7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Sensorimotor Learning of Sound Localization for an Autonomous Robot

Boris Garcia, Mathieu Bernard, Sylvain Argentieri and Bruno Gas
Sorbonne Universités, UPMC Univ. Paris 06, UMR 7222, ISIR, F-75005 Paris, France
CNRS, UMR 7222, ISIR, F-75005 Paris, France

Summary
In the context of robot perception, a new set of methods for self-supervised sensorimotor learning has emerged lately. These methods try to extract robot and environment configuration information from a set of sensorimotor cues, with no use of any *a priori* knowledge. This paper is concerned with such methods, in the context of binaural robot audition. An incremental algorithm is proposed, relying on an auditory evoked behavior which allows a robot to orient its head toward a sound source. During the learning process, this evoked behavior is used in order to gather auditive and proprioceptive data before and after the head has moved to face the sound source. An auditorimotor map can then be constructed. Thereafter, when the source plays again near a set of previously learned configurations, the robot can use the auditorimotor map to infer a motor command that would make it face the source. In other terms, the robot has learned from past sensorimotor experiences how to localize a sound source in the space of its own motor azimuths. In the present article, we offer an experimental validation of the evoked behavior and put to the test an offline version of the algorithm. The auditory evoked behavior implementation being sufficiently accurate, our results show a good localization performance using the learned auditorimotor map.

PACS no. 43.60.Np, 43.66.Pn

1. Introduction

Robotics is a fertile research topic, with a large community working together to endow robotics platform with control, sensory, and information processing since decades. Today a robot is endowed with advanced capabilities, ranging from perception to decision and action, with the aim to make it fully autonomous and adaptable to changes in its environment or in its own body. In this field, Robot Audition is a recent topic, mainly concerned with sound localization, speech enhancement and recognition, generally for human/robot interaction in realistic acoustic conditions. A lot of contributions on these problematics have been proposed in the last 15 years, either rooted in the binaural [1] or array processing paradigm [2].

Historically, most initial contributions to robot audition were concerned with binaural approaches. However, the proposed algorithms exhibit mixed results in realistic conditions. Nevertheless, recent *active* binaural techniques, coupling the robot movement with the induced auditory variations, have demonstrated their effectiveness for sound localization applications [3]. This paper is grounded on such ideas, and proposes an original sensorimotor-based approach to sound source localization. Other works in this field can be cited, all of them aiming at extracting robot and environment configuration information from a set of sensorimotor cues, with no use of any *a priori* knowledge [4, 5, 6, 7]. Among all this work, a binaural auditory learning algorithm has proved during simulation to provide a reliable way for a robot to learn the localization of a sound source in azimuth [8]. This iterative algorithm relies on an auditory evoked behavior allowing a robot head to orient itself toward the sound source. During a step of the learning process, a white noise is played anywhere near the robot head and this evoked behavior is used in order to gather auditorimotor data before and after the head has moved to face the sound source. An auditorimotor map can be constructed on the fly from the concatenation of initial auditory cues and final motor configurations. Thereafter, when the source plays again near a set of previously learned configurations, the robot can use the auditorimotor map to infer a motor command that would make it face the source. In other terms, the robot has learned from past sensorimotor experiences how to localize a sound source in the space of its own motor azimuths. After a sufficient number of learning iterations, the

Figure 1. BinnoBot, a mobile binaural and binocular head.

robot is able to use motor command inference instead of the evoked behavior so as to face the source.

In the present article, we offer an experimental validation of the evoked behavior and put to the test an offline version of this sensorimotor learning algorithm [8]. A robotic platform named BinnoBot has been devoted to this task (Fig. 1). It consists of a mobile binaural head equipped with two (unused) eyes-mimicking video cameras. It is controlled by the action of four motors: two for the neck azimuth and elevation, two for the eyes orientation. Each manipulation was done in an anechoic chamber, real-world scenarios falling to further studies. The auditory evoked behavior implementation being sufficiently accurate, our results show a good localization performance using the auditorimotor map. These results therefore validate the sensorimotor learning algorithm in a robotic context.

