Audio-based robot control from interchannel level difference and absolute sound energy

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

***Abstract*— This paper is a follow-up way to our previous works regarding audio-based control, that is an alternative method for auditory-based robot tasks. Conversely to classic methods oriented towards sound source localization, audio-based control is a sensor-based framework that does not localize the sound source. Instead, auditory features are used as inputs of a closed-loop control scheme. The audio-based control method presented in this paper relies on the sound signal energy measured by two microphones. By combining the interchannel level difference to the acoustic absolute energy level, the control scheme allows positioning the robot with respect to the sound source at a given distance and orientation. Moreover this method has the benefit of a low computation cost, since it only relies on the signal energy measurement. Experimental results conducted on a mobile robot validate the relevance and the robustness of this approach in dynamic and real world conditions.**

I. INTRODUCTION

Aural perception provides rich information that is still under-exploited in robotic applications. Despite a growing community and interest in robot audition, several topics related to hearing sense are still unresolved for real world applications. Among them, sound source localization (SSL) remains a complex task. This task is even more challenging when considering low number of sensors, which is the case of binaural approaches based on two microphones.

Nowadays, binaural SSL methods are generally built upon signal processing and acoustics knowledge, by exploiting acoustic cues in order to mimic the prodigious human aural sense [1][2]. But modeling all variability of acoustic environment and/or perception, that is unique for each individual, is a complex problem to solve. This is especially the case in dynamic environment where the sound source of interest and the robot are also moving. Nonetheless promising results have been obtained in active audition, that couples the motion of the robot to the hearing perception [3][4]. Up to now this approach has been designed in order to improve the localization of the sound source, which is typically the case of [5], where the authors try to minimize the uncertainty of the localization in a feedback loop. These results stress that control methods can also contribute to the complex topic of sound sensing in robotics. From this point of view, we proposed previously in [6] a sensor-based approach relying on the measurement of the Interchannel Time Difference (ITD). This approach allows positioning a mobile robot regarding its current perception of the sound signal, without any need of sound source localization. In this context, this paper explores the use of the acoustic signal energy in a closed-loop control scheme. Indeed, an intuitive way to sense a sound source is to analyze the energy perceived from it. In the binaural approach, the Interchannel Level Difference (ILD) provides information about the direction of the sound source. However when considering free-field installations, only few works, such as [7], exploit this characteristic. This is mainly due to the difficulty to extract consistent ILD in realistic environments for SSL. In the case of head-mounted systems inducing scattering, the ILD generally complements the ITD by providing location information in high frequency ranges [8]. However, for free-field microphones, the ILD is redundant with the ITD while being more sensitive to reverberation [9] and less accurate for distant sound sources. As a consequence most approaches are based on ITD.

By contrast, since the sensor-based control approach is not concerned with localizing the sound source, more freedom is given to exploit acoustic features. Actually with a specific modelling, features known to be less robust to real acoustic conditions can still be utilized. In this paper, we propose a feature modelling based on ILD measurement derived from acoustic properties of sound propagation in free field conditions. The modelling is completed with the use of the absolute sound energy to overcome the lack of consistency and accuracy of the ILD for a distant source. Although the distance cues are not often used in robot audition [10], mainly for accuracy issues, our sensor-based modelling allows us to exploit the absolute level of energy that is directly related to the distance. Thus, by combining these two features, that are the ILD and the absolute level of energy, our control scheme is able to solve the limitations inherent to ILD measurements, by also constraining the distance to the sound source.

As a result, a mobile robot instrumented with a binaural system is able to accurately follow and orient itself towards a moving sound source, without any explicit tracking. The low computation cost of the proposed method, coupled with the robustness and the flexibility of the control framework to dynamic environment, make us believe that our approach can be deployed for real world robotic applications as shown in our experimental results.

The rest of this paper is structured as follows: we first introduce the control framework in Section II. A modelling of the ILD and the energy level are then proposed in Sections III and IV. Section V is devoted to the analysis of the global

[1]Université Rennes I - IRISA, Campus de Beaulieu, 35042 Rennes cedex, France aly.magassouba@irisa.fr

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

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

framework. The paper ends in Section VI with experimental results validating this approach in different and challenging scenarios.

II. CONTROL MODELLING

A. Geometric configuration

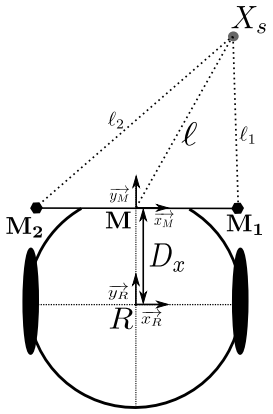


Fig. 1: Robot modelling

Let us consider a non-holonomic unicycle robot, and an omnidirectional sound source $\mathbf{X_s}$ as illustrated in Fig. 1. The robot is controlled in the horizontal plane with two free-field microphones $\mathbf{M_1}$ and $\mathbf{M_2}$, separated by a distance d. Two frames are then defined: $\mathcal{F}_r(\overrightarrow{x_R}, \overrightarrow{y_R})$ attached to the robot, and $\mathcal{F}_m(\overrightarrow{x_M}, \overrightarrow{y_M})$ attached to the microphones. D_x denotes the distance between the center of the robot $\mathbf{R}$ and the midpoint of the microphones $\mathbf{M}$. The sound source $\mathbf{X_s}(x_s, y_s)$, expressed in the microphones frame, is located at a distance ℓ_i from each microphone $\mathbf{M_i}$. We also consider that the sound source is in the front side of the robot (e.g. $y_s > 0$). The robot can be controlled upon two degrees-of-freedom (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 $\overrightarrow{z_R}$.

