Learning the Direction of a Sound Source Using Head Motions and Spectral Features

Antoine Deleforge, Radu Horaud

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Learning the Direction of a Sound Source Using Head Motions and Spectral Features

Antoine Deleforge and Radu Horaud

N° 7529

February 2011

Domaine 4

ISSN 0249-6399 ISRN INRIA/RR--7529--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

Learning the Direction of a Sound Source Using Head Motions and Spectral Features

Antoine Deleforge * and Radu Horaud

Domaine : Perception, cognition, interaction

Équipe-Projet PERCEPTION

Rapport de recherche n° 7529 — February 2011 — 29 pages

Abstract: In this paper we address the problem of localizing a sound-source by combining binaural or monaural spectral features with head movements. Based on a number of psychophysical and behavioral studies suggesting that the problem of spatial hearing is both listener-dependent and dynamic, we propose to address the problem at hand within the framework of unsupervised learning. More precisely, our method is able to retrieve an intrinsic low-dimensional parameterization from the high-dimensional spectral representation of the acoustic input. We address both binaural and monaural spatial localization with both static and dynamic cues. We show that the recovered low-dimensional representations are homeomorphic to the two-dimensional manifold associated with the motor states of a robotic head with two rotational degrees of freedom. We describe the experimental setup and protocols allowing us to gather acoustic data sets with ground truth for both the emitter-to-listener directions and precise head motions. We validate our method using extensive experiments that consist in classifying acoustic vectors from a test set, based on manifold learning with a different training set. Our method strongly contrasts with current approaches in sound localization because it puts forward the role of learning.

Key-words: Sound source localization, dynamic auditory cues, manifold learning, sensorimotor integration.

This work was supported by the European project HUMAVIPS, under EU grant FP7-ICT-2009-247525.

* Corresponding author: Antoine.Deleforge@inrialpes.fr

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Apprendre la Direction d'une Source Sonore en Combinant Mouvements de Tête et Caractéristiques Spectrales.

Résumé : Dans ce papier, nous abordons le problème de la localization sonore en combinant les caractéristiques spectrales monaurales et binaurales des sons à des mouvements de tête. Partant de nombreuses observations psychophysiques et comportementales suggérant que le probléme de l'audition spatiale est à la fois dynamique et dépendante du sujet, nous proposons d'envisager le problème par le biais de l'apprentissage non-supervisé. Plus précisément, notre méthode permet de retrouver une paramétrisation intrinsèque en basse dimension à partir d'une représentation spectrale en haute dimension des données acoustiques. Nous traitons à la fois la localisation binaurale et monaurale, avec des indices statiques ou dynamiques. Nous montrons que les représentations en basse dimension obtenues sont homéomorphiques à la variété bidimensionelle associée aux états moteurs d'une tête robotique dôtée de deux degrées de libertée rotationels. Nous décrivons l'installation et les protocols expérimentaux qui nous ont permis de réunir un ensemble de données acoustiques, précisément annotées à la fois par la direction émeteur-récepteur et les mouvements de têtes. Nous validons notre méthode par des expériences approffondies consistant à classifier les vecteurs acoustiques d'un ensemble test, en se servant d'une variété apprise à partir d'un ensemble d'entraînement différent. Notre méthode contraste fortement avec les approches actuelles en localisation sonore car elle met en avant le rôle de l'apprentissage.

Mots-clés : Localisation sonore, indice audio dynamique, apprentissage de variété, intégration sensori-motrice.

Contents

1 Introduction, Related Work, and Contribution

The humans' instinctive capability of localizing one or several sound sources from the perceived acoustic signals has been intensively studied in cognitive sciences Blauert (1997). Nevertheless, both the existence and the full understanding of a sound localization pathway in the brain are still active topics of research and the exact anatomy and physiology of the auditory cortex is still under debate King and Schnupp (2007). This has also been actively investigated within the framework of *computational auditory scene analysis* (CASA) Wang and Brown (2006). A classical example that illustrates well the difficulty of the problem is the well known *cocktail party problem* (CPP) Cherry (1953); Haykin and Chen (2005): How do listeners manage to decipher speech in the presence of other sound sources, including competing talkers? We note that until today this auditory source separation problem has not received yet a fully satisfactory answer from both neurophysiological and computational perspectives. We believe that finding a proper solution to the problem of *three dimensional* (3D) sound localization is key to fully understanding every day situations which are often analogous to the CPP.

There is behavioral and physiological evidence that human listeners use interaural differences in order to estimate the direction of a sound. Two binaural cues seem to play an essential role, namely the interaural level difference (ILD) and the interaural time difference (ITD). The ITD, measured at the eardrum for broadband stimuli, is approximately constant in the frequency domain and it depends on sound source orientation in approximately the same way from subject to subject. Nevertheless, it is an ambiguous sound localization cue since a number of different sound directions could produce the same ITD value. Alternatively, the ILD is both subject-dependent and frequency-dependent. A number of computational models were developed for robust sound localization and sound tracking based on ITD and ILD Willert *et al.* (2006); Roman and Wang (2008). However, to the extent that both the head and the ears are symmetric, a stimulus presented at any location on the median plane should produce no interaural differences. Similarly, any point off this median plane falls on a *cone of confusion* Woodworth and Schlosberg (1965), upon which the overall interaural differences, either ILD or ITD, are constant. Therefore, the spatial information provided by interaural-difference cues, within a restricted band of frequency, is spatially ambiguous, particularly along a roughly vertical and front/back dimension Middlebrooks and Green (1991).

More elaborate sound localization models incorporate the *head related transfer function* (HRTF) and the *head-related impulse response* (HRIR). For example, azimuth estimation can be done by HRTF data lookup. Based on studying the HRIR parameters of each individual in a database of 45 subjects, Raspaud *et al.* (2010) proposes a generic model for the relationship between azimuth angle and ILDs and ITDs that only depends on one parameter (the distance between the ears). Interesting enough, they experiment with musical signals which have a more complex time-frequency behavior than speech. The idea of HRTF data lookup is also considered in Lu and Cooke (2010) in conjunction with the *direct-to-reverberant energy ratio* (DRR) that is used to measure the distance to a sound source. However, this method can only estimate the

listener-to-source distance with an accuracy of 1m for static sources and 1.5-3.5m for moving sources.

So far we considered a static listener. *Dynamic cues* for sound localization are associated with *head movements*. Based on behavioral data, it has been hypothesized some time ago that head motions might be useful for disambiguating potential confusions generated by the pinna's filter Wallach (1939, 1940). Other psychophysical experiments Thurlow and Runge (1967); Fisher and Freedman (1968); Pollack and Rose (1967) further support the idea that head movements are useful for localization. From a computational point of view Wenzel (1995) evaluated the theoretical contribution of ITDs and ILDs coupled with head motions; it was found that head movements help to solve location confusions considerably. In Muller and Schnitzler (1999a,b, 2001), based on an *acoustic flow* theory and on observations on bats, it is argued that the synthesis of dynamic cues, e.g., frequency and amplitude modulations, could allow the individuals to derive useful temporal cues for sound localization. This was tested in practice Handzel and Krishnaprasad (2002) by placing acoustic sensors in a binaural configuration such that a bio-inspired computational model can be derived. It was showed that the use of acoustic flow under head rotations helped to break the inherent symmetry of the binaural system and thus solve for location ambiguities.

In Walker *et al.* (1998) it is suggested that synthesizing different views (perspectives) by repositioning the pinnae could also break those symmetries. Kneip and Baumann (2008) applied this concept by gathering ITD values from different motor states with a two-microphone device with two degrees of freedom, thus allowing a robot to localize static and continuous sound sources in space with a precision of 10° and 0.5 meters.

More generally, the idea of using deliberate head motions for auditory scene analysis has received little attention from a computational point of view, in spite of psychophysical evidence that humans use dynamic cues for sound localization Blauert (1997). *Dynamic hearing* has recently become an emerging topic in robotics because it enables humanoid robots to localize sounds in order to interact with their environment. One advantage of robots over a static device, is that they can achieve precise goal-directed movements. This can be explored to learn the mapping between sound-source locations and observed acoustic signals for various head movements. This mapping can be learnt in a supervised manner, using a linear regression function as done in Hörnstein *et al.* (2006) or in an unsupervised way using a manifold learning technique as done in Aytekin *et al.* (2008). Indeed, rather than attempting to derive specific models for extracting auditory cues from the acoustic signals, as it has been done in the past, it may be interesting to attempt to learn a parameterization of the spatial information being embedded in the observed data.

In this paper we propose an unsupervised learning method that solves for the single sound-source direction retrieval problem using either *static binaural* or *dynamic monaural* spectral cues. The HRTF is a function that depends on three parameters for sound sources in the far field Otani *et al.* (2009): the frequency f of the emitter, the azimuth ϕ and elevation θ of the line joining the emitter to the listener, i.e., $h(f, \phi, \theta)$. First we show that the ILDs lie on a smooth two-dimensional (non-linear) manifold

embedded in $\mathbb{R}^N$, where N is the number of frequency channels used to represent sounds, and that the ILD-manifold is independent of the spectrum of the emitter. Second we consider the time derivatives of the acoustic inputs which will be referred to as *dynamic acoustic vectors*. These vectors (one for each microphone) may well be viewed as dynamic monaural cues and can be estimated from infinitesimal pan and tilt head motions. We show that although they correspond to the differentiation of a signal in $\mathbb{R}^N$, they lie on a smooth two-dimensional manifold. As is the case with the ILD-manifold, they are independent of the spectrum of the emitter.