This paper is organized as follow. Section 2 introduces the auditory model used as front-end for sound localization, then presents the evoked behavior and the auditorimotor map based localization in their sensorimotor context. Section 3 provides an evaluation of the auditory evoked behavior and of localization accuracy in a set of robotic experiments. Finally section 4 discusses the obtained results and paves the road to further studies.

2. Material and Methods

This section first introduces the front-end auditory model used for binaural sound localization. A sensorimotor definition of sound localization is then presented. The auditory evoked behavior and the auditorimotor map based localization are finally introduced.

2.1. Auditory Model

The bioinspired auditory model presented herein computes cues related to the interaural level difference (ILD), a cue well known to be involved in sound localization in the high-frequencies range [9]. This model includes a pair of artificial pinna providing a spatial directivity suitable for active sound localization [10], a pair of gammatone filterbanks used as a cochlear model [11], transduction and integration steps and finally a binaural stage estimating the ILD.

The cochlear model decomposes the signal captured by a microphone into a set of c frequency channels. Let $x^l(t)$ be the audio signal captured from the left ear and $G = \{G_i\}_{i\in[1,c]}$ be the cochlear filterbank, the left cochlear output is given as $g^l(t) = \{G_i(x^l(t))\}$. The transduction step then converts $g^l(t)$ into a multichannel action potential train $p^l(t)$ by extracting the positive local maxima of the signal, where we have for each channel $i \in [1, c]$:

$$p_i^l(t) = \begin{cases} g_i^l(t) & \text{if } \frac{dg_i^l(t)}{dt} = 0 \text{ and } g_i^l(t) > \tau \\ 0 & \text{else} \end{cases} \quad (1)$$

where τ is the threshold of minimal activity required for an action potential emission. Thresholding deemphasizes the low intensity parts of the cochlear output and is used in Sec. 3 to suppress the motor noise caused by head movements of the robot. The instantaneous energy $s^l(t)$ is then computed from $p^l(t)$ over a constant time window, and we have for each channel $i \in [1, c]$:

$$s_i^l(t) = \sum_{t'=t-T}^{t} p_i^l(t')^2, \quad (2)$$

where T is the integration duration. Once integrated, the signal $s^l(t)$ is undersampled at the frequency $f_s = 2/T$.

The energy $s^r(t)$ is obtained from the right audio signal $x^r(t)$ in the same way. Therefore, given the left and right energies $s^l(t)$ and $s^r(t)$, the ILD signal $s^{ild}(t)$ is finally computed as follow for each channel $i \in [1, c]$:

$$s_i^{ild}(t) = \frac{2s_i^l(t)}{s_i^l(t) + s_i^r(t)} - 1. \quad (3)$$

If the cochlear filterbank activity stays below the threshold τ during the whole time window, that is when we have $s_i^l(t) = s_i^r(t) = 0$, the ILD vector is not defined and we assign the value $s_i^{ild}(t) = 0$. Moreover we have from (3) $s_i^{ild}(t) \in [-1, 1]$ so that $s^{ild}(t)$ provides, for each channel, a normalized estimation of the ILD independent of the input energy. An ILD signal computed from this model and obtained from a binaural recording is presented in Fig. 2.

2.2. Sensorimotor sound localization

The sensorimotor theory [12, 13] considers perception as an interaction between an agent, either a biological or a robotic one, and its environment. In this context it is suggested that the agent analyzes the sensory consequences of its own movements and learns sensorimotor laws that give raise to spatial perception. This

