scattering theory (ST, right).

et al., 2003). AEG and RAEG are the most elementary model. Provided that θ and f stand for the azimuth and frequency of a farfield source (i.e. a source which is far enough to consider planar wavefronts in the vicinity of the robot head, see Section 3.1.1), the (R)AEG approximations of the left and right HRTFs $H_L^{(\mathrm{R})\mathrm{AEG}}(\cdot)$ and $H_R^{(\mathrm{R})\mathrm{AEG}}(\cdot)$ write as

$$\begin{cases} H_L^{(\mathrm{R})\mathrm{AEG}}(\theta, f) = 1, \\ H_R^{(\mathrm{R})\mathrm{AEG}}(\theta, f) = e^{-j\phi(\theta)} = e^{-j2\pi f \tau_{(\mathrm{R})\mathrm{AEG}}(\theta)}, \end{cases} \tag{2}$$

highlighting the fact that the two binaural signals only differ by a delay $\tau_{(\mathrm{R})\mathrm{AEG}}(\theta)$ which is a function of the source angle (note that the left channel has been arbitrarily considered here as the reference, i.e. $H_L^{(\mathrm{R})\mathrm{AEG}}(\cdot)$ and $H_R^{(\mathrm{R})\mathrm{AEG}}(\cdot)$ are defined up to a transfer function $P(f)$ modeling the propagation from the source to the left sensor). The third ST (scattering) model is more involved. Let β be the so-called incidence angle, i.e. the angle between the line from the center of the sphere approximating the head to the source position (r_s, θ_s), and the line from the center of the same head to a measurement point at which the HRTF must be computed. Considering a perfect rigid spherical head, the expression of the diffracted sound pressure wave received at the measurement point allows to write (Duda and Martens, 1998):

$$H^{\mathrm{ST}}(r, \beta, f) = \frac{rce^{-jr2\pi f/c}}{ja^2 2\pi f} \sum_{m=0}^{\infty} (2m+1) P_m\left[\cos(\beta)\right] \frac{h_m(r2\pi f/c)}{h'_m(a2\pi f/c)}, \tag{3}$$

where $H^{\mathrm{ST}}(r, \beta, f)$ is the transfer function linking the sound pressure received at the measurement point and the free-field pressure existing at the head center, with c the speed of sound and a the head radius. $P_m(\cdot)$ and $h_m(\cdot)$ are the Legendre polynomial of degree m and the mth-order spherical Hankel functions respectively, while $h'_m(\cdot)$ denotes the derivative of the function $h_m(\cdot)$. Assuming that the two microphones are antipodally placed on the surface of the sphere, the left and right HRTFs, respectively denoted by $H_L^{\mathrm{ST}}s(r, \theta, f)$ and $H_R^{\mathrm{ST}}(r, \theta, f)$, are then given by, for a sound source located at (r, θ),

$$\begin{cases} H_L^{\mathrm{ST}}(r, \theta, f) = H^{\mathrm{ST}}\left(r, -\frac{\pi}{2} - \theta, f\right), \\ H_R^{\mathrm{ST}}(r, \theta, f) = H^{\mathrm{ST}}\left(r, \frac{\pi}{2} - \theta, f\right). \end{cases} \tag{4}$$

2.2. Binaural and monaural cues for localization

Once the link between the sound source signal to be localized and the two resulting binaural signals has been modeled, it is necessary to focus on the binaural features which can be extracted from these signals to infer localization. These features are first recalled through a short review on sound source localization in humans. Then, the way how these cues can be coupled with the aforementioned HRTFs is investigated.

2.2.1. Sound source localization in humans

About 100 years ago, Rayleigh proposed the *duplex theory* (Rayleigh, 1907), which explains that horizontal localization is mainly performed through two primary auditory cues, namely the *Interaural Level Difference* (ILD) and the *Interaural Time Difference* (ITD). The ILD relates to the difference between the intensity of signals perceived by the right and left ears, due to the head frequency-dependent scattering. Noticeably, if a source emits at a frequency higher than about 750 Hz, then the head and any small-sized element of the face induce scattering, which significantly modifies the perceived acoustic levels, so that the ILD can exceed 30 dB. On the contrary, the ILD is close to 0 dB at low frequencies, as fields whose wavelengths are greater than the head diameter undergo no scattering. This property can clearly been deduced from Fig. 1, where the left and right HRTF amplitudes only significantly differ for frequencies greater than about 800 Hz. The second auditory cue is known as the *Interaural Phase Difference (IPD)*—or its time-counterpart termed *Interaural Time Difference (ITD)*. The ITD is justified by the path difference to be traveled by the wave to reach the ears. It appears on Fig. 1 as a delay between the two HRIR onsets. Note however that the maximum value involved in localization is around 700 μs—i.e. one period of a 1400 Hz sound—due to the ambiguity of IPD values greater than 2π. So, two frequency domains can be exhibited in human horizontal localization, each one involving a distinct acoustic cue. Frequencies under ∼1 kHz are azimuthally localized by means of the IPD, while frequencies above ∼3 kHz exploit the ILD.

It can be straightly inferred that a source emitting from the vertical symmetry plane of the head produces no interaural difference. However, humans are still able to perform localization in such conditions. Indeed, obstacles—including shoulders, head, outer ear, etc.—play the role of scatters which modify the frequency components of acoustic waves. This filtering effect is essential in our ability to determine the elevation of a sound source. Indeed, the sum of all the reflections occurring around the head induces notches into the perceived spectrum, the positions of which are significantly affected by the source elevation (Humanski and Butler, 1988), see Fig. 1. The acoustic feature for vertical localization is thus a spectral cue, termed "*monaural*" as it involves no comparison between the signals perceived at the two ears. Consequently, these notches positions are likely to be used by the brain to infer the elevation. Note that it is not possible, even for humans, to make any difference between a notch caused by the head/outer ear sound diffraction and a notch already present in the source spectrum. Indeed, monaural cues appears to be highly dependent on the source properties, and listening tests demonstrated that humans are more affected by the frequency contents emitted by the sound source than by its true location (Butler and Helwig, 1983).

While the directional aspects of localization have been widely studied, distance perception has received substantially less attention. Generally, it is admitted that like angular estimations, sound source distance can also be inferred from various acoustical properties of the sounds reaching the two ears. Known distance dependent acoustical cues include sound intensity, which unfortunately also depends on the intrinsic source energy, as well as on interaural differences, on the spectral shape and on the Direct-to-Reverberant sound energy Ratio (DRR). So, one can see that nearly all the aforementioned cues, which are used to estimate the angular position of a sound source, are also directly linked to the distance parameter. Actually, human most likely combine these distant-dependent cues together with *a priori* information on the surrounding space and the source of interest so as to get the sensation of a stable distance. The reader interested in this topic will find a comprehensive review in (Zahorik et al., 2005). But one has to keep in mind that human performances in distance discrimination are quite poor. Even under ideal acoustic conditions, the estimated distance appears to be a biased estimate of the actual one, and listening tests have also proven that humans use to significantly overestimate the distance to sources closer than 1 m, while they underestimate distances greater than 1 m (Zahorik et al., 2005).

2.2.2. Auditory models in robotics

As shown in the previous paragraph, the head effect on the perceived sounds is frequency dependent. Such a frequency decomposition of the signals is often implemented with a FFT, while other authors proposed the use of classical bandpass filters, see (Youssef et al., 2012). Nevertheless, how efficient they may be, these methods do not perform a frequency decomposition similar to the human inner ear. *Gammatone filters*, modeling the vibration of the basilar membrane inside the cochlea, have proven well suited to describe the cochlear impulse response in the cat's cochlea (Johannesma, 1972). They also constitute a good approximation of human spectral analysis at moderate sound levels (Patterson et al., 1995). Typical gammatone filters frequency responses are reported in Fig. 3. It can be seen that their bandwidths increase with frequency, in such a way that they represent around 15% of the center frequencies. This is one of the main features of the human auditory system, which confirms our better ability to discriminate low

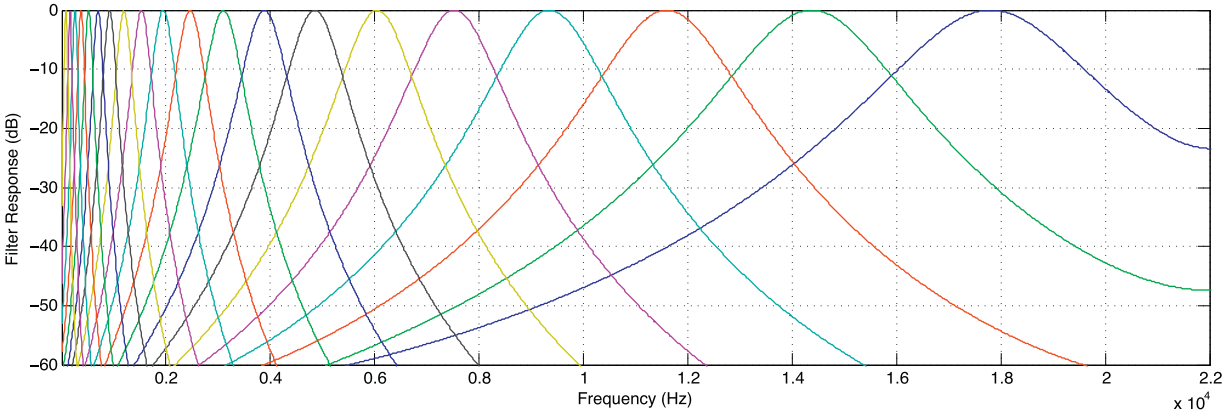


Fig. 3. Typical Gammatone filters frequency responses.

frequencies (Stevens et al., 1937). As a consequence of this decomposition, the representation of any sound signal information is more likely close to the human perception of sounds. As it will be shown in the following, this gammatone frequency decomposition is now very commonly used. Readers interested in more involved auditory models can refer to the Auditory Modeling Toolbox (Søndergaard and Majdak, 2013), available at http://amtoolbox.sourceforge.net.

2.2.3. *Binaural cues for horizontal localization in robotics*

Among all the acoustic features that can be extracted with two microphones, binaural cues are the most often used in robotics. The IPD and/or ILD can indeed lead, with just two microphones and simple computations, to a localization in azimuth. Let $\mathrm{IPD}_{\mathrm{exp}}(f)$ and $\mathrm{ILD}_{\mathrm{exp}}(f)$ term the experimental IPD and ILD extracted from the two signals. There exist numerous ways to estimate these IPD and ILD values: computations in the time or frequency domain, bioinspired models, etc. Readers interested in a review of these approaches can consult (Youssef et al., 2012). Whatever the approach, from these experimental values, the problem is then to determine the position of the emitting sources. This involves a model, expressed either in a mathematical closed-form or as experimental relationships uniting the source attributes (position, frequencies, etc.) and the induced auditory cues.

Considering the AEG or RAEG model represented by Eq. (2), the delay $\tau_{(\mathrm{R})\mathrm{AEG}}(\theta)$ represents the ITD, and its phase counterpart, i.e. the IPD, then verifies (Nakadai et al., 2001)

$$\begin{cases} \mathrm{IPD}_{\mathrm{AEG}}(\theta, f) = 2\pi f \tau_{\mathrm{AEG}}(\theta) = \dfrac{2\pi f a}{c}\cos\theta, & \text{(a)} \\ \mathrm{IPD}_{\mathrm{RAEG}}(\theta, f) = 2\pi f \tau_{\mathrm{RAEG}}(\theta) = \dfrac{\pi f a}{c}\left(\left(\dfrac{\pi}{2} - \theta\right) + \cos\theta\right) & \text{(b)} \end{cases} \tag{5}$$

The main advantage of the first AEG formulation is that the azimuth θ can straightly be approximated by inverting (5a), assimilating $\mathrm{IPD}_{\mathrm{AEG}}(\theta, f)$ to the experimental $\mathrm{IPD}_{\mathrm{exp}}$. However, as already outlined, it cannot describe the effect of a head located between the two microphones, inducing scattering of the sound wave. To better take into account its presence, the RAEG model can be used (note that RAEG is analog to the classical Woodworth–Schlosberg formalization (Woodworth and Schlosberg, 1962)). Indeed, results from Nakadai et al. (2001) show that simulations obtained from this model fit experimental measurements in an anechoic room within the frequency band [500–800 Hz]. But the RAEG model, while being more realistic than AEG, does not fully account for the head waveguide effect. Additionally, both do not provide any meaningful value for the ILD cue, which is accounted for in the theoretical spherical model. Indeed, one has for this ST model

$$\begin{cases} \mathrm{IPD}_{\mathrm{ST}}(r, \theta, f) = arg\left(H_L^{\mathrm{ST}}(r, \theta, f)\right) - arg\left(H_R^{\mathrm{ST}}(r, \theta, f)\right) & \text{(a)} \\ \mathrm{ILD}_{\mathrm{ST}}(r, \theta, f) = 20\log_{10}\dfrac{|H_L^{\mathrm{ST}}(r, \theta, f)|}{|H_R^{\mathrm{ST}}(r, \theta, f)|} & \text{(b).} \end{cases} \tag{6}$$

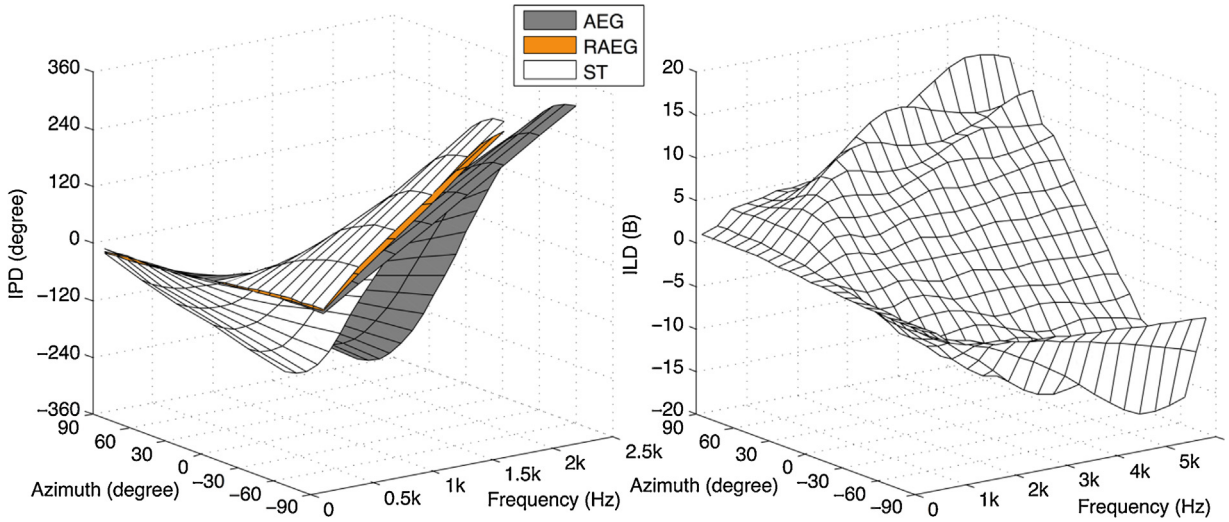


Fig. 4. Comparison of the AEG, RAEG and ST models. (Left) IPD as a function of azimuth θ and frequency. (Right) ILD values for the ST model.

