Sensorimotor Learning of Sound Localization from an Auditory Evoked Behavior

Mathieu Bernard Patrick Pirim Alain de Cheveigné Bruno Gas

***Abstract*— A new method for self-supervised sensorimotor learning of sound source localization is presented, that allows a simulated listener to learn an auditorimotor map from the sensorimotor experience provided by an auditory evoked behavior. The map represents the auditory space and is used to estimate the azimuthal direction of sound sources. The learning mainly consists in non-linear dimensionality reduction of sensorimotor data. Our results show that an auditorimotor map can be learned, both from real and simulated data, and that the online learning leads to accurate estimations of azimuthal sources direction.**

I. Introduction

Sound source localization in animal and human is well known to be a complex task, involving the processing of multiple acoustics cues by an important dedicated neural pathway. Most studies in spatial hearing have considered a static environment, where both the listener and the source are immobile, but active skills involving the listener's own movements are known to contribute reliable cues for sound source localization, for example in distance perception or front-back disambiguation [1], [2]. Among the active processes involved in hearing, the auditory evoked orienting behavior (OB) is a reflex present in human newborns that consists of head and eyes movement toward a sound source [3]. The neural basis of this behavior seems to be hard-wired and to play a role in the subsequent learning of sound localization skills [4]. Auditory evoked behaviors, such as the OB of the barn owl [5] and the phonotaxis behavior of the cricket [6], have been modeled to provide suitable solutions for active audition in autonomous robots [7], [8], [9].

Considering that the brain is initially a naive agent that communicates with the world via an unknown set of afferent and efferent connexions, with no *a priori* knowledge about the space it is immersed in, the sensorimotor approach [10], [11] suggests that the brain analyzes the consequences of its own movements on its sensory perceptions and extracts sensorimotor laws that give it access to the properties of the surrounding space, in addition to its own body. Thus, it has been showed that such a naive agent can recover the dimensionality of physical space without any *a priori* knowledge [12], [13], [14]. Once the space dimensionality is known, the sensorimotor approach can also be applied to the learning of sensory space parametrization from a set of sensorimotor experiences [12], [15], [16], giving rise to spatial perception. The basic assumption is that the sensory space of the agent lies on a low-dimensional manifold whose topology is homeomorphic to the topology of the embodying space. Following this hypothesis, the learning of spatial perception becomes the learning of such a manifold. Methods based on manifold learning have been proposed for auditory localization using supervised linear regression [17], self-organized maps [18] and within the sensorimotor approach using local tangent space alignment [15], [16].

Applying the sensorimotor approach to audition in autonomous systems, we propose in this paper a new method for self-supervised online learning of sound source localization in the azimuthal plan. Contrary to classical methods in sound source localization, where the source position is expressed in term of angle or distance within a Euclidean physical space, the sensorimotor approach links perception and action in an internal representation of space. In this model, introduced in section II, the source azimuth can be estimated actively through the OB or passively, after learning, through an auditorimotor map. Using an online learning system allows the self-supervision and is well designed for autonomous systems, where the variability of the environment is difficultly addressed using supervised learning. In section III, manifolds are computed from both simulated and real sound sources and auditorimotor map localization etimations are evluated during online learning. Finally, section IV discuss limitations and perspectives.

II. Material and Methods

After a definition of the problem of sound localization, the binaural auditory system and the prewired exploratory OB are presented. This OB is based on intensenty cues and is robust enough for robotics usage [7]. Then are presented the learning of the auditorimotor map, the estimation of sound source azimuth and the self-supervision process.

A. Problem Statement

We consider a mobile binaural listener perceiving a sound source localized in space. We define the state e of an environment and $\mathcal{E}$ the manifold of all possible environment states so that $e \in \mathcal{E}$ describe the acoustic properties of the environment and the spatial and spectral properties of the source. The listener is described by its motor state m in a motor space manifold $\mathcal{M}$. Finally, the sensory state s of the listener is included in the manifold $\mathcal{S}$ of the sensory states. The sensory state s of an agent is determined by both its

M. Bernard and P. Pirim are with Brain Vision Systems, Paris, France.

A. de Cheveigné is with the Laboratoire Psychologie de la Perception (CNRS UMR 8158), Sorbonne Paris Cité, Paris, France.

B. Gas is with the Institut des Systèmes Intelligents et de Robotique (CNRS UMR 7222), Université Pierre et Marie Curie, Paris, France.