Consequently, the problem of sound-source localization becomes the problem of learning two-dimensional manifolds for both binaural and monaural cues. An explicit derivation of these continuous manifolds would require an explicit formula for the HRTF. Instead we propose to sample the manifolds by gathering a large number of N-dimensional acoustic observations, each such observation being associated with a different head position (parameterized by pan and tilt angles) and with infinitesimal head motions. Hence the problem of building these manifolds becomes an instance of the non-linear dimensionality reduction problem. One way to solve the latter is to use manifold learning techniques Belkin and Niyogi (2003); Saul and Roweis (2003); Zhang and Zha (2004). We note that, while manifold learning has extensively been used in image analysis and in data mining, it has barely been used for sound localization.

The remainder of this paper is organized as follows. Section 2 describes the experimental setup and the methodology used to gather large training sets of static and dynamic binaural recordings. Section 3 presents a new computational model for audio-motor localization. Section 4 describes in detail techniques used to process collected sounds into training vector sets. Section 5 describes the non-linear dimensionality reduction technique used in practice. The results of our experiments are described and discussed in section 6. Conclusion and directions for future work are presented in section 7.

2 Experimental Setup and Data Acquisition

Existing auditory databases mainly deal with static listeners/emitters. Therefore, one of our first concerns has been to record sounds in the presence of head motions and for various emitter-to-listener directions. We believe that collecting such a large data set of monaural/binaural recording with its associated ground-truth is a contribution in its own right [1]. We opted to place a dummy head onto a robot that can perform fast, silent, and accurate movements with several degrees of freedom. Moreover, the experiments were carried out in real-world conditions, i.e., a room with reverberations and background noise.

[1] All our recordings with associated parameters and ground truth were made publicly available online at `http://perception.inrialpes.fr/~Deleforge/CAMIL_Dataset`.

Figure 1: A binaural dummy head is placed onto a robotic head which can perform precise and reproducible pan and tilt motions (left). The emitter (a loud-speaker) is placed in front of the robot at approximately 2.7 meters (right).

2.1 The Experimental Setup

In all our experiments we used the Sennheiser MKE 2002 dummy-head equipped with a pair of Soundman OKM II Classic Solo microphones which are linked to a computer via a Behringer ADA8000 Ultragain Pro-8 digital external sound card. The head had been mounted onto the University of Coimbra's audiovisual robot head POPEYE[2] with four rotational degrees of freedom: a pan motion, a tilt motion, as well as two additional degrees of freedom for eye vergence Hansard and Horaud (2010). This device was specifically designed to achieve precise and reproducible movements with a very good signal-to-noise ratio. The emitter – a loud-speaker – is placed at approximately 2.7 meters ahead of the robot, as shown on Fig. 1. The loud-speaker's input and the microphones' outputs were handled by two synchronized sound cards in order to simultaneously record and play.

2.2 Recording Audio-Motor Contingencies

Rather than placing the emitter at known 3D locations, we decided to keep the emitter in a fixed reference position and mimic sound directions by rotating the robot head. This allows to record a large data base of sound directions both accurately and automatically. In all our experiments the robot head was positioned in 16,200 *motor states*: 180 *pan rotations* α in the range $\in [-180°, 180°]$ and 90 *tilt rotations* β in the range $\in [-90°, 90°]$. We denote with $M =]-180°, 180°] \times [-90°, 90°]$ the space of all reachable motor states $\boldsymbol{m}_k = (\alpha_k, \beta_k) \in M$. The direct kinematic model of the robot head allows one to easily estimate the simulated position of the emitter in the robot's

[2] `http://perception.inrialpes.fr/POP/`.

frame as a function of pan and tilt:

$$\begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix} = \begin{bmatrix} \cos\beta\cos\alpha & -\sin\alpha & \cos\alpha\sin\beta \\ \cos\beta\sin\alpha & \cos\alpha & \sin\alpha\sin\beta \\ \sin\beta & 0 & \cos\beta \end{bmatrix} \begin{bmatrix} d \\ 0 \\ r \end{bmatrix} + r \begin{bmatrix} \cos\alpha\sin\beta \\ \sin\alpha\sin\beta \\ \cos\beta \end{bmatrix} \quad (1)$$

Notice that this model needs only two parameters: the distance from the tilt-axis to the microphones' midpoint, $r = 0.22m$, and the distance from this midpoint to the emitter, $d = 2.70m$. Indeed, the robot was designed such that the pan-axis passes through the microphones' midpoint.

It is straightforward to notice that while the space M, spanned by (α, β) pairs has a cylindrical topology (a ruled surface homeomorphic to a plane), the space of all possible sound-source positions (x_s, y_s, z_s) approximately lies on a sphere. One can also easily see that several distinct motor states can correspond to the same sound-source position. Therefore, the two spaces have different topologies and are not isomorphic, which means that there is an intrinsic difference between sampling the motor state space – as done in our case – and the sound-source position space.

In addition, with this approach, a recording made at a given motor-state only approximates the sound that would actually be perceived if the source was moved to the corresponding relative position in the room. First, the room *moves* together with the loud-speaker in the robot's frame, which modifies perceived reverberations. Second, the mechanical set up used implies that pan movements induce different sound source displacements in the robot frame depending on the current tilt position. The last approximation will raise issues that are further discussed in Section 6.3.

For each motor state $\boldsymbol{m}_k \in M$, we perform both static and dynamic binaural recordings of artificial reference and random-spectrum sounds emitted by the loud-speaker, as summarized in Table 1. An emitted sound corresponds to:

$$l(t) = K \sum_{i=1}^{N} \omega_i \sin(2\pi f_i t + \phi_i) \qquad t \in [0, 1] \quad (2)$$

where $l(t)$ is the loud-speaker's membrane displacement as a function of time t, $K \in \mathbb{R}$ is the global volume, $F = \{f_1 \dots f_i \dots f_N\}$ is a fixed set of N frequency channels, $\{\omega_i\}_{i=1..N} \in]0, 1]^N$ and $\{\phi_i\}_{i=1..N} \in [0, 2\pi]^N$ are weights and phases associated with each frequency channel. In practice, a set of $N = 600$ frequency channels $F = \{50, 150, 250 \dots, 5950\}$ was used. In order to evaluate the influence of the content of emitted sounds, we recorded both a *reference sound*, i.e., ω_i and ϕ_i are fixed for all motor states, and *random-spectrum sounds*, i.e., ω_i and ϕ_i are drawn from a uniform distribution at each motor state. We used these sounds because of two interesting properties of their spectrograms. First they are rich (many frequency channels represented) which makes them likely to contain rich spatial information. Second they are steady (constant energy at each frequency channel in time) which is crucial to measure the real

influence of head movements on perceived sounds. At each motor state $\boldsymbol{m}_k$ we performed the following recordings with both the reference sound and a random-spectrum sound (each such recording lasts one second):

- A binaural recording with a static listener
- A binaural recording while the listener performs a pan rotation with a constant angular velocity $\dot{\alpha} = 9°/s$
- A binaural recording while the listener performs a tilt rotation with a constant angular velocity $\dot{\beta} = 9°/s$

Table 1: Summary of emitted and recorded sounds in our dataset for a motor state $\boldsymbol{m}_k$.

Monaural sound emitted		Reference Ref/Emitted/0.wav	Random spectrum Rand/Emitted/k.wav
Binaural sound recorded	Motors		
	Static	Ref/Static/k.wav	Rand/Static/k.wav
	Pan left	Ref/Pan/k.wav	Rand/Pan/k.wav
	Tilt down	Ref/Tilt/k.wav	Rand/Tilt/k.wav

3 A Computational Model for Audio-motor Localization

We present now the proposed model for estimating the emitter-to-listener direction based on spectral features and head motions. The dummy head used in our experiments is a fair model of the human head. It is well known that the latter acts as an acoustic filter attenuating the perceived energy of specific frequency channels at each ear and depending on the sound source's 3D position. This attenuation function is commonly referred to as the *head related transfer function* (HRTF): there is a left-ear HRTF and a right-ear HRTF and these functions are specific to each listener, depending on the exact shape of the head, ears, and torso. A recent study Otani *et al.* (2009) showed that for sound-sources in the far field ($> 1.8m$), the HRTF function mainly depends on the source's direction (azimuth and elevation) at a given frequency channel. The influence of the emitter-to-listener distance will therefore not be considered in this work, which only accounts for sound source direction retrieval in the far field. As mentioned in the previous section, our training-set samples the motor-state space rather than the sound-source direction space, and is therefore meant to recover relationships between acoustic inputs and motor states. The actual emitter-to-listener direction can be easily recovered from these motor states using the direct kinematic model (1). For this reason, the dummy head's HRTF functions (left and right microphones) are modeled here by smooth scalar positive functions parameterized by the frequency channel $f \in]0; +\infty[$ and the motor state (α, β):

$$\begin{array}{lccc} h: &]0;+\infty[\times M & \longrightarrow &]0;+\infty[\\ h^L: & (f,\alpha,\beta) & \longmapsto & h^L(f,\alpha,\beta) \\ h^R: & (f,\alpha,\beta) & \longmapsto & h^R(f,\alpha,\beta) \end{array} \tag{3}$$