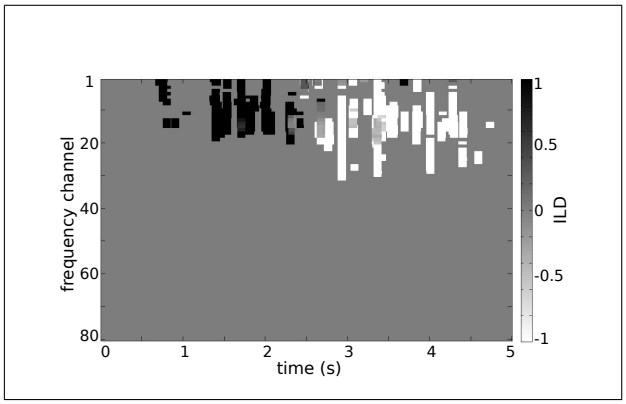

Figure 2. An ILD pattern $s^{ild}(t)$ obtained from a 5 s binaural recording. The sound source is a set of keys moving from left to right of the robot at constant distance and elevation. We use 80 cochlear channels from 100 Hz to 8 kHz (from channel 80 to channel 1 respectively), with $\tau = 0.1$ and $T = 0.1$ s. The threshold τ suppresses the low intensity part of the signal (channels 30-80), the ILD being computed only in the high intensity band (channels 1-30). After integration $s^{ild}(t)$ is sampled at the frequency $2/T = 20$ Hz.

sensorimotor interaction is modeled as an interaction between the environmental space $\mathcal{E}$, the sensory space $\mathcal{S}$, and motor space $\mathcal{M}$ of the agent [4]. In the context of sound localization, a given environmental state $e \in \mathcal{E}$ describes the acoustic properties of the environment as well as the spatial and spectral properties of the source. The state of the agent is described by its motor state $m \in \mathcal{M}$ and its sensory state $s \in \mathcal{S}$. This sensory state s is determined by both the environment and motor states e and m through the so-called sensorimotor law $\Phi(.)$ defined along [4]:

$$s = \Phi(m, e). \tag{4}$$

This sensorimotor law, as well as the environment space are not directly accessible to the agent which has to infer this information from an analysis of its sensorimotor experience. Within this context, source localization can be defined as the estimation of a movement the agent can do to orient itself toward that source [12]. Localization is done in the motor space of the agent and does not rely on any assumption on the physical space. Given a motor space $\mathcal{M}$ and an environment state $e \in \mathcal{E}$, we thus call sound source localization the estimation of the motor state $\tilde{m}$ such as [8]:

$$\tilde{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} |\Phi(m, e) - \Phi(m_{ref}, e_{ref})|, \tag{5}$$

where $|.|$ denotes a given distance metric. The configuration (m_{ref}, e_{ref}) represents a source localized in front of the listener with the head in rest position and corresponds to the most obvious case of localization. The sensory state $s_{ref} = \Phi(m_{ref}, e_{ref})$ is initially unknown and is approximated via evoked behavior experiences (Sec. 2.3). This sensorimotor definition of localization implies a distance minimization and thus a metric on $\mathcal{S}$. The mathematical nature of this sensory space is *a priori* unknown but we assume that $\mathcal{S}$ lies on a differential manifold [4]. Under this hypothesis $\mathcal{S}$ is locally flat and the Euclidian metric can be used for local distance computation.

In the rest of this article we consider a mobile binaural listener perceiving a single stationary sound source. The source emits a white noise at variable azimuth, constant elevation and constant distance in a noise-free anechoic room. The only environmental parameter is thus the azimuthal angle of the source and we have $\mathcal{E} = [-\pi/2, \pi/2]$. The sensory space $\mathcal{S}$ refers to the ILD space. We use the ILD vectors s^{ILD} as described in section 2.1 and thus have $\mathcal{S} = [-1, 1]^c$ with c cochlear channels. Finally the motor space $\mathcal{M}$ describes all the motor commands the robot can generate. Here the BinnoBot is limited to neck rotations in the azimuthal axis and we have $\mathcal{M} = [-\pi/2, \pi/2]$.