Compared to epipolar geometries, the scattering theory exhibits more reliable theoretical expressions of both IPD and ILD as functions of the azimuth θ and distance r, and can thus lead to more reliable identification schemes for localization. It is however important to notice that the accuracy of the approach still depends on the capacity to cope with the room acoustics, which is not always possible in practice. Indeed, if the binaural cues are obtained inside in a real robotics environment including noisy sound sources and reverberation, the results may get very bad: since the models do not capture the distortion due to the room acoustics, the theoretical binaural cues and the measured one cannot fit with each. Nevertheless, the ST formalism was exploited in (Nakadai et al., 2003; Handzel and Krishnaprasad, 2002) to express the pressure field perceived by two microphones symmetrically laid over a sphere, and experimentally tested by Handzel et al. (2003) on a spherical plastic head. But whatever the model, and as outlined in human audition, the observed inappropriateness of IPD (resp. ILD) for high (resp. low) frequencies extends outside the scope of AEG, RAEG and ST strategies, as can be seen in Fig. 4. As mentioned in Section 2.2.1, frequencies above about 1400 Hz lead to IPD values greater than 2π and becomes ambiguous. Noticeably, the human auditory system then relies on the ILD at these frequencies. Indeed Fig. 4 exhibits high ILD values, reaching up to 25 dB for this frequency domain, in the ST model.

2.3. Exploitation in robotics

Historically, most initial contributions to robot audition were rooted into the binaural paradigm. However, as shown in the following, the results remained mixed when facing real-life environments, involving noises and reverberations together with wideband non-stationary sources.

2.3.1. Horizontal localization

In the early 2000s, the use of interaural difference functions for azimuth localization was deeply studied in the framework of the SIG project (Nakadai et al., 2000, 2002). As the robot cover cannot be perfectly isolated from internal sounds, an adaptive filter exploiting the data provided by inner microphones was used to eliminate the motor noises (e.g. ego-noise mentioned in Section 1) from the audio signal perceived by the external pair of microphones. This active auditory system thus allowed to perform measurement during motion (Nakadai et al., 2000), and constituted an interesting improvement over former methods—for instance (Brooks et al., 1999) on the COG humanoid, or (Huang et al., 1997)—based on the *stop-perceive-act* principle. On this basis, an "Active Direction Pass Filter" (ADPF) grounded on the ST model was proposed in (Nakadai et al., 2002) to determine the origin of a sound source and extract it out of a mixture of surrounding sounds. This *model-matching* approach has since been used in a lot of contributions. For a fixed distance r_s, for all frequencies under (resp. above) $f_{\text{th}} = 1500$ Hz, and for each θ, the system computes the theoretical $\text{IPD}_{\text{ST}}(r_s, \theta, f)$ (resp. $\text{ILD}_{\text{ST}}(r_s, \theta, f)$) through (6). Then a cost function d_{IPD} (resp. d_{ILD}) is defined to measure the distance between the measured $\text{ITD}_{\text{exp}}(f)$ (resp. $\text{ILD}_{\text{exp}}(f)$) interaural differences and theoretical ones. The

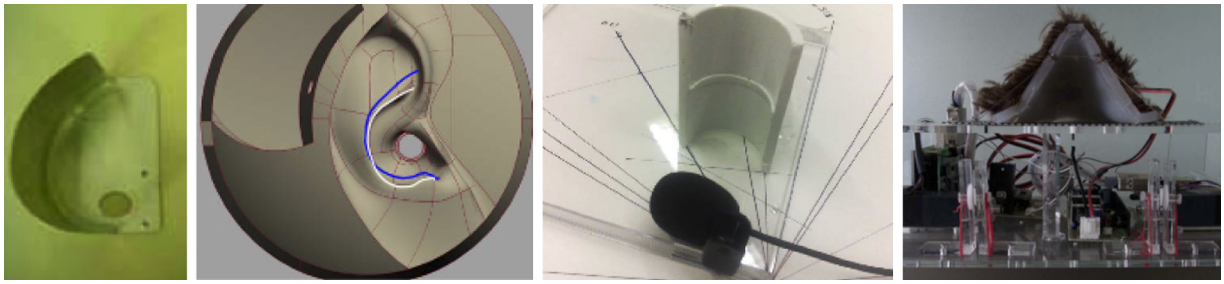


Fig. 5. Some artificial pinnas from the literature. Pictures extracted from (left to right): Kumon et al. (2005), Rodemann et al. (2008), Saxena and Ng (2009), Kumon and Noda (2011). They all share the fundamental asymmetry property.

two distances d_{IPD} and d_{ILD} are then integrated into a belief factor $P_{\text{IPD+ILD}}(\theta)$. The angle $\widehat{\theta}_s = arg\max_{\theta} P_{\text{IPD+ILD}}(\theta)$ is then regarded as the sound source azimuth. The sound source separation performances were also evaluated and compared in Nakadai et al. (2002), depending on the model (RAEG *vs* ST) that relates the source azimuth and the interaural cues. Clearly, the scattering theory provided the best results. So far, these contributions have been among the rare complete binaural audition systems, integrating localization, source separation and recognition. Nevertheless, since HRTFs do not capture the acoustic response of the room where the robot operates, their applicability is generally limited to well-identified environments. One solution could consist in learning the head effect in realistic conditions. Such an idea was successfully assessed in Youssef et al. (2013) through a dedicated neural network able to generalize learning to new acoustic conditions. One can also cite (Berglund and Sitte, 2005), or (Hornstein et al., 2006), where the iCub humanoid robot's head was endowed with two pinnae. The localization is performed by mapping the aforementioned sound features to the corresponding location of the source through a learning method. Another approach is proposed in Kim et al. (2008). Auditory events corresponding to relevant ITD values are gathered into histograms, which are then approximated by Gaussian models whose parameters are identified through the EM method. Peaks in the resulting histogram are then regarded as potential sound azimuths. This allows coping with the multisource case, where multiple sound sources are likely active at the same time. Finally, an implementation of the biology-inspired Jeffress model is proposed in Liu et al. (2008) on a simple robot head endowed with two microphones and stereovision. Interestingly, the ILD pathway is also modeled with a 2D spiking map. The merging of the two interaural maps is also addressed, so as to obtain an efficient sound localization system. The proposed method is shown to share some common well-known properties of the human auditive system, like the ITD maximal efficiency reached when the sound source is in front of the observer. But whatever the approach, ITDs and ILDs can be extracted from the binaural signals in numerous ways: through correlation (shik Kim and Choi, 2009), zero-crossing times comparison (Rodemann et al., 2008), or in the spectral domain (Cavaco and Hallam, 1999). A systematic study of binaural cues, and an analysis of their robustness w.r.t. reverberations is proposed in Youssef et al. (2012). Results show that binaural cues extracted from gammatone filters outperforms other techniques.

2.3.2. *Vertical localization: spectral cues*

As indicated in Section 2.2.1, the elevation of a sound source is mainly related to the positions of notches in the spectra of the perceived signals, which stem from acoustic reflections due to the head and the outer ear. In robotics, quite few authors have developed techniques based on spectral cues. Most of them are based on the scattering induced by an artificial pinnae in charge of collecting the acoustic pressure information and of driving it to microphones. For humans, the specific shape of the pinnae enables a selective spatial amplification of acoustic pressure variations, with a quality factor reaching up to 20 dB. Reproducing such capabilities in robotics is a difficult problem due to the lack of a model of the pinnae shapes which lead to elevation dependent notches. Yet, as a rule of thumb, these shapes must be irregular or asymmetric, and artificial pinnae were proposed in Kumon et al. (2005), Shimoda et al. (2006), Hornstein et al. (2006), Rodemann et al. (2008), or Saxena and Ng (2009).

Fig. 5 shows some of them. A simplified model, inspired by Hebrank and Wright (1974) and based on the superposition of the incident wave with a single wave reflected by the pinnae, enables the prediction of the elevation from the position of notches. Noticeably, these notches, which appear or disappear depending on the elevation, may be hard to detect or may even be blurred by spurious notches induced by destructive interferences coming from acoustic

reflections on obstacles. To solve this problem, Hornstein et al. (2006) introduces the *interaural spectral difference* as the ratio between the left and right channel spectra. While notches may be indistinct in the complex spectra of the two signals, the interaural spectral difference, when interpolated with a 12-degree polynomial, can enable the extraction of their frequency positions. Another solution is proposed in Rodemann et al. (2008). It consists in computing the difference of the left and right energies coming from 100 frequency channels, ranging from 100 Hz to 20 kHz. Strictly speaking, this approach does not involve monaural cues anymore, but allows to obtain spectral cues which are said less sensitive to the source signal frequency content. Concerning the design of the pinnae, a model including a more extensive description of the reflected and diffracted sound waves is proposed in Lopez-Poveda and Meddis (1996). Though it leads to new theoretical expressions of the spectra, it remains hardly valuable for the design of artificial outer ears. Saxena and Ng (2009) also exhibits four different pinnae together with their induced frequency responses for an original work on sound localization from a single microphone. On the other hand, inspired by animals that are able to change the configuration of their pinnae, Kumon and Noda (2011) proposed an *active* ear, which is able to modify its shape to encode elevation (and azimuth).

2.3.3. Distance localization

In the topic of robot audition, distance estimation has been so far based on the triangulation idea. For instance, on the basis of the estimation $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of the azimuth at two distinct positions, triangulation allows to estimate the distance between the robot and the sound source, together with the source azimuth. Generally, this is only possible if the sound source is static in the environment. Some recent works by Markovic et al. (2013) and Portello et al. (2012) proposed a filtering strategy to cope with a possibly moving source. The algorithm mainly relies on ITD to provide an estimation of the source position (r, θ) during the movement of a binaural sensor. Distance estimation is also investigated in Rodemann (2010). Several auditory cues, like interaural differences, sound amplitude and spectral characteristics are compared. Convincing results are shown, exhibiting an estimation error lower than 1 m for a 6 m-far sound source. But the author outlines that its study does not capture the full variability of natural environments. Recent contributions also propose to estimate the DRR. Indeed, it has been shown that distance estimation by humans is more accurate in a reverberant space than in an anechoic one. This estimation is not straightforward: Lu and Cooke (2010) proposes a binaural equalization-cancellation technique, while Vesa (2009) hypotheses the use of the frequency dependent magnitude squared coherence between the left and right signals.

2.4. Conclusion

Using very few microphones, interesting developments have been proposed by the robotics community to provide the robots with a first ability to localize sound sources in their environment. The results are however contrasted. Reproducing the auditory faculty of the human ear is a very difficult problem. First, the exploitation of interaural cues requires a very precise modeling of the perturbations induced by the presence of the head. Second, binaural cues are still very hard to exploit. In any case, an accurate model of the propagation turns out to be essential to finely describe the evolution of auditory cues. Furthermore, all these techniques appear to be very sensitive to variations of the acoustic environment. Most models have been experimentally validated in an anechoic room but cannot be used to accurately localize sounds in real conditions, unless a precise description of the robot's environment is given. But recent *active* variations of the existing algorithms have recently benefited from the additional information brought by the robot motion, thus renewing the interest in binaural approaches to sound localization, see Section 4. Nevertheless, all these difficulties have motivated the robotics community to also envisage localization methods based on a higher number of microphones, possibly benefiting from existing signal processing advances. An overview of these techniques is presented in the next section.