Corresponding author: mathieu.bernard@bvs-tech.com.

environment and motor states e and m through a functional relationship Φ called a sensorimotor law [12]:

$$s = \Phi(m, e). \tag{1}$$

In this paper, we consider a listener placed at a fixed position (m_x, m_y), with $m_x = m_y = 0$, that have a rotation degree of freedom (DOF) in azimuth. We denote $m_\theta \in [-180, 180]$ the orientation of the head, which is modeled as a binaural axis of inter-ear distance $m_d = 0.145$ m, estimation of the mean human value [19]. It should be noted that the knowledge of the inter-ear distance is not required by the proposed method. Moreover the only DOF available is m_θ, therefore giving us the motor manifold $\mathcal{M} = \{m_\theta | m_\theta \in [-180, 180]\}$. The environment is modeled as a 2-dimensional space where a single omnidirectional sound source is emitting a sound signal. In the listener polar coordinate system, the source position is defined by its distance $e_r \in \mathbb{R}$ and azimuth $e_\theta \in [-180, 180]$.

Given a motor manifold $\mathcal{M}$ and an environment state $e \in \mathcal{E}$, we call sound source localization the estimation of the motor state $\tilde{m}$ such as:

$$\tilde{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} |\Phi(m, e) - \Phi(m_0, e_0)|, \tag{2}$$

where $|.|$ denotes a distance metric, $m_0 = m|_{m_\theta = 0}$ and $e_0 = e|_{e_\theta = 0}$. The configuration (m_0, e_0) represents a source localized in front of the listener with the head in the rest position and corresponds to the most obvious case of localization. The sensory state $\Phi(m_0, e_0)$ is initially unknown and is approximated through OB experiences, allowing an estimation of $\tilde{m}$ through the auditorimotor map.

B. Auditory Model

The auditory vectors computed herein contain cues related to the interaural level difference (ILD), a cue well known to be involved in sound localization in the high-frequencies range, where the head shadowing becomes significant and the temporal cues confusing, at least for pure tones [20].

A pair of gammatone filterbanks [21] is used as a cochlear model, decomposing the binaural signal over n^c frequency channels. The filters are designed to approximate the human cochlear basilar membrane linear response at a given frequency [22]. From the binaural filterbank $G = \{G_i^L, G_i^R\}_{i=1..n^c}$, and supposing a binaural acoustic signal $x(t) = (x^L(t), x^R(t))$, we obtain the binaural cochlear output signal $x^G(t) = \{g_i^L(t),\ g_i^R(t)\}_{i=1..n^c}$, where:

$$g_i^L(t) = G_i^L(x^L(t)) \text{ and } g_i^R(t) = G_i^R(x^R(t)). \tag{3}$$

We use $n^c = 30$ channels filterbanks from $f_{min} = 2$ kHz to $f_{max} = 6$ kHz, for which ILD is relevant in humans [20].

Once the binaural cochlear output $x^G(t)$ is computed, it is converted in an action potential train $x^P(t) = \{p_i^L, p_i^R\}_{i=1..n^c}$ by extracting the positive local maxima of the signal, where we have for each channel i:

$$p_i^L(t) = \begin{cases} g_i^L(t) & \text{if } \frac{dg_i^L(t)}{dt} = 0 \text{ and } g_i^L(t) > \tau \\ 0 & \text{else} \end{cases}$$
$$p_i^R(t) = \begin{cases} g_i^R(t) & \text{if } \frac{dg_i^R(t)}{dt} = 0 \text{ and } g_i^R(t) > \tau \\ 0 & \text{else} \end{cases} \tag{4}$$

where τ is the threshold of minimal activity required for an action potential emission. Thresholding deemphasizes the low intensity parts of the cochlear output.

From the computed action potential train, the left and right intensity vectors s^L and s^R are computed as an integration of the squared values of p_i^L and p_i^R:

$$s_i^L(t) = \sum_{t'=t-T}^{t} {p_i^L(t')}^2 \text{ and } s_i^R(t) = \sum_{t'=t-T}^{t} {p_i^R(t')}^2, \tag{5}$$

where T is the integration duration. Once integrated, the signal is undersampled at the frequency $f_s = 2/T$ and the ILD vectors s^{ILD} are finally computed as follow for each channel $i \in [1, n^c]$:

$$s_i^{ILD}(t) = \frac{2 s_i^L(t)}{s_i^L(t) + s_i^R(t)} - 1. \tag{6}$$

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 (6) $s_i^{ILD}(t) \in [-1, 1]$ and $s_i^{ILD}(t) = 0$ if $s_i^L(t) = s_i^R(t)$ so that $s^{ILD}(t)$ provide a normalized estimation of the ILD.

