Binaural auditory interaction without HRTF for humanoid robots: A sensor-based control approach

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

I. Introduction

Until now in robot audition, especially for sound source localization, most of the contributions tackles the issues of robot auditory perception from acoustic, signal processing or physiological perspectives. Typically sound source localization is addressed in the same way as machine hearing, and the techniques developed are not exclusive to robotic context (*e.g* application on tablets). These techniques are attached to solve the question *Where is the sound source ?* that corresponds to a machine hearing problem in the broad sense. The control of the robot and the concept of giving a purpose to the robot are generally neglected. As a result coping with realistic (*i.e.*, dynamic) environments becomes challenging since no feedback from the environment is used.

Nonetheless, unlike other machines, robots are endowed with the motion ability and can move in accordance with a given purpose, such as interacting with the environment. Hence, the previous questioning about the source location can be transformed into *How to reach the sound source?* The latter way to consider the sound source localization problem can only be solved from a robotic perspective, since it requires a voluntary motion of the robot. In this paradigm, already exploited in [1], [3], [2], we thus define the sound source localization problem as a task to be completed by the robot. More explicitly, our approach consists in defining the motion of the robot with respect to the auditory feedback measured from the microphones. Typically this is the principle of the sensor-based control approach. In this short paper, we emphasize the versatility of such approach for robot audition by extending the paradigm already validated on mobile robots (*i.e* free-field microphones) to head-mounted systems.

II. Head-mounted system issues

The sound localization process is mainly based on two cues that are the interaural time difference (ITD) and Interaural level difference (ILD), that provide information about the direction of the source in the azimuth plane of a pair of microphones. The ITD is characterizes the different time for the sound to reach each microphone, while the ILD characterizes the attenuation of the sound energy for microphones at different distances of the source. However these cues are frequency-dependant and are particularly influenced by the scattering effect of the body (*i.e.*, head, pinna and torso) modelled by a non-linear function, the head-related transfer function (HRTF). Such influence characterizes the subjectivity of the sound perception, since each type of anthropomorphic robot or head-mounted system has a particular and unique HRTF. This property increases the level of complexity of sound localization in real conditions, since it requires a perfect modelling of these perturbations in order to retrieve a spatial information of the auditory cues, besides reverberation and noise. Formulating the localization in a sensor-based control framework can solve this problem.

[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

III. Overcoming HRTF constraint with sensor-based control

Let us assume a basic head-turn task in a planar scene. More exactly, the task consists in orienting a robot head towards the given direction of the sound source. In this scenario, the robot is endowed with two microphones, while the sound wave emitted by a source reaches the microphones with an angle α as illustrated in Figure 1. The task expressed in a sensor-based control approach consists in moving the head in the appropriate direction until the signal recorded from the left microphone matches the signal recorded from the right microphone. Having identical signals from both microphones logically implies that the robot is facing the sound source. One immediate advantage of such task modelling lies in the sound source localization that is not required anymore. Thus, since we are not attached to localize the source, our approach can be more robust to inaccurate modelling of the acoustic scene. Indeed in this paradigm the robot is not controlled by the estimated location of the source but rather steered by the auditory properties implied by the final pose of the robot (*i.e.*, similar measurements from both sensors) and the dynamic of the sound features with respect to the motion. These two characteristics are essential keypoints that emphasizes the robustness of the sensor-based approach over the classic sound localization. Actually, the dynamic of the sound features is an element that is independent from the HRTF. Furthermore, the symmetrical property of the robotic structure implies that the perturbation caused by the robot scattering effect on left and right sound recording should be the same when the robot is facing the sound source. As a result, the scattering effect of the head is nullified and the latter task can be completed without any modelling or knowledge of the HRTF.

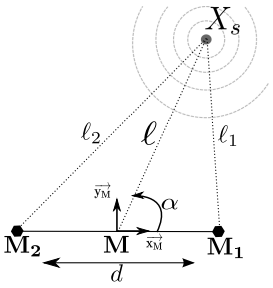

Fig. 1: Geometric configuration of the considered system, that includes a source $\mathbf{X_s}$ emitting a spherical and uniform sound wave, and a pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$.

