Aural Servo: Sensor-Based Control From Robot Audition

Aly Magassouba, Nancy Bertin, *Member, IEEE*, and François Chaumette, *Fellow, IEEE*

***Abstract*—This paper proposes a control framework based on auditory perception. Generally, in robot audition, the motion control of a robot from the sense of hearing relies on sound source localization. We propose in this paper an alternative approach, *aural servo*, which is derived from the sensor-based control framework. In this approach, robot motions are directly connected to the aural perception: The variation of low-level auditory features dictates the motions applied to the robot through a feedback loop. It has the advantage of being robust to spurious measurements and modeling approximations for a low computational cost. This paper presents the theoretical concept of the aural servo framework. Besides a theoretical analysis, the aural servo framework is validated through several experiments on different robotic platforms and under real-world conditions.**

***Index Terms*— Interaural level difference (ILD), interaural time difference (ITD), robot audition, sensor-based control.**

I. Introduction

EXPLOITING the sense of hearing in robotics is still a challenging topic, especially when controlling robot motions with respect to sound source(s). Nowadays, controlling robot motions from this information usually follows a workflow that consists in first, extracting the auditory cues related to the propagation of the sound, second, inferring the sound source location from these cues, and finally, moving the robot according to this location.

In robot audition, a lot of efforts have been dedicated to the sound localization process, that is, the first two steps given above. This focus is explained by the vast literature in signal processing and psychoacoustics supporting this topic. Besides, sound localization concerns applications beyond the robotic context. Actually sound localization can be referred to as a machine hearing problem with applications for hearing aids, conferencing systems or surveillance. In robotics, several applications are based on sound localization. These applications generally consider microphone(s) array setups for more robustness. In this context, controlling robot motions can be used for instance in acoustic monitoring [1] or in search-and-rescue missions where emergency signal can be detected and approached [2]. In human-robot interaction (HRI), motion control gives more naturalness to the interaction (e.g., gazing towards the speaker [3]) or can be used to improve the interaction (e.g., approaching a speaker to hear "better"). Sound localization can also be used as a modality of robot navigation based on acoustic landmarks [4], [5]. Despite this potential, sound localization in realistic environments turns out to be particularly complex, especially when considering binaural setups, as achieved in this paper.

In general, binaural sound localization methods rely on knowledge in psychoacoustics and physiology for extracting auditory cues from an artificial hearing system. Auditory cues such as the interaural time difference (ITD) and the interaural level difference (ILD) are particularly exploited in robot audition [6]–[8], since they provide information about the sound azimuth direction. The efficiency of the localization process strongly depends on the accuracy of these cues and their interpretation into spatial coordinates. The complexity of localization arises from sound perception and more particularly auditory cues that are influenced by each morphology and acoustic conditions (room acoustics, noise, reverberation, and signal frequency). The sensitivity to auditory events is variable for each individual and each location. Real-world configurations that include changing conditions, reverberation or noises, degrade drastically the accuracy of the auditory cues, and by extension the accuracy of the localization process. As of these limitations, when considering binaural setups, only few works addressed real world conditions [9]. Static configurations in controlled environments are usually assumed, although recent developments in the so-called *active audition* [10] tend to address more realistic situations [11]–[13].

In contrast with the approach discussed, we propose in this paper a feedback control system, *aural servo*, linking directly auditory cues to robot control. Instead of following the conventional localization workflow, the motion control is stemmed from the auditory cues variation in a feedback loop that skips the localization step. In this way, the complexity of interpreting auditory cues into spatial coordinates is avoided. In robot audition literature, feedback controllers are seldom used. Besides an application about a Theremin-playing robot [14], feedback controllers have been exploited for robot gaze control in [3], [15], and [16]. In [3], a robot learns the acoustic map space (i.e., relation between ILD and ITD with azimuth and elevation) and uses a feedback-error learning scheme to orient itself towards a sound source. In [15] and [16], the authors present a control scheme derived from an empiric cost function characterizing the relationship between the position of a source and the orientation of the robot head. More recently, [13] developed

Manuscript received May 14, 2017; revised November 2, 2017; accepted December 27, 2017. Date of publication March 20, 2018; date of current version June 6, 2018. This paper was recommended for publication by Associate Editor D. Scaramuzza and Editor C. Torras upon evaluation of the reviewers' comments. *(Corresponding author: Aly Magassouba.)*

The authors are with the Inria, Univ Rennes, CNRS, IRISA, Campus de Beaulieu, Rennes 35042, France (e-mail: aly.magassouba@irisa.fr; nancy.bertin@irisa.fr; francois.chaumette@inria.fr).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. This file contains a video recording of the experiments reported in the paper. The video is very helpful for appreciating the effectiveness of the control framework proposed in the paper, in real world conditions.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2018.2805310

an information-based feedback-loop in order to minimize the uncertainty of localization during an active audition process.

With respect to these works, aural servo consists in a more general approach utilizing the sensor-based control framework. Sensor-based control is nowadays widely developed. The vision and touch senses are the common sensory feedback exploited in this framework. The well-known visual servoing [17] illustrates the wide range of application of such an approach. Robots physical interactions (e.g., grasping tasks) also benefit from the sensor-based approach through proximity [18] or tactile sensors [19].

This paper shows the benefits of using sensor-based control in robot audition. From auditory cues modeling, we develop in this paper several control schemes, based on ITDs, ILDs, and the sound energy level, that can cope with real acoustic conditions in real-time. This paper extends and synthesizes our previous works [20]–[22] by providing new theoretical results (i.e., theoretical stability conditions), evaluating the performance and limits of our approach in comparison to classical localization methods, and by presenting new experimental results, particularly on humanoid robots (from a loudspeaker and by a person directly interacting with the robot). The feedback loop that allows to discard inconsistent measurements and the high resilience of the system to modeling approximations and to punctual erroneous measurements mainly explain the robustness of this approach to real acoustic scenes.

The latter properties and results are emphasized in the rest of this paper that is structured as follows. Section II introduces the main principles of sensor-based control. In Section III, by considering a system endowed with two microphones, the relation between auditory features (ITD, ILD, and the energy level) variation and the microphones motion are characterized through an interaction matrix. The feedback loop is then expressed by a control scheme supported by stability proof in Section IV. The stability conditions of the obtained control schemes allow us to demonstrate the robustness of this approach towards modeling errors. These control schemes are thereafter numerically evaluated and experimentally validated on robots. Experimental results conducted on a mobile robot equipped with free-field microphones and on humanoid robots confirm the relevance of our approach through several positioning and tracking tasks. In Section V, the robot control is performed from individual features, considering free-field microphones. Subsequently more advanced tasks, based on several auditory cues or sound sources, are developed and experimented in Section VI. Ultimately, Section VII addresses the context of humanoid robots.

II. SENSOR-BASED CONTROL

A. Aural Servo Principle

The principle of aural servo consists in controlling the robot motion directly from auditory cues instead of spatial references as performed by sound source localization methods. In this approach, control and sound perception are directly connected. This relation is explicitly developed through the task function formalism [23]. A task consists in a set of auditory measurement conditions to reach. For instance, in a localization approach, it is necessary to extract the azimuth angle of a sound source in order to orient the robot towards this source. In our approach, the same result is obtained by computing a real-time motion that makes ITD measurements converge towards 0. Controlling robot motions in such a manner can be decomposed into three steps. The first step consists in selecting the feature input(s) of the closed loop control system (see Fig. 1), as well as a reference value(s) to be reached. The second step consists in modeling the relationship between the robot motion and the auditory feature(s) variation through the interaction matrix [17]. Finally, the velocity of the robot is computed from the interaction matrix and the current feature measurements.

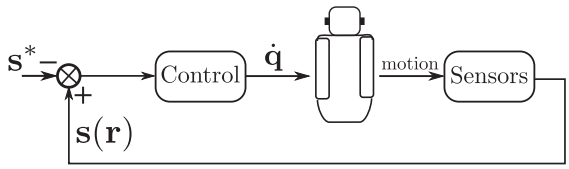


Fig. 1. Aural servo control scheme: A sensor-based approach links the motion of the robot to the sensor measurement $\mathbf{s}(\mathbf{r})$ in a feedback loop until the robot reaches a desired configuration characterized by a demanded measurement $\mathbf{s}^*$.

B. Sensor-Based Control Formulation

The different steps evoked above are formalized in [17] by considering a feature set noted $\mathbf{s}$, that is the input of the closed-loop control system. Once $\mathbf{s}$ is selected, the relation expressed by the interaction matrix $\mathbf{J_s}$ between the feature variation $\dot{\mathbf{s}}$ and the sensor velocity $\mathbf{u}$ is given by

$$\dot{\mathbf{s}} = \mathbf{J_s}\mathbf{u} \tag{1}$$

in which $\mathbf{J_s} \in \mathbb{R}^{k\times n}$ is sized by k the dimension of $\mathbf{s}$ and n the dimension of $\mathbf{u}$. The dimension of $\mathbf{u}$ depends on the linear and angular spatial velocity components that are controlled among $\mathbf{v} = (v_x, v_y, v_z, \omega_x, \omega_y, \omega_z)$. The goal of the closed-loop control system is to minimize the error $\|\mathbf{e}(t)\|$ defined from the task function

$$\mathbf{e}(t) = \mathbf{s}(t) - \mathbf{s}^* \tag{2}$$

where $\mathbf{s}^*$ denotes the desired value of $\mathbf{s}$. Then, a simple control scheme can be designed with a purpose of exponential decoupled decrease of the task function [24]. In this case, the time variation of $\mathbf{e}$ should follow $\dot{\mathbf{e}} = -\lambda\mathbf{e}$, with $\lambda > 0$ a gain that tunes the time to convergence. Then, we obtain

$$\mathbf{u} = -\lambda\mathbf{J_s^+}\mathbf{e} \tag{3}$$

where $\mathbf{J_s^+} \in \mathbb{R}^{n\times k}$ is the Moore–Penrose pseudoinverse of the interaction matrix ($\mathbf{J_s^+} = \mathbf{J_s^{-1}}$ when $\mathbf{J_s}$ is invertible). Nonetheless in real configurations, it is usually impossible to know perfectly $\mathbf{J_s}$ since the interaction matrix may depend on quantities that cannot be directly measured by the sensors. Thus. generally an approximation $\widehat{\mathbf{J_s^+}}$ of $\mathbf{J_s^+}$ is used in (3).

In addition, the results developed in the sequel are also supported by stability proofs. This analysis relies on Lyapunov stability conditions of nonlinear systems [25]. In our configuration, the global asymptotic stability, allowing the system to converge toward the desired configuration whatever its initial position, is guaranteed as soon as [17]

$$\mathbf{J_s}\widehat{\mathbf{J_s^+}} > 0. \tag{4}$$

III. AUDITORY FEATURES FOR AURAL SERVO

This section is dedicated to the modeling of the auditory features that can be used as inputs of the control scheme. Naturally,

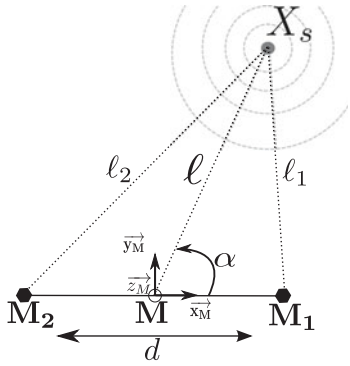


Fig. 2. Geometric configuration of the considered problem, that includes a source $\mathbf{X_s}$ emitting a spherical sound wave, and a pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$.

we have taken inspiration from the vast literature in signal processing and audition to choose the ILD, the ITD and the sound energy level.