B. General framework

A task consists in positioning the robot so that given conditions characterized by the acoustic features are satisfied. The task is performed by considering k features $\mathbf{s}(t)$ extracted from the sound signal and by minimizing the error $\|\mathbf{e}(t)\|$ given by [11]

$$\mathbf{e}(t) = \mathbf{s}(t) - \mathbf{s}^* \tag{1}$$

where $\boldsymbol{s}^*$ denotes the measurements for the desired acoustic features. The time variation of $\mathbf{s}$ is related to the sensors velocity by

$$\dot{\mathbf{s}} = \mathbf{L_s}\mathbf{v} \tag{2}$$

in which $\mathbf{L_s} \in \mathbb{R}^{k\times 3}$ is the interaction matrix sized by k and $\mathbf{v} = (v_x, v_y, \omega_z)$ denoting the spatial linear and angular velocity of the microphones expressed in $\mathcal{F}_\mathbf{M}$. Therefore considering the two-DOF robot previously described, the relationship between $\dot{\mathbf{s}}$ and the control input $\dot{\mathbf{q}}$ is:

$$\dot{\mathbf{s}} = \mathbf{J_s}\dot{\mathbf{q}} \tag{3}$$

where $\mathbf{J_s}$ corresponds to

$$\mathbf{J_s} = \mathbf{L_s}\mathbf{J_r}. \tag{4}$$

$\mathbf{J}_r$ being the robot Jacobian, that is

$$\mathbf{J_r} = \begin{bmatrix} 0 & D_x \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{5}$$

Hence, it is possible to design a control scheme where the robot is controlled with [11]

$$\dot{\mathbf{q}} = -\lambda \widehat{\mathbf{J}_\mathbf{s}^+}\mathbf{e}. \tag{6}$$

In this latter equation $\mathbf{J}_\mathbf{s}^+ \in \mathbb{R}^{2\times k}$ is the Moore-Penrose pseudo-inverse of $\mathbf{J_s}$ and $\lambda > 0$ is a gain that tunes the time to convergence. Generally, an approximation $\widehat{\mathbf{J}_\mathbf{s}^+}$ is considered since it is impossible to know perfectly either $\mathbf{J_s}$ or $\mathbf{J}_\mathbf{s}^+$, as shown in the features modelling in Sections III and IV.

III. INTERCHANNEL LEVEL DIFFERENCE MODELLING

A. ILD estimation

In order to apply the aforementioned control scheme with the ILD, let us assume first that the sound source $\mathbf{X_s}$ generates a signal $a(t)$ that is propagating in a free-field. The sound signal is observed during a time frame of length w. For the rest of this paper we assume that $a(t)$ is a continuous and slowly varying signal regarding w. In this condition, the signal perceived by each microphone $\mathbf{M}_i$ is defined from the spherical propagation equation as:

$$x_i(t) \propto \frac{a(t - \frac{\ell_i}{c})}{\ell_i} \tag{7}$$

where $\frac{\ell_i}{c}$ expresses the sound propagation delay that depends on the sound celerity c. In the following developments, without loss of generality of our approach, we consider an unitary proportional gain in (7). Thus, the energy received by each microphone is approximated by integrating (7) over the time frame w, as follows:

$$E_i = \int_{t=0}^{w} |x_i(t)|^2 \,\mathrm{d}t = \frac{1}{\ell_i^2}\int_{t=0}^{w} a^2(t - \frac{\ell_i}{c})\,\mathrm{d}t \tag{8}$$

The ILD ρ between the two microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ is then calculated from the ratio:

$$\rho = \frac{E_1}{E_2} = \frac{\ell_2^2 \int_{t=0}^{w} a^2(t - \frac{\ell_1}{c})\,\mathrm{d}t}{\ell_1^2 \int_{t=0}^{w} a^2(t - \frac{\ell_2}{c})\,\mathrm{d}t}. \tag{9}$$

Nevertheless with a signal varying slowly during w one can assume that $\int_{t=0}^{w} a^2(t - \frac{\ell_1}{c})\,\mathrm{d}t \approx \int_{t=0}^{w} a^2(t - \frac{\ell_2}{c})\,\mathrm{d}t$. Indeed for a close sound source $\frac{\ell_i}{c} \ll w$, while for a distant sound source $\ell_1 \approx \ell_2$ since $\ell_i \gg d$. Consequently the ILD ρ can be simplified as:

$$\rho = \frac{\ell_2^2}{\ell_1^2} \tag{10}$$

B. ILD modelling

The time variation of ρ is given by:

$$\dot{\rho} = \frac{d}{dt}\left(\frac{\ell_2^2}{\ell_1^2}\right) = 2\frac{\ell_2\dot{\ell}_2\ell_1 - \dot{\ell}_1\ell_2^2}{\ell_1^3} \tag{11}$$

Considering the geometric modelling in Fig. 1, the distances ℓ_i are respectively given by

$$\begin{cases} \ell_1 = \sqrt{(x_s - d/2)^2 + y_s^2} \\ \ell_2 = \sqrt{(x_s + d/2)^2 + y_s^2} \end{cases} \tag{12}$$

Consequently by injecting the latter equation into (11), we obtain:

$$\dot{\rho} = \frac{\dot{x_s}(2x_s + d) + 2y_s\dot{y_s}}{\ell_1^2} - \frac{\dot{x_s}(2x_s - d) + 2y_s\dot{y_s}}{\ell_1^2}\rho. \tag{13}$$

From the kinematic equation:

$$\dot{\mathbf{X}}_s = -\boldsymbol{v_s} - \boldsymbol{\omega_s} \times \mathbf{X_s} \Leftrightarrow \begin{cases} \dot{x}_s = -v_x - \omega_y z_s + \omega_z y_s \\ \dot{y}_s = -v_y - \omega_z x_s + \omega_x z_s \\ \dot{z}_s = -v_z - \omega_x y_s + \omega_y x_s \end{cases} \tag{14}$$

which relates the velocity of a 3-D point $\mathbf{X_s}$ to the sensor spatial velocity $\mathbf{v_s}$, (13) becomes

$$\dot{\rho} = v_x\frac{2x_s(\rho - 1) - d(\rho + 1)}{\ell_1^2} + v_y\frac{2y_s(\rho - 1)}{\ell_1^2} + \omega_z\frac{y_s d(\rho + 1)}{\ell_1^2}. \tag{15}$$