3. Array processing approaches to localization in robot audition

This section deals with the second paradigm mainly used in robot audition: microphone arrays. Contrarily to binaural approaches, where only two microphones are used, array processing relies on multiple microphones, spatially organized along various geometries (such as a line, a circle, a sphere, or the vertices of a cube). Thanks to the redundancy in the signals from the various channels, the acoustic analysis performance and/or robustness can be improved (Van Trees, 2002). Multiple contributions have been proposed in a robotics context, generally concerning source detection and

localization, source separation, and speaker/speech recognition. The reader interested specifically in those topics will find comprehensive reviews in Benesty et al. (2008) and Woelfel and McDonough (2009). But again, this section is entirely devoted to sound source localization, and is focused only on methods used in robotics so far.

In a first subsection, the theoretical aspects of array processing techniques are presented. After a short introduction of the notations and definitions, the basis of the MUSIC method (MUltiple SIgnal Classification), of correlation-based approaches and of beamforming algorithms are carefully depicted. Each technique is momentarily placed out of the scope of robotics, and only a theoretical description of it and of its properties is proposed. Next, in a second subsection, each approach is considered again, but under the angle of its application in robotics: MUSIC, though very powerful, exemplifies the limits imposed by the real time constraint; correlation-based techniques illustrate how information redundancy can enhance the localization accuracy and robustness; beamforming-based methods simplicity makes them ideal candidates for an application in robotics. Thus, the limits of each method in such a robotics context are discussed, and recent contributions in Robot Audition overcoming these limits are highlighted. Finally, a discussion on the constraints brought by the robotic context and on their consequences on the algorithms is proposed in a last subsection.

3.1. Theoretical aspects of array processing in robotics

3.1.1. Notations and definitions

Consider S pointwise sound sources emitting at locations referenced by $\mathbf{r}_s^{\mathrm{s}} = (r_s, \theta_s, \varphi_s)$, $s = 1, \ldots, S$, in a spherical coordinates system. In the following, any monochromatic space-time signal reads as $y(\mathbf{r}, t) = Y(\mathbf{r}, k)e^{jkct}$, with $k = 2\pi f/c$ the wavenumber. In addition, let a microphone array be composed of N identical omnidirectional microphones placed at locations $\mathbf{r}_n^{\mathrm{m}}$, $n = 1, \ldots, N$. Then, the sound signal $m_n(t)$ issued by the S sources and perceived by the nth microphone can be written as

$$m_n(t) = \sum_{s=1}^{S} \frac{\|\mathbf{r}_s^{\mathrm{s}}\|}{\|\mathbf{r}_n^{\mathrm{m}} - \mathbf{r}_s^{\mathrm{s}}\|} \quad s_s^0(t - \frac{\|\mathbf{r}_n^{\mathrm{m}} - \mathbf{r}_s^{\mathrm{s}}\|}{c} + \frac{\|\mathbf{r}_s^{\mathrm{s}}\|}{c}) + b_n(t), \tag{7}$$

where $s_s^0(t)$ terms the fictitious signal perceived at $\mathbf{r} = \mathbf{0}$ and stemming from the single sth source, and the additive noise $b_n(t)$ accounts for parasitic sources in the environment as well as electronic noise in the microphones outputs. So, the Fourier transforms $M_n(k)$, $S_s^0(k)$, $B_n(k)$ of $m_n(t)$, $s_s^0(t)$, $b_n(t)$ satisfy

$$M_n(k) = \sum_{s=1}^{S} V_n(\mathbf{r}_s^{\mathrm{s}}, k) S_s^0(k) + B_n(k), \tag{8}$$

with

$$V_n(\mathbf{r}, k) = \|\mathbf{r}\| \quad e^{jk\|\mathbf{r}\|} \quad \frac{e^{-jk\|\mathbf{r}_n^{\mathrm{m}} - \mathbf{r}\|}}{\|\mathbf{r}_n^{\mathrm{m}} - \mathbf{r}\|} \tag{9}$$

the nth entry of the *array*—or *steering*—*vector* $\mathbf{V}(\mathbf{r}, k) \triangleq (V_1(\mathbf{r}, k), \ldots, V_N(\mathbf{r}, k))^T$. Defining the *source*, *observation* and *noise vectors* by $\mathbf{S}^0(k) \triangleq (S_1^0(k), \ldots, S_S^0(k))^T$, $\mathbf{M}(k) \triangleq (M_1(k), \ldots, M_N(k))^T$ and $\mathbf{B}(k) \triangleq (B_1(k), \ldots, B_N(k))^T$ respectively, (8) can be turned into the matrix form

$$\mathbf{M}(k) = \mathcal{V}(\mathbf{r}_1^{\mathrm{s}}, \ldots, \mathbf{r}_S^{\mathrm{s}}, k)\mathbf{S}^0(k) + \mathbf{B}(k), \tag{10}$$

with $\mathcal{V}(\mathbf{r}_1^{\mathrm{s}}, \ldots, \mathbf{r}_S^{\mathrm{s}}, k) \triangleq (\mathbf{V}(\mathbf{r}_1^{\mathrm{s}}, k)|\ldots|\mathbf{V}(\mathbf{r}_S^{\mathrm{s}}, k))$ the *array matrix*. Note that (9) can significantly be simplified when the distance to the sources tends to infinity, as the wavefronts become planar. This simplification defines the "farfield hypothesis". In the following, quantities related to farfield will be superscripted by the symbol ∞, so that the *farfield array vector* writes as $\mathbf{V}^{\infty}(\theta, \varphi, k) \triangleq (V_1^{\infty}(\theta, \varphi, k), \ldots, V_N^{\infty}(\theta, \varphi, k))^T$, with $V_n^{\infty}(\theta, \varphi, k) = V_n^{\infty}(\mathbf{r}, k) \triangleq \lim_{r \to \infty} V_n(\mathbf{r}, k)$.

From now on, lets consider a linear microphone array, constituted of N microphones located at $z_1, \ldots, z_N$ along the $\mathcal{Z}$-axis. Consequently, because of the rotational symmetry of the problem, all characteristics are invariant w.r.t. the

elevation φ, so that the location vector $\mathbf{r} = (r, \theta, \varphi)$ reduces to $\mathbf{r} = (r, \theta)$. In addition, the nth entry (9) of the array vector becomes

$$V_n(\mathbf{r}, k) = V_n(r, \theta, k) = \frac{r}{\sqrt{r^2 + z_n^2 - 2rz_n \cos\theta}} e^{jkr} e^{-jk\sqrt{r^2+z_n^2-2rz_n\cos\theta}}. \tag{11}$$

In the farfield, (11) particularizes into the well-known expression $V_n^{\infty}(\theta, k) = e^{-jkz_n \cos\theta}$.

3.1.2. The MUSIC method

The MUSIC (MUltiple SIgnal Classification) method, initially proposed in Schmidt (1986), belongs to the so-called "high resolution" approaches because of the sharpness of the conclusions it provides. It is so far one of the most used algorithm in robotics. The pointwise sound sources to be localized are assumed independent, zero-mean stationary, of single frequency k_0, and in number $S < N$. In Eq. (10), the additive noise is assumed zero-mean, stationary, temporally and spatially white, of known equal power on each microphone, and independent of the sources. So, denoting by $\mathcal{I}$ and $\mathcal{O}$ the identity and zero matrices, and $E[\cdot]$ the expectation operator, it is supposed that

$$\mathbf{\Gamma_B} = E[\mathbf{B}\mathbf{B}^H] = \sigma_{\mathcal{N}}^2 \mathcal{I}_{N\times N} \quad \text{and} \quad E[\mathbf{S}^0\mathbf{B}^H] = \mathcal{O}_{S\times N}. \tag{12}$$

As has just been done, the dependencies of variables upon the single involved wavenumber k_0 will be temporarily omitted. MUSIC determines the sources number S together with their ranges and azimuths from the eigen decomposition of the covariance—or interspectral—$N \times N$ matrix $\mathbf{\Gamma_M} = E[\mathbf{M}\mathbf{M}^H]$ relative to the signals perceived at the array.

$$\mathbf{\Gamma_M} = (\mathcal{U}_{\mathcal{S}}|\mathcal{U}_{\mathcal{N}}) \begin{pmatrix} \lambda_1 + \sigma_{\mathcal{N}}^2 & & \mathcal{O} & | & \\ & \ddots & & | & \mathcal{O} \\ & & \lambda_S + \sigma_{\mathcal{N}}^2 & | & \\ -- & -- & -- & & -- \\ & \mathcal{O} & & | & \sigma_{\mathcal{N}}^2 \mathcal{I}_{N-S} \end{pmatrix} (\mathcal{U}_{\mathcal{S}}|\mathcal{U}_{\mathcal{N}})^H, \tag{13}$$

where the real $\lambda_1, \ldots, \lambda_S$ are sorted increasingly $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_S > 0$, $\mathcal{U}_{\mathcal{S}} = (\mathbf{U}_1|\ldots|\mathbf{U}_S) \in \mathbb{C}^{N\times S}$ and $\mathcal{U}_{\mathcal{N}} = (\mathbf{U}_{S+1}|\ldots|\mathbf{U}_N) \in \mathbb{C}^{N\times(N-S)}$. The right eigenvectors $\mathbf{U}_1, \ldots, \mathbf{U}_S$ related to the S greatest eigenvalues $\lambda_1 + \sigma_{\mathcal{N}}^2, \ldots, \lambda_S + \sigma_{\mathcal{N}}^2$ of $\mathbf{\Gamma_M}$ can be shown to span the range of $\mathcal{V}(\mathbf{r}_1^s, \ldots, \mathbf{r}_S^s)$, i.e. the S-dimensional subspace $\mathcal{S}$ of $\mathbb{C}^N$ generated by the steering vector evaluated at the sources locations, henceforth termed *signal space*. In the same way, the range of the matrix $\mathcal{U}_{\mathcal{N}}$ of the $N - S$ remaining eigenvectors—associated to the eigenvalues $\sigma_{\mathcal{N}}^2$— is henceforth termed the *noise space* $\mathcal{N}$. Noticeably, the full eigenvectors matrix $(\mathcal{U}_{\mathcal{S}}|\mathcal{U}_{\mathcal{N}})$ can be selected as orthogonal, i.e. $(\mathcal{U}_{\mathcal{S}}|\mathcal{U}_{\mathcal{N}})^H (\mathcal{U}_{\mathcal{S}}|\mathcal{U}_{\mathcal{N}}) = \mathcal{I}_N$. Consequently, under the aforementioned statistical hypotheses, (13) enables the recovery, from the covariance matrix $\mathbf{\Gamma_M}$, of the number of sources—which is N minus the number of repetitions of $\sigma_{\mathcal{N}}^2$—and of their locations —for their associated steering vectors are orthogonal to $\mathcal{U}_{\mathcal{N}}$. But in practice, $\mathbf{\Gamma_M}$ is not known, as only one time record of $\mathbf{m}(t) \triangleq (m_1(t), \ldots, m_N(t))^T$ is available. One common strategy consists in computing an approximation of this quantity on W time snapshots, e.g. by defining

$$\widehat{\mathbf{\Gamma}}_{\mathbf{M}} = \frac{1}{W} \sum_{w=0}^{W-1} \widehat{\mathbf{M}}_w(k) \widehat{\mathbf{M}}_w^H(k), \tag{14}$$

where $\widehat{\mathbf{M}}_w(k)$ denotes an approximation of $\mathbf{M}(k)$ from a L-point Discrete Fourier Transform (DFT) on the wth time snapshot. Finally, the locations of the sound sources are established by isolating the maximum values of the *pseudo-spectrum*

$$h(r, \theta) = \frac{1}{\mathbf{V}^H(r, \theta) \widehat{\Pi}_{\mathcal{N}} \mathbf{V}(r, \theta)}, \tag{15}$$

where $\widehat{\Pi}_{\mathcal{N}} = \widehat{\mathcal{U}}_{\mathcal{N}} \widehat{\mathcal{U}}_{\mathcal{N}}^H$ is called the *projector onto the noise space* and is estimated through the eigen decomposition of $\widehat{\mathbf{\Gamma}}_{\mathbf{M}}$. All these developments have been obtained for a single frequency k_0. Since most sources of interest in robotics

are not narrowband, broadband extensions must be proposed to cope with realistic scenarios. These will be mentioned in Section 3.2.1.

3.1.3. Localization through correlation

In the same way as a sound reaching our two ears is delayed due to propagation, the spatial sampling performed by a microphone array induces temporal delays, also termed *Time Delay(s) of Arrival*, or TDOAs. The approaches outlined in this section aim at estimating the delay ΔT_{ij} between a pair i, j of microphones constituting the array through the computation of a correlation function $R_{m_i m_j}$. Noticeably, the notions of TDOA and of ITD/IPD—see Section 2.2.1—are fairly similar. Despite ITDs/IPDs are generally devoted to binaural approaches, both quantities account for the same physical reality, and can then be estimated in the same way. Furthermore, some biological models claim that the ITD/IPD interaural cue is determined by the brain through a correlation involving dedicated neuronal delay lines (Purves et al., 2004), sometimes called *coincidence detectors* (Jeffress, 1948). Nevertheless, as the forthcoming TDOA computations as well as their exploitation significantly differ from the functioning of the human auditory system, they have been classified into the array processing approaches. This point of view is discussed further in Section 3.2.2.