A. Scene Configuration

Let us consider a pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$, separated by a distance d, that are embedded on a mobile robot moving in an area free of obstacles (see Fig. 2). A reference frame $\mathcal{F}_m(\mathbf{M}, \overrightarrow{x_M}, \overrightarrow{y_M}, \overrightarrow{z_M})$ is attached to the pair of microphones and originates from its midpoint $\mathbf{M}$. In this frame, the Cartesian coordinates of each microphone are, respectively, $\mathbf{M_1}(\frac{d}{2}, 0, 0)$ and $\mathbf{M_2}(-\frac{d}{2}, 0, 0)$. Then, we consider a point-wise sound source $\mathbf{X_s}(x_s, y_s, 0)$ that, without loss of generality, belongs to the xy-plane parallel to the ground. This hypothesis allows simplifying the analytical developments presented in the sequel while it does not have to be ensured in practice. Besides, $\mathbf{X_s}$ emits continuously an omnidirectional sound wave $a(t)$ in a uniform medium where linear acoustic hypotheses hold. Finally we assume that $\mathbf{X_s}$ is in front of the microphones (i.e., $y_s > 0$). Such an assumption will be overcome afterwards (see Section V-B3). The distances (ℓ_i, ℓ) between $\mathbf{X_s}$ and the pair of microphones are then

$$\begin{cases} \ell_1 = \sqrt{(x_s - d/2)^2 + y_s^2} \\ \ell_2 = \sqrt{(x_s + d/2)^2 + y_s^2} \\ \ell = \sqrt{x_s^2 + y_s^2} \end{cases}. \tag{5}$$

Additionally, α denotes the incident angle of the sound source with respect to the microphones axis. α is also known as the sound direction of arrival (DOA). The sound source position is then characterized by the following relationships:

$$x_s = \ell \cos\alpha, y_s = \ell \sin\alpha \text{ and } \alpha = \text{atan2}\,(y_s, x_s). \tag{6}$$

B. ITD Modeling

Let us now focus on the properties of the ITD τ in the configuration described earlier. From Fig. 2, the sound wave emitted from $\mathbf{X_s}$ reaches each microphone $\mathbf{M_i}$ at a time t_i given by $t_i = \ell_i / c$, in which c is the sound velocity. As a result, the ITD τ between the pair of microphones is

$$\tau = t_2 - t_1 = \frac{\ell_2}{c} - \frac{\ell_1}{c}. \tag{7}$$

In practice ITDs are generally estimated from cross-correlation methods. The method used in this paper is detailed in Section V-B1. Using (7) for recovering the source location defines a half-hyperbola [22]. For distant sound sources, this hyperbola can be approximated by its asymptotes so that the ITD τ becomes

$$\tau = A \cos\alpha \tag{8}$$

where $A = d/c$. Equation (8) expresses the relationship between ITD and DOA under the far-field (FF) assumption. This relationship is commonly exploited by localization methods [26], [27] in order to estimate azimuth angles. However, this assumption can lead to a substantial error in azimuth estimation as the source gets closer to the microphones, since the approximation does not hold anymore.

From the definitions (7) and (8) of the ITD, we can design two interaction matrices characterizing the relationship between the microphones motion and this feature variation. We denote $\mathbf{J}_{\tau_r}$ the interaction matrix obtained from the modeling of τ given by (7). In this case, the time derivative of τ is

$$\dot{\tau} = \frac{1}{c}\left(\dot{\ell}_2 - \dot{\ell}_1\right). \tag{9}$$

By injecting (5) in (9), the derivative of τ develops as

$$\dot{\tau} = \frac{1}{c}\left(\frac{2\dot{x}_s x_s + 2\dot{y}_s y_s + d\dot{x}_s}{2\ell_2} - \frac{2\dot{x}_s x_s + 2\dot{y}_s y_s - d\dot{x}_s}{2\ell_1}\right). \tag{10}$$

By using the well-known kinematic equation [17]

$$\dot{\mathbf{X}}_\mathbf{s} = -\boldsymbol{v_s} - \boldsymbol{\omega_s} \times \mathbf{X_s} \Leftrightarrow \begin{cases} \dot{x}_s = -v_x - \omega_y z_s + \omega_z y_s \\ \dot{y}_s = -v_y - \omega_z x_s + \omega_x z_s \\ \dot{z}_s = -v_z - \omega_x y_s + \omega_y x_s \end{cases} \tag{11}$$

that relates the velocity of a three- dimensional (3-D) point $\mathbf{X_s}$ to the sensor spatial velocity $\mathbf{v}$, (10) becomes

$$\dot{\tau} = v_x \frac{x_s\tau - \frac{A}{2}(\ell_1 + \ell_2)}{\ell_1\ell_2} + v_y \frac{y_s\tau}{\ell_1\ell_2} + \omega_z \frac{\frac{A}{2}(\ell_1 + \ell_2)y_s}{\ell_1\ell_2}. \tag{12}$$

In this equation, we can notice that any motion along v_z, ω_x, or ω_y does not influence τ. Hence, the relevant motions of the microphones are the translations along $\overrightarrow{x_M}$ and $\overrightarrow{y_M}$ axis, and the rotation around $\overrightarrow{z_M}$. The three degrees of mobility of the microphone system are then characterized by $\mathbf{u} = (v_x, v_y, \omega_z)$. Finally, $\mathbf{J}_{\tau_r}$ can be extracted from (12) as

$$\mathbf{J}_{\tau_r} = \begin{bmatrix} \frac{x_s\tau - \frac{A}{2}(\ell_1+\ell_2)}{\ell_1\ell_2} & \frac{y_s\tau}{\ell_1\ell_2} & \frac{\frac{A}{2}(\ell_1+\ell_2)y_s}{\ell_1\ell_2} \end{bmatrix}. \tag{13}$$

Analogously to binaural localization, unknown parameters depending on the source location (x_s, y_s, and ℓ_i) appear in $\mathbf{J}_{\tau_r}$. Knowing that $(x_s, y_s, \ell_i) = f(\ell, \tau)$, an approximated interaction matrix $\widehat{\mathbf{J}_{\tau_r}} = \mathbf{J}_{\tau_r}(\widehat{\ell})$ (with the assumption that τ_r can be measured) is then considered to develop the control scheme.

The FF assumption can also be exploited to design a different interaction matrix based on the link between the DOA α and the ITD τ in (8). This interaction matrix denoted $\mathbf{J}_{\tau_f}$ is given by (see [21] for details)

$$\mathbf{J}_{\tau_f} = \begin{bmatrix} -\frac{\nu^2}{A\ell} & \frac{\tau\nu}{A\ell} & \nu \end{bmatrix} \tag{14}$$

where $\nu = \sqrt{A^2 - \tau^2}$. Similarly to (13), the approximation of the latter interaction matrix is defined with $\widehat{\mathbf{J}_{\tau_f}} = \mathbf{J}_{\tau_f}(\widehat{\ell})$. It should also be noticed that $\mathbf{J}_{\tau_f}$ is naturally linked to $\mathbf{J}_{\tau_r}$. Indeed, under the FF assumption, the following hypothesis $\ell_1 \approx \ell_2 \approx \ell$ holds. Hence, (13) can be rewritten as

$$\mathbf{J}_{\tau_r} \simeq \left[\begin{array}{ccc} \frac{x_s \tau - A\ell}{\ell^2} & \frac{y_s \tau}{\ell^2} & \frac{y_s A \ell}{\ell^2} \end{array} \right]. \tag{15}$$

Subsequently, by replacing τ, x_s and y_s by using (8) and (6), the interaction matrix becomes

$$\mathbf{J}_{\tau_r} \simeq \left[-\frac{A \sin^2 \alpha}{\ell} \quad \frac{A \sin \alpha \cos \alpha}{\ell} \quad A \sin \alpha \right] = \mathbf{J}_{\tau_f} \tag{16}$$

that corresponds to the interaction matrix $\mathbf{J}_{\tau_f}$ when expressing $\sin \alpha$ and $\cos \alpha$ with respect to τ.

C. ILD Modeling

As far as ILD is concerned, under the spherical sound propagation assumption the signal recorded at each microphone is

$$x_i(t) = \frac{a\left(t - \frac{\ell_i}{c}\right)}{\ell_i} \tag{17}$$

where $\frac{\ell_i}{c}$ expresses the sound propagation delay and $a(t)$ is the sound wave defined in the beginning of Section III. By integrating (17) over a frame of length w during which the signal is observed, the energy received by each microphone is defined as follows:

$$E_i = \int_{t=0}^{w} |x_i(t)|^2 \, \mathrm{d}t = \frac{1}{\ell_i^2} \int_{t=0}^{w} a^2 \left(t - \frac{\ell_i}{c}\right) \mathrm{d}t. \tag{18}$$

Equation (18) characterizes the inverse-square law property inherent to spherical sound propagation. The ILD ρ between the two microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ is then computed from

$$\rho = \frac{E_1}{E_2} = \frac{\ell_2^2 \int_{t=0}^{w} a^2\left(t - \frac{\ell_1}{c}\right) \mathrm{d}t}{\ell_1^2 \int_{t=0}^{w} a^2\left(t - \frac{\ell_2}{c}\right) \mathrm{d}t}. \tag{19}$$

Assuming that during w, the recorded sound signal varies little between the two microphones, we can expect that $\int_{t=0}^{w} a^2(t - \frac{\ell_1}{c}) \, \mathrm{d}t \approx \int_{t=0}^{w} a^2(t - \frac{\ell_2}{c}) \, \mathrm{d}t$. Consequently, ρ can be simplified without significant loss of accuracy by

$$\rho = \frac{\ell_2^2}{\ell_1^2}. \tag{20}$$

Using (20) to recover the source location defines a circle (see [28]), apart from the case when $E_1 = E_2$ that defines the perpendicular bisector of the microphones. This circle is centered on the point $(\frac{d}{2}\frac{E_1+E_2}{E_1-E_2}, 0)$ with a radius $c_r = d\left|\frac{\sqrt{\rho}}{1-\rho}\right|$. Such a result leads to ambiguities where no azimuth angle or distance can be directly extracted from ILD cues. This mainly explains why ILD cues are more complex to exploit compared to ITD cues by localization methods even though they provide useful information at high frequencies. Fortunately it is still possible to directly exploit this cue in our case without any additional knowledge since we are not inferring the sound azimuth

From (20) we have

$$\dot{\rho} = \frac{d}{dt}\left(\frac{\ell_2^2}{\ell_1^2}\right) = 2\frac{\ell_2 \dot{\ell}_2 \ell_1 - \dot{\ell}_1 \ell_2^2}{\ell_1^3}. \tag{21}$$

Following similar developments as for the ITD, the interaction matrix $\mathbf{J}_\rho$ can be determined. It is given by [20]

$$\mathbf{J}_\rho = \left[\begin{array}{ccc} \frac{2x_s(\rho-1) - d(\rho+1)}{\ell^2 + \frac{d^2}{4} - dx_s} & \frac{2y_s(\rho-1)}{\ell^2 + \frac{d^2}{4} - dx_s} & \frac{y_s d(\rho+1)}{\ell^2 + \frac{d^2}{4} - dx_s} \end{array} \right]. \tag{22}$$

The approximated interaction matrix ensued from (22) is then $\widehat{\mathbf{J}_\rho} = \mathbf{J}_\rho(\widehat{\ell})$.