By analogy with (2), the interaction matrix $\mathbf{L}_\rho$ related to ρ can then be extracted from the previous equation as:

$$\mathbf{L}_\rho = \begin{bmatrix} \frac{2x_s(\rho-1) - d(\rho+1)}{\ell^2 + \frac{d^2}{4} - dx_s} & \frac{2y_s(\rho-1)}{\ell^2 + \frac{d^2}{4} - dx_s} & \frac{y_s d(\rho+1)}{\ell^2 + \frac{d^2}{4} - dx_s} \end{bmatrix}, \tag{16}$$

where ℓ is the distance between the sound source and the midpoint of the microphones. This matrix contains terms that are unknown, namely the source position x_s, y_s, and the distance ℓ. We then define an approximate interaction matrix given by

$$\widehat{\mathbf{L}_\rho} = \begin{bmatrix} \frac{2\widehat{x}_s(\rho-1) - d(\rho+1)}{\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s} & \frac{2\widehat{y}_s(\rho-1)}{\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s} & \frac{\widehat{y}_s d(\rho+1)}{\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s} \end{bmatrix}, \tag{17}$$

where the approximated parameters are designed below.

C. Approximating $\widehat{\mathbf{L}_\rho}$ parameters

From (9) and (10), the following relationship appears:

$$E_1\ell_1^2 = E_2\ell_2^2. \tag{18}$$

From (12), when $E_1 \neq E_2$, (18) can be rewritten as:

$$(x_s - c_x)^2 + y_s^2 - \frac{E_1E_2d^2}{(E_1 - E_2)^2} = 0 \tag{19}$$

where $c_x = \frac{d}{2}\frac{E_1+E_2}{E_1-E_2}$. This result states that the sound source is located on the circle $\mathscr{C}$ centred on the point $(c_x, 0)$ with a radius $c_r = d\left|\frac{\sqrt{E_1E_2}}{(E_1-E_2)}\right|$. Therefore from the assumption of the front plane working space (i.e. $y_s > 0$), we can set $\widehat{y}_s$ at any value between 0 and c_r, and deduce a corresponding $\widehat{x}_s$ by using (19), and then $\widehat{\ell} = \sqrt{\widehat{x}_s^2 + \widehat{y}_s^2}$. In the case when $E_1 = E_2$, with a similar manipulation as before, (18) reduces to:

$$2dx_s = 0, \tag{20}$$

that corresponds to the bisection of the microphone pair. Thus we can chose in this particular case $\widehat{y}_s > 0$ while $\widehat{x}_s = 0$. However, regarding the interaction matrix, the approximation of these parameters should be refined. Indeed $\mathbf{L}_\rho$ has elements of infinite value as soon as $\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s = 0$. Thus choosing $\widehat{y}_s$ and $\widehat{x}_s$ so that $\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s > 0$ is a good compromise for both cases $E_1 \neq E_2$ and $E_1 = E_2$. This condition is ensured if $\widehat{\ell} > \frac{d}{2}$. A thorough study can also be performed by analyzing the stability properties of such a system. This analysis is generally performed through Lyapunov stability theory, that guarantees the global asymptotically stability of the control scheme as soon as $\mathbf{L}_\rho\widehat{\mathbf{L}_\rho^+} > 0$ [11]. Without making explicit all the developments, the Lyapunov stability is guaranteed for our control scheme when

$$\begin{aligned} f(Z) =& Z^2(4y_s\widehat{y}_s + 4x_s\widehat{x_s}) - 2Zd(x_s + \widehat{x}_s) \\ &+ d^2(1 + y_s\widehat{y}_s) > 0 \end{aligned} \tag{21}$$

where $Z = \frac{\rho-1}{\rho+1}$ and $Z \in]-1, 1[$. One can prove that the quadratic function $f(Z) > 0$ when $sign(\widehat{x_s}) = sign(x_s)$ and $sign(\widehat{y_s}) = sign(y_s)$, which are then the stability conditions that are immediate to ensure in practice.

D. ILD-based positioning task

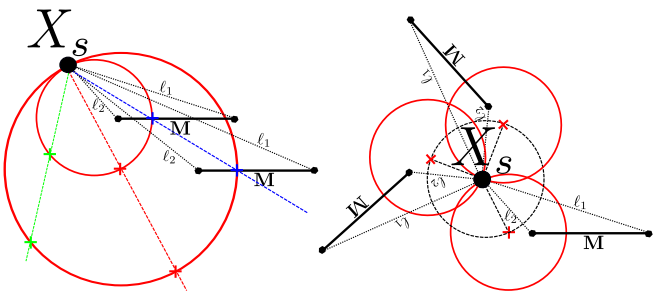
From the interaction matrix in (17), it is then possible to achieve positioning tasks with the control scheme (6). However it is necessary to analyze the behaviour of such a system for a given task. Virtual links [12] provide tools for the understanding of the system behaviour and the motions that should be expected from the control scheme. Relying on this approach, a vector subspace $\mathbf{S}^*$ that represents all motions for which the sensor output $\mathbf{s}$ remains constant is analyzed. This subspace is defined more explicitly as:

$$\mathbf{S}^* = \text{Ker } \mathbf{L_s}. \tag{22}$$

By application on the interaction matrix $\mathbf{L}_\rho$, we obtain:

$$\mathbf{S}^* = \begin{bmatrix} y_s d(1+\rho) & 2y_s(\rho-1) \\ 0 & d(\rho+1) + 2x_s(1-\rho) \\ d(\rho+1) + 2x_s(1-\rho) & 0 \end{bmatrix}. \tag{23}$$

By geometric construction, the motions induced by (23) can be obtained. The first column of this matrix refers to a circular motion around the sound source, with a correct orientation. Indeed, this motion constrains $\mathscr{C}$ to have its center at a constant distance from the sound source as illustrated by Fig.2b. The motion induced by the second column of $\mathbf{S}^*$, represents a translation directed by the circle $\mathscr{C}$. More exactly, the translation of the microphones implies that each point belonging to $\mathscr{C}$ is moved in the direction of the sound source. Therefore for each position with the same ratio ρ, we obtain circles $\mathscr{C}$ that are tangent to the actual position of the sound source (see Fig. 2a). Any linear combination of these two motions is possible which implies that infinite poses exist to complete a given task. These poses can be represented by concentric circles of radius c_r around the sound source. Hence, a typical task involving the use



(a) Translational motion (b) Rotational motion

Fig. 2: Admissible poses of the microphones for a given ρ

of the ILD mainly consists in orienting the robot towards a given direction, similarly to the ITD case [13].