2.3. Auditory evoked behavior

The auditory evoked behavior is a hard-wired reflex allowing the robot to orient its head toward the azimuthal direction of a sound source corresponding to an environment state $e \in \mathcal{E}$. From a given initial motor state $m_{init} \in \mathcal{M}$ and the sensory state $s_{init} = \Phi(m_{init}, e)$ the reflex minimizes the ILD summed over frequency channels through azimuthal rotation of the neck. Calling $s^{ild}_{sum}(t)$ the summed ILD at instant t, we have:

$$s^{ild}_{sum}(t) = \sum_{i=1}^{c} s^{ild}_i(t). \tag{6}$$

In order to lateralize the sound source and to initialize the head motion toward it, the rotation direction k is initiated to the left if $s^{ild}_{sum}(t_0) > 0$ or to the right if $s^{ild}_{sum}(t_0) < 0$, where t_0 is the initial time value. The neck rotation is then done at a constant angular speed and terminates when a change in the sign of $s^{ild}_{sum}(t)$ is detected, i.e. when the head is aligned with the sound source.

After completion of the evoked behavior, the final motor and sensory states m_{end} and s_{end} are obtained and the localization estimation $\tilde{m}$ is given as the total angle of rotation done during the movement:

$$\tilde{m} = m_{end} - m_{init}. \tag{7}$$

Moreover the final sensory state s_{end} gives an approximation of the reference sensory state as introduced in Eq. 5 and we have $s_{end} = \Phi(m_{end}, e) \approx s_{ref} = \Phi(m_{ref}, e_{ref})$. Finally the pair composed by the initial sensory state s_{init} and the motor estimation $\tilde{m}$, obtained after the head has moved, summarizes the sensorimotor experience of the agent when confronted to the environment $e \in \mathcal{E}$.

2.4. Localization on the auditorimotor map

Let the auditory evoked behavior be executed by the robot on n environment states $e_i \in \mathcal{E}$, with $i \in [1, n]$, each one associated with a source at random azimuth. The auditorimotor map A is defined as the set of the pairs of initial sensory states and final motor states obtained for each of the n behavior occurrences. We thus have $A = \{(s_i, m_i) | i \in [1, n]\}$. An online procedure for the learning of the auditorimotor map has been proposed [8]. The present paper focuses on an offline version of the algorithm, where A is given *a priori* to the robot.

Once built, this auditorimotor map can be used by the robot to infer a motor command that would make it face a source of unknown azimuth. Let $s \in \mathcal{S}$ be a sensory state associated with such a source. The estimation of the motor state $\tilde{m}$ associated with s is given by interpolation in A. More precisely, let $K_{\mathcal{S}}(s)$ be the k-nearest neighbors of s in A and $K_{\mathcal{M}}(s)$ their related motor states. $\tilde{m}$ is given from $K_{\mathcal{S}}(s)$ and $K_{\mathcal{M}}(s)$ by inverse distance weighting interpolation, so that:

$$\tilde{m} = \sum_{i=1}^{k} \frac{w_i m_i}{\sum_{j=1}^{k} w_j}, \text{ with } w_i = \frac{1}{|s - s_i|}, \tag{8}$$

where $s_i \in K_{\mathcal{S}}(s)$ and $m_i \in K_{\mathcal{M}}(s)$. The inverse distance weighting ensures that the closest neighbors of s in $K_{\mathcal{S}}(s)$ contribute more importantly to the estimation of $\tilde{m}$.

3. Results

This section presents two experimental results. First we offer an experimental validation of the orientation behavior and the reference sensory state estimation. We then evaluate the localization performance obtained by interpolation in the auditorimotor map.

3.1. Protocol

This experiment being carried in an anechoic chamber, we assume that for a given learning step, the only relevant free parameter that can impact the ILD is the head's relative orientation with respect to the sound source. Randomizing the head's initial azimuth at each step is then equivalent to randomizing the source position. For practical reasons, we chose to implement the former. A single learning step will then look like this:

1. Randomize head position
2. Play white noise using the sound source
3. Save initial sensory and motor states
4. Orient the head towards the source using the auditory evoked behavior
5. Save final sensory and motor states

A sufficient quantity of sensorimotor states will be learned this way, constituting a database usable offline. This database will be split into a learning set and a test set in order to evaluate the efficacy of motor command inference. For each initial test sensation, the k nearest initial sensations in the learning set will be retrieved and an associated interpolated motor command will be computed.