D. Absolute Level of Energy Modeling

Eventually, in order to characterize the relationship between the source location and the absolute level of sound energy, we rely on the sound decay properties. Considering $\mathbf{M}$ as the reference point, the energy received in $\mathbf{M}$ follows the relationship given by (18)

$$E_{\mathbf{M}} = \frac{1}{\ell^2} \int_{t=0}^{w} a^2 \left(t - \frac{\ell}{c}\right) \mathrm{d}t. \tag{23}$$

Equation (23) lets us state that from all points located at the same distance ℓ to the sound source, the same amount of energy $E_{\mathbf{M}}$ is measured. The absolute level of sound energy is then linked to the distance to the sound source, by a proportional gain depending on the intrinsic level of $a(t)$. As of this property, the distance to the sound source cannot be directly extracted from (23), unless the signal emitted $a(t)$ is exactly known. As a consequence, such a feature can hardly be exploited for source localization while we propose in Section VI a way to exploit this cue through aural servo in real situations. To this end, similarly to the ILD case, the interaction matrix $\mathbf{J}_{E_{\mathbf{M}}}$ related to the sound energy perceived in $\mathbf{M}$ can be determined. It is given by (see [20])

$$\mathbf{J}_{E_{\mathbf{M}}} = E_{\mathbf{M}} \left[\begin{array}{ccc} \frac{2x_s}{\ell^2} & \frac{2y_s}{\ell^2} & 0 \end{array} \right]. \tag{24}$$

Then, similarly to the ITD and ILD cases, (24) is approximated through $\widehat{\mathbf{J}_{E_{\mathbf{M}}}} = \mathbf{J}_{E_{\mathbf{M}}}(\widehat{\ell})$.

IV. Basic Tasks

In this section, we consider each auditory cue modeled in the previous section as a single input of the control scheme. We will see that it corresponds to different tasks constraining either the orientation or the range of the sensor. Whether ρ, τ, or $E_{\mathbf{M}}$ is used, the control scheme follows the same form given by (3), in which $\mathbf{J_s}$ has just to be replaced by the corresponding approximated interaction matrix. These matrices contain unknown terms related to the source location. At a first glance, such results are conflicting with the purpose of aural servo. Fortunately it remains possible to approximate these matrices without any knowledge of the source location by relying on the Lyapunov analysis for ensuring the controller convergence.

A. ITD-Based Task

For a task considering an ITD as input feature of the control loop, i.e., $e_\tau = \tau - \tau^*$, the control scheme is given by

$$\mathbf{u} = -\lambda \widehat{\mathbf{J}_\tau^+} e_\tau. \tag{25}$$

This task consists in orienting the microphones towards a particular direction with respect to the source location [21]. First, let us note that whatever the choice of $\mathbf{J}_\tau = \mathbf{J}_{\tau_r}$ or $\mathbf{J}_\tau = \mathbf{J}_{\tau_f}$, the control system is not singular as long as $y_s \neq 0$ (which is equivalent to $\alpha \neq k\pi$, $\forall k \in \mathbb{N}$ with $\mathbb{N} = \{0, 1, 2, ...\}$ and $|\tau| < A$,

and is coherent with our primary assumption $y_s > 0$) and $\ell_i \neq 0$ or $\ell \neq 0$. Any approximations of $\mathbf{J}_\tau$ with $\ell_i \neq 0$ or $\ell \neq 0$ will thus avoid the system to be singular.

More interestingly, these approximations can be better shaped by the Lyapunov global asymptotic stability condition that is obtained when

$$\mathbf{J}_\tau \widehat{\mathbf{J}_\tau^+} > 0. \tag{26}$$

Under the FF assumption, that is when $\mathbf{J}_\tau = \mathbf{J}_{\tau_f}$, the latter condition is written as

$$\frac{\widehat{\ell}^2}{\ell\widehat{\ell}} \frac{\nu^4 + \tau^2\nu^2 + A^2\widehat{\ell}\ell\nu^2}{\nu^4 + \tau^2\nu^2 + A^2\widehat{\ell}^2\nu^2} > 0. \tag{27}$$

We can immediately deduce that $\widehat{\ell} > 0$ and $\nu \neq 0$ is a sufficient condition of stability. In the same vein, similar results are obtained when considering the more rigorous ITD modeling through $\mathbf{J}_\tau = \mathbf{J}_{\tau_r}$. Without making explicit all steps (see [22] for details), the Lyapunov stability is ensured as long as sign $(\widehat{x}_s) =$ sign (x_s), assuming that the sound source is always located in the front side of the microphones (i.e., $y_s > 0$ and $\alpha \in]0; \pi[$).

Both ITD models stress that sufficient conditions for the global asymptotic stability of the system are obtained when the quadrant containing the sound source is known. Given our initial assumption in Section III-A of a sound source located in front of the robot, it is then sufficient to know if the source is located on the left or right side. These conditions are trivial to ensure in practice. Even with rough ITD estimation, an adequate robot motion is computed so that the error converges towards 0, which emphasizes one of the benefits of our approach.

B. ILD-Based Task

For a task characterized by $e_\rho = \rho - \rho^*$, the control scheme is given by

$$\mathbf{u} = -\lambda \widehat{\mathbf{J}_\rho^+} e_\rho. \tag{28}$$

Similarly to the ITD case, this task consists in orienting the microphones towards a particular direction with respect to the source location. To this end, from (22) $\widehat{\mathbf{J}_\rho^+}$ should be approximated so that $\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s > 0$ to avoid any singular configuration. We can also notice that unlike the ITD case, the configuration $y_s = 0$ is not singular. This property will be exploited experimentally in Section V-B3. Furthermore, the selection of $\widehat{x}_s$, $\widehat{y}_s$, and $\widehat{\ell}$ can be eased by taking into account the sound location on the circle described in Section III-C and in [28, Fig.1]. For that, it is possible to set $\widehat{y}_s$ at any value $]0; c_r[$ (c_r being the ILD circle radius as defined in Section III-C), and to deduce a corresponding $\widehat{x}_s$, and then $\widehat{\ell} = \sqrt{\widehat{x}_s^2 + \widehat{y}_s^2}$.

More rigorously, the Lyapunov stability condition is given by

$$\mathbf{J}_\rho \widehat{\mathbf{J}_\rho^+} > 0. \tag{29}$$

This condition is ensured as long as sign $(\widehat{x}_s) =$ sign (x_s). This condition is trivial to guarantee in practice since sign $(\widehat{x}_s) =$ sign (x_s) is obtained directly from ρ since sign $(x_s) =$ sign $(\rho - 1)$. Such a favorable stability condition confirms that the exact source location is not required to ensure the convergence of the controller.

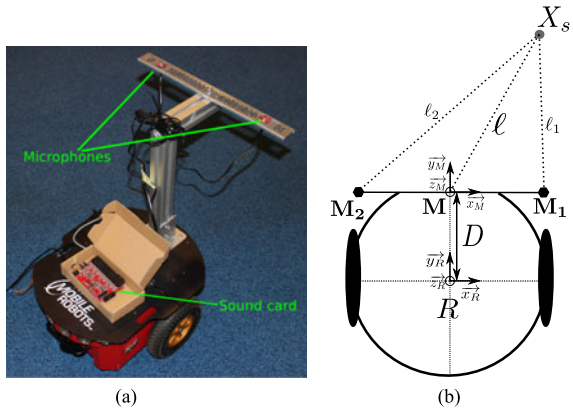


Fig. 3. Modeling of the robotic platform.

C. Energy Level-Based Task

At last, when considering the energy measurement as input of the feedback loop, i.e., $e_{\mathrm{M}} = E_{\mathrm{M}} - E_{\mathrm{M}}^*$ the following control scheme is obtained

$$\mathbf{u} = -\lambda \widehat{\mathbf{J}_{\mathbf{E}_{\mathrm{M}}}^+} e_{\mathrm{M}}. \tag{30}$$

The analysis of the approximated matrix $\widehat{\mathbf{J}_{E_{\mathrm{M}}}}$ is very similar to the ILD case. The sufficient stability conditions is now given by

$$\mathbf{J}_{E_{\mathrm{M}}} \widehat{\mathbf{J}_{E_{\mathrm{M}}}^+} = \frac{\widehat{\ell}^2}{\ell^2} \frac{x_s\widehat{x}_s + y_s\widehat{y}_s}{\widehat{x}_s^2 + \widehat{y}_s^2}. \tag{31}$$

Equation (31) is positive as soon as sign $(\widehat{x}_s) =$ sign (x_s). This result is intuitive since moving along the sound source direction allows controlling the value of E_{M}. However with only E_{M} as input of the control scheme, no information related to the source direction can be inferred. Hence, in order to also control the orientation from the sound source direction, features such as the ITD or ILD should be used alongside with the energy level E_{M}, as it will be done in Section VI. It should also be mentioned that the energy level variation is consistent with the microphones motion only if the mean energy measured over w does not vary with time, which limits the applicability of this approach to continuous stationary signals.

V. Basic Task Results

In this section, we evaluate the control framework developed so far through simulations and experiments in real-world environments. Depending on the feature used in the control scheme, we exhibit the main benefits and limitations for performing a given task.

A. Preliminaries: Robot Modeling and Control Scheme

For the following experiments, we consider a nonholonomic unicycle *Pioneer 3DX* endowed with two omnidirectional microphones as illustrated in Fig. 3. In addition to $\mathcal{F}_m$ attached to the microphones, we define the frame $\mathcal{F}_r(\mathbf{R}, \overrightarrow{x_R}, \overrightarrow{y_R}, \overrightarrow{z_R})$ attached to the robot. D denotes the distance between the center of the robot $\mathbf{R}$ and the midpoint $\mathbf{M}$ of the microphones. The robot can be controlled upon two DOF: The control input $\dot{\mathbf{q}}$ is given by (u, ω), respectively, the translation velocity along $\overrightarrow{y_R}$ and the angular velocity around $\overrightarrow{z_R}$.

From this configuration, the relationship between the variation of a feature $\mathbf{s}$ and the control input $\dot{\mathbf{q}}$ is

$$\dot{\mathbf{s}} = \mathbf{J}\dot{\mathbf{q}} \tag{32}$$

where $\mathbf{J}$ is the feature Jacobian given by $\mathbf{J} = \mathbf{J_s}\mathbf{J_q}$, $\mathbf{J_q}$ being the robot Jacobian. From the model given in Fig. 3, it has the following form

$$\mathbf{J_q} = \begin{bmatrix} 0 & D \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{33}$$

According to (3), the velocity input of the robot is then

$$\dot{\mathbf{q}} = -\lambda\widehat{\mathbf{J}^{+}}(\mathbf{s} - \mathbf{s}^{*}). \tag{34}$$

B. Experiments

Following the analysis performed in the following section, we consider the task of facing a sound source, which can be performed by both ILD and ITD when setting $\rho^* = 1$ or $\tau^* = 0$, respectively.