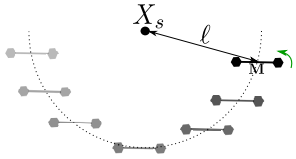
It should also be emphasized that despite the assumption of $y_s > 0$ in the modelling, the front/back ambiguity is inherently solved by our control scheme. Indeed if $y_s < 0$, the motion generated by the control input starts to increase the error $e = \rho - \rho^*$ since the robot moves in opposite direction to the desired configuration. Actually, the error increases until $y_s > 0$. As soon as $y_s > 0$, the situation corresponds to the correct modelling, where the stability is ensured under the trivial condition that $sign(\widehat{x}_s) = sign(x_s)$.

E. ILD accuracy limitation

From simulations conducted on the task of orienting the robot towards the sound source (see Fig. 3), it appears that the ILD-based control is limited to converge accurately in the close neighborhood of the sound source. First, the further the robot from the sound source, the poorer the spatial resolution and the accuracy of the ILD. Indeed, in this situation each ℓ_i becomes large in comparison with d. As a consequence, it comes out that $\ell_1^2 \to \ell_2^2$ since $\ell_2^2 = \ell_1^2 + 2dx_s$. Thus, the energy difference becomes too small and the dynamic of the ILD ρ is not significant anymore: large motions of the robot induced a small change in the ILD as confirmed by the simulations in anechoic conditions. Furthermore, as already stated in [7] or [9], the results exposed in Fig. 3 confirm that the ILD measurement is highly sensitive to reverberation. Systematic bias corrupts ρ when the microphones are far from the sound source, especially for positions close to walls. Indeed, with a reverberation modelling based on image source model [14], p virtual sources can be considered in the scene. Every virtual sound source j emits a sound wave so that each microphone $\mathbf{M_i}$ perceive an additional energy characterized by

$$E_p = \sum_{j=1}^{p} \frac{1}{\ell_{ij}^2} \int_{t=0}^{w} a^2(t - \frac{\Delta_{ij}}{c})\,\mathrm{d}t \tag{24}$$

where $\frac{\Delta_{ij}}{c}$ is the delay induced by the virtual source j. These virtual sound sources interfere with the signal received from the actual sound source. For the sake of simplicity, let us consider the upper bound of this additional energy, by assuming that (24) is a purely constructive sum, and that



$\mathbf{RT_{60}}$ (s)	ℓ (m)			
	0.5	**1**	**2**	**3**
0	< 1	< 1	2.7	6.4
0.05	< 1	3.6	5.6	7.6
0.1	< 1	5.1	7.0	8.4
0.2	< 1	8.8	10.4	21.4

Fig. 3: A simulated task that consists in orienting the robot in the direction of the sound source, from different poses in a 8×6 m^2 room. The final mean error, in degree, is calculated for several reverberation rate (RT_{60}) and distances to the source.

the source is very close to the reflective surfaces (i.e the virtual sources are not time-delayed w.r.t the actual source). In this "worst" case scenario, the measured ILD is modified as follow:

$$\rho = \frac{\frac{1}{\ell_1^2} + \sum_{j=1}^{p} \frac{1}{\ell_{1j}^2}}{\frac{1}{\ell_2^2} + \sum_{j=1}^{p} \frac{1}{\ell_{2j}^2}}. \tag{25}$$

This kind of erroneous measurement has a limited effect on the interaction matrix that is already approximated. Nonetheless, the accuracy of the convergence can be influenced by the error e that is equal to

$$e = \frac{\frac{1}{\ell_1^2} + e_{1p}}{\frac{1}{\ell_2^2} + e_{2p}} - \rho^*. \tag{26}$$

where $e_{1p} = \sum_{j=1}^{p} \frac{1}{\ell_{1j}^2}$ and $e_{2p} = \sum_{j=1}^{p} \frac{1}{\ell_{2j}^2}$. Therefore if e_{1p} and e_{2p} are significant enough w.r.t $\frac{1}{\ell_i^2}$, the error e of the control loop is biased. It can then be deduced that a sufficient condition to obtain accurate ILD measurement (resp. accurate error e) is a high Direct-to-Reverberant Ratio (DRR).

Hence with these limitations, the ILD-based positioning task can be achieved accurately only in the neighbourhood of the sound source. To overcome this limitation, we introduce the sound energy $E_{\mathbf{M}}$ as an additional feature. This feature can be used to control the distance to the sound source by setting a desired energy level. In this way, a positioning task can be performed accurately from an initial configuration relatively far from the source.

