Exploiting the distance information of the interaural level difference for binaural robot motion control

Aly Magassouba[1], Nancy Bertin[2], and François Chaumette[3]

***Abstract*—This paper proposes a control framework allowing to position a robot in range and orientation with respect to a sound source from a binaural system. To this end, a sensor-based control framework is developed upon interaural level difference (ILD) cues. ILD is known to be implicitly related to the sound source azimuth but also to the sound source distance. We emphasize the latter property by introducing the concept of ILD annulus. Then a sensor-based task is designed in order to approach a sound source. This method is validated in simulation and in real world experiments performed on a humanoid robot.**

***Index Terms*—Robot Audition, Sensor-based Control**

I. Introduction

In robot audition, motion control from a binaural setup is generally based on the estimation of the azimuth and/or elevation angles through a localization paradigm. Distance estimation is seldom performed because of the complexity of this process compared to azimuth and elevation estimation. Range positioning is thus generally exploited by systems endowed with an array (*i.e.*, > 2) of microphones, whether by using the interaural time difference (ITD) [1], [2], the interaural level difference (ILD) [3] or both cues [4]. However the latter approach remains out of the scope of this paper that focuses on binaural setups.

Alternatively, in the field of active audition, distance is recovered by fusing acoustic measurements from different positions of the robot, as mostly performed in the state-of-the-art of binaural robot audition [5], [6], [7], [8], [9]. In a different way, we develop in this paper a method that is rather inspired by psycho-acoustics literature related to distance estimation by human listeners. Several studies [10], [11], [12] performed on human subjects demonstrate that binaural cues do not only convey orientation information but also distance. In particular, a correlation between ILD variation and distance has been exhibited: ILD increases rapidly when approaching a sound source in the near-field. In robot audition, such results are exploited in [13], where distance is estimated from a pre-learned sequence of features/distance bins. More specifically, relevant results based on ILD are shown for sound sources located within 2 meters of the microphones. In a sensorimotor approach, [14] developed a supervised learning where an agent learns the relationship between ILD and the state of its motors and then accordingly approaches a sound source. Hence, these studies suggest that the azimuth angle and the distance information can be extracted solely from binaural cues.

These properties are demonstrated and exploited in this paper. Unlike the previous works that are based on empirical results, we characterize analytically the relationship between ILD and distance. Subsequently, we introduce a control framework allowing to position a robot with respect to sound source range. Such a framework builds on our previous work [15] where we developed a binaural control system using ILD, that does not extract any azimuth angle. Based on sensor-based control, our approach allowing to control a robot with respect to (w.r.t.) sound has been experimented on free-field microphones as well as on humanoid robots [16], without Head Related Transfer Function (HRTF) computation. In the latter works, our system was able to control the orientation of a robot by using ILD, and the distance to a sound source from the sound absolute level of energy, under the assumption of a continuous signal. In this paper, we overcome this potential limitation by relying only on ILD to control both orientation and distance. Hence the originality and the novelty of this paper arise from the use of a unique feature for controlling both the orientation and the range of a robot w.r.t. a sound source, in a HRTF-independent framework. Additionally, from ILD that is known to be less time-consuming and complex to process than ITD, we demonstrate that a complex task can be achieved from low- or high-frequency signals and with different acoustic setups.

The remainder of this paper is structured as follows: we first introduce the ILD modeling in Section II. From this modeling, ILD dependence upon distance is demonstrated analytically in Section III. Section IV develops a control scheme allowing to set both range and orientation w.r.t a sound source. Finally, in Section V we validate this approach in simulation and experimentally on a humanoid robot.

II. ILD modeling

A. Geometric configuration

Let us consider the case of a pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$, as illustrated in Fig. 1, embedded on a mobile robot moving in an area free of obstacle. These microphones are separated by a distance d. In this scene, an omni-directional point-wise sound source $\mathbf{X_s}$ that is continuously emitting a

Manuscript received: September, 9, 2017; Revised January, 10, 2018; Accepted January, 30, 2018.

This paper was recommended for publication by Editor Jana Kosecka upon evaluation of the Associate Editor and Reviewers' comments.

[1] Univ Rennes, Inria, CNRS, IRISA, Campus de Beaulieu, 35042 Rennes, France. aly.magassouba@irisa.fr

[2] CNRS, Univ Rennes, Inria, IRISA Campus de Beaulieu, 35042 Rennes, France. nancy.bertin@irisa.fr

[3] Inria, Univ Rennes, CNRS, IRISA Campus de Beaulieu, 35042 Rennes, France. francois.chaumette@inria.fr

Digital Object Identifier (DOI): see top of this page.

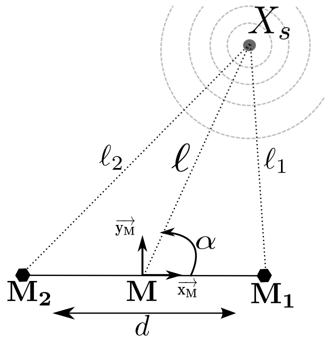


Figure 1: Geometric configuration of the considered problem, that includes a source $\mathbf{X_s}$ emitting a spherical sound wave, and a pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$.

sound wave $a(t)$. From this configuration, we define a frame $\mathcal{F}_m(\mathbf{M}, \overrightarrow{x_M}, \overrightarrow{y_M}, \overrightarrow{z_M})$ attached to the pair of microphones in its midpoint $\mathbf{M}$. The coordinates of each microphone are respectively $\mathbf{M_1}(\frac{d}{2}, 0, 0)$ and $\mathbf{M_2}(-\frac{d}{2}, 0, 0)$. Assuming a planar scene defined by $(\overrightarrow{x_M}, \overrightarrow{y_M})$, $\mathbf{X_s}(x_s, y_s, 0)$ is then located at ℓ_i (resp. ℓ) from each microphone $\mathbf{M_i}$ (resp. $\mathbf{M}$) given by

$$\begin{cases} \ell_1 = \sqrt{(x_s - d/2)^2 + y_s^2} \\ \ell_2 = \sqrt{(x_s + d/2)^2 + y_s^2} \\ \ell = \sqrt{x_s^2 + y_s^2} \end{cases} \quad (1)$$

In this configuration, the microphones are thus endowed with 3 degree-of-freedom (DOF) in the plane defined by $(\overrightarrow{x_M}, \overrightarrow{y_M})$: translation motions with respect to $\overrightarrow{x_M}$, $\overrightarrow{y_M}$ and a rotation motion with respect to $\overrightarrow{z_M}$.

