The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Extracting space dimension information from the auditory modality sensori-motor flow using a bio-inspired model of the cochlea

Charlie Couverture and Bruno Gas

Abstract— **First task robots have to realise is sensing and acting in the environment. Can a robot learn the way it is able to sense and act in the world without any hardwired notions? Is it able to learn it from the only data he has access to, that is high-dimension sensory inputs and motor outputs? This paper presents experimental results obtained on a simulated human listener using a bio-inspired model of the cochlea and real records from human related transfer functions (HRTF). These results show that a naive system that interacts with its environment without knowing the laws governing these interactions can discover information about dimensionality of space. Moreover, the laws determining the sensations of the system as a function of the state of the system and the environment, called the "sensorimotor law", are not simplified as usually in simulations. They are bio-realistic as they are determined by the HRTF recorded on human beings.**

Keywords : sensorimotor contingencies, auditive sensorimotor flow, cochlea, space dimension.

I. Introduction

In mobile robotics perception is essential to achieve task such as navigation, obstacle avoidance, source localisation, or other task which requires to interact with the environment. Currently perception for mobile robotics is mainly passive: the environmental state projects itself on the robot sensors. This projection will be analysed to extract information about environnement state. This approach of perception requires the use of models for the sensors and the environment. An alternative way to this passive explanation of perception is the theory of *sensorimotor contingencies* proposed by O'Regan and Noe [1]. This theory integrates action in perception and explains information extraction by the properties of the dependancies between sensations and action instead of sensors inputs only. The main interest of this kind of theory applied to robotics perception tasks is to find solutions to complex problems in which traditional perception algorithms fail: fast navigation in complex or unstructured environments for instance [2] or autonomous and adaptative comportments in unknown environments (it should be directly of interest for roboticists concerned with unsupervised learning).

Poincaré [3] asked the question of what we can extract from our sensor and motor signals (the sensorimotor contingencies) that can make us understand and move in the surrounding environment. He wrote: *To localize an object simply means to represent to oneself the movements that would be necessary to reach it. It is not a question of representing the movements themselves in space, but solely of representing to oneself the muscular sensations which accompany these movements and which do not presuppose the existence of space*. Following this idea Philipona [4] proposed an algorithm in which the brain of a simulated organism with arbitrary input and output connectivity can infer the dimensionality of the rigid group of the space underlying its input-output relationship, that is the dimension of what the organism will call physical space.

We propose in this article to extend Philipona's work [4] by showing how a system with bio-inspired auditive sensors can deduce dimension of the space from its interaction with its environnement. Aytekin et al. [5] demonstrate quantitatively that the experience of the sensory consequences of its voluntary motor actions allows an organism to learn the spatial location of any sound source. The authors cite the demonstration made by Philipona et al. but their approach assumes that in far field auditory space is two dimensional. By doing that they abandoned the key hypothesis of the brain having absolutely no a priori information about outside physical space (whether it exists at all, whether it has a metric, whether it is euclidean, how many dimensions it possesses). On the contrary our hypothesis is that according to Philipona's and O'Regan work what a biological organism perceives as the dimensionality of space can be inferred without any a priori knowledge from the laws linking the brain's inputs and outputs.

The article is divided into three parts. We first give some understanding keys about the mathematical background. Secondly we describes our system and finally we present our experimentations.

II. Problem formalization

In what follows we consider E as the state of an environment of dimension e and $\mathcal{E}$ as the manifold of all the possible states of the environment so that $E \in \mathcal{E}$. We consider S as the sensory input vector of dimension s of a robotic system taking place in this environment and $\mathcal{S}$ the manifold of all the possible sensory input so that $S \in \mathcal{S}$. We consider M the output vector of the robotic system, that is its motors command vector of dimension m, and $\mathcal{M}$ the manifold of all the possible output of the system with $M \in \mathcal{M}$.