The smoothness assumption of the HRTF will be validated using experimental data in Section 4.2. Given an arbitrarily large set of N frequency channels $F = \{f_1 \dots f_i \dots f_N\}$ we approximately represent a sound with an N-dimensional *frequency vector* $\boldsymbol{x} = (x_1 \dots x_i \dots x_N)^T \in [0,+\infty[^N$, where the i-th element x_i corresponds to the mean intensity of the sound at frequency channel f_i, during a fixed temporal integration window. The way such vectors can be computed in practice from raw sound signals will be detailed in Section 4. With these notations, $\boldsymbol{x}^E$ refers to an *emitted sound* while $\boldsymbol{x}^L$ and $\boldsymbol{x}^R$ refer to sounds recorded by the robot's left and right ears. For a motor-state $(\alpha,\beta) \in M$, the HRTF model leads to the following relationship between emitted and recorded sounds:

$$\begin{cases} x_i^L = h^L(f_i,\alpha,\beta)x_i^E & \forall i \in [1;N] \\ x_i^R = h^R(f_i,\alpha,\beta)x_i^E & \forall i \in [1;N] \end{cases} \tag{4}$$

We will use the term *energy* to refer to the logarithm of a given frequency channel's intensity. We then define an *acoustic input vector* which corresponds to the perceived *energy* at each frequency channel:

$$\boldsymbol{s} = \log \boldsymbol{x} \in \mathbb{R}^N \tag{5}$$

We will use the notations $\boldsymbol{s}^L$ and $\boldsymbol{s}^R$ to denote the acoustic input vectors at the left and right ears, as well as $\boldsymbol{s}^{L,R} = (\boldsymbol{s}^L, \boldsymbol{s}^R) \in \mathbb{R}^{2N}$ to denote their concatenation. Note that this definition is only valid to the extent that $x_i > 0 \ \forall i \in [1;N]$, that is, if all the frequency channels are represented in the emitted and perceived sounds, which was always the case with the recordings described in Section 2 thanks to the specific way sounds were generated (i.e. (2). This is a relatively strong assumption since most of the real world sounds do not have a full spectrum. However, the present model only applies to these recordings, and we will explain in Section 6.4 how they can then be used as training data in order to retrieve the direction of unknown acoustic observations even with a fair amount of missing frequency channels.

By combining (4) and (5) we obtain that $\forall i \in [1;N]$:

$$\begin{array}{l} s_i^L = \log h^L(f_i,\alpha,\beta) + \log x_i^E \\ s_i^R = \log h^R(f_i,\alpha,\beta) + \log x_i^E \end{array} \tag{6}$$

Therefore, $\boldsymbol{s}$ is a multi-valued function that maps the motor-state space M and the emitted sound space $]0;+\infty[^N$ onto the set $\mathcal{S}$ of all possible acoustic input vectors, $\mathcal{S} \subset \mathbb{R}^N$. For a given emitted sound $\boldsymbol{x}^E$, we now consider the set of all possible acoustic input vectors while the robot is allowed to move in all its motor states:

$$\mathcal{S}_{\boldsymbol{x}^E} = \{\boldsymbol{s}(\boldsymbol{x}^E,\alpha,\beta) \mid (\alpha,\beta) \in M\} \tag{7}$$

Under the assumptions that $\boldsymbol{s}$ is a homeomorphism of motor state parameters for a fixed emitted sound, this set lies on a two-dimensional smooth manifold embedded in $\mathbb{R}^N$ and parameterized by (α, β) (cylindrical topology). The existence of such a homeomorphism between $\mathcal{S}_{\boldsymbol{x}^E}$ and M will be experimentally validated in Section 6. For an emitted sound $\boldsymbol{x}^E$, we denote by $\mathcal{S}^L_{\boldsymbol{x}^E}$ and $\mathcal{S}^R_{\boldsymbol{x}^E}$ the left and right *acoustic input manifolds*. We also denote by $\mathcal{S}^{L,R}_{\boldsymbol{x}^E}$ the concatenated acoustic input manifold.

Notice however, that there are as many such manifolds as frequency vectors $\boldsymbol{x}^E$ emitted by the loud-speaker. Learning the structure of such manifolds from acoustic input vectors will only allow to parameterize directions associated with a specific emitted sound, which is of limited interest for sound localization. For this reason we will consider auditory representations that are independent of the emitter's content $\boldsymbol{x}^E$.

3.1 The ILD Manifold

We define ILD vectors by $\boldsymbol{s}^{ILD} = \boldsymbol{s}^L - \boldsymbol{s}^R$. For a motor state $(\alpha, \beta) \in M$ and an emitted sound $\boldsymbol{x}^E \in]0; +\infty[$, it is straight-forward to see from 6 that $\forall i \in [1; N]$:

$$\begin{array}{rcl} s_i^{ILD} & = & \log h^L(f_i, \alpha, \beta) - \log h^R(f_i, \alpha, \beta) \\ & = & \log h^{ILD}(f_i, \alpha, \beta) \end{array} \tag{8}$$

Since the emitted sound component x_i^E cancels out, the ILD vectors do not depend on the sound source content. The *ILD space* corresponding to the set of all ILD vectors from all emitted sounds and all motor states can therefore be written:

$$\mathcal{S}^{ILD} = \{\boldsymbol{s}^{ILD}(\alpha, \beta) \mid (\alpha, \beta) \in M\} \tag{9}$$

Under the assumption that $\boldsymbol{s}^{ILD}$ is a homeomorphism of motor state parameters, this set lies on a two-dimensional smooth manifold, the *ILD manifold*, embedded in $\mathbb{R}^N$, parameterized by (α, β), and independent of the emitted sound. The existence of such a homeomorphism between $\mathcal{S}^{ILD}$ and M will be experimentally validated in Section 6. Therefore, the ILD vectors $\boldsymbol{s}^{ILD} \in \mathbb{R}^N$ may be viewed as content-independent auditory cues for sound localization.

3.2 The Dynamic Acoustic Manifold

So far we considered the static case, i.e., the listener is in a static motor state $\boldsymbol{m} = (\alpha, \beta)$ while emitted sounds are recorded. We consider now the case of a *dynamic listener*. More precisely, we define a motor command c by a tuple $(\dot{\alpha}, \dot{\beta})$, where $\dot{\alpha}$ and $\dot{\beta}$ correspond to constant angular velocities transmitted to pan and tilt motors. In particular, we will later denote:

$$\left\{ \begin{array}{l} c_\alpha = (\dot{\alpha}, 0) \\ c_\beta = (0, \dot{\beta}) \end{array} \right. \tag{10}$$

the motor commands corresponding to pan and tilt head movements at constant velocity, where $\dot{\alpha} = \dot{\beta} = 9^\circ/s$.

An infinitesimal motor displacement $d\boldsymbol{m} = (d\alpha, d\beta)$ during dt corresponds to a tangent vector on the smooth manifold $\mathcal{S}_{\boldsymbol{x}^E}$ of acoustic inputs. If the robot performs any motor command c from the motor state (α, β) in front of a static and steady sound-source emitting $\boldsymbol{x}^E$, it is therefore natural to study the structure of the acoustic input vector's time derivative:

$$\boldsymbol{\tau}(c) = \frac{d\boldsymbol{s}}{dt} \tag{11}$$

This vector will be referred to as a *c-dynamic acoustic vector*. Taking the time derivative of (6) under the smoothness assumption of the HRTF we obtain:

$$\begin{aligned} \tau_i &= \frac{ds_i}{dt} = \frac{\partial s_i}{\partial \alpha}\frac{d\alpha}{dt} + \frac{\partial s_i}{\partial \beta}\frac{d\beta}{dt} \\ &= \frac{\partial \log h(f_i, \alpha, \beta) + \log x_i^E}{\partial \alpha}\frac{d\alpha}{dt} \\ &+ \frac{\partial \log h(f_i, \alpha, \beta) + \log x_i^E}{\partial \beta}\frac{d\beta}{dt} \end{aligned}$$

If we define $h_i = h(f_i, \alpha, \beta)$ and $dh_i/d\boldsymbol{m}$ the gradient of h_i at $\boldsymbol{m} = (\alpha, \beta)$, this can be written in a more compact way as:

$$\tau_i = \frac{1}{h_i}\left(\frac{dh_i}{d\boldsymbol{m}}\right)^\top \dot{\boldsymbol{m}} \tag{12}$$

where we used the fact that $d\log(x_i^E)/d\alpha = d\log(x_i^E)/d\beta = 0$. Since the emitted sound component x_i^E cancels out, the dynamic acoustic vectors do not depend on the sound source's content. The *dynamic acoustic space* $\mathcal{T}(c)$ corresponding to all dynamic acoustic vectors obtained for a given motor command c, all emitted sounds, and all motor states is then defined by

$$\mathcal{T} = \{\boldsymbol{\tau}(\alpha, \beta) \mid (\alpha, \beta) \in M\} \tag{13}$$

Under the assumption that $\boldsymbol{\tau}$ is a homeomorphism of motor state parameters, $\mathcal{T}(c)$ constitutes a two-dimensional smooth manifold, the *c-dynamic acoustic manifolds*, embedded in $\mathbb{R}^N$, parameterized by (α, β), and independent of the emitted sound. The existence of such a homeomorphism between $\mathcal{T}(c)$ and M will be experimentally validated in Section 6. As previously, there will be left- and right-microphone dynamic acoustic manifolds $\mathcal{T}(c)^L$ and $\mathcal{T}(c)^R$, as well as the concatenated dynamic acoustic manifold $\mathcal{T}(c)^{L,R} \subset \mathbb{R}^N$.