B. Geometrical properties of the ILD

In a similar process as in [3], we can shape the different geometrical properties stemmed from ILD. Under the spherical sound propagation assumption the signal recorded at each microphone is

$$x_i(t) = \frac{a(t - \frac{\ell_i}{c})}{\ell_i}, \quad (2)$$

where $\frac{\ell_i}{c}$ expresses the sound propagation delay. By integrating (2) over a frame of length w, the energy received in each microphone is defined as follows:

$$E_i = \int_{t=0}^{w} |x_i(t)|^2 \, \mathrm{d}t = \frac{1}{\ell_i^2} \int_{t=0}^{w} a^2 \left(t - \frac{\ell_i}{c}\right) \mathrm{d}t \quad (3)$$

Equation (3) characterizes the inverse-square law property inherent to a spherical and isotropic sound propagation. The ILD ρ between the microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ is then calculated from the ratio:

$$\rho = \frac{E_1}{E_2} = \frac{\ell_2^2 \int_{t=0}^{w} a^2 \left(t - \frac{\ell_1}{c}\right) \mathrm{d}t}{\ell_1^2 \int_{t=0}^{w} a^2 \left(t - \frac{\ell_2}{c}\right) \mathrm{d}t}. \quad (4)$$

Assuming that during w, the recorded signal varies little between the two microphones, one can consider 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, without significant loss of accuracy, ρ can be simplified as:

$$\rho = \frac{\ell_2^2}{\ell_1^2} \quad (5)$$

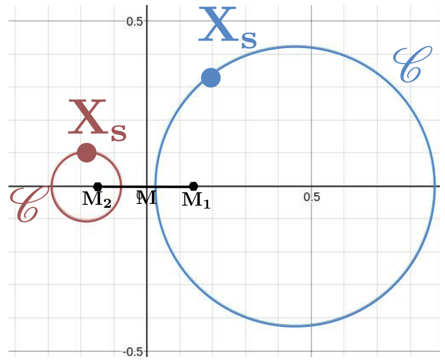


Figure 2: ILD refers geometrically to a circle $\mathscr{C}$ to which the source $\mathbf{X}_s$ belongs.

By developing the expression given by (5) with (1), we obtain the equation of a circle $\mathscr{C}$ characterized by

$$(x_s - c_x)^2 + y_s^2 - \frac{E_1 E_2 d^2}{(E_1 - E_2)^2} = 0, \quad (6)$$

where $c_x = \frac{d}{2} \frac{E_1 + E_2}{E_1 - E_2}$. This result states that $\mathbf{X_s}$ is located on the circle $\mathscr{C}$, illustrated in Fig. 2, centered on the point $(c_x, 0)$ and with a radius c_r given by:

$$c_r = d \left| \frac{\sqrt{E_1 E_2}}{(E_1 - E_2)} \right|. \quad (7)$$

Moreover, (7) implies that the circle $\mathscr{C}$ exists only if $E_1 \neq E_2$. Otherwise the circle degenerates into a straight line that corresponds to the bisector of $\overline{\mathbf{M_1 M_2}}$.

III. ILD-BASED TASK

A. Control scheme

In order to further characterize the properties of ILD, we perform a study from the sensor-based control perspective. This approach aims at minimizing an error e between the current ILD measurement ρ and a desired ILD ρ^*. The control scheme is given by [17]

$$\mathbf{u} = \widehat{\mathbf{L}_\rho^+} \dot{\mathbf{e}}^*, \quad (8)$$

in which $\mathbf{u} = (v_x, v_y, \omega_z)$ is the velocity input of the system, $\dot{\mathbf{e}}^*$ describes a desired behavior of the error, while $\widehat{\mathbf{L}_\rho^+}$ is an approximation of $\mathbf{L}_\rho$ the interaction matrix characterizing the relationship between the microphones motion and the ILD variation. The interaction matrix $\mathbf{L}_\rho$ has been given in [15] as:

$$\mathbf{L}_\rho(x_s, y_s, \ell) = \begin{bmatrix} L_{v_x} & L_{v_y} & L_{\omega_z} \end{bmatrix}, \quad (9)$$

with

$$\begin{cases} L_{v_x} = \frac{2x_s(\rho - 1) - d(\rho + 1)}{\ell^2 + \frac{d^2}{4} - d x_s} \\ L_{v_y} = \frac{2y_s(\rho - 1)}{\ell^2 + \frac{d^2}{4} - d x_s} \\ L_{\omega_z} = \frac{y_s d(\rho + 1)}{\ell^2 + \frac{d^2}{4} - d x_s}. \end{cases} \quad (10)$$

Thus the approximated interaction matrix is $\widehat{\mathbf{L}_\rho^+} = \mathbf{L}_\rho(\widehat{x}_s, \widehat{y}_s, \widehat{\ell})$. We will see in Section V how to approximate these parameters.