1) Features Estimation and Tracking: Estimating the ITD is a topic of research of its own and several methods are available in the signal processing literature. In this paper, we based our estimation on GCC-PHAT [29] technique. The ITD τ is estimated from the maximum peak of the cross-correlation function between the signals $x_1(t)$ and $x_2(t)$ as

$$\widehat{\tau} = \underset{\tau}{\operatorname{argmax}} \max_{l} \sum_{f}^{F} \frac{X_1(f,l)X_2^*(f,l)}{|X_1(f,l)X_2^*(f,l)|} e^{\varphi(\tau)}. \tag{35}$$

$X_1(f,l)$ and $X_2^*(f,l)$ are, respectively, the Fourier transform of $x_1(t)$ and the conjugate of the Fourier transform of $x_2(t)$. In ideal conditions, the maximum peak argument of the latter function corresponds to the ITD of the sound source. In real world conditions, the GCC-PHAT output is not that explicit. Instead of having one dominating peak, which makes immediate the estimation of the actual ITD, several plausible but spurious peaks appear because of reverberation and noise. Therefore, under these conditions, p peaks ($p > 1$) may be considered among which the correct peak has to be found. Furthermore, when considering specific signals such as speech, it is very likely to record sparse data temporally and spectrally. Not all the processed frames contain relevant information, since speech is a nonstationary and intermittent signal. This emphasizes the need of a tracking algorithm to detect the correct ITD. Yet, our approach can provide added value to this tracking problem. One of the benefits of coupling motion and perception lies in predicting the feature variation and, hence, limiting the scope of erroneous measurements. More specifically, the Jacobian matrix $\mathbf{J}_\tau$ can be used to infer the evolution of the tracked ITDs in the next time frame. Given τ as the state x and the velocity $\dot{\mathbf{q}}$ applied to the robot, a local prediction model based on (32) is simply given by

$$\begin{cases} \dot{x}(k) = \widehat{\mathbf{J}}(\tau)\dot{\mathbf{q}} \\ x(k+1) = x(k) + T_e\dot{x}(k) \end{cases} \tag{36}$$

where T_e refers to the sampling time of the control loop. This prediction step also gives useful indications on the source activity. If no ITD measurement "fits" with the prediction, it is very likely that the source is inactive. Hence, the unavailable ITD could be replaced by the predicted one. In our application, the tracking step simply consists in selecting the closest peak to the previous ITD value, by taking into account the local prediction model. However, it should be noted that this approach assumes a correct initialization of the tracker, which can be obtained by selecting the most consistent ITD in the first frames while the robot is not moving.

On the other hand, ILD is straightforward to obtain. Each energy E_1, E_2 is estimated by integrating the recorded signal over a given frame. Therefore, such an approach does not require any tracking method, when assuming a single dominant and continuous sound source in the scene.

2) Facing a Sound Source: For the experiments, two microphones connected to an 8SoundsUSB sound card [30] were used. The sound card operates at a frequency of 48 kHz, and provides windows of 256 samples. The ITD is computed from ten consecutive windows (i.e., $\approx$50 ms) that are subsampled at 16 kHz and low-pass filtered in order to reduce the processing time and to better fit with speech frequency range. The tests were operated in a room with a reverberation time, measured at 1 kHz, RT_{60} $\approx$580 ms. Moreover, the measured SNR was around 25 dB in presence of typical constant and diffuse noise caused by computers and ventilation systems leading to spurious ITD peaks at $\tau = 0$.

TABLE I
EXPERIMENTAL SETTINGS

d	0.31 m
D	0.3 m
c (ITD)	343 m·s^{-1}
$\widehat{\ell}$ (ITD)	1 m
$\lambda(x)$ (ITD)	$5e^{(-4000x)}$
$\widehat{y_s}$ (ILD)	1 m
$\widehat{x_s}$ (ILD)	sign $(\rho - 1) \times 1$ m
λ (ILD)	0.5

The parameters used for the experiments are given in Table I. For the ITD task, a loudspeaker emitted a female speech of 10 s played in loop, while for the ILD task, the sound emitted was a Gaussian white noise. It should be mentioned that the ILD experiment can also be performed considering a speech signal if a speech activity detector was available in order to discard erroneous ILDs when the sound source is not active. Despite our modeling based on a planar scene (see Section III), the loudspeaker position does not ensure such a configuration, since it is at a higher height than the microphones. It should also be mentioned that for the ITD-based task, we used the interaction matrix $\widehat{\mathbf{J}_\tau} = \widehat{\mathbf{J}_{\tau_f}}$ obtained from the FF assumption. The difference between $\widehat{\mathbf{J}_{\tau_r}}$ and $\widehat{\mathbf{J}_{\tau_f}}$ is evaluated later in Section V-C.

For both experiments, the task is successfully achieved (see Fig. 4 and the accompanying video), given an initial arbitrary orientation of the robot with respect to the sound source. At the end of the task, the errors e_τ and e_ρ vanished, while the robot faced the sound source. It should also be mentioned that our approach assumes that the sound source length is long enough so that the robot reaches the desired configuration. This is not always the case depending on the context of the task, especially for HRI. A better tuning of the gain λ to reduce the convergence time, predictions or the use of other modalities (e.g., vision) are

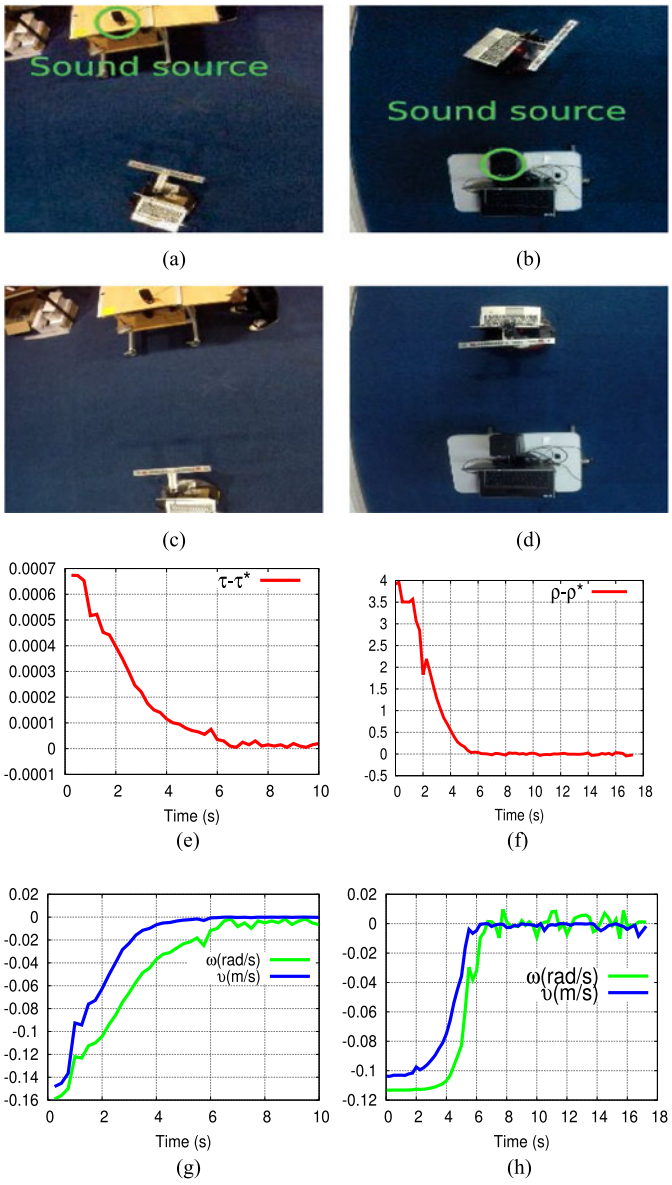


Fig. 4. Robot accurately orients towards the sound source by using ITD (left) or ILD (right) with an exponential decrease of the error. (e) ITD features error, (f) ILD features error, (g) Velocity input, (h) Velocity input.

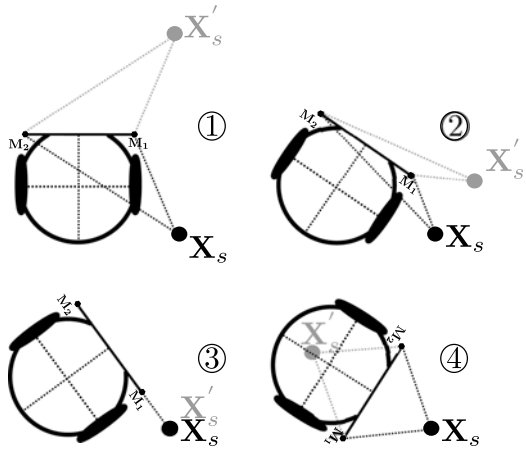


Fig. 5. Facing a sound source located behind: The robot moves with respect to the phantom sound source $\mathbf{X}_s^{'}$ (steps ① to ③) instead of $\mathbf{X}_s$ until reaching a configuration where $y_s > 0$.

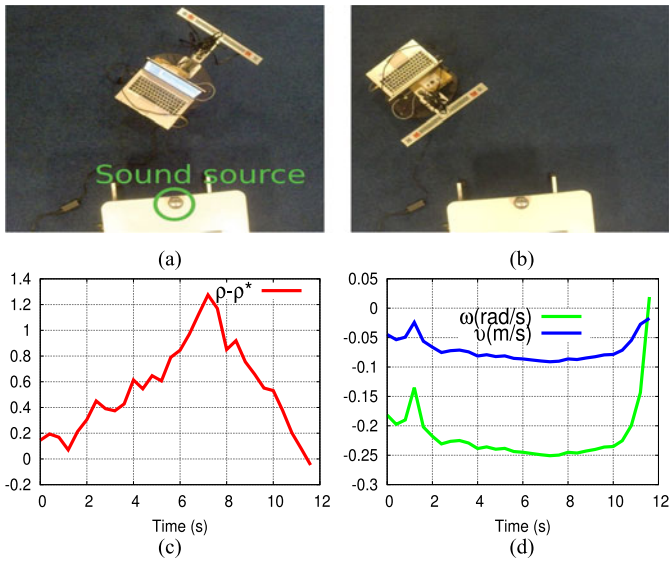


Fig. 6. Front-back ambiguity is inherently solved by the control scheme using ILD. (c) Features error, (d) Velocity input.

solutions to overcome this limit. Nonetheless, we will show in Section VII, that our approach is still adapted to this context.

3) Addressing the Front-Back Ambiguity: Although the previous experiments confirmed the validity of our approach, we assumed that the sound source was in the front of the robot, which can be considered as restrictive. The front-back ambiguity remains an issue for sound source localization that cannot be addressed from binaural cues only. The robot motion is generally used to dissipate this ambiguity. In our approach, the interaction matrices are parameterized with $\widehat{y}_s > 0$ (i.e., $\alpha \in]0;\pi[$). In case the sound source is behind (i.e., $y_s < 0$), the control scheme generates a motion driven by a phantom sound source symmetric to the actual one as illustrated in Fig. 5. As a consequence of this motion, the magnitude of the error $e = s - s^*$ increases. The error increases until $y_s > 0$, from which it will decrease until 0 since the system has entered in the Lyapunov convergence domain.

This analysis has been confirmed by experiments based on ILD, in which we observed the expected behavior of the system. The experiment illustrated by Fig. 6 shows a first phase where the error increases until $y_s > 0$. Subsequently, from this configuration, an exponential decrease of the error is observed until the robot faces the sound source. This result could also be obtained with the ITD. However, in order to cross the singularity $|\tau| = A$ (i.e., $y_s = 0$ discussed in Section IV), the velocity input of the robot should be limited by saturating the controller.