IV. Sound Energy Modelling

A. Energy estimation

From (8), the sound energy received by the point $\mathbf{M}$ is given by

$$E_{\mathbf{M}} = \frac{1}{\ell^2} \int_{t=0}^{w} a^2(t - \frac{\ell}{c})\,\mathrm{d}t. \tag{27}$$

It can be easily proved that the reverberation has less effect for the sound energy than for the ILD. Indeed, for a given task in the worst case discussed above, both $E_{\mathbf{M}}(t)$ and $E_{\mathbf{M}}^*$ include the error e_p caused by the wall reflections. Therefore the energy level error in the control loop is equal to

$$e_{\mathbf{M}} = \left(\frac{1}{\ell^2} + e_p\right) \int_{t=0}^{w} a^2(t - \frac{\ell}{c})\mathrm{d}t - \left(\frac{1}{\ell^{*2}} + e_p^*\right) \int_{t=0}^{w} a^2(t - \frac{\ell^*}{c})\mathrm{d}t \tag{28}$$

However, since the final pose is selected so that the sensors are close to the sound source (*i.e.*, high DRR), one can

deduce that $\frac{1}{\ell^{*2}} \gg e_p^*$. Moreover when the sensors are far from the source $\frac{1}{\ell^2} + e_p < \frac{1}{\ell^{*2}} - e_p^*$. Thus, while approaching to the final pose, we obtain:

$$e_{\mathbf{M}} \underset{s \to s^*}{=} \left(\frac{1}{\ell^2} - \frac{1}{\ell^{*2}} \right) \int_{t=0}^{w} a^2 (t - \frac{\ell}{c}) \, \mathrm{d}t. \tag{29}$$

Thus, the reverberation has a minor effect on this type of feature. From then on, since the energy level in the point $\mathbf{M}$ is not directly available, we just consider the mean value of the energy received by each microphones:

$$E_{\mathbf{M}} = \frac{E_1 + E_2}{2}. \tag{30}$$

B. Energy modelling

Let us consider now the distance from $\mathbf{M}$ to the sound source

$$\ell = \sqrt{x_s^2 + y_s^2}. \tag{31}$$

The time variation of this distance is defined by:

$$\dot{\ell} = \frac{x_s \dot{x}_s + y_s \dot{y}_s}{\ell}. \tag{32}$$

Similarly to the ILD, with the kinematic equation (14), the interaction matrix related to ℓ is easily obtained:

$$\mathbf{L}_\ell = \begin{bmatrix} -\frac{x_s}{\ell} & -\frac{y_s}{\ell} & 0 \end{bmatrix}. \tag{33}$$

Thus with the assumption of a constant signal energy during the time frame w, from (27) the time variation of $E_{\mathbf{M}}$ is

$$\frac{d}{dt} E_{\mathbf{M}} = -2 E_{\mathbf{M}} \frac{\dot{\ell}}{\ell}. \tag{34}$$

Therefore the interaction matrix related to the sound energy perceived in $\mathbf{M}$ is given by

$$\mathbf{L}_{E_{\mathbf{M}}} = -2 \frac{E_{\mathbf{M}}}{\ell} \mathbf{L}_\ell = E_{\mathbf{M}} \begin{bmatrix} \frac{2x_s}{\ell^2} & \frac{2y_s}{\ell^2} & 0 \end{bmatrix}. \tag{35}$$

Once again, an approximation of the latter interaction matrix is used:

$$\widehat{\mathbf{L}_{E_{\mathbf{M}}}} = E_{\mathbf{M}} \begin{bmatrix} \frac{2\widehat{x}_s}{\widehat{\ell}^2} & \frac{2\widehat{y}_s}{\widehat{\ell}^2} & 0 \end{bmatrix}. \tag{36}$$

$\widehat{x}_s$, $\widehat{y}_s$ and $\widehat{\ell}$ are estimated with the parameters given in Section III-D.

C. Energy level-based positioning task

For the interaction matrix $\mathbf{L}_{E_{\mathbf{M}}}$ related to energy level, the kernel that we obtain is:

$$\mathbf{S}^* = \begin{bmatrix} 0 & -y_s \\ 0 & x_s \\ 1 & 0 \end{bmatrix}. \tag{37}$$

The first motion implied by $\mathbf{S}^*$ is a pure rotation. Indeed the distance between $\mathbf{M}$ and the source is invariant w.r.t the orientation of the microphones. The second column illustrates all translations for which the distance to the sound source is unchanged. Namely, it refers to a circular motion around the sound source. Consequently by combining these two types of motion, $\mathbf{S}^*$ describes a circle of radius ℓ around the sound source with unconstrained orientation of the microphones. This result emphasizes that the ILD and the energy level are complementary features. Indeed, in contrast with the energy level, the ILD constrains the orientation of the microphones while the distance is constrained by the energy level.

V. Combining ILD to sound energy

The final interaction matrix $\widehat{\mathbf{L}_{\rho E}}$ that combines the ILD and the sound energy, is obtained by stacking (17) and (36):

$$\widehat{\mathbf{L}_{\rho E}} = \begin{bmatrix} \frac{2\widehat{x}_s(\rho-1) - d(\rho+1)}{\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s} & \frac{2\widehat{y}_s(\rho-1)}{\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s} & \frac{\widehat{y}_s d(\rho+1)}{\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s} \\ \frac{2E_{\mathbf{M}}\widehat{x}_s}{\widehat{\ell}^2} & \frac{2E_{\mathbf{M}}\widehat{y}_s}{\widehat{\ell}^2} & 0 \end{bmatrix}. \tag{38}$$

A. Positioning task

When the energy level is taken into account as it appears in (38), $\mathbf{S}^*$ becomes a rank-one matrix given by

$$\mathbf{S}^* = \begin{bmatrix} y_s \\ -x_s \\ 1 \end{bmatrix}. \tag{39}$$

In this case, the motion described by this subspace refers to a circular motion around the sound source at a distance ℓ with an adequate orientation of the microphones(see Fig.4).