In all this subsection, the link between the two signals measured on the ith and jth microphone of the array (with $i \neq j$ in all the following) is modeled along

$$\begin{cases} m_i(t) = s(t) + n_i(t) \\ m_j(t) = (s * h_{\mathbf{r}})(t) + n_j(t), \end{cases} \tag{16}$$

with * the convolution operator, $s(t)$ the signal received on the ith arbitrarily chosen microphone and originating from the source to be localized, and $h_{\mathbf{r}}(t)$ the deterministic impulse response between the two considered signals. $s(t)$, $n_i(t)$ and $n_j(t)$ are also hypothesized as zero-mean stationary signals. If the signal $s(t)$ propagates from the source to the array in the free field, and without any scatters placed in the vicinity of the microphones, the impulse response $h_{\mathbf{r}}(t)$ only captures the TDOA ΔT_{ij} between the two receivers, i.e. $h_{\mathbf{r}}(t) = \delta(t + \Delta T_{ij}) = \delta_{-\Delta T_{ij}}$. Importantly, ΔT_{ij} can then be directly related to a source azimuth θ thanks to a relation of the form $l_{ij}/c \cos\theta$, with l_{ij} the interspace between the two considered microphones, when working in the farfield. If the two noise signals $n_i(t)$ and $n_j(t)$ are independent of $s(t)$, then the cross-correlation function $R_{m_i m_j}$ comes as

$$R_{m_i m_j}(\tau) = E[m_i(t) m_j(t-\tau)] = (R_{ss} * h_{\mathbf{r}})(-\tau) + R_{n_i n_j}(\tau), \tag{17}$$

with R_{ss} the source autocorrelation function. Since $h_{\mathbf{r}}(t) = \delta_{-\Delta T_{ij}}$, and if the two signals $n_i(t)$ and $n_j(t)$ are independent, then one has

$$R_{m_i m_j}(\tau) = \left(R_{ss} * \delta_{-\Delta T_{ij}}\right)(-\tau) = R_{ss}(\tau - \Delta T_{ij}), \tag{18}$$

bringing to the fore that $R_{m_i m_j}$ is a temporally shifted version of R_{ss}. Since $\forall\tau$, $R_{ss}(\tau) \leq R_{ss}(0)$, then $R_{m_i m_j}$ exhibits a maximum at $\tau = \Delta T_{ij}$. But in practice, as is the case for the MUSIC approach, the cross-correlation $R_{m_i m_j}$ is not known since only one realization of the random signals m_i and m_j is available. The idea is then to build an estimation $\hat{R}_{m_i m_j}$ of the cross-correlation, leading to the definition of the estimated TDOA

$$\widehat{\Delta T}_{ij} = arg_\tau \max\left(\hat{R}_{m_i m_j}(\tau)\right). \tag{19}$$

Many cross-correlation estimators exist in the literature. One of the most used solution in robotics consists in estimating the cross-correlation of filtered versions of the two signals $m_i(t)$ and $m_j(t)$. This is obtained by introducing a function $\Psi(f)$ weighting the frequency contributions of the two signals, the result being brought back in the time domain with an inverse Fourier transform, i.e.

$$\hat{R}_{m_i m_j}(\tau) = \int_{-\infty}^{+\infty} \Psi(f) \hat{S}_{m_i m_j}(f) e^{j2\pi f t} \mathrm{d}f, \tag{20}$$

where $\hat{S}_{m_i m_j}(f)$ denotes the estimate of the cross-power spectral density function of the two signals $m_i(t)$ and $m_j(t)$. Such estimators are known as *generalized cross-correlation* (GCC) techniques in the literature. Various different frequency weights have been proposed, most of them being listed and studied in Knapp and Carter (1976). Among them, one can cite the Roth (Roth, 1971), the Smoothed Coherent Transform (SCoT) (Carter et al., 1973), the Hannan-Thomson

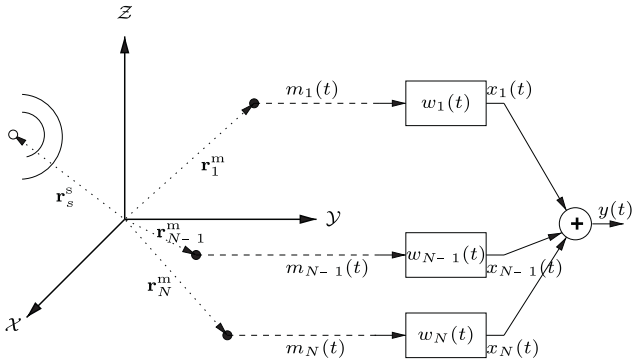


Fig. 6. Basics of beamforming.

(HT) (Hannan and Thomson, 1973), or the Phase Transform (PhaT) processors. This last weighting is by far the most widely used in robotics, and is defined as

$$\Psi_{\mathrm{PhaT}}(f) = \frac{1}{|\hat{S}_{m_i m_j}(f)|}. \tag{21}$$

From this definition, $\hat{R}_{m_i m_j}(\tau)$ then comes as $\hat{R}_{m_i m_j}(\tau) = \int_{-\infty}^{+\infty} e^{j\hat{\phi}(f)} e^{j2\pi ft} \mathrm{d}f$, with $\hat{\phi}(f)$ the phase of $\hat{S}_{m_i m_j}(f)$. In the ideal case when $\hat{\phi}(f) \approx -2\pi f \Delta T_{ij}$, then one gets $\hat{R}_{m_i m_j}(\tau) \approx \delta(t - \Delta T_{ij})$, i.e. the cross-correlation is different from zero only for $\tau = \Delta T_{ij}$, thus proving a very sharp estimation of the TDOA. Nevertheless, since the PhaT operation gives the same importance to all frequencies, it should not be used on narrowband signals, unless if some *a priori* on the frequency bandwidth of interest can be integrated. Such considerations will be discussed in Section 3.2.2.

All the aforementioned approaches mainly rely on a free field model, i.e. the *i*th and *j*th signals only differ in a delay ΔT_{ij}. But as expected, the performances of this delay estimation highly degrade in the presence of reverberations, e.g. when working in a real robotic environment. This appears in the form of estimation outliers, which are all the more frequent as the reverberation time increases (Gustafsson et al., 2003). Additionally, the estimation strategy (19) leads in practice to a TDOA which is a multiple of the sampling frequency T_s, thus limiting the reachable angular resolution. For instance, for two microphones in the freefield spaced 16 cm apart, and a sampling frequency $f_s = 44.1$ kHz, this resolution spans from 3° (for a source facing the robot) to about 18° (for a sound at the left or right of the robot). Some interpolation strategies can nevertheless improve the resolution: interpolation with a parabola or an exponential function (May et al., 2011), or even interpolation of the whole cross-correlation function though *sinc* functions. Finally, as outlined for instance in (21), the correlation processor $\Psi(f)$ must be estimated itself on the basis on an estimation of the cross-power spectral density function $S_{m_i m_j}(f)$. This can be achieved for instance by averaging short-term cross-periodograms (known as Welchs' method (Welch, 1967)), for which the bias and variance have been studied with respect to the overlapping rate or the number of time window used for the estimation (Carter et al., 1973). Similarly, it has been acknowledged that the duration of these windows has a critical effect on the accuracy of the TDOAs ΔT_{ij} extracted from the cross-correlation peaks (Omologo and Svaizer, 1994).

3.1.4. Beamforming based approaches

Among all the methods rooted in signal processing, those based on beamforming are probably the most used in robotics. Their simplicity and low computational cost make them a priori well suited to this context. Yet, as will be shown, their performances strongly depend on the array characteristics, especially on its extent and number of microphones. This subsection then recalls some definitions and generalities on beamforming strategies used in robotics.

The term "beamforming" covers techniques to the combination of the signals coming from an array of discrete sensors, generally in order to focalize it to a specific direction of space $\mathbf{r}_0$. Typically, the signals $m_n(t)$ spatially sampled at the N microphones locations $n = 1, \ldots, N$, are processed by separate linear filters of impulse responses $w_n(\mathbf{r}_0, t)$. These filters are designed in such a way that the sum $y_{\mathbf{r}_0}(t)$ of their outputs is the result of the spatial filtering described

above. This principle is summarized in Fig. 6. On the basis on (8), the time relationship $y_{\mathrm{r_0}}(t) = \sum_{n=1}^{N} w_n(\mathbf{r}_0, t) * m_n(t)$ can be turned into

$$Y_{\mathrm{r_0}}(k) = \sum_{s=1}^{S} D_{\mathrm{r_0}}(\mathbf{r}_s^{\mathrm{s}}, k) S_s^0(k) + \sum_{n=1}^{N} W_n(\mathrm{r_0}, k) B_n(k), \tag{22}$$

with $S_s^0(k)$ the frequency contribution at an arbitrary reference point 0 and due to the sth source, with $W_n(\mathbf{r}_0, k)$ the frequency response of the filter attached to the nth microphone, and

$$D_{\mathrm{r_0}}(\mathbf{r}, k) = \sum_{n=1}^{N} W_n(\mathrm{r_0}, k) V_n(\mathbf{r}, k). \tag{23}$$

This last function of space and time variables is termed *array pattern*, or *beampattern*. It can be assimilated to a transfer function between the signal $s_s^0(t)$ at the arbitrary reference position and caused by the sth source emitting from position $\mathbf{r}$, to the beamformer output $y(t)$, and accounts for the amplification or attenuation of spatial areas. As (23) depends on $V_n(\mathbf{r}, k)$, this definition of the beampattern is valid both in the nearfield and in the farfield. Similarly to (11), the limiting expression $D^{\infty}(\theta, \psi, k) = \lim_{r \to \infty} D_{\mathrm{r_0}}(\mathbf{r}, k)$ can be exhibited when the wavefronts are assumed planar. On this basis, an *energy map* of the environment $E(\mathbf{r}, t)$ is then computed on a time window of length T along

$$E(\mathbf{r}, t) = \int_{t-T}^{t} |y_{\mathbf{r}}(\tau)|^2 \mathrm{d}\tau, \tag{24}$$

the sound sources positions being estimated by detecting the maximum of $E(\mathbf{r}, t)$. Practically, (24) is evaluated on a finite set of potential directions $\mathbf{r}$, see Section 3.2.2.

As already done in (11), and for the sake of simplicity, lets now consider a linear array, made up with N microphones aligned along the $\mathcal{Z}$-axis and having the same interspace d, and whose abscissae z_n verify $z_n = (n - \frac{N+1}{2})d$. Consequently, the array length $L = (N-1)d$. Such an array can be polarized towards a predefined azimuth $\mathbf{r}_0 = \theta_0$ as soon as the filters $w_n(\mathrm{r_0}, t)$ shown in Fig. 6 compensate the delays due to propagation so as to rephase the waves incoming from the DOA $\mathbf{r}_0 = \theta_0$ prior to their summation. Under the planar wavefronts assumption, the transfer functions $W_n(\mathbf{r}_0, k) = W_n(\theta_0, k)$ can be selected as $W_n(\theta_0, k) = e^{jkz_n \cos\theta_0}$, so that the farfield beampattern writes

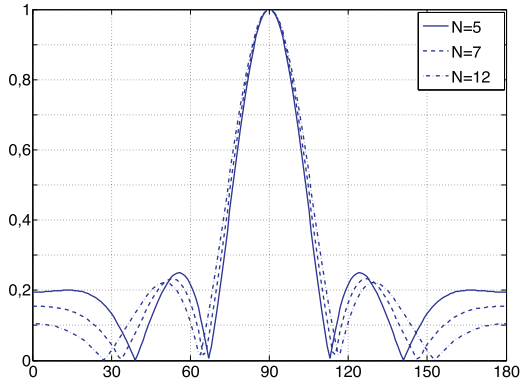
$$D_{\theta_0}^{\infty}(\theta) = \frac{\sin\left(\frac{\pi f}{c} N d(\cos\theta_0 - \cos\theta)\right)}{\sin\left(\frac{\pi f}{c} d(\cos\theta_0 - \cos\theta)\right)}. \tag{25}$$

This so-called *conventional delay and sum beamforming* (DS-BF) strategy is by far the most used in robotics. For instance, Fig. 7(a) shows the module of (25) when considering $\theta_0 = 90°$, $f = 1$ kHz and $L = 0.7$ m. Several comments useful to robotics can be deduced from this farfield array pattern expression in the following configurations:

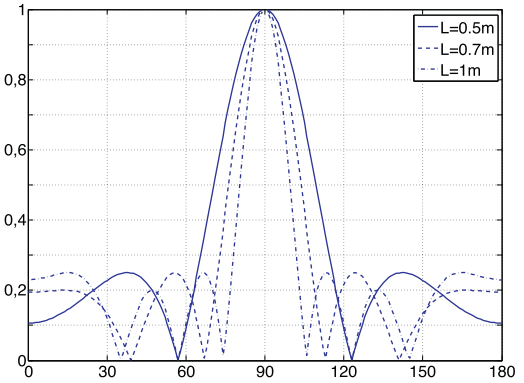
1. variation of the microphones number N, for a fixed array length L and frequency k (or f);
2. change in the length L, for fixed N and k;
3. modification of the frequency k for fixed N and L.

Such a study is fairly classical, see McCowan (2001), Argentieri et al. (2005), and is hereafter summarized for $\theta_0 = 90°$.