B. ILD-based task

Let us focus on the motion implied by a task defined by $\rho = \rho^*$. This analysis is supported by the virtual link approach [18]. In this approach, a vector subspace $\mathbf{S}_\rho^*$ represents the set of admissible microphone motions for which the ILD ρ remains constant:

$$\mathbf{S}_\rho^* = \text{Ker } \mathbf{L}_\rho. \tag{11}$$

To ease this analysis, we consider in the following the interaction matrix $\mathbf{L_X}$ expressed in a frame $\mathcal{F}_m^{'}$ centered on the source $\mathbf{X_s}$. In this case, $\mathbf{L_X}$ expresses the ILD variation when the microphones are moving w.r.t the sound source while $\mathbf{S_X^*}$ represents the set of of admissible sound source motions for which ρ stays unchanged. In a second phase, by symmetry, we can deduce the motions defined in $\mathbf{S}_\rho^*$. Geometrically, the transformation from $\mathcal{F}_m$ to $\mathcal{F}_m^{'}$ is expressed by a translation of a 3D vector $\mathbf{t} = (x_s, y_s, 0)$ and the identity rotation matrix. $\mathbf{L_X}$ is given by

$$\mathbf{L_X} = \mathbf{L}_\rho \mathbf{W} = \begin{bmatrix} \frac{2x_s(\rho-1)-d(\rho+1)}{\ell^2+\frac{d^2}{4}-dx_s} & \frac{2y_s(\rho-1)}{\ell^2+\frac{d^2}{4}-dx_s} & 0 \end{bmatrix}, \tag{12}$$

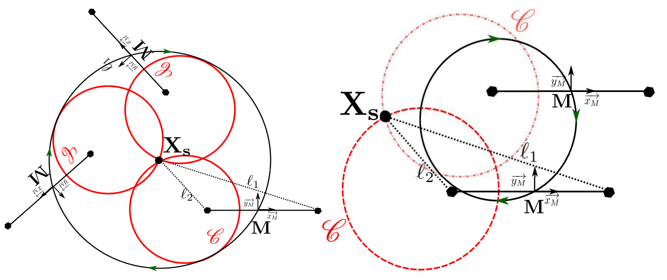
where $\mathbf{W}$ is the spatial motion matrix expressing the relationship between the velocity in $\mathcal{F}_m$ and $\mathcal{F}_m^{'}$. $\mathbf{W}$ is defined as [19]:

$$\mathbf{W} = \begin{bmatrix} \mathbf{R} & [\mathbf{t}]_\times \mathbf{R} \\ 0 & \mathbf{R} \end{bmatrix}. \tag{13}$$

$\mathbf{R}$ and $[\mathbf{t}]_\times$ are respectively the rotation matrix and the skew matrix of the translation vector between $\mathcal{F}_m$ and $\mathcal{F}_m^{'}$. Note that to obtain (12), $\mathbf{W}$ has been adapted to the 3 DOF of our system. Since $\mathbf{L_X}$ is a rank-one matrix, $\mathbf{S_X^*} \in \mathbb{R}^{3\times 2}$ and it is easy to obtain:

$$\mathbf{S_X^*} = \begin{bmatrix} \mathbf{v_{M_1}^*} & \mathbf{v_{M_2}^*} \end{bmatrix} = \begin{bmatrix} 0 & 2y_s(\rho-1) \\ 0 & d(\rho+1)+2x_s(1-\rho) \\ 1 & 0 \end{bmatrix}. \tag{14}$$

The first vector $\mathbf{v_{M_1}^*}$ refers to motion of the sound source rotating around its own center. It can be immediately deduced that, in $\mathbf{S}_\rho^*$, this vector refers to a circular motion of the microphones around $\mathbf{X_s}$, by maintaining the same relative orientation (see Fig. 3a). This motion explicitly characterizes the relation between ILD and the direction since any variation of the microphone orientation w.r.t the sound source would modify the corresponding ILD. More details are given in [15]. As for the second vector $\mathbf{v_{M_2}^*}$, it refers to a motion of the source $\mathbf{X_s}$ on the circle $\mathscr{C}$ (see Fig. 2). Hence the corresponding microphone motion in $\mathbf{S}_\rho^*$ refers to a circular path so that $\mathbf{X_s}$ position varies on $\mathscr{C}$ (see Fig. 3b). By combining these two motions, we obtain an annulus $\mathscr{D}$ characterized by an inner radius r_1 and an outer radius r_2, as depicted in Fig. 4. As a consequence, any linear combination of $\mathbf{v_{M_1}^*}$ and $\mathbf{v_{M_2}^*}$ implies infinite poses for which $\rho = \rho^*$. Geometrically, our control framework defined in Section III-A consists in orienting the robot towards a given direction with respect to the source location while being in or reaching $\mathscr{D}$. Thus this result already lets us envision the relationship between ILD and distance.



(a) (b)

Figure 3: Rotation motion (a) and translation motion (b) implied by $\mathbf{S}_\rho^*$

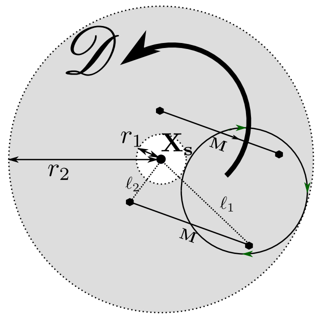


Figure 4: The admissible poses of the microphones for a given ρ belong to an annulus $\mathscr{D}$

C. Relating ILD to distance

In order to gain further insight into the properties of the annulus $\mathscr{D}$, we generalize the task analysis performed previously. For an arbitrary ILD ρ, $\mathbf{M}$ is located in the corresponding annulus $\mathscr{D}$. The boundaries of this annulus can be studied knowing that $r_1 \leq \ell \leq r_2$. The radius r_1 corresponds to the shortest distance between the source position and $\mathbf{M}$, that is when the sound source is located on the intersection of $\mathscr{C}$ and $\overline{\mathbf{M_1M_2}}$ (see Fig. 2). For this configuration, r_1 can be analytically expressed as:

$$r_1 = \left|\frac{d}{2} - \ell_1\right| = \left|\frac{d}{2} - \ell_2\right|. \tag{15}$$