IV. Experimental validation

We consider a robotic system instrumented with a pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ in a area without obstacle. The microphones are separated by a distance d as depicted in Fig. 1. An omni-directional sound source $\mathbf{X_s}$ is continuously emitting. For the following development $\mathbf{X_s}$ is shaped as a point that generates a sound wave uniformly in all directions. It assumes that the medium through which the sound travels is uniform. Furthermore, a frame $\mathcal{F}_m(\overrightarrow{x_M}, \overrightarrow{y_M})$ is attached to the midpoint of the microphones $\mathbf{M}$. Thus, in this frame, the Cartesian coordinates of each microphones are respectively $\mathbf{M_1}(\frac{d}{2}, 0)$ and $\mathbf{M_2}(-\frac{d}{2}, 0)$. 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 $\mathbf{M_i}$. In presence of one sound source only one degree-of-freedom of the robot can be controlled (see [1] and [2]), that is the rotation velocity ω. When controlling the robot through the ITD τ, the control input $\dot{q}$ of the robot is given by [1]

$$\dot{q} = -\lambda \frac{1}{\sqrt{\left(\frac{d}{c}\right)^2 - \tau^2}} (\tau - \tau^*) \tag{1}$$

where λ is a gain that tunes the time to convergence, c the sound celerity and τ^* the desired ITD value. In the task discussed above $\tau^* = 0$, since when facing the source, the sound should reach the two microphones at the same time.

The robot could be also be controlled from the ILD ρ that is a ratio of energy. In this case the control input is simply given by [2]

$$\dot{q} = -\lambda \frac{\ell^2 + \frac{d^2}{4} - d x_s}{y_s d(\rho + 1)} (\rho - \rho^*). \tag{2}$$

In the latter case since the position of the source is unknown, x_s and y_s should be approximated with $\hat{x}_s$ and $\hat{y}_s$. It is shown in [2], that it is enough to ensure that $sign(\hat{x}_s) = sign(x_s)$ and $sign(\hat{y}_s) = sign(y_s)$ to guarantee the stability of the controller. Furthermore the desired pose of the robot is characterized by $\rho^* = 1$, implying that the energy from the left and right microphone recording is the same.

The experiments are carried on the robot *Romeo*, on which we use two microphones that are embedded in the head, with pinna. No knowledge of the HRTF nor modelling of the scattering effect of the robot, are considered in the following results. As expected in both case, the robot is able

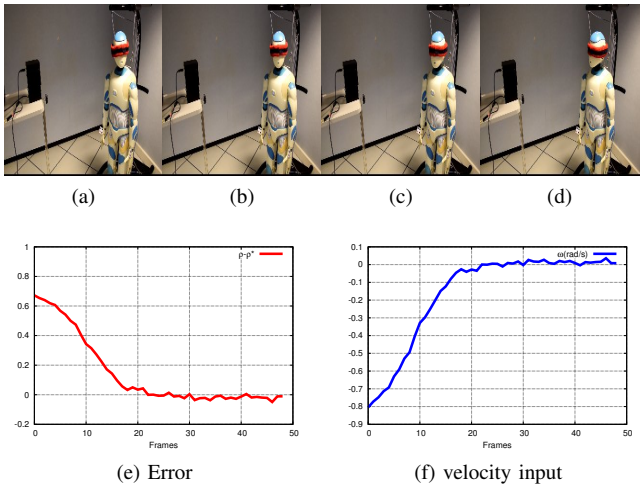

Fig. 2: Head turning task from initial pose in (a) to final pose in (d) using the ILD cue

to accurately orient its head towards the sound source. In the case of ILD exposed in Fig. 2, the considered sound source is a continuous white noise, while for the ITD-based control the source consists in a speech. In the latter case, we used two external microphones instead of the embedded one, in order to cope with the high level of noise in the robot. Nonetheless the perturbation caused by the HRTF is still present. Furthermore, the two solutions proposed can also cope with moving sound source (see the accompanying video), that still remains challenging for localization approaches in real world conditions.

V. Conclusions

This work emphasizes the benefits of using sensor-based control in order to interact through auditory perception in a head-turn task. The task is performed without any knowledge or modelling of the HRTF, which remains a requisite in classic sound localization approach. Exploiting the hearing sense in such scenario can be particularly interesting for human-robot interaction. For now, the proposed framework has been tested on simple configurations, while considering continuous sound sources. A path of improvement may consist in adding visual information, in a multi-modal control scenario, in order to cope with intermittent sound sources.

References

[1] A. Magassouba, N. Bertin, and F. Chaumette. Sound-based control with two microphones. In *IEEE International Conference on Intelligent Robots and Systems*, 2015.

[2] 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*, 2016.

[3] A. Magassouba, N. Bertin, and F. Chaumette. First applications of sound-based control on a mobile robot equipped with two microphones. In *IEEE International Conference on Robotics and Automation,*, pages 2557 – 2562. IEEE, 2016.