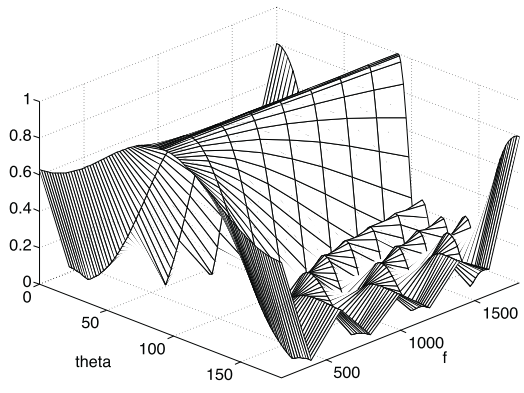
Increasing the number of microphones within a fixed-size array (scenario 1) leads to lower side lobes, see Fig. 7(a). The beampattern corresponding to scenario 2 is shown in Fig. 7(b). The main lobe noticeably gets thinner as the array length increases. As a consequence, it may be necessary to mount a very large array on a robot in order to get a sharp focus towards a given direction of space. Embeddability constraints of course prevent this, and thus limit the resolution of the whole acoustic sensor. Last, the third scenario is presented in Fig. 7(c). Keeping constant the microphones number and interspace, the main lobe width noticeably varies with the frequency f. The spatial resolution at low-frequency is poor, for high-wavelength waves are spatially oversampled by the array. A second phenomenon occurs at high frequencies: these are subject to aliasing, so that multiple replications of the main lobe appear. The spatial sampling of the wave must indeed obey a *Shannon spatial sampling theorem*, in that the maximal microphones interspace d must satisfy $d < d_{\max} = \frac{\lambda_{\min}}{2} = \frac{c}{2 f_{\max}}$, with $f_{\max}$ the maximal frequency in the wavefield. Fig. 7(d) illustrates the aliasing for an antenna made up with $N = 5$ microphones spaced by $d = 17.5$ cm, whose total length is then $L = 0.7$ m.



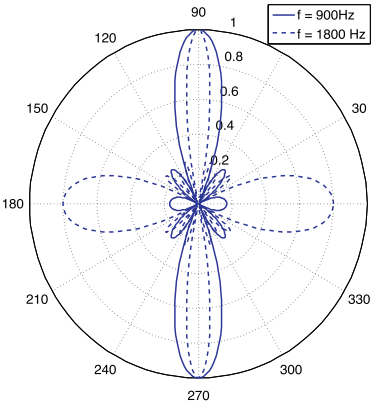
(a) Influence of the microphones number N of an array of fixed length L ($f = 1kHz$, $L = 0.7m$).



(b) Influence of the array length L for a fixed number N of microphones ($f = 1kHz$, $N = 5$).



(c) Normalized beampattern as a function of θ and f ($N = 5$, $L = 0.7m$).



(d) Illustration of the spatial aliasing.

Fig. 7. Different beampatterns of a linear array. (a) & (b) Normalized beampatterns for various values of N and L. (c) & (d) Influence of frequency of a beampattern, for fixed N and L. (a) Influence of the microphones number N of an array of fixed length L ($f = 1$ kHz, $L = 0.7$ m). (b) Influence of the array length L for a fixed number N of microphones ($f = 1$ kHz, $N = 5$). (c) Normalized beampattern as a function of θ and f ($N = 5$, $L = 0.7$ m). (d) Illustration of the spatial aliasing.

This short overview of DS-BF performances demonstrate that being able to precisely focalize in a given direction requires a large array endowed with a lot of microphones, what may not be possible in a robotics context. But even if it were, the resulting beampattern would be still a function of the source frequency, exhibiting a dramatical loss of resolution of the low frequencies. One solution could consist in ignoring such problematic frequency components, by filtering them out. But then it would be difficult to localize any speech signals, where most of the energy spreads from about 300 Hz to 3.3 kHz. Nevertheless, the computational cost of such approaches remains very low, with only N parallel filters running together, making them one of the most used localization techniques in robotics.

3.2. Exploitation in robotics

Now that the theoretical aspects have been overviewed, their applications to robotics are summarized. Following the lines of the above subsections, MUSIC, correlation and beamforming approaches are successively discussed.

3.2.1. MUSIC

As shown in Section 3.1.2, MUSIC consists in computing—for only one frequency k_0—the so-called pseudo-spectrum $h(r, \theta)$ defined in (15), from which the source position is extracted by isolating its maxima. Since the sources of interest in robotics are mainly broadband, the approach needs to be extended to cope with multiple frequencies.

One of the first use of the MUSIC algorithm in robotics is Asano et al. (1999). Therein, an array of $N = 8$ microphones, distributed on the periphery of the robot Jijo-2, enables the localization of vocal sources through an extension of the narrowband method to broadband signals. This extension, named SEVD-MUSIC (for Standard Eigen Value Decomposition-MUSIC), can be seen as "naive", in that it closely follows the lines of the narrowband algorithm. First, the whole frequency range $[k_L -\!- k_H]$ of interest is partitioned into narrow frequency intervals, or "bins", each one centered on k_b, $b = 1, \ldots, B$. The approximation of the covariance matrices $\mathbf{\Gamma_M}(k_1), \ldots, \mathbf{\Gamma_M}(k_B)$ are then computed, following a scheme similar to (14). From the subsequent eigen decomposition of each $\widehat{\mathbf{\Gamma}}_{\mathbf{M}}(k_b)$, separate pseudo-spectra $h_b(r, \theta)$ are determined, $b = 1, \ldots, B$. The localization consists in isolating the maxima of the *average pseudo-spectrum* $h_{\mathrm{Av}}(\cdot)$

$$h_{\mathrm{Av}}(r, \theta) = \frac{1}{B}\sum_{b=1}^{B} h_b(r, \theta). \tag{26}$$

Such a broadband extension is still in use in recent works (Ishi et al., 2011, 2013).

More recently, MUSIC received more attention from roboticists in order to deal with realistic scenarios possibly involving loud noise sources. In such a case, it can be difficult to easily identify the noise and signal spaces from the eigenvalue decomposition of the array cross-correlation matrix $\mathbf{\Gamma}_M$. For this reason, the GEVD-MUSIC (Generalized Eigen Value Decomposition-MUSIC) is proposed (Nakamura et al., 2011). It consists in defining an additional freely-tunable correlation matrix $\mathbf{\Gamma}_N$ for the frequency k_0, and solving the new GEVD problem

$$\mathbf{\Gamma}_M \mathbf{U}_n = \lambda_n \mathbf{\Gamma}_N \mathbf{U}_n, \tag{27}$$

where $\mathbf{U}_n$ and λ_n depict the generalized eigenvectors and eigenvalues of the $(\mathbf{\Gamma}_M, \mathbf{\Gamma}_N)$ matrix pencil respectively. The rest of the algorithm remains identical, since solving (27) allows to determine the noise and signal spaces, and then the computation of the pseudo-spectrum. Again, this operation is conducted along all frequency bins, to get an average pseudo-spectrum, along (26). The choice of the correlation matrix $\mathbf{\Gamma}_N$ is free, but selecting $\mathbf{\Gamma}_N = \mathbf{\Gamma}_B = E[\mathbf{B}\mathbf{B}^H]$ is a common choice which whitens the noise-related eigenvalues, and thus significantly eases the definition of the noise and signal spaces in the presence of loud noise sources. Interestingly, an extended, adaptive, version of GEVD-MUSIC has been proposed recently in Okutani et al. (2012). Called iGEVD-MUSIC (for incremental GEVD-MUSIC), it consists in incrementally estimating the correlation matrix $\mathbf{\Gamma}_N = E[\mathbf{B}\mathbf{B}^H]$ as a function of the current time frame. It then allows the use of the MUSIC algorithm in outdoor applications with drones, for which the level of the involved noises (ego-noise of the drone itself, and wind sound) is loud and especially dynamic (Furukawa et al., 2013). But SEVD, GEVD or iGEVD approaches all suffer from the same problem: they are computationally expensive, with a high calculation cost for subspaces decomposition and for the pseudo-spectrum determination, both being generally performed on a frame-by-frame basis for real-time operations.

As a solution, the GSVD-MUSIC (Generalized Singular Value Decomposition-MUSIC) is proposed in Nakamura et al. (2012). As indicated by its name, it mainly relies on a generalized singular value decomposition, which consists in determining the left and right singular vectors $\mathbf{U}_l$ and $\mathbf{U}_r$ respectively, together with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ such that

$$\mathbf{\Gamma}_N^{-1} \mathbf{\Gamma}_M = \mathbf{U}_l \Lambda \mathbf{U}_r^H. \tag{28}$$

Once this decomposition is performed, the algorithm remains identical, with the left singular vectors and their corresponding singular values being used for the separation between the signal and noise spaces (Nakamura et al., 2012). At the end, GSVD is shown to be computed almost 3 times quicker than GEVD, which is a critical improvement for real-time applications. But again, such decomposition has to be performed for each frequency bin of interest. The contribution (Argentieri and Danès, 2007) exploits the idea of alignment as per Wang and Kaveh (1985), thus constituting a Coherent Broadband source localization algorithm (CB-MUSIC). Basically, the idea is to make the noise and signal spaces identical along all frequency bins through so-called *focalization matrices* $T(\mathbf{r}, k_b)$ verifying $T(\mathbf{r}, k_b)\mathcal{V}(\mathbf{r}, k) = \mathcal{V}(\mathbf{r}, k_0)$, with k_0 an arbitrary reference frequency. This way, the array vector at any frequency k is transformed into its value at frequency k_0. Thanks to this property, a unique correlation matrix gathering all

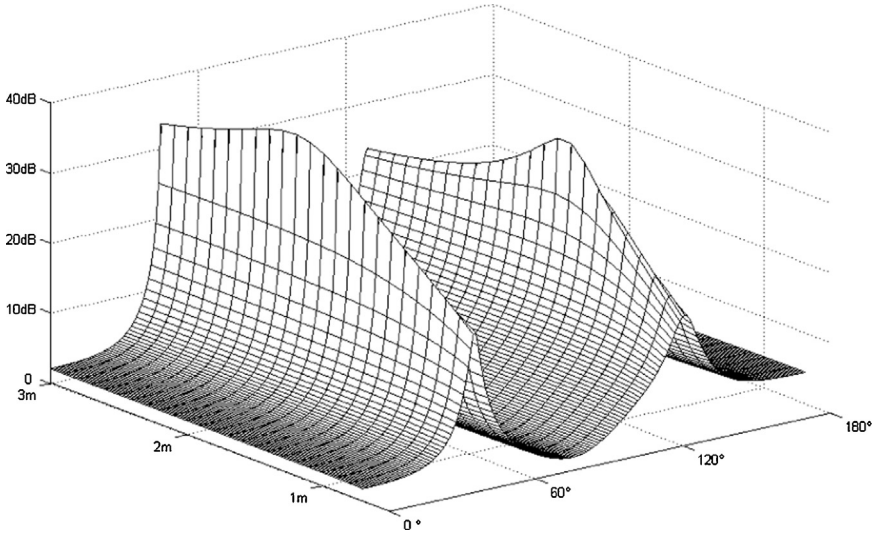


Fig. 8. Typical MUSIC pseudo-spectrum, for two sources in the nearfield of a linear array.

the information along all frequency bins can be defined. Its generalized eigenvalue decomposition then allows the identification of the signal and noise spaces, and thus of the MUSIC pseudo-spectrum. In comparison with the other aforementioned approaches, only one generalized eigenvalue decomposition is necessary, thus limiting the computation cost of the method. Its implementation in a coherent beamspace paradigm is proposed along the lines of Ward and Abhayapala (2004) and an original constructive method is proposed to the synthesis of focalization matrices, in a convex optimization setup. Besides, the approach is able to deal with reverberant robotics environments, since the statistical independence of the sources together with their mutual independence w.r.t. the noise can be relaxed.

All the approaches result in the computation of a pseudo-spectrum function. Such a function is depicted in Fig. 8, for a linear array and two independent sources placed respectively at $(r, \theta) = (2\,\text{m}, 60^\circ)$ and $(1\,\text{m}, 120^\circ)$.

As expected, the two very sharp peaks can be seen at the exact sources positions. But computing the pseudo-spectrum at each candidate source position can result in a high computational costs. A hierarchical strategy with a coarse-to-fine approach is proposed in Nakamura et al. (2012) to solve this issue. Another problem is the need of a careful tuning of the threshold applied on the MUSIC spectrum for peak detection, whose value will highly depends on the reverberation time and the number of sources in the environment. As a solution, Otsuka et al. (2011) proposed an automatic parameter estimation technique relying on the variational Bayesian framework. Another hidden point concerns the source number, which must be known *before* identifying the noise and signal spaces, and thus before determining the broadband pseudo-spectrum. A Bayesian approach to this problem is proposed in Asano et al. (2013), relying on a Markov chain Monte Carlo (MCMC) method. It allows to jointly estimate the number of sources and their locations in realistic reverberant conditions. An information-theoretic approach, grounded on statistical identification—namely the Minimum Akaike Information Criterion Estimate (MAICE) defined in Akaike (1974)—and relying on Wang and Kaveh (1985), Wax and Kailath (1985) has also been proposed in Danès and Bonnal (2010). In addition to its sound theoretical bases, it has a very low computational cost and requires no prior threshold definition. Interestingly, the whole coherent beamspace MUSIC+MAICE detection and estimation has been implemented on a system-on-a-programmable-chip architecture (Lunati et al., 2012).

