Multimodal Sound Localization for Humanoid Robots Based on Visio-Auditive Learning

Karim Youssef, Sylvain Argentieri, and Jean-Luc Zarader

Abstract— This paper deals with sound source localization in a humanoid robotics context. Classical binaural localization algorithms often rely on the following process: first, binaural cues are extracted from the left and right microphone/ear signals; next, a model is exploited to infer the possible localization of the sound source. Such a method thus requires an accurate modeling of the head acoustic shadowing, or precise Head-Related Transfer Function measurements. In order to avoid these last complicated steps, we propose in this paper an original multimodal sound source localization method. The relationship between binaural auditory cues and the position of the sound source to be located in an image is learned by a partially-connected neural network. This approach has a higher resolution and is less complex than state-of-the art techniques. Simulations and experimental results are shown, demonstrating the effectiveness of the proposed method. A very accurate azimuth estimation is provided, while elevation requires additive cues to be more efficiently approximated.

I. INTRODUCTION

Robot Audition has reached today an undeniable level of scientific maturity, thanks to various works in the last decade dealing with sound source localization, speaker or speech recognition, source extraction, etc. While artificial audition was first envisioned as a simple extension of signal processing and/or acoustic algorithms, it has brought to the fore specific and original robotics-related constraints as well as the new field of artificial binaural audition. Binaural audition consists in exploiting only two microphones in order to perform the aforementioned auditive tasks. In that sense, such a system is frequently considered as biomimetic, and is naturally well-suited to Humanoid Robots. But the use of only two microphones is shown to be extremely awkward when considering real acoustic environments, involving irrelevant sound sources, reverberations or noises coming from the robot itself. As a consequence, binaural audition still suffers from a lack of robustness. Considering the specific sound localization task, one of the first ideas dedicated to robotics was to elaborate closed-form models of the binaural sensor. One can cite [12] or [6], where the Auditory Epipolar Geometry (AEG) or the Scattering Theory (ST) are suggested to obtain the interaural cues which make the localization possible. Another solution consists in identifying the Head-Related Transfer Function (HRTF), which provides the impulse response linking the emitted signal to the two (left and right) perceived signals as a function of the source position. Inverting this HRTF thus allows the estimation of this position [8]. But all these

K. Youssef, S. Argentieri and J.-L. Zarader are with UPMC Univ. Paris 06 and ISIR (CNRS UMR 7222), F-75005, Paris, FRANCE `name@upmc.fr`

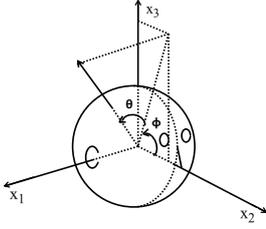solutions are not so robust against changes in the robot's environment, as shown in [12]. Nevertheless, very promising results have been already proposed in the field of binaural speaker recognition as we showed in [19] and [3], or binaural sound localization [17], [11], [14].

In comparison, Humans can perform impressive auditive tasks in very difficult acoustic conditions. Actually, each of us has learned his own HRTF and unconsciously extracts adaptive auditive characteristics from the two ears. But how is this learning performed? Among all human sensing capabilities, Vision is of particular interest, and recent works hypothesize some visually guided auditory adaptation processes for sighty subjects [1]. In the same way, [20] shows that the human auditive system may require vision to optimally calibrate elevation-related auditory cues. These work seem to demonstrate that audition requires other sensing capabilities to become more adaptive and robust. As a consequence, coupling vision and sound together with a learning-based approach might provide a good solution to obtain effective auditive capabilities. Recent contributions have been proposed in this field. One can cite [13], where the system can acquire sound localization abilities through repetition of movements and visual perception, or [10] where a method that mixes auditive and visual information for sound localization in reverberant environments is proposed.

This paper deals with binaural sound source localization. We propose here to combine a multimodal data acquisition, relying on binaural auditory cues and visual markers, with a learning of the relation between what the robot "hears" and "sees" by a neural network. This paper is divided into four sections. The first section is devoted to the presentation of the method, and to the evaluation of the performances in a supervised learning context. The second section presents the multimodal learning of the source position. Learning results are carefully described, in order to capture how vision and audition are related. Experimental results are introduced in a third section. A discussion of the method and ongoing works are then proposed in the fourth section. Finally, a conclusion ends the paper.

II. SUPERVISED LEARNING OF THE SOURCE POSITION

The proposed localization system is presented in this section. It mainly relies on a neural network that will be able to estimate the source position on the basis of interaural cues extracted from the two perceived signals. This position is expressed in function of both azimuth and elevation angles θ and ϕ respectively (see Figure 1). Contrarily to many studies that disregard one of the angles (mainly the elevation),

Fig. 1. Azimuth θ and elevation ϕ with respect to the head.

like [11], [16], [10] or [18], this work aims at estimating both angles at the same time. As a first step, in this section only, the proposed learning algorithm is directly supervised, as the two angles are supposed to be precisely known during the learning step. In the next section, they will not be required thanks to a multimodal approach.

A. Details of the proposed system

Like all learning-based approaches, the system requires a database which is here made of binaural auditory cues. The way they are extracted is first carefully described. Next, the functioning of the proposed neural network is presented. Finally, the database itself is depicted.