The sensory input is determined by both the environment state and the current motor state so that there is a functionnal relationship between the manifolds $\mathcal{E}$, $\mathcal{M}$ and $\mathcal{S}$ that is called "sensorimotor law":

$$S = \Phi(M, E) \tag{1}$$

C.Couverture is with the ISIR laboratory, UPMC Univ Paris 06, 4, place Jussieu, 75005, Paris Charlie.Couverture@isir.upmc.fr

B.Gas is with the ISIR laboratory, UPMC Univ Paris 06, 4, place Jussieu, 75005 Paris Bruno.Gas@upmc.fr

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

It means that we are not using a motor control approach $M = \phi(S)$, but on the contrary an approach based on the observation of sensory consequences of motor commands. We shall focus on the tangent space $\{dS\}$ of S at some point $S_0 = \phi(M_0, E_0)$. Following Philipona's argumentation two natural subspaces $\{dS\}_{dE=0}$ and $\{dS\}_{dM=0}$ can be identified in $\{dS\}$:

$$dS = \left.\frac{\partial \phi}{\partial M}\right|_{(M_0,E_0)} .dM + \left.\frac{\partial \phi}{\partial E}\right|_{(M_0,E_0)} .dE \qquad (2)$$

such that $\{dS\} = \{dS\}_{dE=0} + \{dS\}_{dM=0}$ where $\{dS\}_{dE=0}$ is the vector subspace of sensory input variations due to a motor change only and $\{dS\}_{dM=0}$ the input variations due to an environment change only. More precisely $\{dS\}_{dE=0}$ and $\{dS\}_{dM=0}$ are the tangent spaces at the point S_0 of manifolds $\phi(E_0, \mathcal{M})$ and $\phi(\mathcal{E}, M_0)$ of sensory inputs obtained through variations of respectiveley M only and E only. Let $C(M_0, E_0)$ be the intersection of $\{dS\}_{dE=0}$ and $\{dS\}_{dM=0}$. A non empty intersection means that their exists some perceptual changes arround S_0 that can be obtained equally either from dE or from dM, that is the so called *compensable movements* of Poincaré. Our main objective is to determine the dimension d of the space of the compensated movements.

One has:

$$\begin{aligned} \dim(C(E_0, M_0)) \quad &= \dim(\{dS\}_{dM=0}) \\ &+ \dim(\{dS\}_{dE=0}) \\ &- \dim(\{dS\}) \end{aligned} \qquad (3)$$

which can be writen as $d = p + e - b$ where $d = \dim C(E_0, M_0)$ that is the dimension of the compensable movements. When the robot is stationary dimension of subspace $\{dS\}_{dM=0}$ gives the number e of variables necessary for a local description of the environment. When the environment is stationary dimension of subspace $\{dS\}_{dE=0}$ gives the number p of variables necessary to describe the variations of input signals due to robot motions. When both M and E vary certain input changes can be obtained either from dE or dM such that the dimensionality of $\{dS\}$ is lower than the sum of dimension of $\{dS\}_{dM=0}$ and $\{dS\}_{dE=0}$ (see eq. 3). We used Principal Component Analysis (PCA) to estimate those numbers e, p and b.

In Philipona's simulation [4], [6] perceptual modalities were both visual, tactile and auditive but with simple simulated sensors. In this article we propose to use much more realistic datas and sophisticated sensors but only in the field of auditory sensing with the pending question: can an organism identify the dimension of the auditory space throught its interaction with it? Giving a positive answer to the previous question Aytekin et al. [5] assumed that the auditory space dimension were known. That is two-dimensional because only two parameters are necessary to identify a sound source location in auditory space in the far field, as we consider an isotropic sound source. As Aytekin et al. we carry on Philipona's work by adapting the example given by Poincaré for visual perception of space to the auditory perception of space. But our task is not to localize sound source position. We only want to determine auditory space dimension since we consider this question as a preliminar.