C. Auditory Evoked Orientation Behavior (OB)

The OB is a hard-wired auditory evoked behavior allowing the listener to orient its head toward the azimuthal direction of a sound source corresponding to an environment state e. From given initial motor state $m_{init} \in \mathcal{M}$ and sensory state $s_{init} = \Phi(m_{init}, e)$ the OB minimizes the ILD sum signal $s_{sum}^{ILD}(t)$ through azimuthal rotation, where we have:

$$s_{sum}^{ILD}(t) = \sum_{i=1}^{n^c} s_i^{ILD}(t). \tag{7}$$

In order to lateralize the sound source and to initialize the reflex motion toward it, the rotation direction k is given as $k = 1$ (to the left) if $s_{sum}^{ILD}(t_0) > 0$ and $k = -1$ (to the right) if $s_{sum}^{ILD}(t_0) < 0$, where t_0 is the initial time value. The motor command is then initialized at a constant angular speed of 60 deg.s^{-1} and terminates when a change in the sign of $s_{sum}^{ILD}(t)$ is detected, that is when the mean ILD cross a zero value and the head have been aligned to the sound source.

After completion of the OB, the final motor and sensory states m_{end} and s_{end} are obtained and the source azimuth relatively to the initial listener position is given as the total angle of rotation done during the OB. Furthermore we have:

$$\begin{aligned} s_{end} &= \Phi(m_{end}, e) \\ &= \Phi(m_0 + \delta m, e_0 + \delta e), \end{aligned} \tag{8}$$

where e_0 and m_0 as in (2). In this simulation the motor state error δm is due to the sampling error of a constant rotation controlled by the sign of $s_{sum}^{ILD}(t)$. The environment state error δe can be due in a complex environment to the effects of environment reverberation or source non-stationarity.

D. Self-Supervised Auditorimotor Map Learning

The proposed method for the self-supervised auditorimotor map learning, based on the OB, is composed of three functional elements that are the construction of the auditorimotor map, the estimation of a sound source azimuth location from the map and a self-supervision process allowing the validation of the estimation, and its correction if required.

1) Auditorimotor map learning: Suppose that the OB has been executed on n auditory experiences corresponding to n different environment states and let $S_{init} = \{s_{init,i}\}_{i\in[1,n]}$, $S_{end} = \{s_{end,i}\}_{i\in[1,n]}$ and $M_{end} = \{m_{end,i}\}_{i\in[1,n]}$ be respectively the set of the initial sensory states, final sensory states and final motor states. The auditorimotor map A links a low-dimensional representation R_{init} of S_{init} to the set M_{end}, so that $A = \{a_i\}_{i\in[1,n]}$, with $a_i = (r_{init,i}, m_{end,i})$.

Here we compute the low-dimensional representation R_{init} using the Laplacian eigenmaps non linear dimensionality reduction technique [23], [24]. In what follows we call P this manifold learning procedure such as $R_{init} = P(S_{init})$. The Laplacian eigenmaps compute a low-dimensional representation of the data in which the distances between a data point and its k-nearest neighbors in S_{init} are minimized, thus preserving local properties of the data in R_{init}. Based on a k-neighborhood graph and using the spectral graph theory, the distance minimization is defined as an eigenproblem. We use the implementation proposed in [24] and a neighborhood order $k = 12$. Moreover we suppose that the space dimensionality has previously been estimated [13][14], and we therefore fix it to $dim(R) = 2$, corresponding to movement and sound localization in the horizontal plane (see III-A).