3.2.2. Correlation-based approaches

TDOA estimation. The application to robotics of correlation-based techniques presented in Section 3.1.3 is common since the beginning of Robot Audition. While the very first approaches were very naive, i.e. estimation of the TDOA by detecting the zero crossing points in the signals (Cavaco and Hallam, 1999), the standard cross-correlation $R_{m_i m_j}$ defined in (17) has been used in a lot of works. In Okuyama et al. (2002), the intercorrelation is computed in order to infer the TDOAs between four microphones disposed on the vertices of a tetrahedron. The originality comes from the selection of the observation window: rather than computing the intercorrelation on the whole duration of the signals,

a plain thresholding enables the detection of echoes-free temporal zones, onto which the TDOAs are determined. One can also cite (Luo et al., 2010), where a 4 microphones array is used to track a sounding docking station outside the field of view of the robot. A slightly different application of the standard cross-correlation is proposed in Bando et al. (2013), were the sound emitted by loudspeakers placed on the surface of a snake robot is used to estimate its posture using TDOA. Another traditional use of the standard cross-correlation consists in estimating the TDOA at the output of a filterbank (Youssef et al., 2012). Such an idea has been extensively used in a binaural context, where the filterbank is made of gammatone filters (see Section 2.2.2). This results in a TDOA function of the frequency, which is in essence analog to the IPD cue (shik Kim and Choi, 2009).

As already mentioned in 3.1.3, other strategies to cross-correlation computation exist. Among them, GCC techniques with the PhaT weighting function (GCC-PhaT) is by far the most used in a robotics context. Its high temporal resolution in TDOA estimation justifies this choice, while it is known that this processor is highly sensitive to the length of the time windows used to estimate the cross-power spectral density function involved in (21). One can cite for instance (Mungamuru and Aarabi, 2004), or (Wang et al., 2004) where the PhaT processor is exploited on a 24 evenly spaced microphones array fitted on the 3.2 m-long walls of a room which is visited by a tour-guide robot. More recent use of the PhaT approach can be cited: Kim et al. (2008) where a triangular 3-microphone array is used to infer source location from short time observations so as to cope with the movement of the sound source or the robot; Badali et al. (2009) presents an evaluation of various real-time sound localization approaches from a cubical 8 -microphone array in which GCC-PhaT is compared with beamforming techniques; Hilsenbeck and Kirchner (2011) proposes a robust approach to the acoustic perception of the presence of people from a pair of microphones. But GCC-PhaT only takes into account the phase of the perceived signals in the intercorrelation computation, giving the same importance to each frequency. As such a weighting does not differentiate the source and noise frequencies, the overall sensitivity of the method to noise is increased and voice localization becomes harder. As a solution, Valin et al. (2006) defines an alternative processor which penalizes the frequencies at which the signal-to-noise ratio is low. This *Reliability-Weighted Phase Transform* (RWPhaT) strategy results in a new adaptive frequency weight $\Psi(f)$. This GCC strategy is still used in Fréchette et al. (2012), on a 8-microphone array embedded on the Spartacus robot, to show the efficiency of a complete artificial audition system for speech recognition and dialogue management. Different adaptations of the PhaT processor have also been proposed. In Hu et al. (2009), an eigen structure-based GCC is outlined, based on the eigenvalue decomposition of the microphones auto-correlation matrix. Results show that the proposed processor exhibits less outliers than the traditional GCC-PhaT. In Liu and Shen (2010) and Liu et al. (2013), the GCC-PhaT-$\rho\gamma$ is proposed to deal with small SNR and large reverberation situations. Results demonstrate improvements w.r.t. the PhaT approach in terms of angular localization error, be the robot at rest or moving.

From TDOA to localization. Once the TDOAs have been computed by one of the above methods, the problem of localizing the source from their values must be addressed. For instance, consider a dipole in the farfield, made up with two microphones separated by a distance d_{ij}. In this planar wavefront case, the most direct approach to the determination of the azimuth θ_s consists in inverting the formula $\Delta T_{ij} = \frac{d_{ij}}{c}\cos\theta_s$. This basic geometric rule is used in Murray et al. (2005), the computed azimuths being involved into a neural network based sound source tracker. The same strategy is used in Luo et al. (2010), or Hilsenbeck and Kirchner (2011). Following the same lines, one can deduce the Cartesian coordinates $\mathbf{r}^{\mathrm{s}} = (u, v, w)$ of a source from the known positions $\mathbf{r}_n^{\mathrm{m}} = (x_n, y_n, z_n)$, $n = 1, \ldots, N$, of the microphones constituting an array. If the propagation occurs in free space, the wavefronts impinging on the microphones are nested spheres centered on the source. Under the assumption that each wavefront supports only one microphone and that the distance d from the source to the first receptor is related to the TDOAs ΔT_{1n} between the 1st and every nth microphone, the following holds:

$$\forall n \in [1, \ldots, N], \quad (x_n - u)^2 + (y_n - v)^2 + (z_n - w)^2 = (d + c\Delta T_{1n})^2. \tag{29}$$

After some manipulations, a matrix equation follows which leads to the unknowns (u, v, w, d). This method is proposed in Mahajan and Walworth (2001) to measure the time of flight of ultrasonic waves. Note that the antenna must hold at least 4 microphones in the planar case, 5 in the 3D case, otherwise the system is underdetermined. Unfortunately, the involved matrices may be ill-conditioned, so that very close TDOA values may lead to significantly different position estimates. This is why Valin et al. (2003) proposes a simpler model, analogous to the one used in Okuyama et al.

(2002) and assuming planar waves. Noticing that the unit vector $\boldsymbol{\nu} = (u', v', w')$ pointing to the source—assumed to be at infinite distance—and the vector $\mathbf{r}_{ij} = \mathbf{r}_j^{\mathrm{m}} - \mathbf{r}_i^{\mathrm{m}}$ connecting the ith microphone to the jth one satisfy

$$\forall n \in [1, \ldots, N], \quad \boldsymbol{\nu}^T \mathbf{r}_{n1} = c\Delta T_{n1}, \tag{30}$$

u', v', w' can be obtained through the resolution of a least square problem. In this approach, the matrix to be inverted depends solely on the microphones positions, and can therefore be tuned so as to improve its conditioning. Moreover, once the sensor geometry is fixed, the inverse matrix is constant and can thus be put in memory to reduce the necessary computations for localization. Nevertheless, the underlying propagation model assumes that planar wavefronts impinge on the antenna. As a solution, Hu et al. (2009) recently proposed a generic extension of (30) to deal with the nearfield case, thus being able to estimate the distance to the source. Finally, a novel geometric formulation of the sound localization problem through TDOA is proposed in the very recent contribution (Alameda-Pineda and Horaud, 2014), where an algebraic analysis and a global optimization solver are proposed for arbitrarily-shaped non-coplanar microphone arrays.

Binaural vs. correlation-based array approaches As already stated in Section 3.1.3, the spatial sampling performed by a microphone array induces temporal delays which are similar to the notions of interaural time differences (ITDs) traditionally used in a binaural context. But does it mean that binaural techniques presented in Section 2 are actually a subset of array processing approaches? Even technical papers in robot audition remain quite ambiguous on this question. For instance, it is true that the steering vector, capturing the sources propagation and microphone array geometry, is a notion close to the HRTF accounting for the body effect on the binaural signals. Concerning the specific case of correlation-based TDOA/ITD estimation, GCC techniques are indeed exploited in both world. But while they have been used to estimate TDOAs in an array context—with the underlying assumption that the signals sensed by each microphone constituting the array only differ by a time delay, their exploitation in a binaural context is quite systematically performed through a frequency dependent analysis, see Section 2.2.2. Moreover, a systematic study of binaural cues proposed in Youssef et al. (2012) demonstrates that ITDs estimated with GCC after a gammatone frequency decomposition provides a better cue for sound localization than the same ITDs computed after a standard FFT analysis. More generally, psychoacoustics effects and their modelization in a binaural context highly differ from the more traditional processing exploited with microphone arrays. The reader interested in this topic can also consult (Katz and Noisternig, 2014) where a complete comparison of ITD estimation methods is proposed.

3.2.3. Beamforming

Among all the aforementioned strategies to sound source localization, beamforming remains probably the most exploited one in robotics. As recalled in Section 3.1.4, beamformers are mainly designed to electronically polarize an array towards some specific DOA, and then to scan several directions of interest. An *acoustic energy map* can then be computed along (24), which is expected to be maximum at the actual sources DOAs. This strategy has been mainly coupled with Delay-And-Sum Beamformers (DS-BF) in a lot of contributions. Interestingly, the computational cost of DS-BF has been addressed in Valin et al. (2004) and Valin et al. (2007) in two ways: first, the energy map is computed in the frequency domain through cross-correlations; next, the needed successive polarizations are performed towards directions defined by a recursive uniform icosahedron grid laid on a sphere. Other DOAs discretizations can be envisaged, depending upon the sensor shape and the number of test points, which lead to a tradeoff between the necessary computing power and the targeted resolution. But this conventional DS-BF strategy suffers from a lack of resolution in the polarization of low frequencies, together with the need of a high number of microphones, as demonstrated in Section 3.1.4. An example of such a DS-BF energy map for a short-length linear microphone array is shown in Fig. 9 (top-left) when trying to localize two speakers uttering from the azimuth 60° and 120°. Large main lobes regularly appear in the energy map, the two sound sources being then hardly spatially separable. Such a problem is often mentioned in the literature: in Tamai et al. (2004), where an array of 128 microphones spreaded into a room is used, the authors proposed to filter out all the frequencies below 500 Hz; the reference (Mattos and Grant, 2004) gets close conclusions when simulating the 8-microphone antenna implemented on the small mobile platform EvBoy II: while the beampattern main lobe is thin enough for frequencies over 1 kHz, frequencies below 800 Hz cannot be exploited for localization; even with a three-ring 32-microphone array. Tamai et al. (2005) shows that the bad array directivity at low frequencies and the aliasing effect at low wavelengths conducts the localization to be performed only for frequencies between 1 and 2 kHz. More recent works still highlight this frequency limitation. For instance, Sasaki

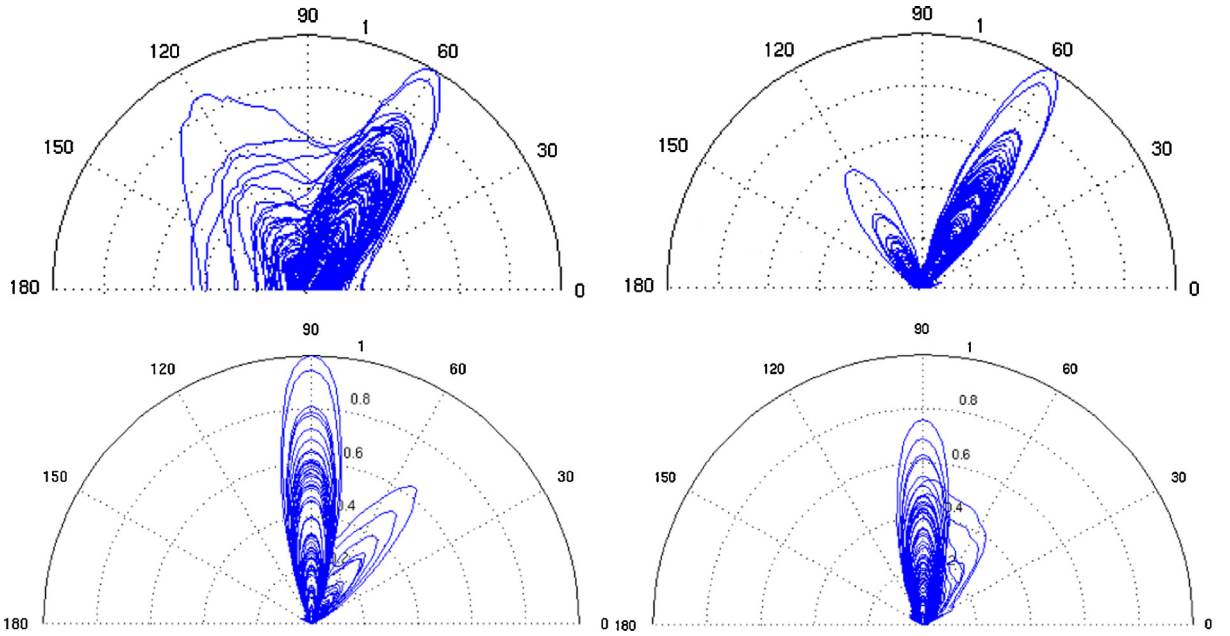


Fig. 9. Acoustic energy maps of the environment (one curve per time snapshot) when using conventional beamformer (top left) or farfield frequency-invariant beamformer (top right and bottom left), see Argentieri et al. (2006). When used in the nearfield, a farfield frequency-invariant beamformer conducts to distorted energy maps (bottom right).

et al. (2013) stated that if two sound sources are close to each other, false positive detections appear in the proposed system because of the wide directivity of DS-BF.

Different solutions have been proposed so far to deal with the low frequencies bad directivity of DS-BF approaches. In (Sasaki et al., 2013), an additional tracking step is introduced to reject the false detections. In Valin et al. (2004), a probabilistic post-filtering of the acoustic energy map is performed, based on two simple short-term and mean-term estimators. Because of the temporal smoothing of the localization, a satisfactory robustness is achieved w.r.t. the actuators noise together with a reasonable computational complexity. An other solution consists in optimizing the array geometry so as to improve the consecutive beampattern. For instance, an evaluation index—relying on beampattern mainlobe width and sidelobes level measurements—is defined in Sasaki et al. (2011) and Sasaki et al. (2012) so as to optimize the placement of 64 microphones over a 350-mm-diameter sphere. A valuable alternative may also consist in the synthesis of frequency-invariant broadband beamformers, as argued in Argentieri et al. (2006). Simulations of realistic scenarios entailing a 8-microphone linear array conclude to a significant improvement in the consequent acoustic maps, so that sources with close DOAs can be distinguished (see Fig. 9). Importantly, it is also established that the localization of sources emitting in the nearfield—e.g. at proximal human—robot interaction distance—is distorted if it entails a frequency-invariant beamformer designed under the farfield assumption. An original nearfield frequency-invariant array pattern synthesis method is thus proposed, under the knowledge of the source range.