4) Addressing the Case of a Moving Sound Source: Eventually, we tackled the problem related to a moving sound source. Considering robot audition state-of-the-art, dealing with moving sources is challenging from sound source localization, since it requires to track and to model the source motions. By contrast, aural servo is a closed-loop control that induces flexibility and reactivity to any modification of the acoustic perception, thanks to the real-time feedback. This is verified by experiments in

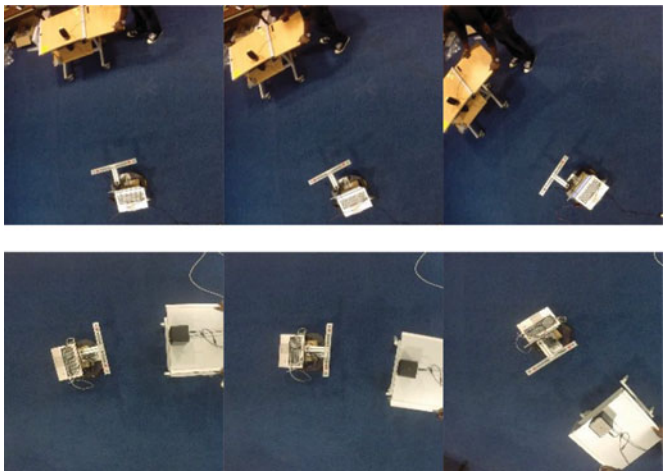


Fig. 7. Both frameworks, based on ITD (top) or ILD (bottom), can cope with a moving sound source.

which the sound source is moved laterally and arbitrary while being maintained in the *auditory fovea* of the microphones by the robot motion. Similar results are obtained for both tasks as illustrated in Fig. 7. Although the robot behavior remains satisfactory, we noticed during these experiments a small tracking error characterized by a small delay between the robot motion and the source motion (see the accompanying video). A better gain (λ) tuning and advanced control strategies integrating the potential source motion would certainly improve this result [17].

C. Evaluation

We showed that the aural servo approach is suitable for real-world applications and is able to address situations including front-back ambiguity or a moving sound source. In this part, we evaluate our system through simulations for studying task repeatability and providing ground truth references to assess the performance of our approach. All the following evaluations have been performed by considering a speech sound source.

1) Influence of the Distance to the Sound Source: The first evaluation concerns the achievement of the ITD and ILD tasks with respect to the sound source distance. More specifically, we evaluate the influence of near-field (NF) and FF zone. The simulation environment designed from *Roomsimove* [31] consists of a anechoic room. The positions of the sound source are defined so that α varied by 18° at given distances ℓ within the range $[18°, 162°]$. Since the task mainly consists in orienting the robot w.r.t the source, we controlled only the rotational velocity ω_z. This task is repeated from FF ($\ell = 5$ m) and NF distances ($\ell = 0.5$ m) for two different tasks corresponding to $\alpha^* = 0°$ (i.e, face the sound source) and $\alpha^* = 45°$. It should be noted that for the second task, τ^* and ρ^* are measured by positioning the robot in the desired configuration. We also considered for comparison a classical localization approach (**loc**) based on the estimation of the azimuth angle [derived from (8)] and a feedback control loop based on this localization. In the latter case, the control scheme (3) uses the interaction matrix $\mathbf{J}_\alpha$ that can be easily derived from (16) (see [21] for details), and the azimuth angle α as input.

The results are illustrated in Table II in which the error of the first task with $\alpha^* = 0°$ and the second task with $\alpha^* = 45°$ (bracketed values) are given. These simulations show that both control schemes using $\mathbf{J}_{\tau_r}$ and $\mathbf{J}_{\tau_f}$ give excellent results for both the NF and the FF. From the stability perspective, the actual interaction matrix related to the FF assumption $\mathbf{J}_{\tau_f}$ is just an approximation of $\mathbf{J}_{\tau_r}$, as already demonstrated in Section III-B, so that it can be written $\mathbf{J}_{\tau_f} = \widehat{\mathbf{J}_{\tau_r}}$.

TABLE II
ABSOLUTE ERROR OF THE FINAL AZIMUTH ANGLE FOR THE NF AND THE FF FOR TASKS WITH $\alpha^* = 0°$ AND $\alpha^* = 45°$ (BRACKETED)

		α (degrees)								
		18	36	54	72	90	108	126	144	162
NF	J_{τ_f}	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$
	J_{τ_r}	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$
	J_ρ	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$
	J_α	$\frac{*}{(6.5)}$	$\frac{*}{(6.2)}$	$\frac{*}{(6.4)}$	$\frac{*}{(5.9)}$	$\frac{*}{(6.9)}$	$\frac{*}{(7.5)}$	$\frac{*}{(7.3)}$	$\frac{*}{(7.3)}$	$\frac{*}{(7.0)}$
	loc	$\frac{7.1}{(7.1)}$	$\frac{8.2}{(8.2)}$	$\frac{6.3}{(6.3)}$	$\frac{3.9}{(3.9)}$	$\frac{*}{(*)}$	$\frac{3.9}{(3.9)}$	$\frac{6.8}{(6.8)}$	$\frac{8.2}{(8.2)}$	$\frac{7.1}{(7.1)}$
FF	J_{τ_f}	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$
	J_{τ_r}	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$
	J_ρ	$\frac{3.9}{(2.7)}$	$\frac{4.2}{(*)}$	$\frac{3.9}{(1.8)}$	$\frac{2.9}{(2.6)}$	$\frac{*}{(3.1)}$	$\frac{*}{(3.5)}$	$\frac{2.5}{(3.3)}$	$\frac{3.7}{(4.0)}$	$\frac{4.3}{(3.9)}$
	J_α	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$
	loc	$\frac{*}{(*)}$	$\frac{1.1}{(1.1)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{*}{(*)}$	$\frac{1.1}{(1.1)}$	$\frac{*}{(*)}$

Note that $*$ refers to a nonsignificant error ($e < 1°$)

Hence, without violating the stability conditions of $\mathbf{J}_{\tau_r}$, it is ensured that a control scheme based on $\mathbf{J}_{\tau_f}$ will converge under NF conditions. This outcome is also valid for $\widehat{\mathbf{J}_{\tau_f}}$ with the same stability constraints. Consequently, an ITD task can be performed accurately at any distance despite a FF assumption. Conversely to this method, the localization approach based on the FF assumption is degraded in the FF, while the closed-control loop using $\mathbf{J}_\alpha$ performs as good as $\mathbf{J}_{\tau_r}$ or $\mathbf{J}_{\tau_f}$ only for the task when $\alpha^* = 0$. This emphasizes the benefits of the closed-loop over classical localization that only estimates azimuth angle. A localization-based feedback loop is more robust to the FF approximation for the task where $\alpha^* = 0°$, since the robot orientation from which $\alpha = 0°$ can be measured is the same in the NF and FF. However, this is not the case for $\alpha \neq 0$, which explains the deteriorated results for the second task. Compared to aural servo, a localization-based feedback loop then exhibits several drawbacks. First, it assumes that the azimuth angle can be extracted, which can be time consuming and complex (e.g., head-mounted systems or ILD-based localization) although recent developments in sound localization [8], [32] aim to overcome this limitation with the late breakthrough of machine learning techniques. Furthermore, as it appears for the task $\alpha = \alpha^* = 45°$, any modeling approximation (e.g., FF or intermicrophone distance) immediately affects the positioning task when $\alpha^* \neq 0°$. These results clearly demonstrate the advantage to control robots at the scale of auditory cues, such as aural servo, instead of angular estimation through localization approaches.

At last, from the results based on $\mathbf{J}_\rho$ we can infer that the ILD task is accurately performed only in the FF. Increasing the distance between the microphones and the source degrades the accuracy. This is caused by the limitations inherent to ILD definition $\rho = \ell_2^2/\ell_1^2$. Since $\ell_2^2 = \ell_1^2 + 2dx_s$, when the robot is far from the sound source, ℓ (respectively each ℓ_i) becomes

large in comparison with the intermicrophone distance d. As a result, it comes out that $\ell_1^2 \to \ell_2^2$ and $\rho \to 1$. Thus, the energy difference between the two microphones becomes too small to be used. This result also confirms that ILD measurements are more relevant for a wide intermicrophones distance, which is unfortunately not compatible with the robotic context. By contrast, and as it will be exploited in Section VII head-mounted systems are less affected by this limitation thanks to the head shadowing effect.

2) Influence of Reverberation: The second set of evaluations concerns the robustness to reverberation by considering several reverberation times $\mathrm{RT}_{60} \in [0; 0.6]$ s at a fixed distance of 2 m. For the ITD case, with a high level of SNR (i.e., 30 and 25 dB) the task is performed accurately in all cases with an error below 5° for the final microphones orientation [see Fig. 8(a)]. A more exhaustive study emphasizes the ability of our approach to cope with erroneous/missing measurements, that are very likely to occur in a real scenario because of reverberation and speech pauses. A measurement is considered erroneous if the error between the actual ITD and the estimated ITD leads to an error of 5° or more in the corresponding DOA α. Fig. 8(b) illustrates the average rate of erroneous/missing measurements. As expected, this rate of erroneous ITDs increases with a higher level of noise and reverberation. But, the control scheme is still able to complete the task by using the prediction derived from (36) (i.e., the prediction is used when no consistent measurements are obtained), even for cases where around 20% of the measurements are missing or erroneous. This result shows the effectiveness of our method to cope with punctual inaccurate measurements. In parallel, a similar evaluation shows that the ILD task is much more affected by reverberation. In particular, the combined effect of distance and reverberation drastically decreases the task accuracy, as detailed in Fig. 8(c). In practice, early reflections that can be modeled as virtual sound sources adding up to the actual sound source signal recorded by each microphone, directly impacts the ratio ρ and thus the error e_ρ.

3) Influence of Noise: In the last set of evaluations, the noise effect is studied by adding a diffuse noise in the recorded signal. The noise signal follows a normal distribution with a mean value of 0 and a variance depending on the desired SNR. Unsurprisingly, the results are deteriorated as noise increases. This effect is much more pronounced for the ITD task [see Fig. 8(a)] where most of the failures are related to a tracking issue caused by a wrong initial ITD measurement (e.g., $\alpha(t=0) = \pi/2$). This result fits with the method chosen for the ITD calculation (GCC-PHAT), known to be less robust to noise. Approaches reducing the noise level would undeniably improve these results. As for ILD tasks, the influence of noise is limited as depicted in Fig. 8(c), since the final error remains below 5° when there is no reverberation. Apart from extreme level of noise (SNR ≤ 0 dB), any diffuse noise adds up similarly on each microphone measurement, unlike reverberation, which limits the impact on ρ and the task error e_ρ.

D. Discussion

To sum up this set of evaluations, we can notice that aural servo methods are robust to modeling approximations, which is exemplified by the ITD task based on the FF assumption. Likewise, rough features estimation are well supported by these methods. Such results explain the robustness of our approach in the real-world scenarios presented in Section V-B. Additionally, we can notice that the ITD-based method and the ILD-based method complement each other on several aspects. On one hand, ITD task is robust to reverberation and varying distances (NF/FF) to the sound source. However, it requires a robust tracking and is particularly affected by noise. On the other hand, ILD task does not require any tracking, but is suitable to single-sourced environments. Reverberations, that can be modeled by several virtual sources, particularly degrade the task accuracy, unlike noise. Furthermore, ILD task is efficient only in the NF area.