2) Source localization: Suppose that the OB have been executed on n auditory experiences corresponding to n different environment states and that the manifold R_{init} and the auditorimotor map A has been computed as above. The finding of $\tilde{m}$ as formulated in (2) from a new experiment $s \in \mathcal{S}$ is done through neighborhood relationship in A. Firstly the new sensory state s is projected on the manifold R_{init}, giving us an estimation of its low-dimensional value $\tilde{r} = P_e(s)$, where P_e is an out-of-sample extension allowing the projection of new points on the existing manifold [24], [25]. Let $K_{\tilde{r}} = \{r_i | r_i \in R\}_{i=[1,k]}$ be the set of the k-nearest neighbors of the point $\tilde{r}$ in the manifold R_{init} and $K_{\tilde{m}} = \{m_i | m_i \in M_{end}\}_{i=[1,k]}$ be the set of their corresponding motor states in A such as $a_i = (r_i, m_i)$. Estimation of $\tilde{m}$ from $K_{\tilde{r}}$ and $K_{\tilde{m}}$ is computed by inverse distance weighing interpolation [26] such as:

$$\tilde{m} = \sum_{i=1}^{k} \frac{w_i m_i}{\sum_{j=1}^{k} w_j}, \text{ with } w_i = \frac{1}{|\tilde{r} - r_i|}. \tag{9}$$

Considering a manifold R_{init} composed of very few points or an auditory perception s quite different from previously learned experiences, the projection $\tilde{r} = P_e(s)$ can fall as an outlier whithin an area of R_{init} with no near neighbors, inducing an irrelevant estimation of $\tilde{m}$. Denoting $d_k(r)$ the mean distance of a point r to its k-nearest neighbors in R_{init}, the projection $\tilde{r}$ is detected as an outlier in R_{init} using the maximum normed residual test [27] on the mean neighborhood distance d_k. $\tilde{r}$ is considered as an outlier if:

$$\frac{|d_k(\tilde{r}) - \mu(d_k)|}{\sigma(d_k)} > v_{crit}(n, \alpha), \tag{10}$$

where $\mu(d_k)$ and $\sigma(d_k)$ are respectively the mean and the standard deviation of d_k for all the points in R_{init}. $v_{crit}(n, \alpha)$ is a critical value that depends of the number n of points in R_{init} and the significance level α allowing the tuning of the sharpness of the outlier detection [27] (see III-B). If the projection $\tilde{r} = P_e(s)$ is detected as outlier, the *a priori* estimation $\tilde{m}$ is rejected and the OB is initiated for an *a posteriori* localization. The final states s_{end} and m_{end} given by the OB, along with the initial sensory state s, are added to their respective sets S_{end}, M_{end} and S_{init}. The manifold is then updated with the new set S_{init}, *i.e.* the Laplacian is recomputed and the auditorimotor map updated, adding a new experience in the manifold.

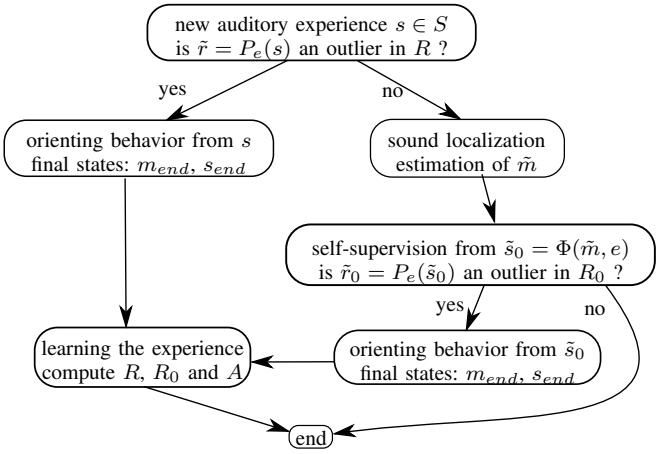


Fig. 1. The successive steps composing the self-supervised auditorimotor map learning algorithm. See text for details.

3) Self-supervision: Following (8), the final sensory state $s_{end} = \Phi(m_{end}, e)$ given by the OB is an estimation of the reference sensory vector $\Phi(m_0, e_0)$. We denote $R_0 = \{r_{0,i}\}_{i\in[1,n]}$ the set of the projected final sensory states such as $r_{0,i} = P_e(s_{end,i})$ and we note r_0 the mean value of the points in R_0. As expressed in (2) the point r_0 therefore represent the low-dimensional estimation of the reference point $\Phi(m_0, e_0)$ (see III-B).