4 From Sound Signals to Acoustic Vectors

We now describe the methodology that we use to process raw sound signals collected during experiment, i.e, section 2, and to transform these signals into the acoustic input vectors (6), the ILD vectors (8) and the dynamic acoustic vectors (12).

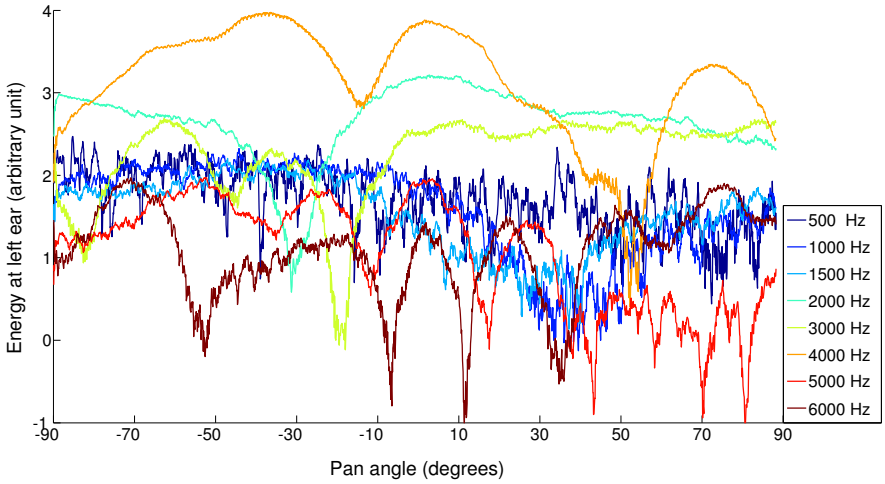
4.1 Spectrograms

The model described in section 3 requires to represent sounds both in the time and frequency domains. A commonly used representation in CASA is to use gamma-tone filter banks Wang and Brown (2006). Although these filters are known to model the human auditory system quite well, they also generate overlaps in the intensities of frequency channels that are unwished in the framework of our approach. For this reason, we chose to employ simpler spectrograms. These spectrograms are computed using a sliding discrete Fourrier transform of the raw signal within a specified time window, in order to capture the temporal variation of the sound spectrum: They discretize signals both in time and frequency. The way this discretization is achieved is critical and must be carefully tuned.

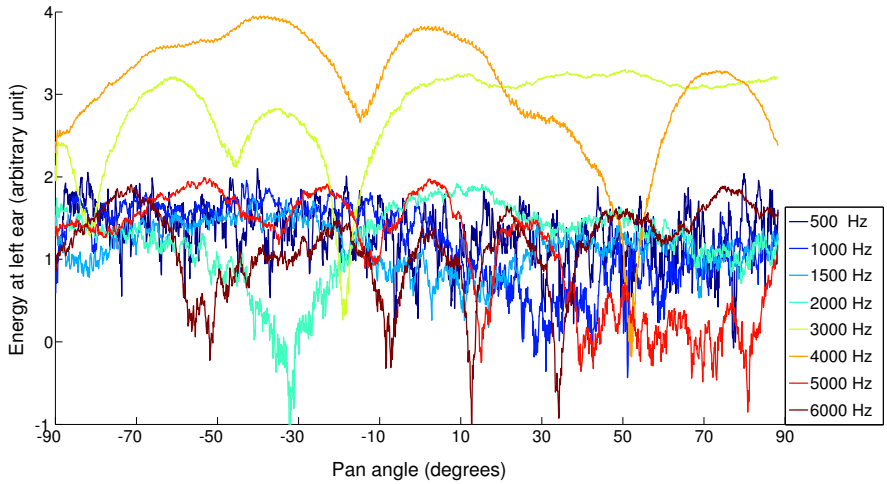
Two crucial parameters are to be considered for temporal discretization: the temporal integration and the frame shift. The temporal integration is the length of the time window inside which the discrete Fourrier transform is computed. For the Fourrier transform to be meaningful, the temporal integration should be at least twice larger than the largest sound period considered. On the other hand, the notion of *instantaneous frequency* will be lost for too large time windows if the sound varies too much. The frame shift parameter corresponds to the delay between two time windows. The smaller the frame shift, the higher the resolution of the spectrogram in time, at the cost of a higher computational burden.

The discretization of spectrograms in the frequency domain mainly relies on three parameters: the lowest and highest frequency channels, and the number of channels. One can choose to spread frequency channels within this range either linearly or logarithmically. In practice, a logarithmic scale should be preferred because it accounts better for harmonics generated by the discrete Fourrier transform, and it coincides with the neural encoding of frequency channels in the human auditory system. There are several ways to compute intensity values at a given frequency channel, the simplest one being to average the signal in the frequency domain within a neighboring window. However, experiments showed that using the maximal intensity inside each window yields more stable spectrograms, by dealing with small frequency fluctuations due to the sampling approximation or possible Doppler effect.

A good tradeoff between computational cost and spectrogram precision is achieve with a temporal integration of 200ms and a frame shift of 10ms. Fig. 2(a) and (b) show some *time-energy curves* (logarithm of the spectrogram intensity values at a given frequency channel) obtained with these parameters while performing a 180 degrees pan



(a) Time-energy curves for an emitted random sound



(b) Time-energy curves for another emitted random sound

Figure 2: This figure shows the variation of the energy perceived by the left microphone for various frequency channels while the static loud-speaker emits two different random sounds, e.g, (a) and (b), and while the listener performs a 180° rotation from left to right at constant velocity c_α in (10).

movement leftwards at constant velocity, i.e, c_α in (10), in the presence of a static loud-speaker that emits two distinct random spectrum sounds as defined in (2). These curves call for several remarks.

First, time-energy curves corresponding to the lowest frequency channels (500 Hz, 1000 Hz and 1500 Hz) do not exhibit any coherent variation with respect to the pan

angle, they are highly discontinuous and they are not invariant to the emitted sound. This can be explained by the fact that the acoustic filter of the dummy-head only acts on sounds with wavelength below the head diameter ($\approx$ 18cm), which corresponds to frequency channels higher than 1900 Hz. This correlates well with numerous psychophysical and behavioral observations suggesting that the ILD is mainly responsible for sound localization in the high frequency domain, whereas the ITD is rather used for the low frequency domain Middlebrooks and Green (1991).

Second, time-energy curves above 2000 Hz are invariant with respect to the emitted random sound, up to an additive constant, independently of their own and other channel's energy. This shows that perceived energies of frequency channels are independent from each other and only depend on the motor-state parameter α up to an additive constant. This may account for an experimental validation of the HRTF acoustic model (3) as well as (6), (8) and (12). Moreover and despite a slight noise, they appear to be generally continuous and differentiable with respect to the motor-state parameter: This comforts the assumption that the HRTF is a smooth function of the motor-state parameters, i.e., section 3. Based on these observations, we used the following settings in order to compute the ILD and dynamic acoustic vectors: There are $N = 400$ frequency channels ranging from 2000 Hz to 6000 Hz in logarithmic scale with a temporal integration of 200 ms and a frame shift of 10 ms.

Finally, it is worthwhile to notice from Fig. 2 that there are peaks and notches of the HRTF at specific frequencies and angles. These extrema are well known and commonly referred to as *spectral features* in the psycho-acoustical literature. Several experiments on human subjects suggest that they are involved in vertical sound localization (e.g. Hebrank and Wright (1974), Greff and Katz (2007)). In our case, the existence of such extrema corresponds to important local distortions in the studied manifold, which implies the need of a relatively dense sampling of the motor-state space (2-3 degrees between two adjacent points) for manifold learning to work well in practice. The complex shape of the curves also justifies the use of non-linear dimensionality reduction techniques rather than linear methods.

4.2 Acoustic Input and ILD Vectors

Using the temporal and spectral parameters just mentioned, records which are one second long, static and monaural will result in a spectrogram of 81 frequency vectors, each such vector corresponding to $N = 400$ frequency channels. As shown in the left plot of Fig. 3, the perceived energy, when the head remains static, is relatively stable in time. This energy is averaged over a period of one second to obtain the N-dimensional acoustic input vectors $\boldsymbol{s}^L$, $\boldsymbol{s}^R$ as well as the ILD vector $\boldsymbol{s}^{ILD} = \boldsymbol{s}^L - \boldsymbol{s}^R$, e.g., (6) and (8).