Furthermore, knowing the following relationships

$$\frac{\ell_2}{\ell_1} = \sqrt{\rho} \text{ and } \ell_1 + \ell_2 = d, \tag{16}$$

each ℓ_i with respect to ρ is given by

$$\ell_2 = \frac{d\sqrt{\rho}}{1+\sqrt{\rho}} \text{ and } \ell_1 = d - \frac{d\sqrt{\rho}}{1+\sqrt{\rho}}. \tag{17}$$

By injecting (17) into (15), we obtain

$$r_1 = \left|\frac{d}{2} - \frac{d\sqrt{\rho}}{1+\sqrt{\rho}}\right| = \frac{d}{2}\left|\frac{1-\sqrt{\rho}}{1+\sqrt{\rho}}\right|. \tag{18}$$

With a similar reasoning r_2 corresponds to the configuration where the sound source is the furthest from $\mathbf{M}$. From (4) and (7) , c_r is given by

$$c_r = d\left|\frac{\sqrt{\rho}}{1-\rho}\right|, \tag{19}$$

so that we can express r_2 as

$$r_2 = r_1 + 2c_r = \frac{d}{2}\left|\frac{1-\sqrt{\rho}}{1+\sqrt{\rho}}\right| + 2d\left|\frac{\sqrt{\rho}}{1-\rho}\right|. \tag{20}$$

From these results, one can emphasize the following properties.

1) Uniqueness of the annulus: First, we assume that the source $\mathbf{X_s}$ is located on the right side (*i.e.*, $x_s > 0$ and $\rho > 1$). In this configuration where $\rho \in]1;\infty[$, we obtain that $r_2 \in]\infty; \frac{d}{2}[$ and $c_r \in]\infty; 0[$ are monotonic decreasing functions of ρ. On the other hand when $\rho < 1$ and $x_s < 0$, the inner radius $r_1 \in]\frac{d}{2}; 0[$ is a monotonic decreasing function, the width $c_r \in]0; +\infty[$ is a monotonic increasing function, while the outer radius $r_2 \in]\frac{d}{2}; +\infty[$ is a monotonic increasing function. These results implies that an arbitrary $\rho \in]1; +\infty[$ or $\rho \in]0; 1[$ is characterized by a unique annulus $\mathscr{D}$ shaped by r_1, r_2 and c_r. Furthermore, since $c_r(\rho) = c_r(\frac{1}{\rho})$, $r_1(\rho) = r_1(\frac{1}{\rho})$ and $r_2(\rho) = r_2(\frac{1}{\rho})$, for (ρ_1, ρ_2) so that $\rho_1 = 1/\rho_2$, the same annulus is obtained but with a different microphone orientation (*i.e.*, a rotation by π).

This uniqueness property implies that ILD evolves with distance. Indeed, by positioning the microphones in different annuli, while maintaining the same orientation w.r.t. the sound source ($\rho \neq 1$), different ρ are measured. This property is also consistent with Fig. 3a implying constant ILD measurement for a fixed orientation w.r.t the sound source only if the distance stays constant.

2) Accuracy of the distance information: From the preceding results, we analyze how accurate the distance can be estimated, that is how narrow $\mathscr{D}$ is, since we have $r_1 \leq \ell \leq r_2$. As stated previously, c_r is a monotonically increasing function for $\rho \in]0; 1[$, while it is a monotonically decreasing function for $\rho \in]1; \infty[$. As a result, for very dissimilar energy level between $\mathbf{M_1}$ and $\mathbf{M_2}$, that is equivalent to eccentric positions of $\mathbf{X_s}$ w.r.t the interaural axis, $\mathscr{D}$ is narrow. In this configuration distance can be estimated fairly accurately. However, for $\mathbf{X_s}$ located near the bisector of $\overline{\mathbf{M_1M_2}}$ (*i.e.*, auditory fovea), the distance information becomes inaccurate since the width of $\mathscr{D}$ increases. Likewise, for distant sound sources implying $\ell_1^2 \to \ell_2^2$ and $\rho \to 1$, the same limitation can be emphasized. This limitation is exacerbated when $\rho = 1$ since we get $r_1 = 0$ and $r_2 = \infty$ that lead to an infinite area $\mathscr{D}$. These results are illustrated in Fig. 5 showing the evolution of the parameter c_r with respect to ILD.

Interestingly, all these properties have also been observed on human listeners [20], [11], [21], [22]. For instance, the experiments conducted in [11] showed that human listeners were able to fairly estimate distance using ILD for nearby and eccentric sound positions, while around the auditory fovea, this capacity was lost. Furthermore, we also notice the strong similarity between the $2c_r$ curve in Fig. 5 and the averaged measured ILD across human subjects at low frequency with respect to distance in ([20], Fig. 4). Although we considered until now free-field setups, such a similarity is not really surprising: at low frequency, the head attenuation effect is reduced and ILD tends to be similar as in free-field condition. Analytically, having the source distance "exactly" varying as $2c_r$ curve implies that the sound source is located near the outer radius of $\mathscr{D}$ (*i.e.*, $\ell \approx r_2$) since $r_1 \ll 2c_r$. In such a configuration the distance to the sound source can then be accurately estimated. More importantly, the curves presented in [20] for higher frequencies exhibit similar shape as in Fig. 5. Such a similarity emphasizes that at low or high frequencies, for head-mounted setups or free-field setups, the variation of ILD (or derivative) remains unchanged. This interesting outcome is certainly one of the keypoint of our approach: we can apply our control system to different platforms independently of the signal frequency since the interaction matrix $\mathbf{L}_\rho$ used to control the robot is directly extracted from ILD derivative.

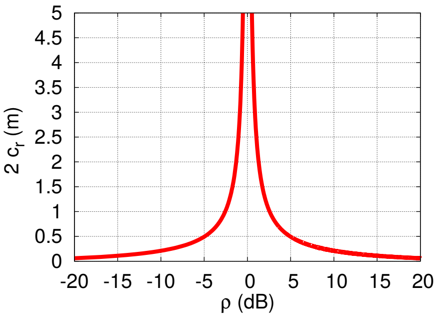


Figure 5: Annulus width with respect to ILD value ($d = 0.3$ m).