Once $\tilde{m}$ have been estimated, the self-supervision allows the system to check for localization errors after a movement of the listener to the motor state $\tilde{m}$, and to correct the localizaton estimate if required. After movement to the motor state $\tilde{m}$ the resulting sensory vector is projected on R_{init} and we call $\tilde{r}_0$ this projection so that $\tilde{r}_0 = P_e(\Phi(\tilde{m}, e))$. The validation consists of an outlier detection of $\tilde{r}_0$ in the dataset R_0. As in (10), the detection is done using the maximum

normed residual test but, instead of the distance d_k, we use the mean distance of the points in R_0 to their mean r_0. Error correction is applied to outliers, as defined above.

4) Integrative algorithm: The algorithm integrating the different elements presented above is showed in Fig. 1 and is executed for each new auditory experience. To be noted that the learning of a new point is done only if the estimation of $\tilde{m}$ has failed, that is if the point is not represented in the map. Moreover we have previously supposed an existing manifold R_{init} containing n points. Precisely the minimal number of points required for the computation of the manifold is equal to the neighborhood order k. A condition is thus added before the estimation of $\tilde{r}$ so that if $n \leq 2k$ the OB is systematically launched and the system accumulates a minimal amount of data before the first auditorimotor map learning.

III. Results

This experimental section is composed of two parts. Manifold learning on both real and simulated data is presented first, followed by a detailed description of the auditorimotor map learning algorithm during successive iterations. Tab. I summerizes the parameters common to all the experiments.

A. Manifold Learning on Real and Simulated Data

Auditory vectors used for manifold learning are generated from stationary broadband random spectrum sound sources as in the CAMIL dataset [16], with a modified normalization. An emitted sound x^e sampled at the frequency $f^e = 20$ kHz is given in function of time as:

$$x^e(t) = \frac{\sum_{i=1}^{n^e} \omega_i \sin(2\pi f_i t + \phi_i)}{\sum_{i=1}^{n^e} \omega_i}, \qquad (11)$$

where $F = \{f_i\}_{i=1..n^e}$ is a set of n^e fixed frequencies, $\{\omega_i\}_{i=1..n^e} \in [0,1]^{n^e}$ and $\{\phi_i\}_{i=1..n^e} \in [0,2\pi]^{n^e}$ are uniform random weights and phases associated with each frequency. We use a set of $n^e = 600$ uniformly distributed frequencies from $f^e_{min} = 50$ Hz to $f^e_{max} = 6$ kHz.

Sound sources are placed here at fixed distance $e_r = 2.7$ m and random azimuth $e_\theta \in [-180, 180]$, for which localization is confronted with front-back ambiguity [20].

TABLE I
Parameters reference table.

Source		
f^e	sample frequency (kHz)	20
n^e	number of sinusoids	600
f^e_{min}	min frequency (kHz)	5.10^{-2}
f^e_{max}	max frequency (kHz)	6
Listener		
n^c	number of cochlear frequency channels	30
f_{min}	min center frequency (kHz)	2
f_{max}	max center frequency (kHz)	6
τ	cochlear output threshold	10^{-7}
T	integration duration (s)	10^{-2}
Auditorimotor map		
d_r	low-dimension for manifold learning	2
k	neighborhood order	12

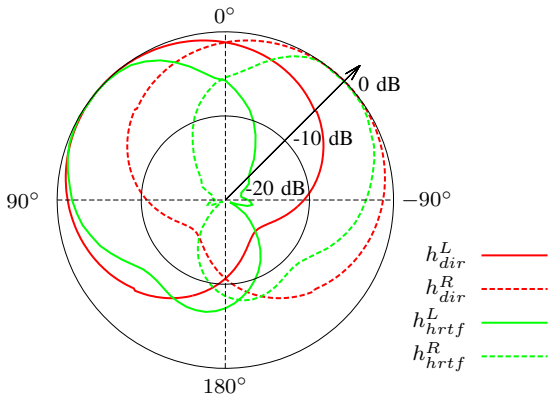


Fig. 2. Outer ears azimuthal directivity. Normalized root mean square response of the two binaural outer ear filters (h^L_{dir}, h^R_{dir}) and (h^L_{hrtf}, h^R_{hrtf}) in function of the listener-to-emitter direction θ (in deg) for a reference sound signal $x^e(t)$ as in (11). Listener front direction correspond to $\theta = 0°$.