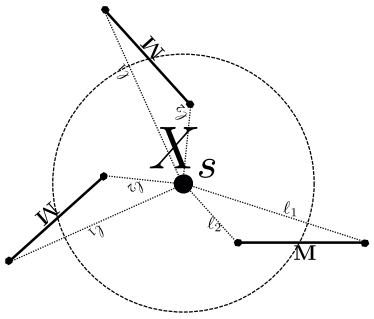


Fig. 4: Admissible poses of the microphones considering the sound energy level and the ILD as acoustic features

Consequently with a good selection of the desired features value, it is possible to reach a pose around the sound source by only considering the measurement of the energy and the ILD. When the robot is far from the sound source, the ILD measurement is just a rough approximation of the actual direction of the source. However as the robot is moving closer, the ILD measurement is refined, and the motion of the robot becomes more accurate. Hence, from rough and approximated features measurements, a virtuous cycle is created from the closed-loop control, that permits to achieve difficult tasks (see Section VI). Moreover, it should be emphasized that no complex signal processing, nor filtering are required for the control scheme. Additionally, the features measurements and the control scheme are independent from any tracking method, as long as the sound source of interest is predominant in the environment. Thus the computation cost of such a system is drastically decreased, when compared to classic SSL methods. Actually, the most complex calculation operation is the computation of the inverse of the Jacobian matrix $\mathbf{J_s} \in \mathbb{R}^{2\times 2}$.

B. Control scheme

For the 2-DOF robot described in Section II, the control input of the robot is

$$\dot{\mathbf{q}} = -\lambda \widehat{\mathbf{J}^{-1}_{\rho \mathbf{E}}} \mathbf{e} \tag{40}$$

with

$$\widehat{\mathbf{J}_{\rho \mathbf{E}}} = \begin{bmatrix} \frac{2\widehat{y}_s(\rho-1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} & \frac{D_x(2\widehat{x}_s(\rho-1)-d(\rho+1))+\widehat{y}_s d(\rho+1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} \\ \frac{2E_{\mathbf{M}}\widehat{y}_s}{\widehat{\ell}^2} & \frac{2D_x E_{\mathbf{M}}\widehat{x_s}}{\widehat{\ell}^2} \end{bmatrix}. \tag{41}$$

The determinant of this Jacobian matrix is:

$$|\widehat{\mathbf{J}_{\rho \mathbf{E}}}| = \frac{2E_{\mathbf{M}}\widehat{y}_s d(D_x - \widehat{y}_s)(1+\rho)}{\widehat{\ell}^2(\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s)}. \tag{42}$$

Apart from the degenerate cases already mentioned previously ($\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s = 0$ and $\widehat{\ell}^2 = 0$), singularities of the control system could appear as soon as $\widehat{y}_s = D_x$ or $\widehat{y}_s = 0$. In these particular situations, the determinant of the matrix would be equal to 0, and the interaction matrix would not be invertible. Hopefully, in practice the sound source position is approximated so that $\widehat{y}_s > D_x$.

VI. Experimental results

A. Experimental setup

The experiments were conducted on a Pioneer 3DX robot with two omnidirectional microphones as illustrated on Fig. 6. These microphones were connected to a sound card 8SoundsUSB [15] that processes the signal in real time. The sound card operates at a sampling frequency of 48 kHz, and provides frames of 256 samples. The sound energy is computed from 10 consecutive windows frames (*i.e.*, 50 ms). The global processing time of one iteration of the control scheme including the sound recording is around 60 ms. The experiments and acoustics conditions are detailed in the next sections. The parameters given in Fig. 6 were used for all experiments. An adaptive gain $\lambda(x)$ in which x refers to the infinity norm of the error $\mathbf{e}$ is used to smooth the robot motion. The accompanying video to this paper illustrates the experiments.

d	0.31 m
D_x	0.3 m
$\widehat{y}_s$	1 m
$\widehat{x}_s$	sign$(\rho-1)$ $\times 1$ m
$\lambda(x)$	$0.45e^{(-1.5x)}$

Fig. 6: Experimental settings

B. Typical positioning tasks

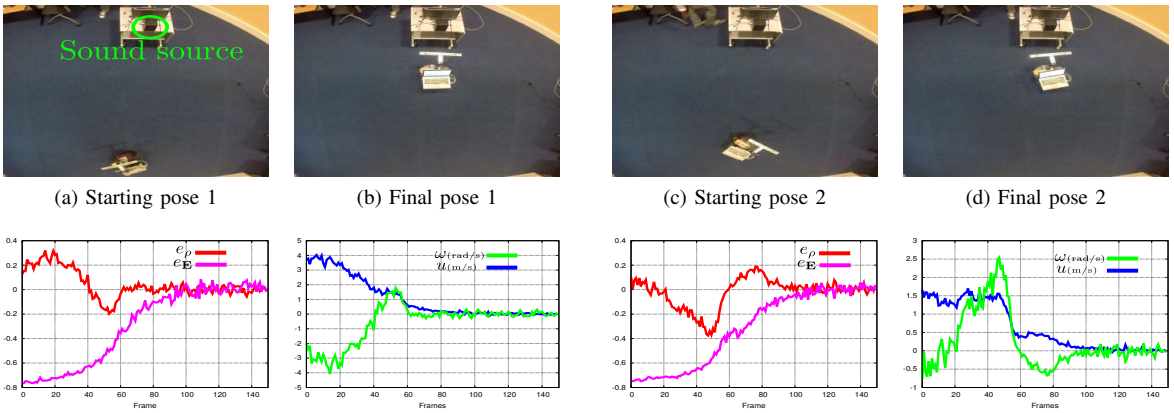
The first tests were conducted in a room characterized by a reverberation time $RT_{60} \approx 580$ ms. The sound source is a loudspeaker emitting a white Gaussian noise. The desired energy level is learned by placing the robot at 80 cm in front of the robot, while the desired ILD is set so that $\rho^* = 1$. The SNR at the desired pose is around 20 dB. The loudspeaker being directional, the admissible poses of the microphones were not in a circular configuration as stated in Fig. 4, but they were shaped by the sound radiation property of the speaker. First we showed the consistence of our approach with a static sound source, by starting the robot from different poses. As illustrated in Fig. 5, the control scheme allows to reach a pose satisfying the desired features from various initial poses[1]. Actually, as long as the difference of energy between the microphones is perceptible at the initial pose, the control scheme is able to position the robot in a desired configuration. A fine tuning of λ or w can also improve the sensitivity of the control scheme to small difference in the energy level.