3) Repercussion on ILD-based tasks: The former properties affect the task for which $\rho = \rho^*$ by requiring the microphones to be located in $\mathscr{D}$. Indeed, since the area of $\mathscr{D}$ is infinite when $\rho = 1$, a task $\rho = \rho^* = 1$ mainly consists in facing the sound source as exploited in [15]. This task does not constrain distance to the sound source (this is also clear from (9) and (10) since the component of $\mathbf{L}_\rho$ related to v_y is equal to 0 when $\rho = 1$). However for ρ^* implying highly dissimilar energy level between $\mathbf{M_1}$ and $\mathbf{M_2}$, the ILD-based task can only be achieved at a given range to the sound source, since c_r (resp. r_2) decreases. Such a task consists in reaching $\mathscr{D}$ with a given orientation. For instance, with $d = 0.3$ m and $\rho \equiv \pm 4$ dB the width of the annulus is $2c_r \approx 0.63$ m. In practice a distance $0 < \ell = r_1 < \frac{d}{2}$ (*i.e.*, $\mathbf{X_s}$ located on $\overline{\mathbf{M_1M_2}}$) is not realistic. Hence we can refine the distance range to 0.15 m $< \ell \leq 0.66$ m . As a result, a robot solves a task implying $\rho = \rho^* \equiv 4$ dB by reaching a pose below 0.66 m to the sound source if it starts from a larger distance.

IV. Control Strategy

A. Exploiting range and azimuth information

From the task analysis performed in the previous section, one can exploit the range information contained in ILD that is characterized by the requirement to reach $\mathscr{D}$. In this manner, by setting ρ^* implying dissimilar energy values (*i.e.*, a narrower $\mathscr{D}$), a robot endowed with a pair of microphones can be driven towards a pose close to the sound source of interest. Unfortunately, such control tasks orient the microphones so that $\mathbf{X_s}$ is in an eccentric position. In robot audition, this pose is not particularly suitable for interacting with or tracking a sound source. Facing and/or maintaining a sound source in the auditory fovea are more relevant tasks. But on the other hand, for the task where $\rho^* = 1$, the distance information cannot be exploited as already discussed. Approaching and facing a sound source are then conflicting objectives, that cannot be processed simultaneously using ILD only. We thus propose to control both degrees of freedom, azimuth and range, through a sequence of tasks. More exactly, we consider first controlling the distance to the sound source $\mathbf{X_s}$ and then facing $\mathbf{X_s}$. Such an approach can be performed simply by using different ρ^* in the control scheme.

B. Control scheme

In order to develop this control strategy, the control input (8) is decomposed into three successive phases given by:

$$\begin{cases} 1)\mathbf{u}_1(t) = \omega_z = L_{\omega_z}^{-1}\dot{e_1}^* \text{ while } e_1 > e_{tr} \\ 2)\mathbf{u}_2(t) = (v_x, v_y) = [L_{v_x}, L_{v_y}]^+\dot{e_2}^* \text{ while } e_2 > e_{tr2} \\ 3)\mathbf{u}_3(t) = \omega_z = L_{\omega_z}^{-1}\dot{e_3}^* \end{cases} \tag{21}$$

where e_{tri} are the error thresholds that condition the switching time t_{si} between two successive tasks. More specifically the different phases induced by these tasks are:

1) Task 1: The first phase consists in controlling the initial microphone orientation by considering ω_z only. The orientation w.r.t the sound source is controlled by decreasing the error $e_1 = \rho - \rho_1^*$. Naturally we choose $\rho_1^* \neq 1$ for exploiting the distance information during the second task.

2) Task 2: The second phase consists in moving towards the sound source. In this phase ρ_2^* is selected in such a way that the microphones reach a pose at ℓ^* from the sound source. In order to avoid any orientation variation that could lead to poses with different ℓ^*, the control input is restricted to v_x and v_y and decreases the error $e_2 = \rho - \rho_2^*$. In this way the microphones keep their initial orientation given by the first task.

3) Task 3: Eventually, the microphones are moved to face the sound source with $\rho_3^* = 1$. To prevent the system from any backward motion induced by the task characterized by an error $e_3 = \rho - \rho_3^*$, once again, only ω_z is involved.

C. Task continuity

In order to ensure a smooth transition between the three successive steps, it is necessary to ensure the continuity of the different tasks at each switching time t_{si}. To this end, we use the task sequencing formalism proposed in [23]. More prosaically, the first task is defined by $\dot{e_1}^*(t) = -\lambda e_1(t)$ with λ characterizing the time to convergence. The second and third tasks are determined by

$$\dot{e_i}^*(t) = -\lambda e_i(t) + \nu_i(t) \tag{22}$$

where the additional term $\nu_i(t)$ is given by

$$\nu_i(t) = \exp^{-\mu(t-t_{si})}\left(\dot{e}_{i-1}^*(t_{si}) + \lambda e_i(t_{si})\right). \tag{23}$$

μ tunes the length of the transition between two successive tasks. In (22), $\nu_i(t)$ gradually vanishes so that $\dot{e_i}^*(t)$ smoothly varies from $\dot{e}_{i-1}^*(t_{si})$ to $-\lambda e_i(t)$. The parameters tuning of the control scheme (ρ_i^*, μ) is detailed in the next section for the experimental results.

V. Results

A. Simulation results

We first demonstrate and analyze the validity of our approach in simulation. In an environment created from *Roomsimove* [24], we consider anechoic conditions in which several positioning tasks are performed with respect to a speech source. The microphones are initially located at 3 m $<$ ℓ $<$ 3.6 m as illustrated in Fig. 6. From the speech signal, we extracted ILD every 100 ms. For more simplicity, ILD is computed as the ratio of absolute signal energy (without frequency filtering) received at each microphone (see (4)). For the first task the desired ratio is set to $\rho_1^* = 1.2 \equiv 0.8$ dB, while the second task corresponds to $\rho_2^* = 1.7 \equiv 2.3$ dB and the third to $\rho_3^* = 1$. By referring to Fig. 5 and r_2 in (20), it can then be expected to reach a pose at $\ell < 1.14$ m.