VI. Advanced Tasks

In this section, we study tasks composed of a set of auditory cues. This allows introducing more constraints on the robot desired pose, and is done either with a single (see Section VI-A) or with multiple sound sources (see Section VI-B).

A. Approaching a Sound Source

1) Global Approach: In this part, we consider the task of approaching a sound source. Both the orientation of the robot and the distance to the sound source are controlled. To perform this task, we consider coupling ILD to the level of energy E_M through (34) leading to $\mathbf{e} = (\rho - \rho^*, E_\mathrm{M} - E_\mathrm{M}^*)$. Although ILD is accurate only in the NF, one can overcome this limitation by using the level of energy to regulate the distance. In parallel, ILD provides an azimuth angle information through the sign of $\widehat{x}_s$ that allows approximating the interaction matrix $\widehat{\mathbf{J}_{E_\mathrm{M}}}$. The interaction matrix $\widehat{\mathbf{J}_{\rho E}}$ combining the ILD ρ to the energy level E_M is then obtained by stacking $\widehat{\mathbf{J}_\rho}$ and $\widehat{\mathbf{J}_{E_\mathrm{M}}}$ as follows:

$$\widehat{\mathbf{J}_{\rho E}} = \begin{bmatrix} \frac{2\widehat{x}_s(\rho-1)-d(\rho+1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} & \frac{2\widehat{y}_s(\rho-1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} & \frac{\widehat{y}_s d(\rho+1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} \\ \frac{2E_\mathrm{M}\widehat{x}_s}{\widehat{\ell}^2} & \frac{2E_\mathrm{M}\widehat{y}_s}{\widehat{\ell}^2} & 0 \end{bmatrix}. \quad (37)$$

Similarly to the basic tasks described in Section IV, to avoid any singular configuration, the approximated parameters of $\widehat{\mathbf{J}_{\rho E}}$ should be set so that $\widehat{\ell} \neq 0$ and $\widehat{\ell}_i \neq 0$. Then, by analyzing the set of poses for which $\mathbf{e} = 0$ through the null space of $\mathbf{J}_{\rho E}$, it can be demonstrated that the task consists in reaching a circle of radius ℓ, centered on the sound source, with a given orientation (see [20]).

2) Experimental Results: The task is designed with $\rho^* = 1$ and E_M^* measured 50 cm in front of the source. The sound source corresponds to a white Gaussian noise. Moreover, since the energy level in $\mathbf{M}$ is not directly available, we approximate E_M as the mean value of the energy received by each microphone with $E_\mathrm{M} \approx (E_1 + E_2)/2$. As illustrated in Fig. 9, the task can be completed from initial poses relatively far from the sound source (>3 m). Actually, as long as the difference of energy between the microphones is perceptible at the initial pose (i.e., $\rho(t) \neq 1$ for $\alpha \neq \pi/2$), the control scheme is able to position the robot in a desired configuration.

Such a task can be referred to as a coarse-to-fine approach. When the robot is far from the sound source, the orientation control from ILD is rough because of the lack of resolution of this cue. Yet, as the robot moves closer to the sound source, the ILD measurement is refined, and the robot motion becomes more accurate. From this result and knowing that aural servo can cope with a moving sound source, we addressed the task of following

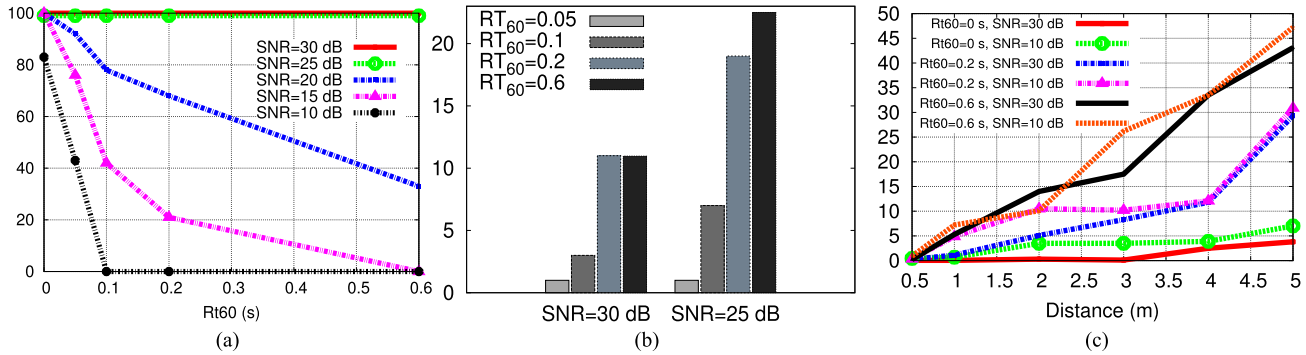


Fig. 8. ITD-based task and ILD-based task evaluation. (a) Rate of ITD-based tasks successfully achieved. (b) Rate of missing/erroneous measurements among achieved ITD-based tasks. (c) Mean error (degree) of ILD-based tasks.

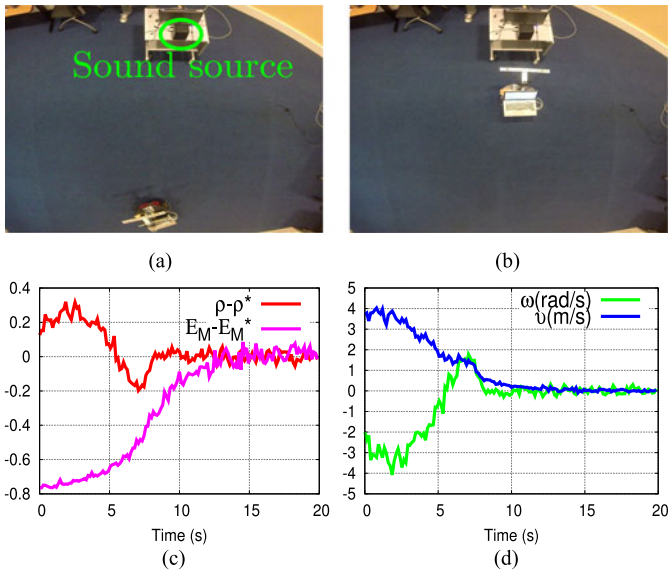


Fig. 9. Positioning task based on ILD and energy level measurement. (c) Features error, (d) Velocity input.

Fig. 10. Different applications following a sound source: (left) Indoor navigation, (right) Cooperative task with a UAV.

a sound source for two different applications detailed in [20] and illustrated in Fig. 10 and in the accompanying video. First, we developed a navigation system, showing that our approach is robust and flexible to changing environments. In this experiment, despite changing level of noise and reverberation, a robot was able to navigate continuously in various indoor environments by following a sound source. A cooperative task has also been developed with a unmanned aerial vehicle (UAV) guiding a mobile robot through the noise produced by its propellers.

TABLE III
FINAL ABSOLUTE MEAN ERROR, IN DEGREE, AND THE FINAL RANGE ERROR (BRACKETED VALUES), IN CM, ARE CALCULATED FOR SEVERAL REVERBERATION TIMES (RT_{60}) AND DISTANCES TO THE SOURCE

RT_{60} (s)	ℓ (m)							
	0.5		1		2		3	
0	<1	(1.47)	<1	(1.47)	<1	(1.48)	<1	(1.47)
0.05	<1	(1.77)	<1	(1.77)	<1	(1.78)	<1	(1.8)
0.1	<1	(1.94)	<1	(2.05)	<1	(1.95)	<1	(1.96)
0.2	<1	(3.90)	<1	(3.90)	<1	(3.96)	<1	(3.98)

3) Evaluation: The control scheme has been evaluated in simulation too. The simulated environment and conditions are strictly identical to the evaluation performed in Section V-C. As previously, the task consists in facing the sound source at a given distance. Hence, ρ^* was set to 1, while E^*_{M} was measured in anechoic conditions when $\rho = 1$ and $\ell = 0.5$ m. This task is performed from several starting distances ℓ and reverberation times.

The results are summarized in Table III, where the mean absolute error of orientation and the mean absolute error of the range positioning are reported. These results are impressively accurate as opposed to the ILD-only control framework. From all starting poses, even in the FF and with reverberation, the robot is always able to face the sound source accurately as the errors reported are all below 1°. It can also be noted that the energy level is little affected by reverberation, since the range error varies only from 1 to 4 cm. Furthermore, interestingly, the energy level is consistent over changing conditions. Despite an initial measurement E^*_{M} performed under anechoic conditions, the range error remains lower than 5 cm for a reverberation level $\mathrm{RT}_{60} = 0.2$ s. Of course, more accurate results could be obtained if E^*_{M} was measured in the same acoustic conditions as the real environment. These excellent properties of robustness and flexibility to echoic and changing environments substantially explain the satisfactory results obtained with real experiments.

B. ITD-Based Task With Two Sound Sources

1) Global Approach: A robot can also be controlled with cues extracted from several sound sources. This approach is exemplified in this section by considering two sound sources. We use ITD cues in the error vector $\mathbf{e}$ of the control scheme (34):

$\mathbf{e} = (\tau_1 - \tau_1^*, \tau_2 - \tau_2^*)$. Indeed as stated in Section V, ILD cues are not suitable to deal with several sources, since each source contribution is smeared through the signal integration process. Conversely, each source ITD can be detected through the peaks of the cross-correlation function. Hence, by considering two sound sources $\mathbf{X}_{\mathbf{s}_1}$ and $\mathbf{X}_{\mathbf{s}_2}$ in the scene, the interaction matrix $\widehat{\mathbf{J}_{\boldsymbol{\tau}}}$ related to τ_1 and τ_2 is given by

$$\widehat{\mathbf{J}_{\boldsymbol{\tau}}} = \begin{bmatrix} -\frac{\nu_1^2}{A\widehat{\ell}_1} & \frac{\tau_1\nu_1}{A\widehat{\ell}_1} & \nu_1 \\ -\frac{\nu_2^2}{A\widehat{\ell}_2} & \frac{\tau_2\nu_2}{A\widehat{\ell}_2} & \nu_2 \end{bmatrix} \tag{38}$$

where $\nu_i = \sqrt{A^2 - \tau_i^2}$. Once again, any singular configuration is avoided when setting in $\widehat{\mathbf{J}_{\boldsymbol{\tau}}}$, $\ell_i \neq 0$. From the null space of $\mathbf{J}_{\boldsymbol{\tau}}$ and the analysis of the set of poses for which $\mathbf{e} = 0$, it can be demonstrated that the task related to this control scheme consists in reaching a pose on the circumscribed circle characterized by $\mathbf{X}_{\mathbf{s}_1}$, $\mathbf{X}_{\mathbf{s}_2}$, and the position of $\mathbf{M}$ when $\mathbf{e} = 0$.

This approach can also be extended to more sound sources. By considering three or more sound sources, all the three DOF of the robot are constrained and the task consists in reaching a unique pose defined by each τ_i^*. This configuration is more thoroughly studied in [21]. Such an approach could be particularly interesting for multirobot control tasks, where each robot would emit a distinct sound signal as in [33].