In the second part of the experiment, we considered a moving sound source. As shown in Fig. 7, the robot is able to follow accurately the sound source, with exactly the same control scheme and without any knowledge about the motion of the sound source. As expected, when the robot is far from the sound source, it can be observed on Fig. 7e that the ILD is not always accurate. Indeed we can notice some abrupt changes in the ILD error curve. But as soon as the robot gets closer to the sound source, the ILD value is corrected, and the task can be correctly completed.

C. Robustness and flexibility in long range navigation

In the second experiment, we conducted a long range navigation test. Starting from the previous room, we moved the source through different environments of the laboratory with the robot pursuing the sound source in real time. Thus several acoustic conditions were encountered during the navigation as described in Fig. 8. The SNR varied from 20 dB to 13 dB while the reverberation rate changed from 580 ms to 880 ms depending on the location. The ambient noise was mainly caused by the ventilation systems and a server room in ⓓ. Nonetheless the robot never lost the track of the sound source despite dynamic and challenging conditions. Indeed the rooms and corridors crossed by the robot have different shapes, and are built with different materials and clutter. For instance the corridor ⓔ is narrow and produces strong first order echoes. Besides the echoes are not necessarily symmetric, because of the office doors that were open. In ⓕ different surrounding materials (glass door, metallic door of the elevator, walls...) were producing several types of echoes that could disturb the control scheme. Consequently, from this experiment it can be emphasized that the proposed control scheme is robust and flexible enough to deal with indoor real world environment. Indeed, since our method does not rely on any tracking or filtering of the signal, there is no tuning nor parameter dependent on the acoustic environment. Thus our approach is robust to environment changes.

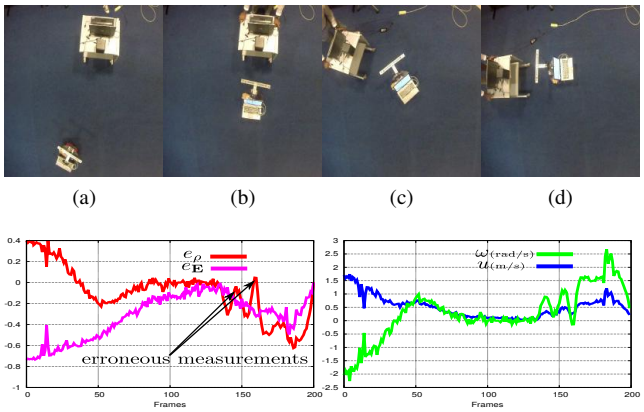
[1]For space reasons, the experimental results involving only the ILD, as well as the front/back ambiguity, are not discussed in the paper, but are presented in the accompanying video.



(a) Starting pose 1 (b) Final pose 1 (c) Starting pose 2 (d) Final pose 2

(e) Features error (f) Velocities input (g) Features Error (h) Velocities input

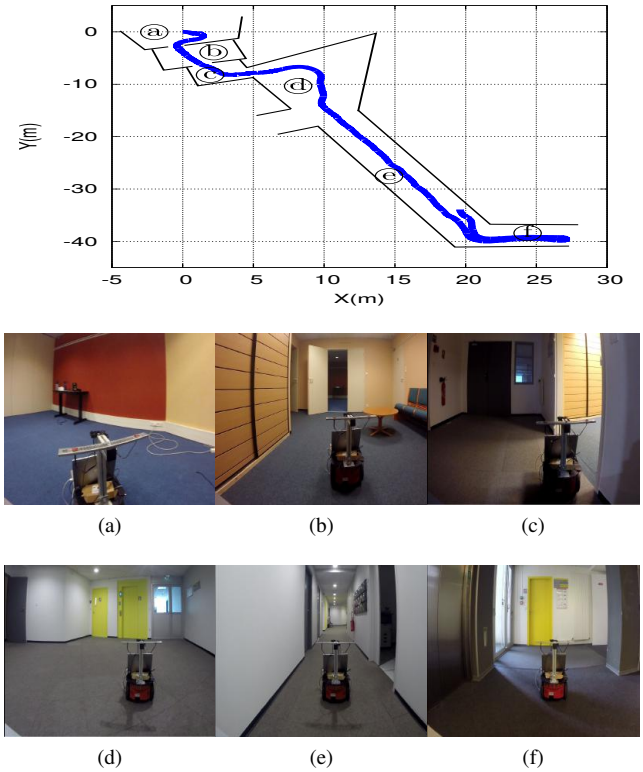
Fig. 5: Typical positioning tasks from two different starting poses



(a) (b) (c) (d)

(e) Features error (f) Velocities input

Fig. 7: Following a moving sound source



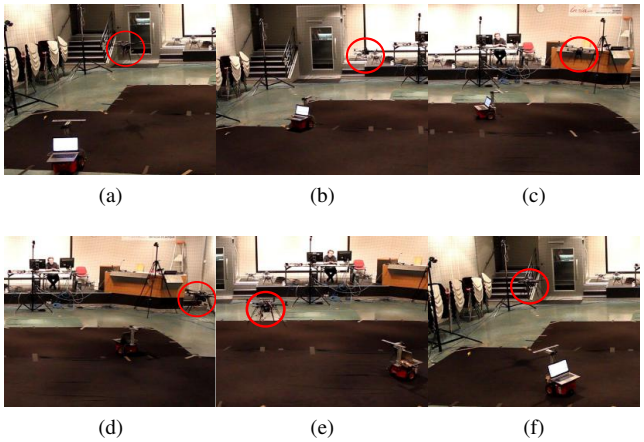
(a) (b) (c)

(d) (e) (f)

Fig. 8: Odometry data from the navigation task in indoor environment.The acoustic conditions for each location are respectively: ⓐ $RT_{60} \approx$ 580ms SNR$\approx$ 20dB, ⓑ $RT_{60} \approx$ 620ms SNR$\approx$ 20dB, ⓒ $RT_{60} \approx$ 680ms SNR$\approx$ 16dB, ⓓ $RT_{60} \approx$ 880ms SNR$\approx$ 13dB, ⓔ $RT_{60} \approx$ 700ms SNR$\approx$ 18dB, ⓕ $RT_{60} \approx$ 620ms SNR$\approx$ 17dB.