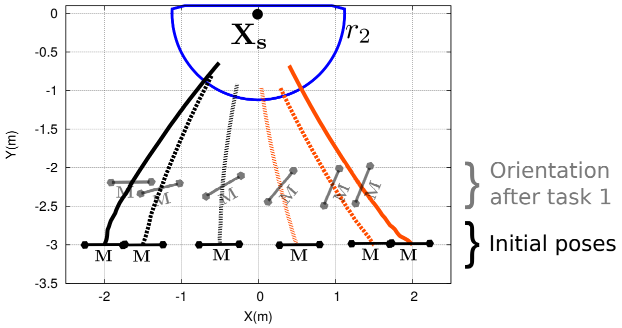
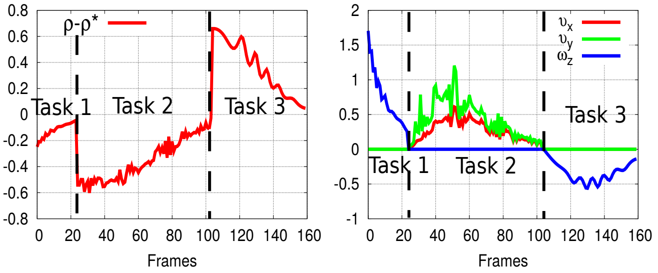


Figure 6: Simulation results: the trajectories obtained from different initial configurations lead to a final pose where $\ell < r_2$.

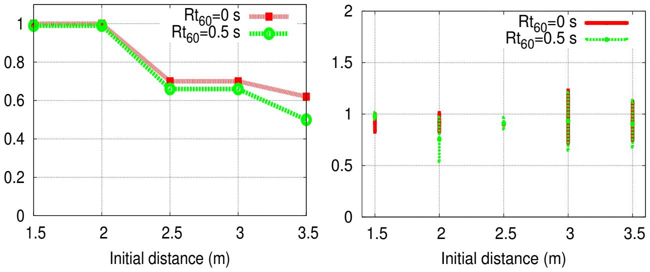


(a) Error (b) Velocity input

Figure 7: Simulation results: typical data from one of the generated trajectory. For each task, the error successfully converges towards 0.

From Fig. 6, the six positioning tasks are all correctly achieved from the different starting poses. As detailed in the example in Fig. 7, in all phases of each task, the error $\rho - \rho_i^*$ decreases to 0. Furthermore, the final poses correspond to our initial expectation since we obtain $\ell < r_2$ at the end in all cases. However these distances from the sound source are not the same for all simulations. Indeed depending on the orientation of the microphones w.r.t. the sound source, the value ρ_2^* is reached at slightly different distances ℓ. For the same orientation it is expected to reach the same distance to the sound source for the second task as illustrated in Fig. 3a. However, since at the end of the first task the microphones have different orientations w.r.t the sound source there is a variation in final pose distance ℓ.

In a second phase, we analyze the accuracy of our system under varying initial distances, and reverberation times. It should be determined if the positioning task is correctly achieved and how accurate the final pose is. In the same simulation environment as before, we conducted several tasks from an initial distance varying distance $\ell = \{1.5, 2, 2.5, 3, 3.5\}$ m, from 15 different positions for each ℓ. We reproduced these simulations in anechoic conditions and with a reverberation time RT$_{60}$ $= 0.5$ s. On Fig. 8a, we represented the rate of successful tasks. Unsurprisingly, we notice that reverberation deteriorates the performance of our system, especially for distant positions. Likewise, the poor resolution of ILD for far distance affects our system performance. Indeed with an inaccurate orientation after the first task, sometimes the microphones could not reach a pose where $\rho_2 = \rho_2^*$ during the second task. This phenomenon mainly explains the drop in performance of our system for $\ell > 2$ m.



(a) Rate of successful tasks

(b) Final pose mean and deviation in meters of the final pose from the sound source

Furthermore we characterized the accuracy of our system (among the tasks that are correctly achieved), by plotting the mean position and deviation for each starting distance ℓ. As illustrated in Fig. 8b the accuracy of our system remains satisfactory. Indeed, when starting from the same pose, the deviation remains below 28 cm even for distance above 3 m while reverberation varies from 0 to 0.5 s for a fixed starting distance. The accuracy is particularly good in the near field ($\ell \leq 2$ m) with a deviation below 20 cm, which is acceptable for most real world applications. In comparison to distance estimation studies like [13] or [25], our system performs better, even though the evaluation condition and the purpose of these works are slightly different. Even when considering the whole set of different starting pose ℓ, our results still remains comparable to these studies. At last, we studied the effect of the sound type on the task accuracy for different microphone distance $d = \{0.3, 0.15\}$ m, with $\ell = 2$ from 15 different poses. We used a speech signal (a), a white noise signal (b), a pure tone of 100 Hz (c), a pure tone of 10 kHz (d) and a pure tone of 10 kHz with half amplitude (e). It should be noted that in the case of $d = 0.15$ m the desired ILD ρ_1^* and ρ_2^* are respectively measured in the desired pose of the end of the task 1 and calculated in order to obtain a similar annulus as for $d = 0.3$ m. For all these signals, the task was always successfully achieved as already observed on the previous study (see Fig. 8a). For these tasks, the mean final distance to the sound source is given in Table I.