3.3. Conclusion

Using microphones array, promising results have been exhibited by the robotics community to localize sound sources. But those array processing techniques are also probably the most sensitive to the four constraints mentioned in Section 1 and raised by the robotics context. Indeed, the embeddability constraint limits the overall size of the array, while it has been shown that the larger array, the better localization results. For instance, beamforming strategies are very sensitive to this parameter, see Section 3.1.4. In the same vein, real-time and frequency constraints preclude the use of naive broadband extensions of existing algorithms in the literature. This has been illustrated with the MUSIC algorithm in Section 3.2.1, whose only the more involved extensions are able to cope with broadband sources (Argentieri and Danès, 2007) and limited computational resources. The frequency constraint also has a great influence on the array beampatterns, thus limiting their applicability in realistic scenario. Finally, the existence of reverberations or self-noises

must be carefully addressed. Some solutions have been proposed for MUSIC (by incrementally estimating the noise correlation matrix), or for correlation-based approaches (by adapting the GCC weighting as a function of the SNR). This in turn clearly questions all the developments proposed in the signal processing and Acoustic Community in light of the robotics context, thus emphasizing the multidisciplinary nature of robot audition.

4. Conclusion

Sound source localization methods developed in the robotics community for the past 15 years have been reviewed in this paper. They can be partitioned into two classes. On the one hand, binaural techniques aim at reproducing artificially the human auditory system. The difficulty to exploit elementary acoustic cues has been underlined, together with the fundamental role of the head in the localization process. Though several propagation models have been proposed in the literature, the most basic of them are not sufficient to explain experimental measurements in an anechoic room. On the other hand, array processing techniques involving a larger number of microphones turn out to be intrinsically more accurate and robust. Different approaches were presented and their relevancy to robotics was discussed. The extension of the high-resolution MUSIC method to broadband signals requires special care to cope with computational resources and the presence of noise in the environment. Correlation approaches to localization lead to accurate conclusions, yet they mainly assume planar wavefronts in order to limit the algorithmic complexity. Last, due to their versatility and low cost, beamforming based strategies are the most often used. However, the focalization of conventional beamformers is limited at low frequencies, so that alternative beamforming methods may be needed, and extreme care must be taken when dealing with nearfield sources. Importantly, a lot of the aforementioned contributions have been integrated within open software frameworks and hardware, making the most advanced approaches accessible to non-experts in the field. The most advanced solutions are the $\mathcal{HARK}$ (HRI-JP audition for robots with Kyoto University) software (Nakadai et al., 2010), the ManyEars framework (Grondin et al., 2013), or the EAR (Embedded Audition for Robotics) system (Bonnal et al., 2010; Lunati et al., 2012).

Most of the cited works in this paper have only focused on the auditory scene analysis from a static view of the world. This idealized situation greatly eases the problem, while it is clear that speech and hearing takes place in a world where none of the static assumptions hold (Cooke et al., 2007). This is exactly what makes *Robot Audition* a robotics problem on its own: the intrinsic mobility of robots can be exploited to improve the analysis of the scene. Actually, the robotics Community has not extensively addressed this *active audition* topic, while it constitutes one of the most promising issue in embodied audition. Indeed, the first contributions in this field clearly showed that the movements of a mobile robot can improve the performance of localization, be the algorithm based on a binaural sensor (Nakadai et al., 2001) or a microphone array (Wang et al., 1997). More recent contributions have now clearly demonstrated how the motion can be exploited together with the induced changes in the auditory perception to improve the analysis, especially in the binaural framework. In this vein, Kneip and Baumann (2008) proposed a binaural sound localization system relying on the intersection of successive "cones of confusion" related to ITD measurements during a head movement. Martinson et al. (2011) also proposed to exploit the motion to dynamically reconfigure an array made of multiple microphones embedded on mobile robots to improve the sound localization. One can also mention Kumon et al. (2010)—who proposed a motion planning system whose objective is to maximize the effectiveness of a speech recognition module, or Bernard et al. (2010, 2012)—where the sound localization problem is rewritten in terms of a sensorimotor approach, with experiments made on the Psikharpax rat robot from the CNRS ROBEA research program and the European FP7-ICT-IP ICEA (Integrating Cognition Emotion and Autonomy) project. Stochastic filtering has also emerged as a valuable framework for sound localization and tracking during robot movement (Lu and Cooke, 2011). Recent contributions have proven the effectiveness of the approach in an active binaural experimental context, with the ability to cope with intermittent moving sources in the presence of false measurements (Portello et al., 2012; Markovic et al., 2013). Importantly, all these works raise one classical problem when working in this active paradigm: the noise coming from the robot itself during the movement—the so-called "ego-noise"—can alter the auditory perception. An illustrative extreme case of this problematic can be seen in Furukawa et al. (2013), where a multirotor UAV, endowed with a microphone array, has to face nonstationary ego-noise emitted during its flight while performing a sound source localization task. Ego-noise cancellation is still under investigation. The most promising solutions to this problem are based on noise patterns, which are collected into ego-noise databases, and then subtracted from the signals according to the generated movement (Ince et al., 2009, 2011). But since the velocity and/or acceleration of each joint is not always correlated with the emitted ego-noise, other approaches have been proposed. Among them, one can cite (Even et al.,

2009)—relying on multichannel Wiener filters, or (Tezuka et al., 2014)—where non-negative matrix factorization is proposed as a solution to the distortions brought by the spectral subtraction method.

Providing robots with efficient and robust auditory functions will keep on being an exciting challenge during the forthcoming years. Many problems which are considered as solved elsewhere have been renewed by the difficulties raised by the robotics context. The growing number of international projects dedicated to embodied audition clearly demonstrates the interest in this topic. Among them, one can cite the BINAAHR project (BINaural Active Audition for Humanoid Robots—French/Japan project funded by ANR and RSJ, ended in 2013), or the two just starting FP7 European projects EARS (Embodied Audition for RobotS) and TWO!EARS (Reading the world with TWO!EARS). All these forthcoming developments will be a source of stimulating discussions between the scientific communities of acoustics, signal processing, robotics, but also physiology and psychoacoustics. We then hope that this survey of existing approaches to the "low-level" stage of sound source localization will motivate new researchers to join the fertile field of Robot Audition.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723.

Alameda-Pineda, X., Horaud, R., 2014. A geometric approach to sound source localization from time-delay estimates. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (6), 1082–1095.

Algazi, V.R., Duda, R.O., Thompson, D.M., Avendano, C., 2001. The CIPIC HRTF database. In: IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 99–102.

Argentieri, S., Danès, P., 2007. Broadband variations of the music high-resolution method for sound source localization in robotics. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2009–2014.

Argentieri, S., Danès, P., 2007. Convex optimization and modal analysis for beamforming in robotics: theoretical and implementation issues. In: European Signal Processing Conference, pp. 773–777.

Argentieri, S., Danès, P., Souères, P., 2005. Prototyping filter-sum beamformers for sound source localization in mobile robotics. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 3551–3556.

Argentieri, S., Danès, P., Souères, P., 2006. Modal analysis based beamforming for nearfield or farfield speaker localization in robotics. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 866–871.

Argentieri, S., Portello, A., Bernard, M., Danès, P., Gas, B., 2013. Binaural systems in robotics. In: Blauert, J. (Ed.), The Technology of Binaural Listening. Springer, Berlin/Heidelberg/New York, NY, pp. 225–253 (Chapter 9).

Asano, F., Asoh, H., Matsui, T., 1999. Sound source localization and signal separation for office Robot Jijo-2. In: IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems, pp. 243–248.

Asano, F., Asoh, H., Nakadai, K., 2013. Sound source localization using joint Bayesian estimation with a hierarchical noise model. IEEE Trans. Audio Speech Lang. Process. 21 (9), 1953–1965.

Badali, A., Valin, J.-M., Michaud, F., Aarabi, P., 2009. Evaluating real-time audio localization algorithms for artificial audition in robotics. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2033–2038.

Bando, Y., Mizumoto, T., Itoyama, K., Nakadai, K., Okuno, H., 2013. Posture estimation of hose-shaped robot using microphone array localization. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3446–3451.

Benesty, J., Chen, J., Huang, Y., 2008. Microphone Array Signal Processing, Springer Topics in Signal Processing. Springer.

Berglund, E., Sitte, J., 2005. Sound source localisation through active audition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 509–514.

Bernard, M., N'Guyen, S., Pirim, P., Gas, B., Meyer, J.-A., 2010. Phonotaxis behavior in the artificial rat Psikharpax. In: International Symposium on Robotics and Intelligent Sensors (IRIS), Nagoya, Japan, pp. 118–122.

Bernard, M., Pirim, P., de Cheveigne, A., Gas, B., 2012. Sensorimotor learning of sound localization from an auditory evoked behavior. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 91–96.

Bonnal, J., Argentieri, S., Danès, P., Manhès, J., Souères, P., Renaud, M., 2010. The EAR project. J. Robot. Soc. Jpn. 28 (1), 10–13 (Special issue "Robot Audition").

Brooks, R., Senior, T., Uslenghi, P., 1999. The COG project: building a humanoid robot. In: Nehaniv, C. (Ed.), Computations for Metaphors, Analogy, and Agents. Springer-Verlag, pp. 52–87.

Butler, Helwig, C.C., 1983. The spatial attributes of stimulus frequency in the median sagittal plane and their role in sound localization. Am. J. Otolaryngol. 4 (73), 165.

Carter, G., Knapp, C., Nuttall, A., 1973. Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing. IEEE Trans. Audio Electroacoust. 21 (4), 337–344.

Carter, G.C., Nuttall, A.H., Cable, P., 1973. The smoothed coherence transform. Proc. IEEE 61 (10), 1497–1498.

Cavaco, S., Hallam, J., 1999. A biologically plausible acoustic azimuth estimation system. In: Third International Workshop on Computational Auditory Scene Analysis of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 78–87.

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

Cooke, M., Lu, Y.-C., Lu, Y., Horaud, R.P., 2007. Active hearing, active speaking. In: International Symposium on Auditory and Audiological Research, Helsingor, Denmark, pp. 33–46.
Danès, P., Bonnal, J., 2010. Information-theoretic detection of broadband sources in a coherent beamspace music scheme. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1976–1981.
Duda, R., Martens, W., 1998. Range dependence of the response of a spherical head model. J. Acoust. Soc. Am. 104 (5), 3048–3058.
Even, J., Sawada, H., Saruwatari, H., Shikano, K., Takatani, T., 2009. Semi-blind suppression of internal noise for hands-free robot spoken dialog system. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 658–663.
Fréchette, M., Letourneau, D., Valin, J., Michaud, F., 2012. Integration of sound source localization and separation to improve dialogue management on a robot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2358–2363.
Furukawa, K., Okutani, K., Nagira, K., Otsuka, T., Itoyama, K., Nakadai, K., Okuno, H., 2013. Noise correlation matrix estimation for improving sound source localization by multirotor UAV. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3943–3948.
Grondin, F., Ltourneau, D., Ferland, F., Rousseau, V., Michaud, F., 2013. The ManyEars open framework. Auton. Robots 34 (3), 217–232.
Gustafsson, T., Rao, B., Trivedi, M., 2003. Source localization in reverberant environments: modeling and statistical analysis. IEEE Trans. Speech Audio Process. 11 (6), 791–803.
Handzel, A.A., Andersson, S.B., Gebremichael, M., Krishnaprasad, P., 2003. A biomimetic apparatus for sound-source localization. In: IEEE Conference on Decision and Control, vol. 6, pp. 5879–5884.
Handzel, A.A., Krishnaprasad, P.S., 2002. Biomimetic sound-source localization. IEEE Sens. J. 2, 607–616.
Hannan, E., Thomson, P., 1973. Estimating group delay. Biometrica 60 (2), 241–253.
Haykin, S., Chen, Z., 2005. The cocktail party problem. Neural Comput. 17 (9), 1875–1902.
Hebrank, J., Wright, D., 1974. Spectral cues used in the localization of sound sources on the median plane. J. Acoust. Soc. Am. 56 (6), 1829–1834.
Hilsenbeck, B., Kirchner, N., 2011. Listening for people: exploiting the spectral structure of speech to robustly perceive the presence of people. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2903–2909.
Hornstein, J., Lopes, M., Santos-Victor, J., Lacerda, F., 2006. Sound localization for humanoid robots – building audio-motor maps based on the HRTF. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1170–1176.
Hu, J.-S., Yang, C.-H., Wang, C.-K., 2009. Estimation of sound source number and directions under a multi-source environment. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 181–186.
Huang, J., Ohnishi, N., Sugie, N., 1997. Separation of Multiple Sound Sources by using Directional Information of Sound Source, vol. 1. Springer, Tokyo.
Humanski, R.A., Butler, R.A., 1988. The contribution of the near and far ear toward localization of sound in the sagittal plane. J. Acoust. Soc. Am. 83 (10), 2300.
Ince, G., Nakadai, K., Rodemann, T., Hasegawa, Y., Tsujino, H., Imura, J., 2009. Ego noise suppression of a robot using template subtraction. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 199–204.
Ince, G., Nakadai, K., Rodemann, T., Imura, J., Nakamura, K., Nakajima, H., 2011. Incremental learning for ego noise estimation of a robot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 131–136.
Ishi, C., Liang, D., Ishiguro, H., Hagita, N., 2011. The effects of microphone array processing on pitch extraction in real noisy environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 550–555.
Ishi, C., Even, J., Hagita, N., 2013. Using multiple microphone arrays and reflections for 3D localization of sound sources. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3937–3942.
Jeffress, L.A., 1948. A place theory of sound localization. J. Comp. Physiol. Psychol. 41, 35–39.
Johannesma, P.I.M., 1972. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In: IPO (Ed.), Symposium on Hearing Theory. , pp. 58–69.
Katz, B.F., Noisternig, M., 2014. A comparative study of interaural time delay estimation methods. J. Acoust. Soc. Am. 135 (6), 3530–3540.
Kim, C.-T., Choi, T.-Y., Choi, B., Lee, J.-J., 2008. Robust estimation of sound direction for robot interface. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 3475–3480.
Kim, H.-D., Komatani, K., Ogata, T., Okuno, H., 2008. Design and evaluation of two-channel-based sound source localization over entire azimuth range for moving talkers. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Nice, France, pp. 2197–2203.
Knapp, C., Carter, G., 1976. The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. 24 (4), 320–327.
Kneip, L., Baumann, C., 2008. Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis. J. Acoust. Soc. Am. 124 (5), 3108–3119.
Kohlrausch, A., Braasch, J., Kolossa, D., Blauert, J., 2013. An introduction to binaural processing. In: Blauert, J. (Ed.), The Technology of Binaural Listening. Springer, Berlin/Heidelberg/New York, NY, pp. 1–32 (Chapter 1).
Kumon, M., Noda, Y., 2011. Active soft pinnae for robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 112–117.
Kumon, M., Shimoda, T., Kohzawa, R., Iwai, Z., 2005. Audio servo for robotic systems with pinnae. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 885–890.
Kumon, M., Fukushima, K., Kunimatsu, S., Ishitobi, M., 2010. Motion planning based on simultaneous perturbation stochastic approximation for mobile auditory robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 431–436.
Liu, H., Shen, M., 2010. Continuous sound source localization based on microphone array for mobile robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4332–4339.
Liu, J., Erwin, H., Wermter, S., 2008. Mobile robot broadband sound localisation using a biologically inspired spiking neural network. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2191–2196.