III. System description

A. General overview

Our system simulates the audition of multiple sound sources by a simplified human listener as shown on Fig. 1. It means that the simulated listener is able to operate movements of its own head and that it perceives the sensations due to the sound sources in the environment. The sources are placed on a 1-meter diameter sphere centered on the head. The position of a source with relation to the reference frame requires two angles, θ and ϕ. The orientation of the head with relation to the reference frame requires three parameters α, β and γ. The acoustic sensory inputs received by the listener are computed as follow. Firstly signal of each sound source is filtered by the auricle of each ear of the listener. The filter function operated by the outer ear is a function of the direction of the incomming signal and is modelized by the Head Related Transfer Function (HRTF). Secondly the signal filtered by the HRTF as it appears at the entrance of the canal is encoded by the cochlea. The model of the cochlea that we use consists of an array of independent bandpass filters. The energy output of each filter of both ears is computed on fixed length frames and constitute the elements of the sensory input vector S. The HRTF phase information (available to humans up to 3 kHz) is ignored to simplify analysis.

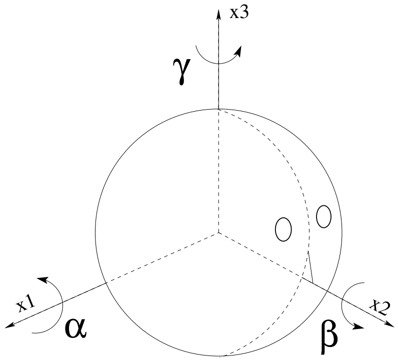

Fig. 1. View of the head of our simulated listener with the angles of rotation as parameters for its movements.

As described in section II $\mathcal{E}$ is the manifold of all the possible states of the sound sources (time, frequency and spatial properties). The motor states (e.g. governing the head orientation) are elements of the manifold $\mathcal{M}$. All the acoustic sensory inputs received at the ears are parts of the manifold $\mathcal{S}$.

Our goal is to estimate the intrinsic dimension s of the sensory input vector S. To achieve this our simulated listener moves or the environment moves or both while sensory inputs S are recorded. Dimension s of sensory input is estimated by computing a Principal Component Analysis (PCA) on the covariance matrix of all the iterations of S.

In the following sections we describe our simulated listener with more details.

B. Head-Related Transfer Function

The Head Related Transfer Function describes the interaction between the sound of the source and the outer ear, the head and the torso of the listener. Due to the physical geometry of the body of the listener the HRTF depends on the direction of the incomming sound. Fig. 2 presents left and right HRTF for a source positioned in front of the listener. For a given source at azimuth θ and elevation ϕ the acoustic responses $S_l(f)$ and $S_r(f)$ received at the left and right ears respectively are:

$$\left\{ \begin{array}{l} S_l(f) = A(f).H_l(f,\theta,\phi) \\ S_r(f) = A(f).H_r(f,\theta,\phi) \end{array} \right. \tag{4}$$

where the H_is represent the HRTF of left and right ears and $A(f)$ represents the spectral composition of the source.

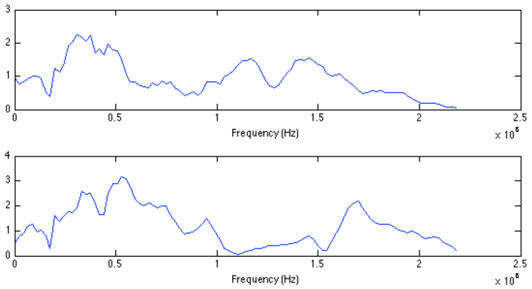

Fig. 2. Left (top) and right (bottom) HRTF corresponding to a source situated in front of the listener

The CIPIC HRTF Database [7] is a public-domain database of high-spatial-resolution HRTF measurements for 45 different subjects, including the KEMAR mannequin (a tool used to analyze how sound waves interact with the human body) with both small and large auricle. The database includes 1250 measurements of head-related impulse responses for each subject. These "standard" measurements were recorded at 25 different interaural-polar azimuths (from $-80°$ to $+80°$) and 50 different interaural-polar elevations (from $-45°$ to $+230.625°$) (see [7] for additional details).

Because we want to estimate the tangent space dimension (eq. 3) at a point S_0 we need to limit the spatial motor actions to small head movements arround a fixed head orientation. As a consequence HRTFs need to be interpolated if we want to retrieve a particular azimuth and elevation which is not one of the values present in the database. We used for that a two dimensional linear interpolation method based on matlab conventional interpolation functions.