	(a)	(b)	(c)	(d)	(e)
$d = 0.3$	1.004	1.077	1.022	1.004	1.004
$d = 0.15$	0.891	1.112	0.992	1.023	1.026

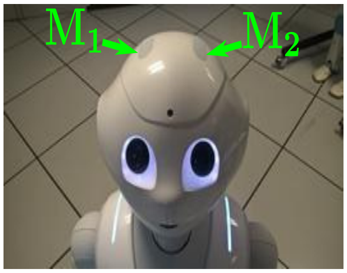
Table I: Mean distance to the sound source in meter for a speech signal (a), a white noise signal(b), a pure tone of 100 Hz(c), a pure tone of 10 kHz (d) and a pure tone of 10 kHz with half amplitude (e).

First we can emphasize that our approach is robust to the signal type, since the results are similar for the speech signal, the white noise signal and the pure tone signals. More importantly, this study confirms that our approach is frequency independent as already hypothesized in Section III. The pure tones at different frequencies lead to the same results and mean final distance. Moreover, we noticed that changing d does not affect the task accuracy since ρ_1^* and ρ_2^* are adapted to this new configuration, although the variation in the final poses is higher with $d = 0.15$ m but still remains acceptable. This is not surprising since with shorter inter-microphone distance the lack of ILD resolution is higher. Such results stress a major advantage of our approach: it can be used on different setups independently of the signal type and frequency, which allowed us to experiment it on a humanoid robot as described in the next section.

B. *Experimental results*

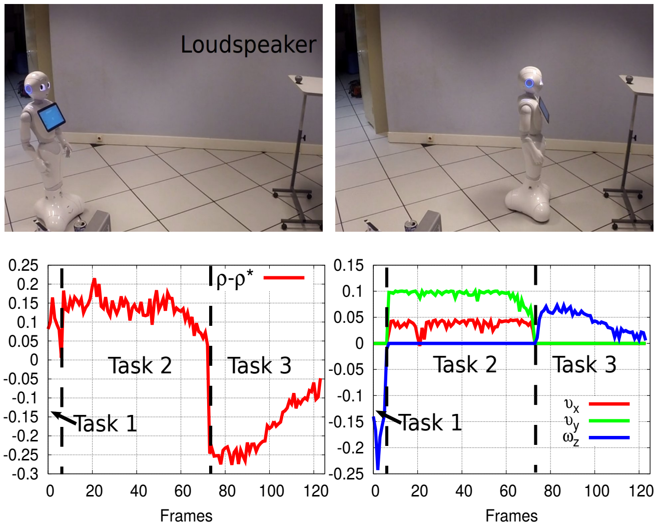
The control framework has also been experimented in real world conditions. We directly implemented our approach on the humanoid robot *Pepper* without accounting for its specific HRTF. Indeed even if our model is based on free-field sound propagation, we already demonstrated the versatility of the audio-based control approach that can be freely applied on free-field or head-mounted systems (see [26] for details). On this robot we used two microphones as depicted in Fig. 8, separated by a distance $d \approx 8$ cm. In the following experiments we control the holonomic base of the robot through $\mathbf{u} = (v_x, v_y, \omega_z)$. With a reverberation time RT$_{60}$ > 1 s in the room, a loudspeaker was playing continuously a white Gaussian noise. The different parameters used for the experiments are detailed in Fig. 8. It should also be mentioned that ρ_1^* and ρ_2^* have been obtained by direct measurement, while μ has been fixed empirically to 5λ. From the robot orientation corresponding to $\rho_1 = \rho_1^*$, the interaction matrix $\widehat{\mathbf{L}_\rho}$ was roughly approximated with constant values $\widehat{x}_s = 0.5\widehat{y}_s$ that corresponded to a motion in the sound source direction. $\widehat{\ell}$ was defined in consequence with (1).

In a first experiment illustrated by Fig. 9 the robot is located at $\ell \approx 2.5$ m to the sound source. During this experiment the robot initially oriented itself in order to measure ρ_1^*. Thereafter, the robot moved towards the sound source until reaching a pose at $\ell \approx 0.8$ m to the loudspeaker. Finally the robot turned towards the sound source. This behavior is stressed by the velocity input of the robot. Likewise the error converges towards 0 for each of the phases described above. It should also be noted that the error curve does not exactly reaches 0 at the end of the different tasks because of the robot that cannot handle very small velocity inputs.



d	0.08 m
ρ_1^*	0.83
ρ_2^*	0.6
ρ_3^*	1
$\widehat{y}_s$	3 m
$\widehat{x}_s$	sgn($\rho-1$)×1.5 m
λ	1
μ	5λ

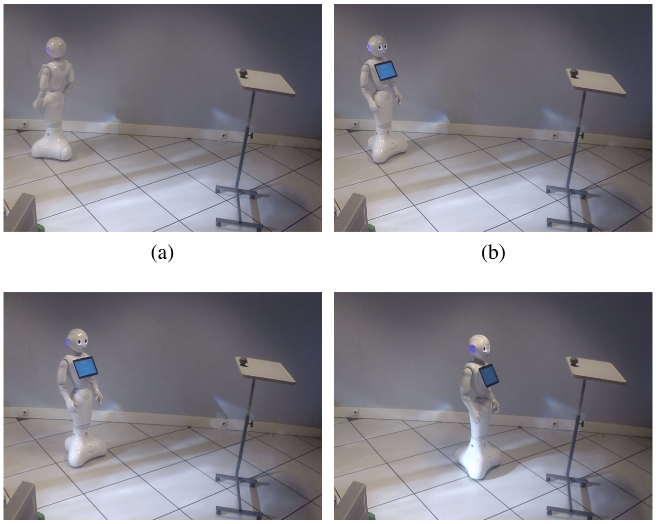
Figure 8: Experimental settings



(c) Features error (d) Velocity input

Figure 9: *Pepper* approaches the sound source through three tasks for which $e_i = \rho - \rho_i^*$ that allow controlling the orientation and the distance to the sound source.

We also repeated this experiment for a different configuration where we started with the sound source being located behind the robot. As depicted in Fig. 10 such a configuration is correctly tackled by the control scheme. As already developed in our previous work [15], the control scheme is able to address the front-back ambiguity during the first task. Afterwards, the following tasks are addressed similarly to the previous experiment.