2) Experimental Results: For this application, we used the settings given in Table I. The approximated distances in (38) were $\widehat{\ell}_1 = \widehat{\ell}_2 = 1$ m. In this experiment, besides the female speech we added a second sound source corresponding to a burst of white Gaussian noise of 25 ms followed by 25 ms of silence played in loop. This time, the objective was to reach a pose where $\tau_1^* = -\tau_2^*$ with $\alpha_1^* \equiv 50^\circ$. In addition to a tracking algorithm, the control scheme stresses the need of a correct labeling of each ITD. The goal is to associate to each τ_i the desired τ_i^* so that the task can be correctly completed. Fortunately, this labeling problem is trivial to solve in our configuration. Indeed, if we consider the working space as the half plane in front of the microphones, the ordinality of $\tau_i(t)$ and τ_i^* is the same. Namely if $\tau_1^* < \tau_2^*$ then $\tau_1(t)$ should be smaller than $\tau_2(t)$.

Fig. 11 illustrates the experimental results: Starting from a pose around 3 m away from the sources, the error of the measured ITDs successfully converges to 0 while the robot follows a straight and smooth trajectory. When evaluating in simulation the control scheme, we obtain similar results as in the case of a single source [see Fig. 8(a) and (b)]. However the results are less robust than with a single source in real experiments. The robustness decreases notably because of the issue related to track each source ITD when considering noncontinuous signals. Although the prediction scheme given in (36) improves the tracking, with intermittent sound sources, some plausible ITDs issued from the echoes of the active sound sources may be associated to the ITD of the inactive sound sources. For this kind of configuration, it is then necessary to take into consideration the number of active sound sources.

VII. Application on Humanoid Robots

So far, we considered free-field sound propagation. In this final section, we extend our approach by addressing the context of humanoid robots. Binaural localization on humanoid robots is the closest configuration to biological auditory systems, but at the same time probably the most challenging configuration for robot audition. The combination of acoustic perturbations (reverberation, noise) and head scattering effects, characterized in the head related transfer functions (HRTFs), creates challenging conditions to be resolved by localization methods.

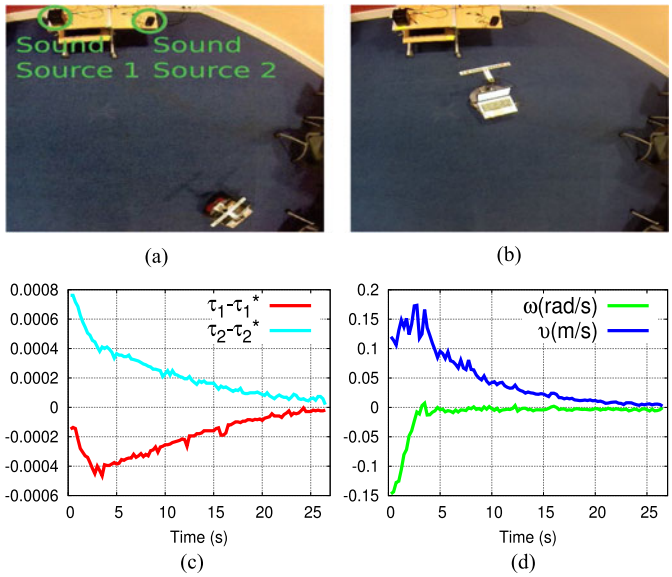


Fig. 11. With two sound sources, the robot reaches a pose satisfying the given bearing conditions. (c) Features error, (d) Velocity input.

A. Versatility of Aural Servo Framework

Without measuring or modeling the HRTFs related to a given humanoid robot, it is clear that our initial acoustic models are not valid anymore. Yet, it is still possible to apply the aural servo framework in this context. As depicted throughout this paper, the core principle of aural servo is based on the auditory feature variation. This characteristic has been stressed in Section V where we showed that the same control scheme could be applied in the NF or in the FF despite an imperfect acoustic model. These results let us hypothesize that the variation of auditory cues could also be robust to the individual variability of sound cues perception (i.e., HRTFs). In order to assess this hypothesis, we conducted a set of experiments based on ILD features. The experiments consisted in observing the variation of ILD with respect to the source motion from two different humanoid robots, *Romeo* and *Pepper* designed by Softbank Robotics. For both robots, we considered only two microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ separated by a distance d, as depicted in Fig. 12. The inter-microphones distances are, respectively, $d_{\text{Pepper}} \approx 0.07$ m and $d_{\text{Romeo}} \approx 0.12$ m. During the experiments, a white Gaussian noise was continuously emitted from a loudspeaker. This loudspeaker was moved in a circular motion around the robots head (i.e., at a constant distance) so that the DOA α varies from 0 to π. The ILD ρ computed from $\mathbf{M_1}$ and $\mathbf{M_2}$ was then measured all along the loudspeaker motion. These experiments were conducted in a highly reverberant environment where $\text{RT}_{60} > 1$ s. In addition, the same task was conducted in simulation in an anechoic free-field environment where $\mathbf{M_1}$ and $\mathbf{M_2}$ are separated by a distance $d = 0.12$ m and $d = 0.07$ m. The results are given in Fig. 12 where the absolute value of $\rho_{\text{dB}} = 10\log_{10}(\rho)$ is plotted in order to facilitate the analysis. First, it can be noticed

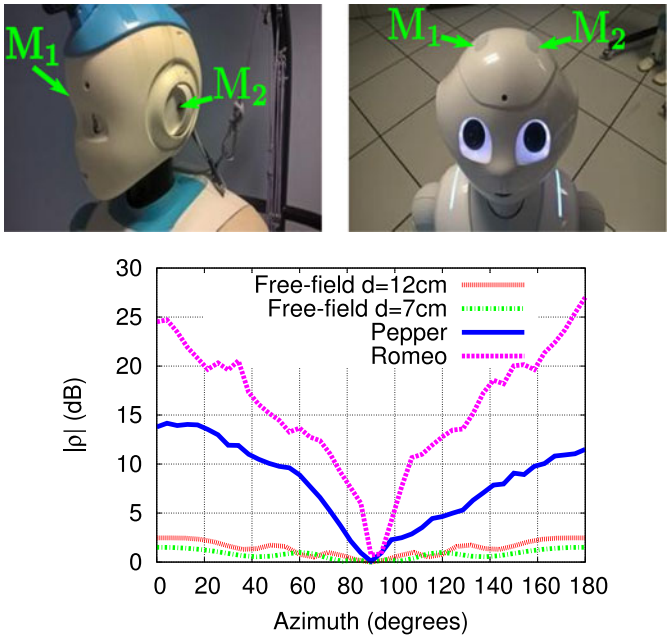


Fig. 12. ILD variation measurement on *Romeo* (top-left) and *Pepper* (top-right).

that all configurations share the same "**V**" shape that reflects the symmetry of ILDs with respect to the microphones perpendicular bisector. An ILD is maximal for the most eccentric position of a source, while its minimum value ($\approx$0 dB) is reached when this source is in the auditory fovea. Despite different robot structures and acoustic conditions, the variation property of ILD cues is preserved. By contrast, the intrinsic values of ρ_{dB} are drastically changed depending on the auditory setup considered. On one hand, any azimuth estimation method would need to model the influence of HRTFs, while on the other hand, with our approach, tasks consisting in facing a sound source can be performed from a free-field acoustic model. As a consequence, the complexity and the computation cost of our method is drastically decreased compared to classical localization methods. The same results are obtained when considering ITD cues: The minimum absolute ITD value is obtained in the auditory fovea, while eccentric positions lead to higher values. The following experimental results support these conjectures.

B. Experimental Results

The first experiment was carried out on *Romeo* by using the two microphones $\mathbf{M_1}$ and $\mathbf{M_2}$. The room acoustics corresponds to the conditions described earlier ($\text{RT}_{60} > 1$ s). We conducted separately two tasks that consisted in facing a sound source by using ITD and ILD cues. The sound emitted from a loudspeaker corresponded to a white Gaussian noise for the ILD-based task, while a speech signal was used for the ITD-based task. In ILD case, we used a white Gaussian noise in order to avoid the speech activity detection (which is out of the scope of this paper), since the residual noise of the microphones caused by the robot ventilation system could lead to erroneous measurements when the sound source is not active. It should also be mentioned that two external microphones, fixed on each pinna, were used for the ITD-based task due to the high level of internal noise of the robot. The results depicted in Fig. 13 are, respectively, based on

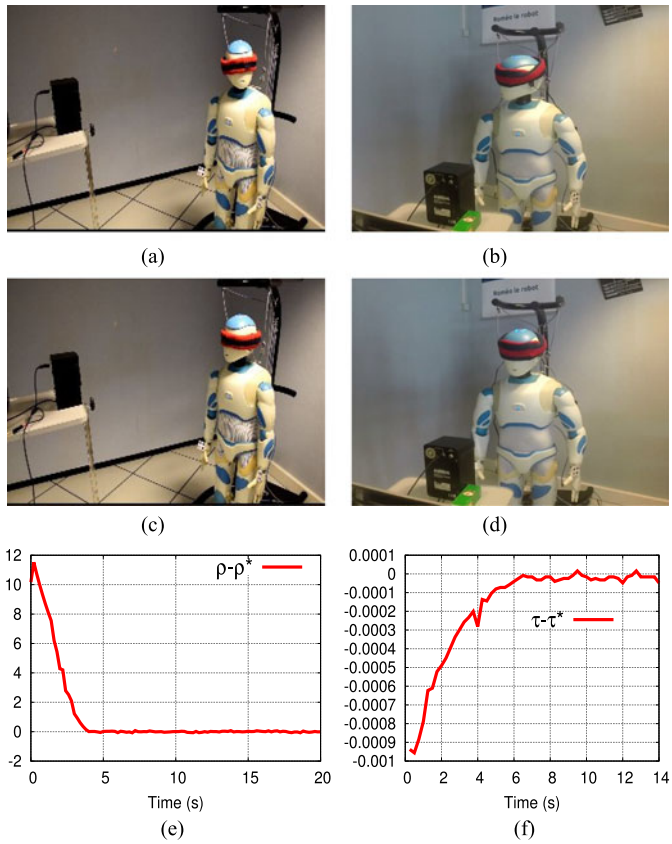


Fig. 13. Gaze control of *Romeo* from ILD measurements (left) and ITD measurements (right). (e) ILD features error, (f) ITD features error.

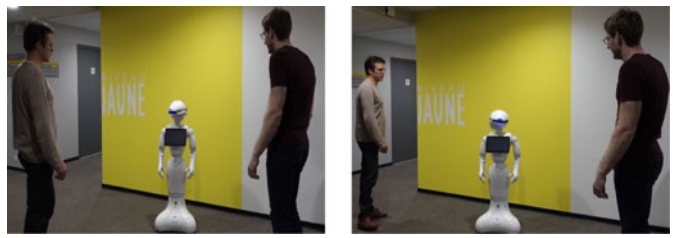


Fig. 14. Gaze control of *Pepper* from ITD with real users.

the control schemes defined in (25) and (28), that become

$$\omega_{\text{itd}} = -\lambda \frac{1}{\nu}(\tau - \tau^*) \tag{39}$$

for the ITD case and

$$\omega_{\text{ild}} = -\lambda \frac{\ell^2 + \frac{d^2}{4} - d\widehat{x}_s}{\widehat{y}_s d(\rho + 1)}(\rho - \rho^*) \tag{40}$$

for the ILD case, where the control input ω corresponds to the angular velocity that sets the orientation of the robot head. As predicted, the gazing task was correctly achieved in both cases despite a free-field propagation model. The error curve, for both tasks, follows an exponential decrease while the robot accurately faces the sound source once the error vanished. The case of a moving sound source has also been addressed similarly to the free-field case. These results are given in the accompanying video.