4.3 Dynamic Acoustic Vectors

When the listener performs head motions, one is faced with the problem of dynamic recording in order to estimate the time derivative of the acoustic vectors, namely the

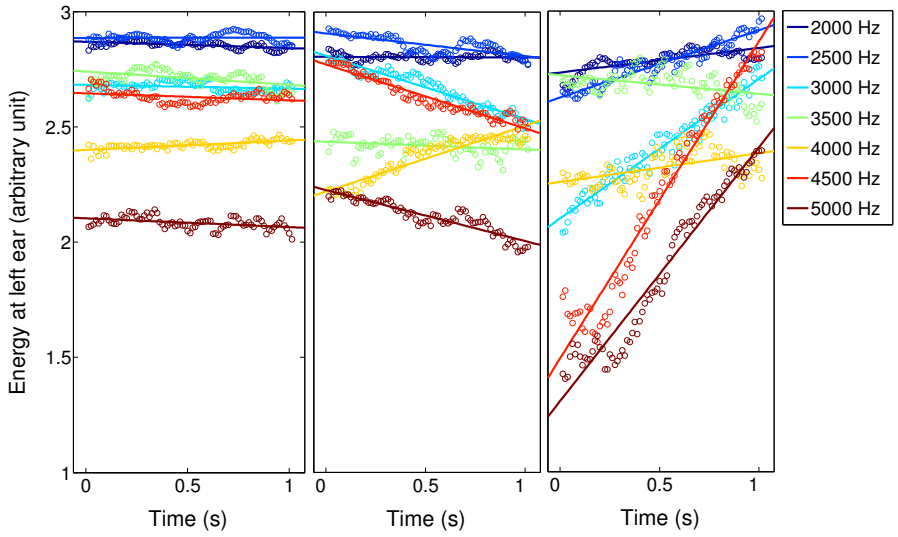


Figure 3: This figure shows the variation of the energy perceived at the left ear, during one second, and for various frequency channels when a random-spectrum sound x^E is emitted, i.e., (2). Circles represent recorded energy values. Lines correspond to least-square error linear interpolations of the energy variation. The figure illustrates the following situations: The head is static (left), the head performs a motor command c_α (middle), and finally the head performs motor command c_β (right).

dynamic acoustic vectors $\boldsymbol{\tau}(c)$ (12). As previously, we compute the spectrograms and we take the logarithm thus yielding $N = 400$ time-energy curves associated with each recording. Fig. 3-middle and -right show that this leads to a significant variation over time of the perceived energy, at each frequency channel. We used linear regression to locally approximate this temporal variation:

$$\left\{ \begin{array}{rcl} s_i^L(t) & = & a_i^L t + b_i^L \\ s_i^R(t) & = & a_i^R t + b_i^R \end{array} \right. \tag{14}$$

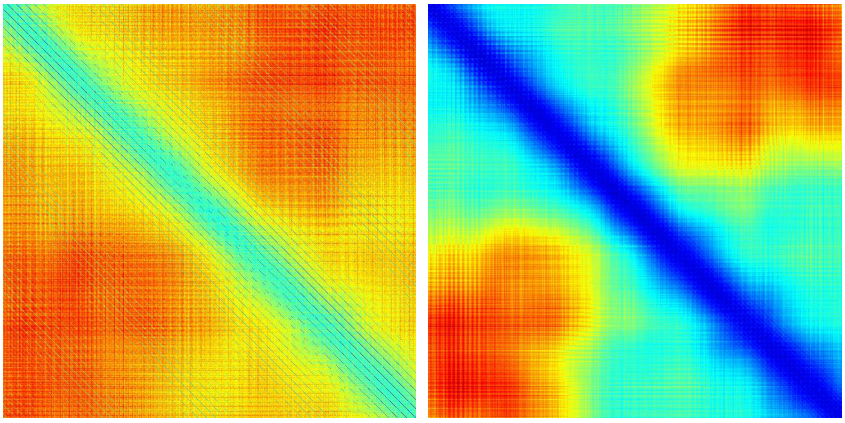
where a_i^L and a_i^R are the local slopes of the perceived energies s_i^L and s_i^R at frequency f_i. We finally define dynamic acoustic vectors using the slopes of these local linear approximations:

$$\boldsymbol{\tau}^L(c) = (a_1^L \dots a_i^L \dots a_N^L) \tag{15}$$
$$\boldsymbol{\tau}^R(c) = (a_1^R \dots a_i^R \dots a_N^R) \tag{16}$$

4.4 Content-Independent Spatial Auditory Cues

A preliminary experiment was performed aimed at the validation of the assumptions that the ILD vectors and the dynamic acoustic vectors are both invariant with respect

to the emitted sound's spectrum. For that purpose we recorded sounds emitted by a static loud-speaker while the listener was undergoing a complete 180° pan rotation in 90 steps of 2° each, i.e., 90 pan positions $\alpha \in [-90^\circ, +90^\circ]$. At each one of these pan positions we recorded 30 random-spectrum sounds, both with a static and a dynamic listener, as explained in Section 2.2. This resulted in the computation of $2700 = 90 \times 30$ left-ear and right-ear acoustic input vectors $\boldsymbol{s}^L$ and $\boldsymbol{s}^R$, as well as an equal number of concatenated acoustic input vectors $\boldsymbol{s}^{L,R}$, ILD vectors $\boldsymbol{s}^{ILD}$, and left- and right- c_α-dynamic acoustic vectors $\boldsymbol{\tau}^L(c_\alpha)$ and $\boldsymbol{\tau}^R(c_\alpha)$10.



(a) Acoustic input vectors $\boldsymbol{s}^{L,R}$

(b) ILD vectors $\boldsymbol{s}^{ILD}$

(c) left-ear c_α-dynamic acoustic vectors $\boldsymbol{\tau}^L(c_\alpha)$

(d) right-ear c_α-dynamic acoustic vectors $\boldsymbol{\tau}^R(c_\alpha)$

Figure 4: This figure shows 2700×2700 pairwise Euclidean distance matrices between vectors obtained at 90 different motor states (90 pan values $\alpha \in [-90^\circ, +90^\circ]$) and with a loud-speaker emitting 30 different random spectrum sounds, i.e., (2). Vectors are sorted left-right and bottom-down with respect to their corresponding pan value. The distance varies from zero (dark blue) to higher values (dark red) [*This figure is best seen in color*].

We computed the pairwise distances between these vectors. As it may be seen in Fig. 4-(a), the acoustic input vectors $\boldsymbol{s}^{L,R}$ are highly content-dependent and they cannot provide proper sound localization information. Alternatively, the zero-distance stripe around the 2700×2700 matrices' diagonals in Fig. 4-(b), -(c) and -(d) show that the ILD and acoustic dynamic vectors are well suited for sound localization. Indeed, these zero-distance diagonal stripes correspond to pairwise distances between vectors that are associated with similar motor states: This validates experimentally the computational model outlined in Section 3.

5 Manifold Learning Via Dimensionality Reduction

The experimental setup (Section 2) and associated sound processing techniques (Section 4) allow to collect large sets of high-dimensional spectral-feature vectors with ground-truth spatial localization. In principle, although these feature vectors are computed in a high-dimensional space, they should lie on low-dimensional manifolds that are parameterized by motor-state parameters. In the absence of an explicit model, we propose to recover a low-dimensional manifold parameterization through a dimensionality reduction technique.

If this manifold corresponds to a low-dimensional linear subspace of the high-dimensional spectral-feature vector space, linear methods such as principal component analysis, or one of its numerous variants, can be used. Nevertheless, the model developed in Section 3 postulates that the sought manifolds are nonlinear. Several manifold learning algorithms have been recently developed, most notably including kernel PCA Scholkopf *et al.* (1998), ISOMAP Tenenbaum *et al.* (2000), local-linear embedding (LLE) Saul and Roweis (2003), or Laplacian eigenmaps (LE) Belkin and Niyogi (2003). Both kernel PCA and ISOMAP can be viewed as generalizations of multidimensional scaling (MDS) Cox and Cox (1994) and are based on computing the top eigenvalues and eigenvectors of a Gram matrix. They require the computation of either all the pairwise geodesic distances Tenenbaum *et al.* (2000) or of all the pairwise kernel dot-products, which may be computationally expensive. Alternatively, methods like LLE and LE only require the computation of pairwise similarities between each vector and its neighbors. For large data sets, such as ours, this results in very sparse matrix solvers.

In spite of the elegancy of such methods as LLE and LE, which are widely used in machine learning applications, we preferred to use the local tangent space alignment method (LTSA) proposed in Zhang and Zha (2004). This method approximately represents the data in a lower dimensional space by minimizing a reconstruction error in 3 steps. First, a local neighborhood around each point is computed. Second, for each such point, local coordinates of its neighboring points on the local tangent space are approximated using PCA. Third, these local coordinates are optimally aligned to construct a global low-dimensional representation of the points lying on the manifold. As already noticed in Aytekin *et al.* (2008), the main advantage of LTSA over spectral graph methods, is its ability to work well with non-compact manifold subsets with

boundaries, such as is the case with our data. LTSA's only free parameter is the integer k corresponding to the neighborhood size around each point. We experimented with different values of k and we noticed that, in our case, e.g, 16,200 vectors of dimension 400, the choice of k was not critical. In practice we implemented a fast version of the kNN algorithm which, in conjunction with sparse matrix solvers, allowed us to implement an efficient non-linear dimensionality reduction method based on LTSA.