(a) (b)

(c) (d)

Figure 10: The control framework is able to address the front-back ambiguity: during the first task the robot is correctly oriented. Thereafter the tasks 2 and 3 can be completed.

VI. Conclusion

We developed in this paper a control strategy allowing to approach a sound source from a binaural system by measuring only ILD. For lateral sound sources, in addition to the azimuth angle, a relationship between ILD and distance appears as demonstrated analytically in this paper with the concept of ILD annulus. This analysis is consistent with observations made on human listeners. From this relationship, a set of audio-based tasks has been designed to control the orientation and the distance to a sound source: a first task orients the microphones, the second task consists in approaching the sound source at a given distance, while in the last task the microphones are controlled to face the sound source. This framework has been validated, in real environments, on a humanoid robot that is able to approach a sound source. Notably, the validity of our approach has been confirmed for different configurations independently of the signal frequency/type or the experimental setup (head-mounted system or free-field microphones). Furthermore, our approach does not necessitate any explicit sound source localization. Developing this kind of methods to control the distance to a sound source presents several advantages. Compared to a distance cue such as the direct-to-reverberant ratio, the ILD estimation is easier and less time consuming to process. In parallel, this cue is adapted to non-stationary signals unlike the sound level, another distance cue that requires to "know" the sound source. Additionally, compared to ITD the orientation control is processed with low complexity since ILD is easier to extract, although ITD is more accurate for long distances. Despite these promising results, the developed control system still lacks of accuracy for the range positioning because of the poor spatial resolution of ILD when starting from a distance greater than 3 m. If a greater accuracy and performance are needed it should be possible to combine ILD and ITD (since both cues can easily be estimated when considering one sound source) to improve the system. A different path of improvement would be to exploit in the same phase the orientation and distance cues in order to obtain a more natural and smooth motion for the robot.

References

[1] 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," *Robotics and Autonomous Systems*, vol. 27, no. 4, pp. 199–209, 1999.

[2] J-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2003, vol. 2, pp. 1228–1233.

[3] S. Birchfield and R. Gangishetty, "Acoustic localization by interaural level difference," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2005, vol. 4, pp. iv–1109.

[4] W. Cui, Z. Cao, and J. Wei, "Dual-microphone source location method in 2-d space," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2006, vol. 4, pp. IV–IV.

[5] T. Nakadai, K.and Lourens, H G. Okuno, and H. Kitano, "Active audition for humanoid," in *AAAI/IAAI*, 2000, pp. 832–839.

[6] A. Portello, P. Danès, and S. Argentieri, "Active binaural localization of intermittent moving sources in the presence of false measurements," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 3294–3299.

[7] H-D Kim, K. Komatani, T. Ogata, and H G Okuno, "Binaural active audition for humanoid robots to localise speech over entire azimuth range," *Applied Bionics and Biomechanics*, vol. 6, no. 3-4, pp. 355–367, 2009.

[8] K. Nakadai, H G Okuno, and H. Kitano, "Exploiting auditory fovea in humanoid-human interaction," in *AAAI/IAAI*, 2002, pp. 431–438.

[9] L. Kneip and C. Baumann, "Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3108–3119, 2008.

[10] A J. Kolarik, B. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss," *Attention, Perception, & Psychophysics*, vol. 78, no. 2, pp. 373–395, 2016.

[11] D S. Brungart and W M. Rabinowitz, "Auditory localization of nearby sources. head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1465–1479, 1999.

[12] N. Kopčo and BG. Shinn-Cunningham, "Spatial unmasking of nearby pure-tone targets in a simulated anechoic environment," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2856–2870, 2003.

[13] T. Rodemann, "A study on distance estimation in binaural sound localization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 425–430.

[14] M. Bernard, P. Pirim, A. De Cheveigné, and B. Gas, "Sensorimotor learning of sound localization from an auditory evoked behavior," in *IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 91–96.

[15] A. Magassouba, N. Bertin, and F. Chaumette, "Audio-based robot control from interchannel level difference and absolute sound energy," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2016, pp. 1992–1999.

[16] A. Magassouba, N. Bertin, and F. Chaumette, "Binaural auditory interaction without HRTF for humanoid robots: A sensor-based control approach," Workshop on Multimodal Sensor-based Control for HRI and soft manipulation, IROS'2016, October 2016.

[17] F. Chaumette and S. Hutchinson, "Visual servo control. i. basic approaches," *IEEE Robotics and Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.

[18] F. Chaumette, P. Rives, and B. Espiau, "Classification and realization of the different vision-based tasks," *Visual Servoing*, vol. 7, pp. 199–228, 1993.

[19] G. Hager, W-C Chang, and S. Morse, "Robot feedback control based on stereo vision: Towards calibration-free hand-eye coordination," in *IEEE International Conference on Robotics and Automation*. IEEE, 1994, pp. 2850–2856.

[20] N. Kopčo and BG. Shinn-Cunningham, "Effect of stimulus spectrum on distance perception for nearby sources a," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1530–1541, 2011.

[21] P. Cochran, J. Throop, and WE Simpson, "Estimation of distance of a source of sound," *The American journal of psychology*, vol. 81, no. 2, pp. 198–206, 1968.

[22] BG. Shinn-Cunningham, S. Santarelli, and N. Kopčo, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1627–1636, 2000.

[23] N. Mansard and F. Chaumette, "Task sequencing for high-level sensor-based control," *IEEE Transactions on Robotics*, vol. 23, no. 1, pp. 60–72, 2007.

[24] 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.

[25] Ya-C Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1793–1805, 2010.

[26] A. Magassouba, *Aural servo: towards an alternative approach to sound localization for robot motion control*, Ph.D. thesis, Université Rennes 1, 2016, https://hal.inria.fr/tel-01426710v3.