C. Cochlear coefficients

The signal filtered by the HRTF is the signal of the source as it appears at the entrance of the auditory canal of the ear. According to Patterson's model [8] the cochlea is seen as tonotopically organized filters from high frequencies at the base of the cochlea to low frequencies at the apex. In Patterson's model the bandwidth of each cochlear filter is described by an Equivalent Rectangular Bandwidth (ERB) [9] using a gamma-tone filter [10]. A critical band or ERB filter models the signal that is present within a single auditory nerve cell or channel. Cochlear channels are spaced so that each filter overlaps its neighbors by the same amount (see Fig. 3). We used the gamma-tone filter implementation of Malcolm Slaney's Auditory Toolbox [11], [12]. The lowest frequency center was 100Hz and the highest was 18 kHz. We used 40 filters for each ear, leading to a 80 coefficients vector, each of them representing the energy computed on the filter output signal (see Fig. 4 for an example of four filter responses to a gaussian noise input signal). Fig. 5 presents the 40 cochlear coefficients of the left ear computed as the white noise signal of a sound source is filtered by the left HRTF corresponding to a position of the source right in front of the listener.

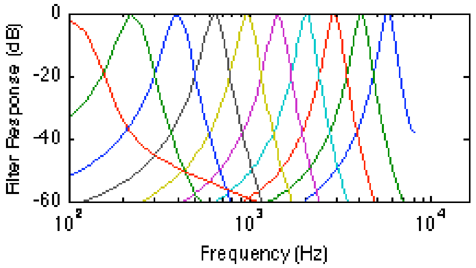

Fig. 3. An example of a gamma-tone filter bank with 10 filters from 100 Hz to 8000 Hz.

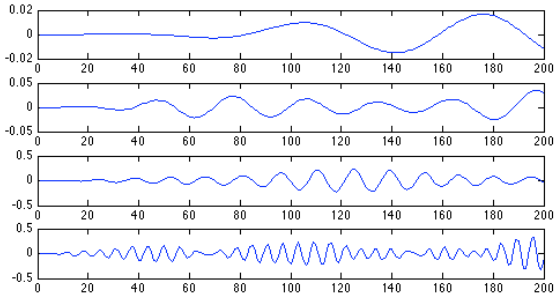

Fig. 4. Output of filters 8, 15, 22 and 30 (top to bottom) of the cochlea we use for our experiments. Input signal is a white gaussian noise.

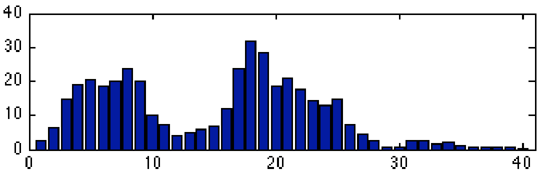

Fig. 5. Cochlear coefficients of the left ear computed for a gaussian noise presented as the signal of a sound source positioned in front of the listener.

D. Estimation of the dimension of the movement.

Let us consider $\{dM\}$ as a set of 10 motor commands and $\{dE\}$ a set of 10 environmental positions (which are in our case sound source positions). Our objective is to observe whether head movements dimension can be extracted from the perceptual flow, e.g. variations of auditive sensations due to the relative movement between the head and the source.

Motors commands or sound sources movements produce 10 sensory inputs changes each: $\{dS\}_{dE=0}$ and $\{dS\}_{dM=0}$. We want to compute eigenvalues of the covariance matrix of these sample sets. Eigenvalues should fall into two classes : significant values (non-zero values) and unsignificant values (near zero). The number of significant eigenvalues gives an estimation of the intrinsic dimension of the embedded samples. We used the same method as Philipona in [4] to distinguish these two classes. Finding the boundary between the two classes can be done by ordering the λ_i in decreasing order, and locating the value of i such that the ratio between λ_i and λ_{i+1} is largest (see fig.6). More precisely, we used:

$$dim = \arg\max \frac{Eigv(i)}{Eigv(i+1)} \tag{5}$$

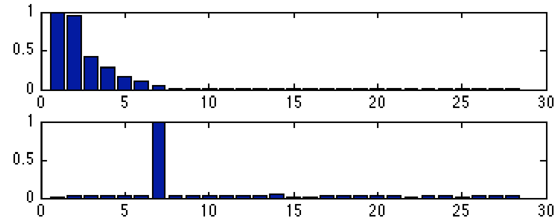

Fig. 6. Top: eigenvalues obtained for 2-D movement (normalized scale). First seven values are significant while others are not. Bottom: ratio between eigenvalue i to eigenvalue $i+1$, where i is the eigenvalue index (normalized scale). Here the estimated dimension is 7.

IV. Experimental results

We report here 4 experiments we have conducted to show that intrinsic dimension of the sensory inputs corresponds to the dimension of the action of the environment or the simulated listener. In other words, we aim at showing that the sensorimotor law, describing the interaction between a system and its environment, has properties that permit to retrieve information about the environment without any model of it.

A. Movement of the head in one direction. One source of white gaussian noise.

The first experiment consists in moving the head in one direction, while there is one fixed sound source in the auditory space. The head moves 10 times by 1° by steps of $0.1°$ around its initial position. Thus we have a set $\{dM\}$ of ten infinitesimal movements resulting in $\{dS\}$ sensory responses available to estimate their intrinsic dimension. The experience is repeated for several positions of the source belonging to the interval $[-25°; 25°]$ for θ and $[-50°; 50°]$ for ϕ with steps of 2°. We have reported on table I the rate of good estimation. This experimentation has been realised three times, one for each direction α, β and γ. The result obtained for the β direction movement of the head is weaker than for the two others. Fig. 7 presents the map of sound source positions which have been correctly estimated, e.g. $dim = 1$. One can notice that the β movement is the rotation of the head around the axis passing by the centre of the head and the position of the source at azimuth 0 and elevation 0. As a consequence, movements of the head when the source is near the central position result in very small perceptual variations which could explain why dimension is badly estimated for source positions around the center of the rotation. Fig. 8 presents the first 10 eigenvalues extracted from the covariance matrix of the 80 cochlear coefficients obtained for ten rotation movements of the head in the direction α for a position of the source. The dimension is correctly estimated ($dim = 1$).

TABLE I

Results of dimension estimation for a movement of the head in one direction with one motionless source.

Dim.	direction		
	α	β	γ
1	**98.1%**	**92.0814%**	**99.3967%**
2	1.81%	3.2428%	0.6033%
3	0%	1.810%	0%
4	0%	1.5083%	0%
5	0.08%	1.3575%	0%
6+	0%	0%	0%

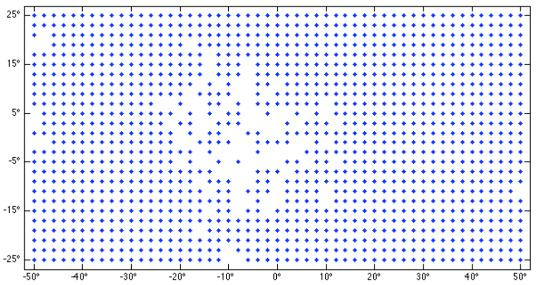

Fig. 7. Map of the positions of the source for a movement of the head in the direction β. Points represent position for which estimated dimension is 1.

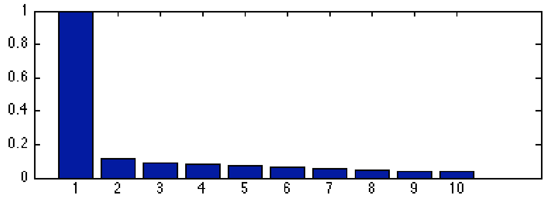

Fig. 8. Eigenvalues obtained for a 1-D movement in α. Estimated dimension of the movement is 1. Source is positioned at $(7, 13)$.