Nevertheless, we introduced two modifications to the original LTSA algorithm. First, LTSA uses kNN to determine neighboring relationships between points, yielding neighborhoods of identical size k over the data. This has the advantage of always providing connected neighborhood graphs but it can easily lead to inappropriate connexions between points, especially at boundaries or in the presence of outliers. A simple way to overcome these artifacts is to implement a *symmetric* version of kNN, by considering that two points are connected if and only if each of them belongs to the neighborhood of the other one. Comparisons between the outputs of standard and symmetric kNN are showed in Fig. 5. Although symmetric kNN solves connexions issues at boundaries, it creates neighborhood of variable sizes, and in particular some points might get disconnected from the graph. Nevertheless, it turns out that detecting such isolated points is an advantage because it may well be viewed as a way to remove outliers from the data. In our case the value of k was set manually; in practice any value in the range $[10, 20]$ yielded satisfying results.

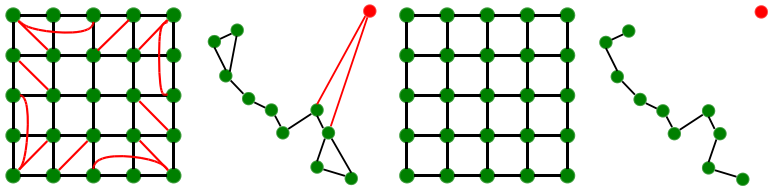


Figure 5: Illustration of the differences between standard kNN ((a) and (b)) and symmetric kNN ((c) and (d)) using two examples in 2D. (a) and (c) show the graphs obtained from an *image-like* arrangement of points with boundaries ($k = 4$), while (b) and (d) shows the behavior of the two kNN algorithms in the presence of an outlier ($k = 2$).

Second, in the standard LTSA algorithm the target low-dimension used to represent the data corresponds to the dimension *common* to all the local tangent spaces; this dimension could be estimated during the second step of LTSA based on PCA. Notice however that it is very unlikely that the dimensions estimated like this, using the k-neighbourhoods of the high-dimensional points, would yield the same values for the whole data set. Alternatively, we rely on the dimension predicted by the computational model developed in this paper, namely that the dimension of the acoustic manifolds should be equal to 2. Therefore, we retained the top two eigenvalue-eigenvector pairs in the second step of LTSA. Moreover, we would like to represent manifolds which are, in principle, homeomorphic to the 2D surface of a cylinder. The best way to visualize such a 2D curved surface is to represent it in the 3D Euclidean space and to

visualize the 3D points lying on that surface. For this reason, we retain the three largest eigenvalue-eigenvector pairs in the global alignment stage of LTSA (third step), such that the extracted manifolds can be easily visualized.

6 Computational experiments and results

The method presented above allows to build two-dimensional acoustic manifolds in a completely unsupervised way and to obtain an *intrinsic parameterization*, e.g, using LTSA. Indeed, our computational model predicts that these manifolds are homeomorphic to the surface of a cylinder parameterized by the motor-state variables (α, β). This means in practice that each manifold point corresponds to an emitter-to-listener sound direction. However, rather than representing these directions in the *extrinsic* 3D world space, it is represented *intrinsically* on one of the acoustic manifolds. Once such a manifold has been learned, it can be used as a training data set in a classification framework to find the direction of a sound emitted from an unknown position.

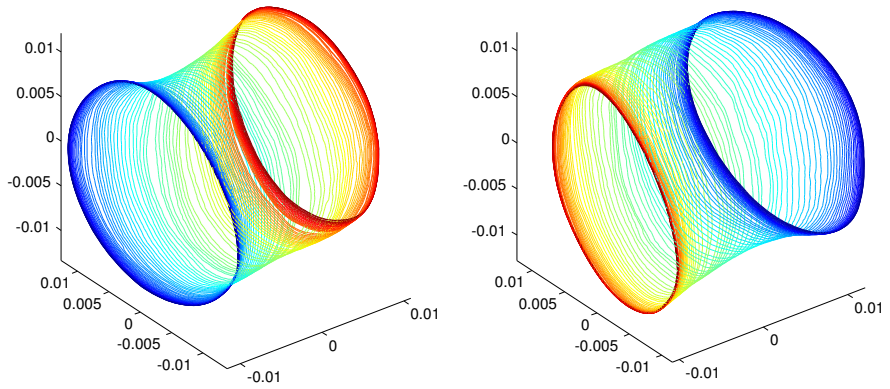
The experimental setup and data collection protocols described in Section 2 allows us to establish a one-to-one association between the manifolds extracted from the acoustic data with LTSA and the *ground-truth* motor-state values. We will refer to such an association as an *audio-motor map*. An audio-motor map proves the existence of a homeomorphism between the motor-state space and an acoustic manifold (i.e. Section 3) if the following three criteria are verified:

- The map has the same topology as the motor-state space (i.e. cylindrical)
- Level-set lines associated with the *intrinsic* (α, β) manifold parameterization should not cross each other
- The ordering of the points along a level-set line must be the same for the acoustic and ground-truth manifolds

We implemented a simple method to represent audio-motor maps that allows a qualitative verification of these three criteria. In the lower-dimensional representation of data obtained with LTSA, we link points associated to the same ground truth tilt values (tilt level-set lines) such that they form parallel closed curves onto a manifold homeomorphic to a cylinder.

6.1 The Manifold of Acoustic Inputs

We start by analyzing the acoustic input manifolds associated with the left and right microphones, $\mathcal{S}^L_{\boldsymbol{x}_0}$ and $\mathcal{S}^R_{\boldsymbol{x}_0}$, in the presence of a reference emitted sound $\boldsymbol{x}^E_0$. Fig. 6(a)-(b) shows audio-motor maps obtained using acoustic input vectors $\boldsymbol{s}^L$ and $\boldsymbol{s}^R$ as well as 16,200 motor states of the robot head. These maps qualitatively validate the existence of a homeomorphism between $\mathcal{S}^L_{\boldsymbol{x}_0}$ and $\mathcal{S}^R_{\boldsymbol{x}_0}$ on one side and the the motor state



(a) Left ear - $\mathcal{S}^L_{\boldsymbol{x}^E_0}$ (b) Right ear - $\mathcal{S}^R_{\boldsymbol{x}^E_0}$

Figure 6: These audio-motor maps correspond to the left- and right acoustic input manifolds for a reference emitted sound. This illustrates the fact that the initial 400-dimensional vectors lie on cylinder-like surfaces. The colored curves connect data points with the same ground-truth tilt values, ranging from $\beta = -90°$ (dark blue) to $\beta = +90°$ (red), ordered with their ground-truth pan values ranging from $\alpha = -180°$ to $\alpha = +180°$.[*This figure is best seen in color.*]

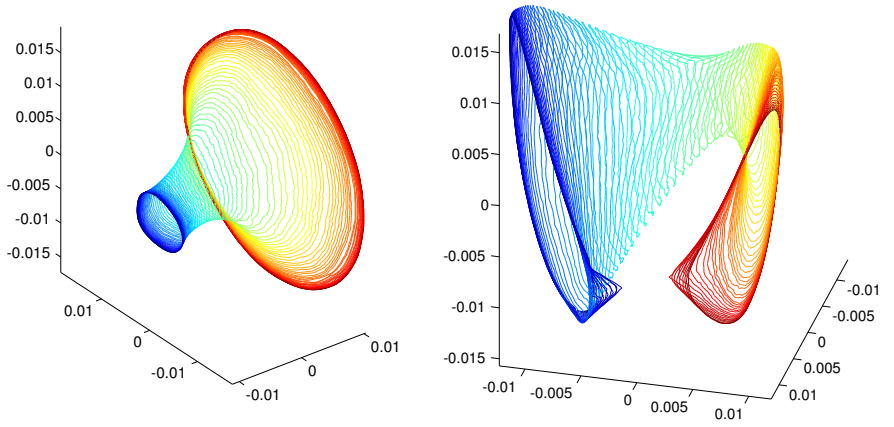
space M on the other side. Although it proves that the acoustic input manifolds have the expected structure, such maps are of limited interest in practice, because of their dependency on the sound-source content, as outlined in Section 3.

6.2 The ILD Manifold

To turn our attention now to content-independent spaces. Fig. 7(a) shows the result obtained using input vectors $\boldsymbol{s}^{ILD}$ corresponding to the 16,200 motor states of the robot; These ILD vectors were obtained with a different random-spectrum sound (i.e. (2)) emitted at each motor state. Although all the records were made in the presence of different sounds, the resulting audio-motor map qualitatively validates that the ILD manifold $\mathcal{S}^{ILD}$ is homeomorphic to the motor state space M.