In order to stress the relevance of spectral cues for front-back disambiguation, manifolds are learned from real and simulated data, all consisting of random spectrum sources but differing in the transmission channel and the contained cues. Therefore real data comes from the CAMIL dataset [16], a collection of binaural dummy head recordings that include room reverberation, head shadowing and outer ear filtering. Contrary to these recordings, the simulated data are emitted in a simulated anechoic environment and filtered by one of the two outer ears model used here. Firstly, h^L_{dir} and h^R_{dir} are purely directive filters and their responses are independent of the source spectrum and therefore does not provide any spectral cues. Secondly, h^L_{hrtf} and h^R_{hrtf} consist of head-related transfer function (HRTF) measurements of a dummy head endowed with human-like pinna [28] including auditory spectral cues. Although they differ in their spectral cues, the two outer ear filters have comparable properties in azimuthal directivity, as shown in Fig. 2. In more details, and considering the emitter position relatively to the left and right ears, we note the listener-to-emitter left and right distance and azimuth d^L, d^R and θ^L, θ^R respectively. The left and right perceived sounds x^L and x^R are function of the listener-to-emitter angle and distance and correspond to:

$$x^L(t) = \frac{h^L(\theta^L) * x^e(t)}{d^L} \text{ and } x^R(t) = \frac{h^R(\theta^R) * x^e(t)}{d^R}, \qquad (12)$$

where $h^L(\theta)$ and $h^R(\theta)$ represent the left and right outer ear filters and $x^e(t)$ the input signal as defined in (11).

Manifolds learned from ILD vectors in the three different configurations are shown Fig. 3. The results show that the manifolds are able to retrieve the left-right direction of a sound source for the three proposed configurations. Moreover this experiment show that intensity cues alone are not sufficient for front-back disambiguation and that spectral cues from the HRTF are needed for a complete disambiguation. Note that the dimensionality of the input space for manifold learning is equal to the 30 frequency channels of the gammatone filterbank output. A comparable

(a) Simulation. Filters (h^L_{dir}, h^R_{dir}). (b) Simulation. Filters (h^L_{hrtf}, h^R_{hrtf}). (c) Dummy head CAMIL recordings.

Fig. 3. Manifold learning and front-back disambiguation. Three manifolds learned with Laplacian eigenmaps from a set of 2000 sound sources in the azimuthal range $[-180, 180]$ (color bar). (a) From simulated data and directive ear filters, the left-right direction of the source is well represented but the front-back position is ambiguous. (b) Adding spectral cues to the simulated auditory vectors leads to a complete front-back disambiguation. (c) The same is true for real data from the CAMIL dataset. For simulated data the integration duration is equal to 10^{-2} s whereas it is equal to 10^{-1} s for real data.

study [16] based on short-term Fourier analysis and ILD vectors show that a minimum of 40 frequency channels is sufficient for manifold learning in azimuth and elevation, with a mean angular error about 2 degrees.

B. Orienting Behavior and Auditorimotor Map Learning

Considering the configuration of simulated sound sources and directive outer ear filters h^L_{dir} and h^R_{dir} described above, this set of experiments evaluates the learning and the evolution of localization error in the azimuth range $[-90, 90]$. Firstly, given an environment state $e \in \mathcal{E}$ and a motor state $m \in \mathcal{M}$, we define the absolute localization error δm as the absolute difference between the real azimuth of the source and the head angle corresponding to the OB final motor state so that $\delta m = |e_\theta - m_\theta|$. The mean error δm obtained from a set of 1000 OB final motor states is equal to 0.46 deg (standard deviation of 0.31) and the 2D representation of the initial and final sensory states of these different OB experiments are showed Fig. 4. The manifold is learned from the initial states through Laplacian eigenmaps. The final states, projected on the manifold with the out-of-sample extension, are all projected on a region indicating a centered sound source, thus providing an experimental evidence of the existence of a reference state as introduced in (2).

Focusing now on the auditorimotor map learning algorithm, we executed it on 400 iterations corresponding to 400 different auditory experiences in the azimuthal range $[-90, 90]$ with the outer ears filters (h^L_{dir}, h^R_{dir}). Fig. 5 presents the detailed evolution of the learning process for two outlier significance levels $\alpha = 0.01$ and $\alpha = 0.05$. The value of α fix a compromise between the learning speed and the learning precision and it is show that the auditorimotor map is learned with success, with a mean localization error about 1 degree for $\alpha = 0.05$ after about 200 iterations. Related approachs that learn an auditory or auditory-visual reflex on a robotic model require from 2000 to 10000 iterations to converge [5], [9], [18].