B. Movement of the head in one dimension. Two sound sources

The second experiment consists in moving the head in one direction, as described previously, but with two sound sources of white gaussian noise instead of one. The dimension is estimated for each combination of positions of both sources. The experiment is repeated for the three possible movements of the head α, β and γ. This experiment aims at showing that even if two sources are present, the “dimension” of the signal is still related to the dimension of the movement and not to the number of sources. As the sources are static and the head is moving, the relative movements of the sound sources with relation to the head are the same. So the movement can be described by only one parameter and the expected result of the estimated dimension is $dim = 1$. We placed the first source in four random positions while the second source takes positions in the intervals $[-25°; 25°]$ for elevation and $[-50°; 50°]$ for azimuth with steps of 2° for each position of the first source. Table II presents the obtained results. The dimension has been estimated for the 5304 combinations of both sources positions. For the movement of the head in the direction α, we obtain good estimation of the dimensions (83%). For direction β, 85% of the sound source positions are correctly estimated. The movement of the head in the direction γ induced a weaker score than other directions. Only 66% of the positions give rise to good estimates while 32% give an incorrect estimate ($dim = 2$). The results were totally different between one position of the first source ($(-3°, 8°)$ leading to 10% of good estimates) and the three others (leading each of them to 86.20%, 83.86% and 86.20% of good estimates).

TABLE II

ESTIMATION RESULTS FOR MOVEMENTS OF THE HEAD IN ONE DIRECTION WITH TWO SOURCES OF NOISE.

Dim.	direction		
	α	β	γ
1	**83.2956%**	**85.0867%**	**66.7044%**
2	16.6101%	14.6305%	32.8997%
3	0.0943%	0.1320%	0.3959%
4	0%	0.0566%	0%
5	0%	0.0943%	0%
6+	0%	0%	0%

C. Movement of the head in two directions with one source of white noise.

This experiment consists in moving the head in two directions, while one source of white gaussian noise is present. It aims at showing in what extent the algorithm can estimate 2D movements with only one sound source. The head makes a movement of 1° in one direction then 1° in another direction by steps of 0.1°, starting with the “nose” pointing to azimuth 0 and elevation 0. Table III reports results for positions of the sound source within $[-25°; 25°]$ for elevation and $[-50°; 50°]$ for azimuth with steps of 2°. Total number of estimations is 1326. Expected result is $dim = 2$ because two parameters are sufficient to describe the movements of the head with relation to the source. One can see that the ratio of well estimated dimension is of the same order that the one obtained for 1D movements of the head (see table II).

TABLE III

ESTIMATED DIMENSION FOR A HEAD MOVEMENT IN TWO DIRECTIONS WITH ONE SOURCE OF NOISE.

Dim.	% of result
1	11.4630%
2	**83.8612%**
3	4.6757%
4+	0%

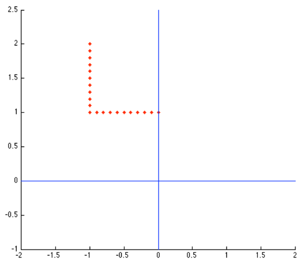

Fig. 9. Successive positions of the source with relation to the head for a 2-directions movement of the head. Vertical and horizontal lines are azimuth 0 and elevation 0.

D. Movement of the head in two directions with two sources of noise.

This experiment is similar to the one with a movement of the head in one direction and two sources of white noise, except that the head makes a movement in two directions. Table IV presents the results.

TABLE IV

ESTIMATED DIMENSION FOR A HEAD MOVEMENT IN TWO DIRECTIONS WITH TWO SOURCES OF NOISE.

Dim.	% of result
1	6.8627%
2	**61.7647%**
3	28.3560%
4	2.8658%
5	0.1508%

E. Independant movement of two sources of white noise. Head is motionless.

TABLE V

RESULTS OBTAINED FOR EXPERIMENTS 1 AND 2.