6.3 The Dynamic Acoustic Manifold

The dynamic acoustic spaces $\mathcal{T}^L(c)$, $\mathcal{T}^R(c)$ and $\mathcal{T}^{L,R}(c)$ can be obtained from pan and tilt motor commands c_α and c_β (10) in the presence of a random-spectrum sound source (i.e. (2)). Experimentally, we noticed that the concatenated c_α-dynamic acoustic vectors $\boldsymbol{\tau}^{L,R}(c_\alpha)$ cannot be used to build an audio-motor map because the symmetric kNN algorithm leaves out too many disconnected points. This can be explained by



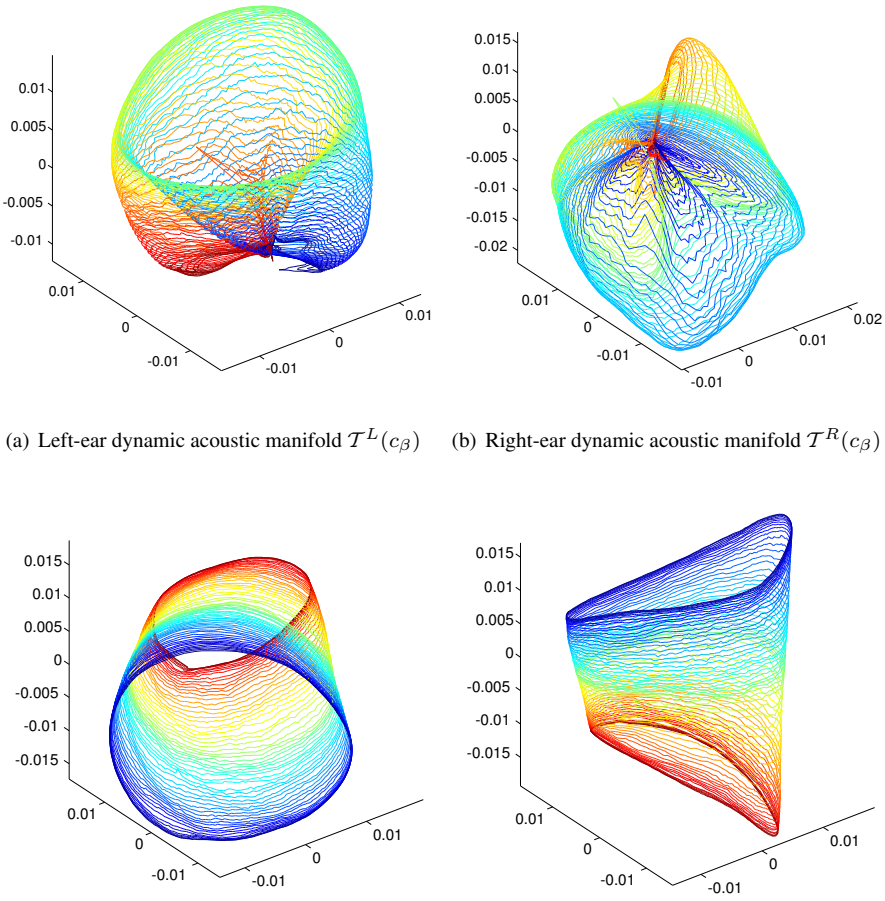
(a) ILD manifold $\mathcal{S}^{ILD}$

(b) Concatenated dynamic acoustic manifold $\mathcal{T}^{L,R}(c_\beta)$

Figure 7: Binaural manifolds built using either the ILD vectors (a) or the concatenated c_β-dynamic acoustic vectors (b). The colored curves connect data points with the same ground-truth tilt values, ranging from $\beta = -90°$ (dark blue) to $\beta = +90°$ (red), ordered with their ground-truth pan values ranging from $\alpha = -180°$ to $\alpha = +180°$.[*This figure is best seen in color*.]

the fact that the mechanical set up used implies that pan movements induce different sound source displacements in the robot frame depending on the current tilt position, as mentioned earlier in Section 2. One should notice, however, that tilt positions are not conditioned by pan positions and hence one can compute an audio-motor map from concatenated c_β-dynamic acoustic vectors $\boldsymbol{\tau}^{L,R}(c_\beta)$. Fig. 7(b) shows that the manifold thus obtained, although distorted, qualitatively verifies the three criteria for a homeomorphic mapping. Therefore, the c_β-dynamic acoustic manifold $\mathcal{T}^{L,R}(c_\beta)$ can be thought of as being homeomorphic to the motor state space M.

One may argue that the concatenated dynamic acoustic manifolds may be of little interest since the ILD manifold provides similar results, and this without the need of any head motion. Nevertheless, dynamic data remains extremely interesting because it can by used in conjunction with a *single* microphone thus yielding *monaural* manifolds, one for each microphone. Fig. 8(a)-(b) show audio-motor maps obtained from monaural $c_\beta-$dynamic acoustic vectors $\boldsymbol{\tau}^{L}(c_\beta)$ and $\boldsymbol{\tau}^{R}(c_\beta)$. These plots reveal important distortions of the manifolds due to some confusions between records associated with high and low tilt positions. To overcome this problem, we removed from the training data set those input vectors corresponding to high and low tilt values. Fig. 8(c)-(d) shows the manifolds obtained from a subset of 10,800 motor states in the range $\alpha \in [-180°, 180°]$ and $\beta \in [-60°, 60°]$. These results qualitatively validate the existence of a homeomorphism between a subset of $\mathcal{T}^{L}(c_\beta)$ and $\mathcal{T}^{R}(c_\beta)$ on one side and a subset of the motor state space M on the other side.



(a) Left-ear dynamic acoustic manifold $\mathcal{T}^L(c_\beta)$ (b) Right-ear dynamic acoustic manifold $\mathcal{T}^R(c_\beta)$

(c) Truncated left manifold (d) Truncated right manifold

Figure 8: Monaural dynamic manifolds obtained with the whole data set (a), (b) and with a truncated data set (c), (d). First row: all the 16,200 motor states are used. Second row: a subset of 10,800 motor states corresponding to $\alpha \in [-180°, 180°]$ and $\beta \in [-60°, 60°]$ is used. The color conventions are the same as above. [*This figure is best seen in color.*]

6.4 Sound Localization and Missing Frequencies

All these results show that both binaural static cues (i.e. ILD vectors) and monaural dynamic cues (i.e. c_β-dynamic acoustic vectors) may be used to retrieve the direction of a sound source independently of its content. Indeed, they experimentally prove the existence of homeomorphisms between our content-independent acoustic manifolds and the motor-state manifold M. Therefore, the problem of localizing an unknown

sound source from a few auditory observations is equivalent to classifying a sample from a test-set based on training using another data set. In that sense, a major contribution of this paper is to view sound localization as a classification problem, whereas the vast majority of standard approaches has used close-form solutions in order to recover spatial (3D) parameters either from ITDs, from ILDs, or from both.

In practice we implemented a nearest neighbor classifier to assess the validity of our sound localization framework based on content-independent acoustic vectors. On one side, the training set $\tilde{\mathcal{S}}^{ILD}$ contains ILD vectors which are estimated using all the 16,200 motor states and with the speaker emitting a single reference sound, i.e., (2) with fixed parameters. On the other side, the test set $\mathcal{S}^{ILD}$ is composed of ILD vectors estimated while the speaker is emitting different random-spectrum sounds drawn from a uniform distribution at each state, i.e., (2). Given an ILD vector $\boldsymbol{s}_i^{ILD}$ from the test-set, we find its nearest neighbor $\tilde{\boldsymbol{s}}_i^{ILD}$ in the training-set and we use its associated ground-truth motor state $(\tilde{\alpha}_i, \tilde{\beta}_i)$ to infer the spatial direction of the sound-source using the direct kinematics of the robot head (1). This allows to quantitatively evaluate our method by direct comparison of the ground-truth motor-state of the tested ILD vector, i.e., (α_i, β_i) with the motor state $(\tilde{\alpha}_i, \tilde{\beta}_i)$ found with the classification method just described. One can therefore define the absolute angular error E_i as follow:

$$E_i = |\tilde{\alpha}_i - \alpha_i| + |\tilde{\beta}_i - \beta_i| \tag{17}$$

We computed this mean absolute angular error for all the ILD vectors in the test set. The same approach was used to retrieve sound source directions from monaural dynamic cues, using monaural c_β-dynamic acoustic vectors from 10,800 motor states corresponding to $\alpha \in [-180°, 180°]$ and $\beta \in [-60°, 60°]$ (see Section 6.3). Table 2 summarizes the mean absolute angular errors obtained using both static binaural cues and monaural dynamic cues.

Table 2: Mean absolute angular error over 16,200 test ILD vectors and 10,800 test monaural c_β-dynamic acoustic vectors.

Test data set	$\mathcal{S}^{ILD}$	$\mathcal{T}^L(c_\beta)$	$\mathcal{T}^R(c_\beta)$
Mean error	$0.0064°$	$0.0826°$	$0.0841°$

It should however be noted that this only allows to retrieve the emitter-to-listener direction under the assumption that the observed acoustic vectors, whether binaural or monaural, have a full spectrum, namely provided that the 400 frequency channels are significantly represented in the perceived sounds (see Section 3). This is a relatively strong assumption considering that many real-world sounds only have a sparse spectrum (e.g. human voices or music instruments).

Consequently, we simulated perceived sounds with sparse spectrum and tested the validity of our method using sound with missing frequency channels. We used our method with a test-set full-spectrum vector, i.e., all 400 frequency channels are significantly present in that vector. Then we projected this full-spectrum sound onto an n-dimensional subspace $\mathbb{R}^n$ of $\mathbb{R}^N$ ($n < N$), where the subset F_n of represented

frequencies was randomly selected: $F_n = \{f_{i_1}, \dots f_{i_k}, \dots f_{i_n}\} \subset F$ (see Section 3). Finally, we compute the nearest neighbor of this *sparse* acoustic vector in its associated sparse training set. In practice, in the presence of an unknown sparse-spectrum auditory observation, the projection could be done by applying a threshold on perceived energies s_i^L and s_i^R to remove under-represented frequency channels. The mean absolute angular error obtained with this approach was computed for the different cues, while varying the number of represented frequencies. We repeated this experiment with both ILD vectors and dynamic vectors.