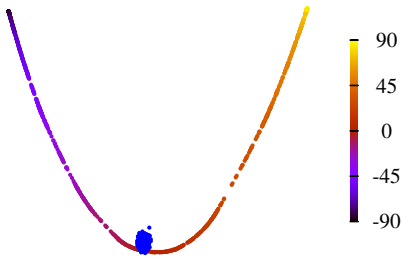
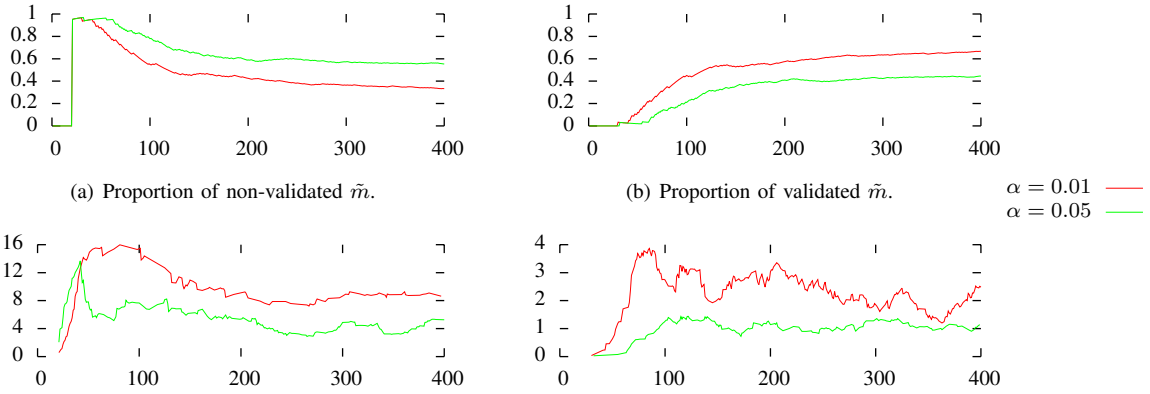


Fig. 4. Manifold learning and orienting behavior. A manifold learned with Laplacian eigenmaps computed from a set of 1000 auditory vectors of azimuth in $[-90, 90]$ corresponding to OB initial sensory states (color bar) and projection of OB final sensory states on the manifold (blue cluster).

IV. Discussion

The algorithm introduced in this paper is presented in a simple form. Some further improvements should significantly increase its performances. Firstly, the manifold is learned on the initial sensory states only, so that including the intermediate states in the learning process should reduce the amount of auditory experiences required to build the auditorimotor map. Secondly, the reduction dimension was done with Laplacian eigenmaps, a method requiring the memorization of the high-dimensional set of points S_{init} for the update of the auditorimotor map. This is costly in terms of memory storage and seriously reduces the plausibility of such a learning process in a biological system. This problem can be addressed using a different manifold learning algorithm, such as self-organized maps [17]. Finally the estimation of the motor state $\tilde{m}$ is given here as a simple inverse distance interpolation whereas, at each point, the manifold is shaped along privileged directions, so that adding a directional factor in the interpolation function might improve the estimation [26]. The approach is used here in a simple context and could be extended to more complex motor and sensory spaces, for exemple to phonotaxis behavior [7], temporal auditory cues (well known as ITD), visuo-auditory environments [9] and adaptation to sensory alterations [5]. The ILD based auditory model and orienting behavior presented here have been implemented on a robotic platform [7], and manifolds have successfully been learned from real data, so the method presented here should address promising applications in active perception for autonomous systems.



(a) Proportion of non-validated $\tilde{m}$.

(b) Proportion of validated $\tilde{m}$.

(c) Localization error for non-validated $\tilde{m}$ (in deg).

(d) Localization error for validated $\tilde{m}$ (in deg).

Fig. 5. Self-supervised auditorimotor map learning. Over 400 auditory experiences, evolution of the learning process and prediction error for different outlier significance levels α. (a) Proportion of non-validated $\tilde{m}$ estimations (*i.e.* % of 'yes' in the second outlier detection). (b) Proportion of validated $\tilde{m}$ estimations (% of 'no'). (c - d) Localization error for non-validated and validated $\tilde{m}$ estimations. The proportion of direct OB (*i.e.* % of 'yes' in the first outlier detection), not plotted here, is about 6 % and almost occurs during the firsts iterations.