Liu, H., Fu, Z., Li, X., 2013. A two-layer probabilistic model based on time-delay compensation for binaural sound localization. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 2705–2712.
Lopez-Poveda, E., Meddis, R., 1996. A physical model of sound diffraction and reflections in the human concha. J. Acoust. Soc. Am. 100 (5), 3248–3259.
Lu, Y.-C., Cooke, M., 2010. Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources. IEEE Trans. Audio Speech Lang. Process. 18 (7), 1793–1805.
Lu, Y.-C., Cooke, M., 2011. Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners. Speech Commun. 53 (5), 622–642.
Lunati, V., Manhes, J., Danès, P., 2012. A versatile system-on-a-programmable-chip for array processing and binaural robot audition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 998–1003.
Luo, R., Huang, C., Huang, C., 2010. Search and track power charge docking station based on sound source for autonomous mobile robot applications. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1347–1352.
Mahajan, A., Walworth, M., 2001. 3-D position sensing using the differences in the time-of-flights from a wave source to various receivers. IEEE Trans. Robot. Automat. 17 (1), 91–94.
Markovic, I., Portello, A., Danès, P., Petrovic, I., Argentieri, S., 2013. Active speaker localization with circular likelihoods and bootstrap filtering. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2914–2920.
Martinson, E., Apker, T., Bugajska, M., 2011. Optimizing a reconfigurable robotic microphone array. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 125–130.
Mattos, L., Grant, E., 2004. Passive sonar applications: target tracking and navigation of an autonomous robot. In: IEEE International Conference on Robotics and Automation (ICRA), vol. 5, pp. 4265–4270.
May, T., van de Par, S., Kohlrausch, A., 2011. A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Trans. Audio Speech Lang. Process. 19 (1), 1–13.
McCowan, I.A., 2001. Robust Speech Recognition Using Microphone Arrays. Queensland University of Technology, Australia (Ph.D. thesis).
Middlebrooks, J.C., Green, D.M., 1991. Sound localization by human listeners. Ann. Rev. Psychol. 42, 135–159.
Mungamuru, B., Aarabi, P., 2004. Enhanced sound localization. IEEE Trans. Syst. Man. Cybernt. B 34 (3), 1526–1540.
Murray, J., Wermter, S., Erwin, H., 2005. Auditory robotic tracking of sound sources using hybrid cross-correlation and recurrent networks. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 891–896.
Nakadai, K., Lourens, T., Okuno, H., Kitano, H., 2000. Active audition for humanoids. In: National Conference on Artificial Intelligence (AAAI), Austin, TX, pp. 832–839.
Nakadai, K., Lourens, T., Okuno, H., Kitano, H., 2000. Active audition for humanoid. In: 17th National Conference on Artificial Intelligence, pp. 832–839.
Nakadai, K., Okuno, H., Kitano, H., 2001. Epipolar geometry based sound localization and ex traction for humanoid audition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), vol. 3, pp. 1395–1401.
Nakadai, K., Okuno, H.G., Kitano, H., 2002. Auditory fovea based speech separation and its application to dialog system. In: IEEE/RSJ International Conference on Intelligent Robots and System (IROS), vol. 2, pp. 1320–1325.
Nakadai, K., ichi Hidai, K., Okuno, H.G., Kitano, H., 2002. Real-time speaker localization and speech separation by audio-visual integration. In: IEEE International Conference on Robotics and Automation (ICRA), vol. 1, pp. 1043–1049.
Nakadai, K., Okuno, H.G., Kitano, H., 2002. Real-time sound source localization and separation for robot audition. In: International Conference on Spoken Language Processing (ICSLP), pp. 193–196.
Nakadai, K., Matsuura, D., Okuno, H.G., Kitano, H., 2003. Applying scattering theory to robot audition system: robust sound source localization and extraction. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), vol. 2, pp. 1147–1152.
Nakadai, K., Takahashi, T., Okuno, H.G., Nakajima, H., Hasegawa, Y., Tsujino, H., 2010. Design and implementation of robot audition system HARK open source software for listening to three simultaneous speakers. Adv. Robot. 24 (5–6), 739–761.
Nakamura, K., Nakadai, K., Asano, F., Ince, G., 2011. Intelligent sound source localization and its application to multimodal human tracking. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 143–148.
Nakamura, K., Nakadai, K., Ince, G., 2012. Real-time super-resolution sound source localization for robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 694–699.
Okutani, K., Yoshida, T., Nakamura, K., Nakadai, K., 2012. Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3288–3293.
Okuyama, F., Takayama, J., Ohyama, S., Kobayashi, A., 2002. A study on determination of a sound wave propagation direction for tracing a sound source. In: Proceedings of the 41st SICE Annual Conference, vol. 2, pp. 1102–1104.
Omologo, M., Svaizer, P., 1994. Acoustic event localization using a crosspower-spectrum phase based technique. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. 273–276.
Otani, M., Ise, S., 2006. Fast calculation system specialized for head-related transfer function based on boundary element method. J. Acoust. Soc. Am. 119 (5 Pt 1), 2589–2598.
Otsuka, T., Nakadai, K., Ogata, T., Okuno, H.G., 2011. Bayesian extension of music for sound source localization and tracking. In: International Conference on Spoken Language Processing (Interspeech), pp. 3109–3112.
Otsuka, T., Ishiguro, K., Sawada, H., Okuno, H., 2012. Unified auditory functions based on Bayesian topic model. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2370–2376.
Patterson, R.D., Allerhand, M., Giguère, C., 1995. Time domain modelling of peripheral auditory processing: a modular architecture and a software platform. J. Acoust. Soc. Am. 98, 1890–1894.

Portello, A., Danès, P., Argentieri, S., 2012. Active binaural localization of intermittent moving sources in the presence of false measurements. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3294–3299.
Purves, D., Augustine, G.J., Fitzpatrick, D., Hall, W.C., LaMantia, A.-S., McNamara, J.O., Williams, S.M., 2004. Neuroscience. Sinauer Associates.
Rayleigh, L., 1907. On our perception of sound direction. Philos. Mag. 13 (74), 214–232.
Rodemann, T., Ince, G., Joublin, F., Goerick, C., 2008. Using binaural and spectral cues for azimuth and elevation localization. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2185–2190.
Rodemann, T., 2010. A study on distance estimation in binaural sound localization. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 425–430.
Roth, P.R., 1971. Effective measurements using digital signal analysis. IEEE Spect. 8 (4), 62–70.
Søndergaard, P., Majdak, P., 2013. The auditory modeling toolbox. In: Blauert, J. (Ed.), The Technology of Binaural Listening. Springer, Berlin/Heidelberg, pp. 33–56.
Sasaki, Y., Hujihara, T., Kagami, S., Mizoguchi, H., Oro, K., 2011. 32 channel omnidirectional microphone array design and implementation. J. Robot. Mechatr. 23, 378–385.
Sasaki, Y., Kabasawa, M., Thompson, S., Kagami, S., Oro, K., 2012. Spherical microphone array for spatial sound localization for a mobile robot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 713–718.
Sasaki, Y., Hatao, N., Yoshii, K., Kagami, S., 2013. Nested IGMM recognition and multiple hypothesis tracking of moving sound sources for mobile robot audition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3930–3936.
Saxena, A., Ng, A., 2009. Learning sound location from a single microphone. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 1737–1742.
Schmidt, R., 1986. Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. 34 (3), 276–280.
shik Kim, H., Choi, J., 2009. Binaural sound localization based on sparse coding and SOM. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2557–2562.
Shimoda, T., Nakashima, T., Kumon, M., Kohzawa, R., Mizumoto, I., Iwai, Z., 2006. Spectral cues for robust sound localization with pinnae. In: IEEE International Conference on Intelligent Robots and Systems (IROS), pp. 386–391.
Stevens, S.S., Volkman, J., Newman, E., 1937. A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. 8 (3), 185–190.
Tamai, Y., Kagami, S., Mizoguchi, H., Amemiya, Y., Nagashima, K., TachioTakano, 2004. Real-time 2 dimensional sound source localization by 128-channel huge microphone array. In: IEEE International Workshop on Robot and Human Interactive Communication, pp. 65–70.
Tamai, Y., Sazaki, Y., Kagami, S., Mizoguchi, H., 2005. Three ring microphone array for 3D sound localization and separation for mobile robot audition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 903–908.
Tezuka, T., Yoshida, T., Nakadai, K., 2014. Ego-motion noise suppression for robots based on semi-blind infinite non-negative matrix factorization. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 6293–6298.
Valin, J.-M., Michaud, F., Rouat, J., L'etoumeau, D., 2003. Robust sound source localization using a microphone array on a mobile robot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), vol. 2, pp. 1228–1233.
Valin, J.-M., Michaud, F., Hadjou, B., Rouat, J., 2004. Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. In: IEEE International Conference on Robotics and Automation (ICRA), vol. 1, pp. 1033–1038.
Valin, J.-M., Michaud, F., Rouat, J., 2006. Robust 3D localization and tracking of sound sources using beamforming and particle filtering. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), p. IV.
Valin, J.-M., Michaud, F., Rouat, J., 2007. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. Robot. Auton. Syst. 55 (3), 216–228.
Van Trees, H.L., 2002. Optimum Array Processing, Vol. IV of Detection, Estimation, and Modulation Theory. John Wiley & Sons, Inc.
Vesa, S., 2009. Binaural sound source distance learning in rooms. IEEE Trans. Audio Speech Lang. Process. 17 (8), 1498–1507.
Wang, H., Kaveh, M., 1985. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. IEEE Trans. Acoust. Speech Signal Process. 33, 823–831.
Wang, F., Takeuchi, Y., Ohnishi, N., Sugie, N., 1997. A mobile robot with active localization and discrimination of a sound source. J. Robot. Soc. Jpn. 15 (2), 223–229.
Wang, Q.H., Ivanov, T., Aarabi, P., 2004. Acoustic robot navigation using distributed microphone arrays. Inf. Fusion 5 (2), 131–140.
Ward, D.B., Abhayapala, T.D., 2004. Range and bearing estimation of wideband sources using an orthogonal beamspace processing structure. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. 109–112.
Wax, M., Kailath, T., 1985. Detection of signals by information theoretic criteria. IEEE Trans. Acoust. Speech Signal Process. 33 (2), 387–392.
Welch, P., 1967. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. 15 (2), 70–73.
Wierstorf, H., Geier, M., Spors, S., 2011. A free database of head related impulse response measurements in the horizontal plane with multiple distances. In: Audio Engineering Society Convention.
Woelfel, M.C., McDonough, J., 2009. Distant Speech Recognition. Wiley, Chichester.
Woodworth, R., Schlosberg, H., 1962. Experimental Psychology. Holt, Rinehart/Winston, NY, pp. 349–361.
Youssef, K., Argentieri, S., Zarader, J.-L., 2012. Towards a systematic study of binaural cues. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1004–1009.
Youssef, K., Argentieri, S., Zarader, J.-L., 2013. A learning-based approach to robust binaural sound localization. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2927–2932.
Zahorik, P., Brungart, D.S., Bronkhorst, A.W., 2005. Auditory distance perception in humans: a summary of past and present research. Acta Acust. Unit. Acust. 91 (3), 409–420.