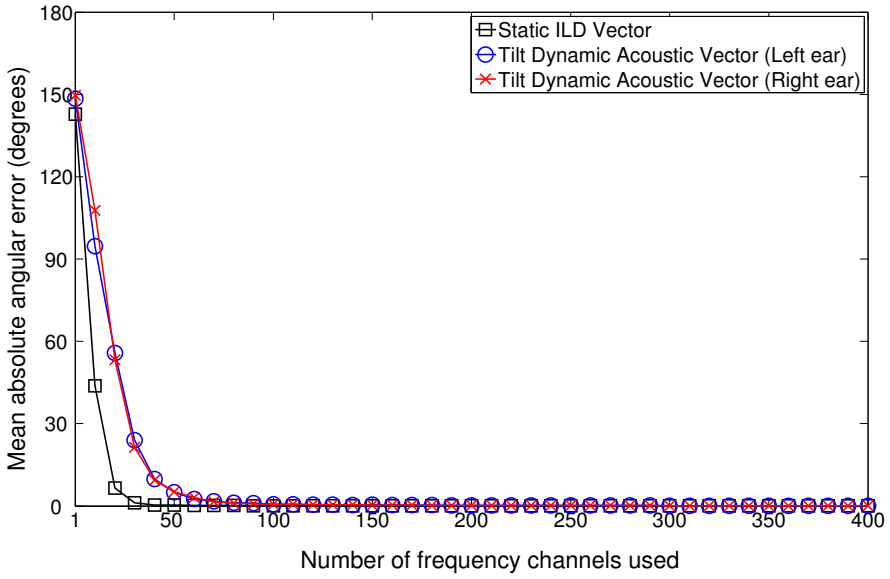


Figure 9: Mean absolute angular error with respect to the number n of frequency channels represented in the perceived sound, using ILD or $c_\beta-$dynamic acoustic vectors.

Fig. 9 summarizes the influence of the number of represented frequency channels on the mean absolute angular error. One can see that using ILD vectors, the error is below 2° for sparse spectra only containing 40 frequency channels, while using dynamic vectors one needs at least 80 channels in order to achieve the same accuracy. This result correlates with several psychophysical and behavioral studies suggesting that the accuracy of vertical sound localization by humans is weaker for narrow-band sound sources than for broad-band sound sources Roffler and Butler (1968), Gardner and Gardner (1973), Butler and Helwig (1983).

7 Conclusion

Computational sound localization has long been addressed using static acoustic features such as ILD and ITD. Based on a number of psychophysical studies suggesting

that sound localization could be both a listener-dependent and a dynamic problem, we proposed a novel unsupervised learning approach making use of high-dimensional information available with binaural and monaural dynamic auditory data, i.e., data gathered with a listener that moves its head while recording sounds. Our method is able to retrieve an intrinsic spatial parameterization of training sets of acoustic data in the presence of a single emitting source. This parameterization can then be associated with the ground-truth motor parameters of the listener, thus allowing to infer the direction of unknown auditory observations. Results obtained with our approach put forward manifold learning as a powerful tool for learning sound localization with robotic heads. They also quantitatively support several psychophysical studies implying that the use of head movements in combination with the spectral richness of perceived sounds could improve monaural sound localization. In addition, the idea that the HRTF can be viewed as a function of motor-states, and more generally that the meaning of sensory inputs can be learned in terms of their corresponding motor actions rather than their corresponding external parameters (i.e. source's position) strongly supports sensorimotor theories of human development notably put forward in Poincaré (1929), Held and Hein (1963), and more recently in ORegan and Noe (2001).

In the future, we plan to test the robustness of our approach for sound localization while varying external factors, that is, by moving the emitter to various positions around the robot, with a large variety of emitted sounds, in different rooms with different environmental conditions. Furthermore, we believe that a promising direction of investigation towards concrete applications would consist in using the proposed model for the task of sound-source separation. Indeed, our technique allows to obtain low-dimensional intrinsic parameterizations of unknown acoustic observations even with a fair amount of missing frequencies. We believe that these parameterizations could be used to optimally cluster the spectrum of an auditory observation according to the emitters' location, and therefore separate competing sound sources.

References

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

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, **15**(6), 1373–1396.

Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press.

Butler, R. A. and Helwig, C. C. (1983). The spatial attributes of stimulus frequency in the median sagittal plane and their role in sound localization. *American Journal of Otolaryngology*, **4**, 165–173.

Cherry, E. C. (1953). Some experiment on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, **25**(5), 975–979.

Cox, T. and Cox, M. (1994). *Multidimensional Scaling*. Chapman and Hall, London.

Fisher, H. G. and Freedman, S. J. (1968). The role of the pinna in auditory localization. *J. Audit. Res.*, **8**, 15–26.

Gardner, M. B. and Gardner, R. S. (1973). Problem of localization in the medial plane: effect of pinnae cavity occlusion. *Journal of the Acoustical Society of America*, **53**, 400–408.

Greff, R. and Katz, B. F. (2007). Perceptual evaluation of HRTF notches versus peaks for vertical localisation. In *The Nineteenth International Congress on Acoustics*.

Handzel, A. A. and Krishnaprasad, P. S. (2002). Biomimetic sound-source localization. *IEEE Sensors Journal*, **2**, 607–616.

Hansard, M. and Horaud, R. P. (2010). Cyclorotation models for eyes and cameras. *IEEE Transactions on System, Man, and Cybernetics–Part B: Cybernetics*, **40**(1), 151–161.

Haykin, S. and Chen, Z. (2005). The cocktail party problem. *Neural Computation*, **17**, 1875–1902.

Hebrank, J. and Wright, D. (1974). Spectral cues used in the localisation of sound sources on the median plane. *Journal of the Acoustical Society of America*, **56**.

Held, R. and Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *J. Comp. Physiol. Psych.*, **56**(5), 872–876.

Hörnstein, J., Lopes, M., Santos-victor, J., and Lacerda, F. (2006). Sound localization for humanoid robots building audio-motor maps based on the hrtf. contact project report. In *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1170–1176.

King, A. J. and Schnupp, J. W. H. (2007). The auditory cortex. *Current Biology*, **17**(7), 236–239.

Kneip, L. and Baumann, C. (2008). 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*, **124**, 3108–3119.

Lu, Y.-C. and Cooke, M. (2010). Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources. **18**, 1793–1805.

Middlebrooks, J. C. and Green, D. M. (1991). Sound localization by human listeners. *Annual Review of Psychology*, **42**, 135–159.

Muller, R. and Schnitzler, H.-U. (1999a). Acoustic flow perception in cf-bats: properties of the available cues. *Journal of the Acoustical Society of America*, **105**, 2959–2966.

Muller, R. and Schnitzler, H.-U. (1999b). Acoustic flow perception in cf-bats: properties of the available cues. *Journal of the Acoustical Society of America*, **105**, 2959–2966.

Muller, R. and Schnitzler, H.-U. (2001). Computational assessment of an acoustic flow hypothesis for cf-bats. In *Computational models of auditory function*. IOS Press.

Otani, M., Hirahara, T., and Ise, S. (2009). Numerical study on source-distance dependency of head-related transfer functions. *Journal of the Acoustical Society of America*, **125**(5), 3253–61.

ORegan, K. J. and Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, **24**, 939–1031.

Poincaré, H. (1929). *The Foundations of science; Science and Hypothesis, the Value of Science, Science and Method*. New York: Science Press. Translated from French by Halsted. Original title: La Valeur de la Science, 1905.

Pollack, I. and Rose, M. (1967). Effect of head movement on the localization of sounds in the equatorial plane. *Percept. Psychophys.*, **2**, 591–596.

Raspaud, M., Viste, H., and Evangelista, G. (2010). Binaural source localization by joint estimation of ild and itd. **18**(1), 68–77.

Roffler, S. K. and Butler, R. A. (1968). Factors that influence the localization of sound in the vertical plane. *Journal of the Acoustical Society of America*, **43**, 1288–1259.

Roman, N. and Wang, D. (2008). Binaural tracking of multiple moving sources. **16**(4), 728–739.

Saul, L. and Roweis, S. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, **4**, 119–155.

Scholkopf, B., Smola, A. J., and Muller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299–1319.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.

Thurlow, W. R. and Runge, P. S. (1967). Effect of induced head movements on localization of direction of sounds. *Journal of the Acoustical Society of America*, **42**, 480–488.

Walker, V. A., Peremans, H., and Hallam, J. C. T. (1998). One tone, two ears, three dimensions: A robotic investigation of pinnae movements used by rhinolophid and hipposiderid bats. *Journal of the Acoustical Society of America*, **104**, 569–579.

Wallach, H. (1939). On sound localization. *Journal of the Acoustical Society of America*, **10**, 270–274.

Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *J. Exp. Psychol*, **27**, 338–368.

Wang, D. and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. IEEE Press.

Wenzel, E. M. (1995). The relative contribution of interaural time and magnitude cues to dynamic sound localization. In *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*.

Willert, V., Eggert, J., Adamy, J., Stahl, R., and Koerner, E. (2006). A probabilistic model for binaural sound localization. **36**(5), 982–994.

Woodworth, R. S. and Schlosberg, H. (1965). *Experimental Psychology*. Holt.

Zhang, Z. and Zha, H. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, **26**(1).

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399