V. Conclusion

This paper applied the sensorimotor approach to sound source localization and proposed a new method for auditorimotor map learning based on an auditory evoked behavior. Based on intensity cues computed by a binaural auditory system, this behavior allows a naive agent to learn a low-dimensional auditorimotor map and to estimate the azimuthal position of new auditory experiences. Our results shows that the auditorimotor map is learned with success and provide accurate estimation of azimuthal sources direction.

References

[1] L. Kneip and C. Baumann, "Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis." *Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3108–3019, 2008.

[2] Y.-C. Lu and M. Cooke, "Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners," *Speech Communication*, vol. 53, no. 5, pp. 622–642, 2011.

[3] R. Kearsley, "The newborn's response to auditory stimulation: A demonstration of orienting and defensive behavior," *Child Development*, vol. 44, no. 3, pp. 582–590, 1973.

[4] D. Muir, R. Clifton, and M. Clarkson, "The development of a human auditory localization response: A U-shaped function," *Canadian Journal of Psychology*, vol. 43, no. 2, pp. 199–216, 1989.

[5] M. Rucci, G. Edelman, and J. Wray, "Adaptation of orienting behavior: from the barn owl to a robotic system," *IEEE Transactions on Robotics and Automation*, vol. 15, no. 1, pp. 96–110, 1999.

[6] A. Horchler, R. Reeve, B. Webb, and R. D. Quinn, "Robot phonotaxis in the wild: A biologically inspired approach to outdoor sound localization," *Advanced Robotics*, vol. 18, no. 8, pp. 801 – 816, 2004.

[7] 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*, 2010, pp. 118–122.

[8] S. Andersson, A. Handzel, V. Shah, and P. Krishnaprasad, "Robot phonotaxis with dynamic sound-source localization," in *IEEE International Conference on Robotics and Automation*, vol. 5, 2004, pp. 4833–4838.

[9] L. Natale, "Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head," *Robotics and Autonomous Systems*, vol. 39, no. 2, pp. 87–106, 2002.

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

[11] J. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behavioral and brain sciences*, pp. 939–1031, 2001.

[12] D. Philipona, J. O'Regan, and J. Nadal, "Is there something out there? Inferring space from sensorimotor dependencies," *Neural Computation*, vol. 15, pp. 2029–2049, 2003.

[13] C. Couverture and B. Gas, "Extracting space dimension information from the auditory modality sensori-motor flow using a bio-inspired model of the cochlea," in *IEEE International Conference on Intelligent Robots and Systems*, no. 1, 2009, pp. 2742–2747.

[14] A. Laflaquiere, S. Argentieri, B. Gas, and E. Castillo-Castaneda, "Space dimension perception from the multimodal sensorimotor flow of a naive robotic agent," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1520–1525.

[15] M. Aytekin, C. Moss, and J. Simon, "A sensorimotor approach to sound localization," *Neural Computation*, vol. 20, pp. 603–635, 2008.

[16] A. Deleforge and R. Horaud, "Learning the direction of a sound source using head motions and spectral features," *INRIA Research Report*, no. 7529, 2011.

[17] J. Hörnstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots-building audio-motor maps based on the HRTF," in *IEEE International Conference on Intelligent Robots and Systems*, 2006, pp. 1170–1176.

[18] E. Berglund, J. Sitte, and G. Wyeth, "Active audition using the parameter-less self-organising map," *Autonomous Robots*, vol. 24, no. 4, pp. 401–417, 2008.

[19] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.

[20] J. Blauert, *Spatial Hearing*. MIT Press, 1997.

[21] R. Patterson, I. Nimmo-Smith, and J, "An efficient auditory filterbank based on the gammatone function," *Institute of Acoustics on Auditory Modelling Report*, no. December, pp. 14–15, 1987.

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

[23] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[24] L. Van der Maaten, E. Postma, and H. Van den Herik, "Dimensionality reduction: a comparative review," *Tilburg University Technical Report*, no. 2009-005, 2009.

[25] Y. Bengio, J. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering," in *Advances in Neural Information Processing Systems*, vol. 16, 2003, pp. 177–184.

[26] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," *ACM national conference*, pp. 517–524, 1968.

[27] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, pp. 1–21, 1969.

[28] W. Gardner, "HRTF measurements of a KEMAR dummy-head microphone," *Massachusetts Institute of Technology*, vol. 97, no. 6, pp. 3907–3908, 1994.