Dim.	exp.1	exp.2
1	2.07%	2.9586
2	**68.6391%**	13.6095%
3	17.4556%	5.6213%
4	11.2426%	**46.1538%**
5	0.5917%	31.6568%
6+	0%	0%

The first experiment consists in estimating the dimension of the signal recorded by our simulated listener when two sources of white noise are present and move in one direction each. The head is motionless and its “nose” is pointing to azimuth 0 and elevation 0. The first source is initially positioned at azimuth $10°$ and elevation $6°$ and makes a move of $1°$ in one direction then in another direction with steps of $0.1°$, while the second source is initialy positioned in an area delimited by $[-50°; 50°]$ for azimuth and $[-25°; 25°]$ for elevation with step of $2°$. The experiment is repeated for each position of the second source within this area. Dimension has been estimated 1326 times. The expected dimension is 2 because two parameters are required to describe the variations of the cochlear coefficients occuring due to the movement of the sources. In other words, two parameters are enough to describe the position of the two sources as they move in only one direction each. In more than 68%, the dimension estimated is 2.

The second experiment is similar to the first one, except that the movement described by the sources is in two directions. The expected dimension is 4 because 4 parameters are enought to describe the position of the sources. Thus, the variations induced by the movement should require four parameters to be described. Table V presents results obtained for experiments 1 and 2. In 46% of the positions of the second source, the dimension is well estimated. But for 54% of the positions, a wrong dimension is estimated. We postulate that this is due to the too large amplitude of movements.

V. Conclusions — perspectives

We used a model of the human cochlea and records of head-related transfert function of a human subject to simulate the audition by a listener of moving sources. We used an algorithm to extract spatial information from the auditive modality flow. The conducted experiments have shown that property of the action of a simulated system can be retrieved from the sensory inputs only. This is of direct interest for roboticists since this approach of perception does not rely on a model of the sensor, of the environment, or of the morphology of the system. Not having a model to update but on the contrary look for information every time the agent need it, permits to have up-to-date datas.

Moreover, as there is no assumption on the structure of the datas, multimodal fusion is potentially easier to achieve. Input sensory datas from each modality are all related to action of the agent, and to the state of the environment in a coherent way, e.g. a movement of the head has consequences on auditive, visual, tactile, proprioceptive inputs.

We are currently working on estimation of the dimension with multimodal signals. Future results will give us the opportunity to show wether the Poincaré hypothesis concerning space dimension can be verified from real audio and video datas.

References

[1] J. K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behav. Brain. Sci.*, vol. 24, no. 5, pp. 939–973; discussion 973–1031, 2001.

[2] D. Lhomme-Desages, C. Grand, F. Ben Amar, and J.-C. Guinot, "Doppler-based ground speed sensor fusion and slip control for a wheeled rover," *IEEE/ASME Transactions on Mechatronics*, vol. PP, no. 99, pp. 1–9, 2009.

[3] H. Poincaré, *The foundations of science; Science and hypothesis, the value of science, science and method.*, . (G.B. Halsted, trans. of La Valeur de la science, Ed. New York: Sciences Press., 1929.

[4] D. Philipona, J. K. O'Regan, and J.-P. Nadal, "Is there something out there? inferring space from sensorimotor dependencies," *Neural Comp.*, vol. 15, no. 9, pp. 2029–204, 2003.

[5] M. Aytekin, C. F. Moss, and J. Z. Simon, "A Sensorimotor Approach to Sound Localization," *Neural Comp.*, vol. 20, no. 3, pp. 603–635, 2008.

[6] D. Philipona, J. K. O'Regan, J.-P. Nadal, and O. J.-M. Coenen, "Perception of the structure of the physical world using unknown sensors and effectors," *Advances in Neural Information Processing Systems*, vol. 15, 2004.

[7] V. Algazi, R. Duda, R. Morrisson, and D. Thompson, "The cipic hrtf database," *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to audio and Acoustics*, pp. 99–102, 2001.

[8] R. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," *Advances in Speech, Hearing and Language*, vol. 3, pp. 547–563, 1996.

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

[10] K. Robenson, R. Patterson, and J. Holdsworth, "Complex sounds and auditory images," in *Processings of Auditory Physiology and Perception*, vol. 9, 2004, pp. 429–446.

[11] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Apple Computer, Tech. Rep., 1993.

[12] ——, "A matlab toolbox for auditory modeling work," Interval Research Corporation, Tech. Rep., 1998.