In the following experiments, we use $c = 30$ frequency channels from $f_{min} = 2$ kHz to $f_{max} = 6$ kHz, for which ILD is relevant in humans [9]. In order to suppress motor noise, the transduction threshold is set to $\tau = 10^{-7}$. The integration duration for ILD computation is set to $T = 10$ ms. Finally the interpolation in the sensorimotor space is done with a neighborhood order $k = 12$.

3.2. Auditory evoked behavior

The auditory evoked behavior has been tested thoroughly and the experimental parameters relevant to the optimization of its precision have been brought out. The reflex behavior precision showed indeed dependency on several tweakable parameters. Three main components affect the overall behavior precision:

- The quantity $s_{sum}^{ild}(t)$ based upon which the algorithm decides whether to stop the movement, is a sum of several frequency band contributions. It follows that the number of filters in the gammatone banks impacts the reliability of $s_{sum}^{ild}(t)$ as a cue: the fewer filters the less localization information. We estimated that under our experimental conditions, we needed to have at least 25 gammatone filters in order for the summed ILD to show the same profile as the ILD computed on the mere acquired signal.
- The ILD time integration T used to smooth out quick variations for robustness concerns, induces a phase shift in the ILD as a function of time. This delays the zero-crossing event that triggers the motor stop command by some duration Δt, producing a reflex overshoot. The magnitude of this effect can be circumscribed by fine-tuning the ILD integration duration, keeping in mind that there is a trade-off here between robustness and precision [14].
- The overshoot given in degrees depends directly on the motor speed. Indeed, for a given Δt, other things being equal, a higher motor speed allows the head to move somewhat more before it is required to stop. This means the delay can always be compensated by reducing the reflex speed, at the expense of the global experiment duration.

That said, these parameters could be set in such a way that the vast majority of the final azimuths do not overshoot by more than $\pm 5°$, still allowing to learn 700 points in the manifold in about one hour. After a run, we were able to establish statistics on 700 final azimuths. The mean final azimuth is $-1.61 \pm 2.77°$.

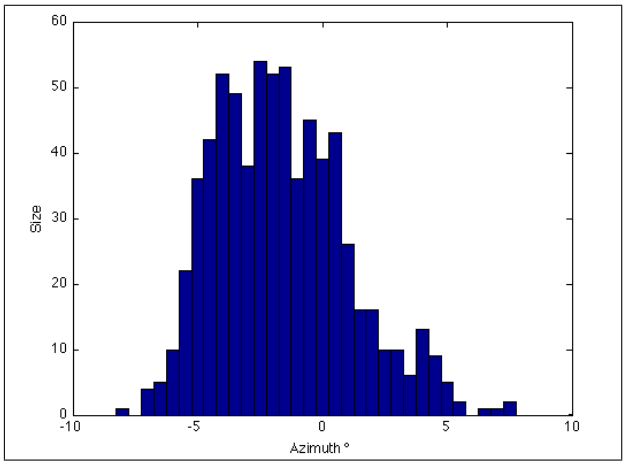

Figure 3. Final azimuths histogram obtained from 700 runs of the auditory evoked behavior.

The azimuths distribution underlying the histogram in Fig. 3 seems bell-shaped.

3.3. Localization on the auditorimotor map

A learning session is composed of 700 initial and final sensorimotor states. This data set is split into two distinct sets. A learning set containing 600 samples is used to build the auditorimotor manifold. A test set gathering 100 samples is devoted to assess the benefits of the interpolation method when confronted to states that have not been learned.