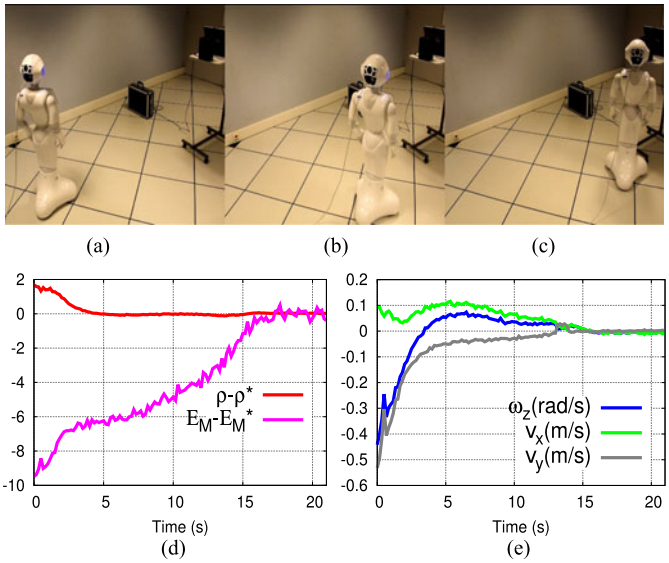


Fig. 15. *Pepper* approaches the sound source and reaches a pose satisfying $\rho = \rho^* = 1$ and $E_{\mathrm{M}} = E^*_{\mathrm{M}}$. (d) Features error, (e) Velocity input.

The gaze control (based on ITD) was also successfully tested with real users on *Pepper*. In the accompanying video (also in Fig. 14), our gaze control system allowed to focus in real-time on the speaker even when the latter was moving without any tracking. This experiment illustrates that our approach can be adapted to HRI context (despite the limitation pointed in Section V-B2), and does not particularly require a long signal length thanks to the limited computational cost. Several extensions of this experiment can be envisioned, such as application for intelligent cameras for conferencing systems, extension to 3-D scenes by adding more microphones in order to control the elevation, or coupling aural servo with vision to get an even more robust solution.

Finally, we applied on *Pepper* the control framework based on ILD and the energy level as input features. In this case, we are not attached to control the robot head, but rather its holonomic base. The interaction matrix $\widehat{\mathbf{J}_{\rho E}}$ given in (37) was directly used to control the robot with

$$\mathbf{u} = -\lambda \widehat{\mathbf{J}_{\rho E}}^{+} \mathbf{e}, \tag{41}$$

where $\mathbf{u} = (v_x, v_y, \omega_z)$. The task consisted in approaching a loudspeaker playing a white Gaussian noise. For this purpose, the desired ILD was set to $\rho^* = 1$. E^*_{M} was measured experimentally when the robot was located at $\ell \approx 0.5$ m from the loudspeaker. The results are given in Fig. 15. Similarly to the experiments performed in Section VI-A, *Pepper* was able to reach a pose satisfying the given task and even follow the loudspeaker when it started to move (see the accompanying video). These experiments particularly emphasize the generality of our approach that does not depend on the type of robot. The same type of control scheme is used on Pepper, Romeo, or on Pioneer.

VIII. Conclusion

In this paper, we introduced the concept of aural servo. The contributions are two-fold. First, we provided a complete modeling and theoretical analysis of our approach for several auditory features. Second, we developed experimental results in real environments, on various robots (mobile and humanoid robots) and situations. In details, we studied the ILD, the ITD, and the absolute sound energy level. The modeling of these cues let us control and position a robot with respect to a sound source without localizing it. We demonstrated theoretically and experimentally that ITD and ILD cues allow us to control the orientation of a robot, while the sound energy level regulates the distance to a sound source. More advanced motion controls have also been developed through a homing task from several sound sources ITDs and a navigation task from ILD and the sound energy. Globally, the relevance of aural servo has been assessed by the higher robustness of this approach to modeling approximations (e.g., FF assumption, free-field propagation, etc.) compared to classical localization methods as well as the robustness to adverse acoustic conditions, such as high and fluctuating reverberation times. Similarly our approach is general since it does not depend on the robot type, as confirmed by our different experiments.

Such results open several perspectives for robot audition applications in the fields of navigation, conferencing systems or HRI. For instance, it could be interesting to combine several auditory features (ILD, ITD, direct-to reverberant ratio, spectral notches) and to fuse auditory cues with other sensing modalities (e.g., vision) in a more human-like system. In such a configuration, cues like spectral notches naturally extend our approach to 3-D configurations by providing elevation information.

References

[1] J. Even, N. Kallakuri, Y. Morales, C. Ishi, and N. Hagita, "Creation of radiated sound intensity maps using multi-modal measurements onboard an autonomous mobile platform," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 3433–3438.

[2] M. Basiri, F. Schill, P. U. Lima, and D. Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 4737–4742.

[3] L. Natale, G. Metta, and G. Sandini, "Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head," *Robot. Auton. Syst.*, vol. 39, no. 2, pp. 87–106, 2002.

[4] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie, "A model-based sound localization system and its application to robot navigation," *Robot. Auton. Syst.*, vol. 27, no. 4, pp. 199–209, 1999.

[5] I. Marković and I. Petrović, "Speaker localization and tracking with a microphone array on a mobile robot using von mises distribution and particle filtering," *Robot. Auton. Syst.*, vol. 58, no. 11, pp. 1185–1196, 2010.

[6] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2003, vol. 2, pp. 1147–1152.

[7] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.

[8] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 2927–2932.

[9] A. Deleforge, R. Horaud, Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 718–731, Apr. 2015.

[10] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. Assoc. Adv. Artif. Intell.*, 2000, pp. 832–839.

[11] A. Portello, P. Danès, and S. Argentieri, "Active binaural localization of intermittent moving sources in the presence of false measurements," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 3294–3299.

[12] I. Kossyk, M. Neumann, and Z.-C. Marton, "Binaural bearing only tracking of stationary sound sources in reverberant environment," in *Proc. IEEE-RAS Int. Conf. Hum. Robots*, 2015, pp. 53–60.

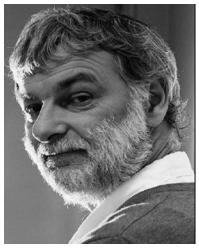
[13] G. Bustamante, P. Danès, T. Forgue, and A. Podlubne, "Towards information-based feedback control for binaural active localization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 6325–6329.
[14] A. Alford, S. Northrup, K. Kawamura, K. Chan, and J. Barile, "A music playing robot," in *Proc. Conf. Field Service Robots*, 1999, pp. 29–31.
[15] M. Kumon, T. Sugawara, K. Miike, I. Mizumoto, and Z. Iwai, "Adaptive audio servo for multirate robot syst." in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2003, vol. 1, pp. 182–187.
[16] M. Kumon, T. Shimoda, R. Kohzawa, I. Mizumoto, and Z. Iwai, "Audio servo for robotic system with pinnae," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2005, pp. 1881–1886.
[17] F. Chaumette and S. Hutchinson, "Visual servoing and visual tracking," in *Springer Handbook of Robotics*. New York, NY, USA: Springer, 2008, pp. 563–583.
[18] B. Espiau, J.-P. Merlet, and C. Samson, "Force-feedback control and non-contact sensing: A unified approach," in *Proc. 8th CISM-IFTOMM Symp. Theory Practice Robots Manipulators*, 1990.
[19] P. Sikka, H. Zhang, and S. Sutphen, "Tactile servo: Control of touch-driven robot motion," in *Experimental Robotics III*. New York, NY, USA: Springer, 1994, pp. 219–233.
[20] A. Magassouba, N. Bertin, and F. Chaumette, "Audio-based robot control from interchannel level difference and absolute sound energy," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 1992–1999.
[21] A. Magassouba, N. Bertin, and F. Chaumette, "Sound-based control with two microphones," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 5568–5573.
[22] A. Magassouba, "Aural servo: Towards an alternative approach to sound localization for robot motion control," Ph.D. dissertation, Université Rennes 1, Rennes, France, 2016, [Online]. Available: https://hal.inria.fr/tel-01426710v3
[23] C. Samson, B. Espiau, and M. Le Borgne, *Robot Control: The Task Function Approach*. London, U.K.: Oxford Univ. Press, 1991.
[24] B. Espiau, F. Chaumette, and P. Rives, "A new approach to visual servoing in robotics," *IEEE Trans. Robot. Autom.*, vol. 8, no. 3, pp. 313–326, Jun. 1992.
[25] J.-J. E. Slotine, *et al.*, *Applied Nonlinear Control*. vol. 99, Englewood Cliffs, NJ, USA: Prentice-Hall, 1991.
[26] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2003, vol. 2, pp. 1228–1233.
[27] K. Nakadai, H. G. Okuno, and H. Kitano, "Epipolar geometry based sound localization and extraction for humanoid audition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2001, vol. 3, pp. 1395–1401.
[28] S. Birchfield and R. Gangishetty, "Acoustic localization by interaural level difference," in *Proc. IEEE Int. Conf. Acoust., Speech Sig. Process.*, 2005, vol. 4, pp. iv-1109–iv-1112.
[29] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acous., Speech Sig. Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.
[30] Abran-Côté *et al.*, Eight sound USB. 2012. [Online]. Available: http://eightsoundsusb.sourceforge.net
[31] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.
[32] T. May, N. Ma, and G. J. Brown, "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *Proc. IEEE Int. Conf. Acous., Speech Sig. Process.*, 2015, pp. 2679–2683.
[33] M. Basiri, F. Schill, D. Floreano, and P. U. Lima, "Audio-based localization for swarms of micro air vehicles," in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2014, pp. 4729–4734.

Aly Magassouba received the Ph.D. degree in signal processing from University of Rennes, France, in 2016.

He conducted his dissertation in Irisa and Inria Rennes laboratories within Lagadic group, where he was a Postdoctoral Researcher. Since 2017, he has been a Postdoctoral Researcher with NICT, Kyoto, Japan. His research interest include visual servoing, acoustics, signal processing and spoken language understanding.

Nancy Bertin (M'06) received the Ph.D. degree in signal processing from Télécom ParisTech, Paris, France, in 2009.

Since 2010, she has been a CNRS permanent Researcher with the PANAMA group (Univ Rennes, Inria, CNRS, IRISA), where her research interests include audio scene analysis, source separation and compressed sensing of acoustic fields, with a particular methodological emphasis on sparse and cosparse representations.



François Chaumette (M'02–SM'09–F'13) received the Ph.D degree in computer science from University of Rennes, France, in 1990. Since 1990, he has been with Inria, Irisa, Rennes, France. His research interests include robotics and computer vision, especially visual servoing and active perception.

Dr. Chaumette is currently on the Editorial Board of *International Journal of Robotics Research*, Founding Senior Editor of IEEE ROBOTICS AND AUTOMATION LETTERS, and Senior Editor of IEEE TRANSACTION ON ROBOTICS.