D. Cooperative application

The framework has also been tested in a cooperative task involving two robots. This time, instead of using the loudspeaker to generate the sound, we used the propellers of an unmanned aerial vehicle (UAV) as the sound source. Indeed most UAVs are known to be noisy. However, in a classic SSL scheme, this sound is considered more as a disturbance than a feature to exploit [16][17], while our approach takes advantage of this noise. In this experiment an UAV (mikrokopter MK-Quadro), remotely controlled, led the unicycle ground robot just by the sound naturally emitted from the propellers. Nonetheless it should be noted that the sound emitted by the UAV was not stationary nor omnidirectional. Indeed the UAV was oscillating during the flight and the sound was produced by the four propellers of the UAV. Nevertheless, despite these unfavourable conditions, the ground robot was able to follow the UAV, even if the motion of the robot was less smooth than in the previous experiments. This basic experimental scenario confirmed that our approach is relevant and suitable for cooperative tasks involving several robots. Furthermore, the control scheme based on only two microphones and the low computation cost of the framework make us believe that this kind of control scheme can be embedded on different type of robots. More evolved and complex tasks could then be achieved involving cooperation between aerial and ground robots or formation control of swarm robots.



(a) (b) (c)

(d) (e) (f)

Fig. 9: An example of cooperative application where an aerial robot (circled in red) is leading a ground robot with the sound emitted by its propellers

VII. CONCLUSION

In this paper, we proposed a framework to control robot motions with respect to a sound source in a binaural context. The control scheme is a sensor-based approach relying on the measured energy of the sound signal, without any localization of that sound source. Two auditory features are used, the interchannel level difference (ILD) and the signal energy. The ILD allows orienting the robot towards the sound source while the signal energy controls the distance between the robot and the sound source. By combining these two features the robot is able to follow a moving sound source, despite erroneous ILD measurements, as confirmed in the experiments. Furthermore, this sensor-based framework is directly performed on the raw measurements without tracking, filtering or signal enhancement. As a result, the computation cost of the control scheme is small, less than 10 ms in our experiments. The different experiments described in this paper emphasized the robustness of the control scheme towards reverberation, variability of the environment and front/back ambiguity. The different use cases showed the applicability of the method:

- A mobile robot navigated by pursuing a sound source through indoor environment while facing different level of reverberation and noise. This kind of tasks has potential applicability to a wide a range of field such as security patrolling or delivery services.
- An aerial robot has been able to guide a ground robot only with the sound of the propellers. In this case, applications in the field of multi-robot and cooperative tasks, or search and rescue mission can be emphasized.

To the best of our knowledge, no other prior work has been able to complete such tasks in real world conditions, while using only two microphones.

Globally this work shows the benefits of using control methods, besides acoustic and signal processing. This type of approach could help fulfilling the requirements for robots endowed with hearing sense that are: embeddability, real time processing, adaptivity to broad environment, robustness to noise and reverberation. For all these reasons, we do believe that the sensor-based approach is a fertile path that could open new horizons to robot audition.

Meanwhile, ongoing work concerns the extension of this approach to other type of signals. For now, one constraint of the framework is the use of continuous sound signal. It should be interesting to extend this method to intermittent sound sources such as speech. In the same vein, our approach could be extended to support multiple sound sources, by including a system of sound classification for instance. An application of this method on a head-mounted system is also intended in a close future. The promising preliminary results obtained without modelling the scattering effect of the head, makes us believe that our approach could be particularly suitable for humanoid robots.

ACKNOWLEDGMENT

The authors would like to thank Thomas Bellavoir for his technical assistance for the experiments involving the UAV.

REFERENCES

[1] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots-building audio-motor maps based on the hrtf," in *IROS'2006*, pp. 1170–1176.

[2] K. Youssef, S. Argentieri, and J-L. Zarader, "Towards a systematic study of binaural cues," in *IROS'2012*, pp. 1004–1009.

[3] I. Markovic, A. Portello, P. Danes, I. Petrovic, and S. Argentieri, "Active speaker localization with circular likelihoods and bootstrap filtering," in *IROS'2013*, pp. 2914–2920.

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

[5] G. Bustamante, P. Danes, T. Forgue, and A. Podlubne, "Towards information-based feedback control for binaural active localization," in *ICASSP'2016*, pp. 6325–6329.

[6] A. Magassouba, N. Bertin, and F. Chaumette, "First applications of sound-based control on a mobile robot equipped with two microphones," in *ICRA'2016*, pp. 2557–2562.

[7] S. Birchfield and R. Gangishetty, "Acoustic localization by interaural level difference," in *ICASSP'2005*. IEEE, 2005, vol. 4, pp. iv–1109.

[8] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ild and itd," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2010.

[9] W. Cui, Z. Cao, and J. Wei, "Dual-microphone source location method in 2-d space," in *ICASSP'2006*, vol. 4, pp. IV–IV.

[10] T. Rodemann, "A study on distance estimation in binaural sound localization," in *IROS'2010*, pp. 425–430.

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

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

[13] A. Magassouba, N. Bertin, and F. Chaumette, "Sound-based control with two microphones," in *IROS'2015*, pp. 5568–5573.

[14] J B Allen and D A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[15] D. Abran-Côté et al., "Usb synchronous multichannel audio acquisition system," Tech. Rep., 2014.

[16] M. Basiri, F. Schill, D. Floreano, and P. U Lima, "Audio-based localization for swarms of micro air vehicles," in *ICRA'2014*, pp. 4729–4734.

[17] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H. G Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor uav," in *IROS'2013*, pp. 3943–3948.