3.3.1. Manifold learning

The sole free parameter in the experiment that can affect the ILD, is the orientation of the head relative to the source. The sensation manifold is then expected to be intrinsically one dimensional. In order to verify this statement, a dimensionality reduction operation is carried out on the manifold using a Principal Component Analysis (PCA), keeping the first two components accounting for the most data variance. Sensations projected to the plane this way (see Fig. 4) show a thin curvilinear profile, confirming the monodimensional nature of the auditory manifold.

By furthermore representing the movement associated to each initial sensation using a color scale, we get a grasp of the smooth nature of this manifold. Two near sensations correspond to two near motor commands. The final sensations can now be projected to the same plane thanks to the PCA coefficients computed before. Doing this we obtain the blue cluster in Fig. 4. This cluster is dense and localized on the manifold, identifying clearly enough the reference sensory state.

3.3.2. Inferring motor commands

For each sample in the test set, we were able to infer a motor command that would make the robot face the source. In order to visually assess the effectiveness of the inference algorithm, the robotic head was first

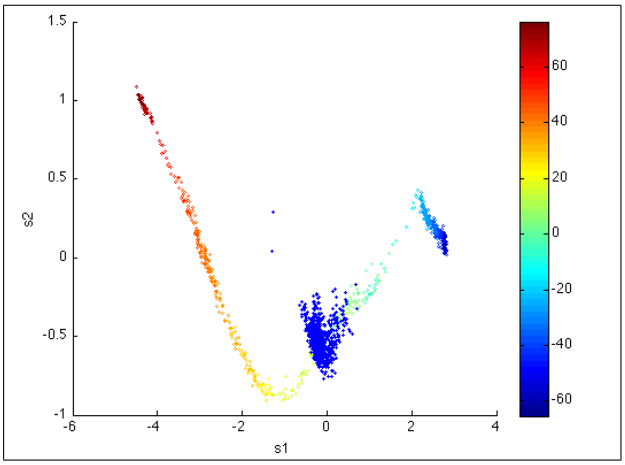

Figure 4. Auditorimotor manifold learned by BinnoBot on 600 sensorimotor states. The high dimensional auditory manifold has been projected in the plane by means of a PCA. Each initial sensory state s_{init} is associated to the rotational movement (in degrees) necessary to face the sound source (color scale). The deep blue cluster corresponds to projected final sensory states s_{end}.

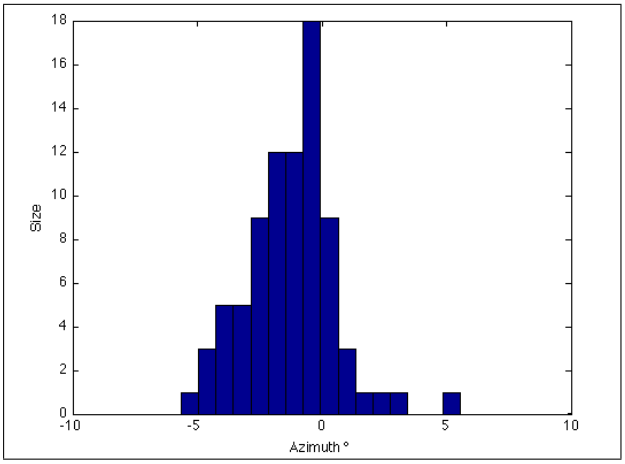

Figure 5. Final azimuths histogram using the interpolation method after the learning process on 600 sensorimotor states.

oriented according to the test initial position, then executed the inferred motor command. The head is able to return to the reference position each time, facing the source. The set of final azimuthal positions is obtained and subjected to the same statistical analysis we did in subsection 3.2.

The mean final azimuth is $-1.21 \pm 1.77°$. The azimuth distribution underlying the histogram of Fig. 5 seems indeed much sharper than the one from Fig. 3. This lower standard deviation indicates a significant precision enhancement. Furthermore, motor commands could be executed at full speed, as speed is not a precision limiting factor anymore.

4. Discussion

In essence, we learned from our results that a robot devoted to a sound localization task can take advantage of previous auditorimotor experience, so as to build an internal representation of its auditory space, during a manifold learning phase. This representation can then be used to localize a sound source, showing improved overall performance as compared to the reflex behavior. Indeed, the orientation movement is carried more precisely when arising from motor command inference based on previous learning. Motor command inference also presents the advantage of delivering a localization result after a constant short amount of time, enabling the robot to work at full speed, showing great reactivity. During the learning phase, there is a prerequisite for the source to be stationary and immobile while the orientation movement is carried. On the contrary, after the learning process, a brief sound can trigger a quick orientation behavior, as a unique sound frame is needed to infer a motor command. These acquired properties endow the robotic platform with the ability to react quickly and precisely to external auditory stimuli, rendering it suitable for sound source localization and tracking duties.

The sound source used for the learning phase was emitting continuous white noise, which made the whole inference process require that precise spectral category in order to work correctly. Further efforts are intended to limit the adverse effects of spectral content, by playing more elaborate sounds during the learning process. Experiments also need to be carried out in a non anechoic environment so as to confront the learning model to more diverse and realistic scenarios. An online version of the algorithm is currently showing good results in simulation and an experimental validation is scheduled soon.

5. Conclusion

Our results show that in a sufficiently simple environment, a naive agent is able to learn a representation of its auditory space, and to use it afterwards to localize a sound source in its own azimuthal motor space. No environment and robot model were needed here, the agent only had access to sensorimotor cues to construct its sensorimotor map. To some extent, the agent was able to learn auditory perception and gained additional performance by doing so. The reflex behavior used during the learning process was outperformed by the motor command inference algorithm in several ways. Inferring motor commands using past sensorimotor experience allows a fast and precise response to a stimulus.

Acknowledgement

This research has been partially supported by EU FET grant TWO!EARS, ICT-618075.

References

[1] S. Argentieri, A. Portello, M. Bernard, P. Danès, and B. Gas. Binaural systems in robotics. In J. Blauert, editor, *The Technology of Binaural Listening*, chapter 9, pages 225–253. Springer, Berlin–Heidelberg–New York NY, 2013.

[2] C.T. Ishi, J. Even, and N. Hagita. Using multiple microphone arrays and reflections for 3d localization of sound sources. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 3937–3942, Nov 2013.

[3] I. Markovic, A. Portello, P. Danes, I. Petrovic, and S. Argentieri. Active speaker localization with circular likelihoods and bootstrap filtering. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 2914–2920, Nov 2013.

[4] D. Philipona, J.K. O'Regan, and J.-P. Nadal. Is there something out there? Inferring space from sensorimotor dependencies. *Neural computation*, 15(9):2029–49, 2003.

[5] M. Aytekin, C.F. Moss, and J.Z. Simon. A sensorimotor approach to sound localization. *Neural Computation*, 20(3):603–635, 2008.

[6] A. Laflaquiere, S. Argentieri, O. Breysse, S. Genet, and B. Gas. A non-linear approach to space dimension perception by a naive agent. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 3253–3259, Oct 2012.

[7] A. Deleforge, F. Forbes, and R. Horaud. Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds. *International Journal of Neural Systems*, 2014.

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

[9] J. Blauert. *Spatial hearing*. MIT Press, 1997.

[10] M. Bernard, S. N'Guyen, P. Pirim, B. Gas, and J.A. Meyer. Phonotaxis behavior in the artificial rat Psikharpax. In *International Symposium on Robotics and Intelligent Sensors*, pages 118–122, Nagoya, Japan, 2010.

[11] B. Glasberg and B. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138, 1990.

[12] H. Poincaré. *The Foundations of science*. The Science Press, 1921.

[13] J.K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–1031, 2001.

[14] M. Bernard. *Audition active et integration sensorimotrice pour un robot autonome bioinspire*. PhD thesis, Ecole doctorale SMAER, 2014.