1) Auditory cues extraction: The auditory characteristics are obtained thanks to the following successive computation steps, see Figure 2. First, a sound source is selected. A whitened discrete gaussian noise with sampling frequency $f_s = 44.1$kHz is exploited in all the following, as its spectrum spreads over a wide frequency band. Next, the CIPIC HRTF database [2] is used to simulate the acoustic scattering induced by the head and leading to the left and right perceived signals (see II-A.3). $N_{\text{filter}} = 20$ gammatone filters, defined by Patterson et al. in [15], are then exploited to reproduce the cochlear frequency filtering (see Figure 3 (top)). Their central frequencies $f_c(i), i \in [1, N_{\text{filter}}]$ range from 100Hz to about $f_s/2$. Finally, the auditory cues are computed on the basis of the 2×20 resulting signals. We propose here to work with the following classical interaural cues, each of them being computed on $N = 1024$-point time windows according to the following equations:

- ILD (Interaural Level Difference):

$$\text{ILD}(f_c(i)) = 20 \log_{10} \frac{E_l(f_c(i))}{E_r(f_c(i))}, \qquad (1)$$

where $E_l(f_c(i))$ and $E_r(f_c(i))$ respectively represent the left and right cochlear filter output powers corresponding to the ith gammatone response centered at frequency $f_c(i), i \in [1, N_{\text{filter}}]$;
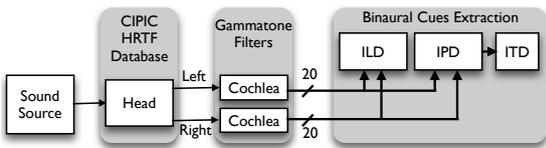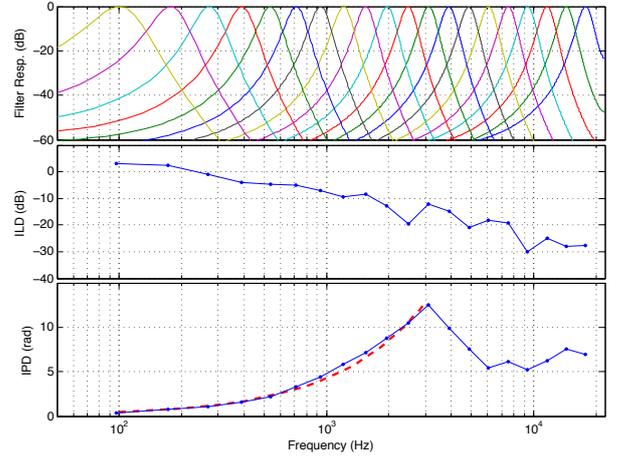

Fig. 2. Auditory cues extraction diagram.

Fig. 3. Auditory cues as a function of the frequency f, for a sound source in the median plane, and azimuth $\theta = 75°$. (Top) Gammatone filter responses in dB. (Middle) ILD value, computed at each cochlear center frequency; values for $f < 1.5$kHz are neglected. (Bottom) IPD values (solid line), computed at each cochlear center frequency; values for $f > 3$kHz become inconsistent. The ITD estimate, expressed in term of phase, is represented in red (dashed line).

- IPD (Interaural Phase Difference):

$$\text{IPD}(f_c(i)) = 2\pi f_c(i)\tau_{lr}(f_c(i)), \text{with} \qquad (2)$$
$$\tau_{lr}(f_c(i)) = k/f_s \text{ and } k = \arg_n \max(R_{lr}^{(i)}[n]),$$

where $R_{lr}^{(i)}[n] = \frac{1}{N}\sum_{m=0}^{N-n-1} l_i[m+n]r_i[m]$ is the biased estimate of the cross-correlation function between the two signals $l_i[n]$ and $r_i[n]$ originating from the ith left and right gammatone filters respectively;

- ITD (Interaural Time Difference):

$$\text{ITD} = \frac{1}{2\pi}\mathbf{f}^+ \, \text{IPD}(\mathbf{f}), \qquad (3)$$

where $(.)^+$ denotes the Moore-Penrose pseudo-inverse, $\mathbf{f} = (f_c(1), f_c(2), \ldots, f_c(N_{\text{filter}}))^T$ and $\text{IPD}(\mathbf{f}) = (\text{IPD}(f_c(1)), \ldots, \text{IPD}(f_c(N_{\text{filter}})))^T$. Consequently, the ITD value is actually obtained by a least square operation performed on the IPD.

So, each source position can now be described by a vector containing N_{filter} ILD values, N_{filter} IPD values and 1 ITD value. But it is known that ILD presents very small values for small frequencies related to large wavelengths. Consequently, the ILD is only computed for $f_c > 1.5$kHz. In the same vein, the index of the cross-correlation $R_{lr}^{(i)}[n]$ maximum becomes ambiguous for high frequencies and results in a false IPD estimation (see Figure 3 (bottom)). So, the IPD is only taken into account for $f_c < 3$kHz only. In this frequency band, the phase is almost linear, thus justifying the least-square method used to evaluate the ITD (see dashed line in Figure 3 (bottom)). Finally, 13 ILD, 12 IPD and 1 ITD values are involved in the vector being presented as input of the neural network.

2) Neural network characteristics and learning algorithm: In this section, a feed-forward neural network with a partial connections structure is employed. It has a 15-cell hidden
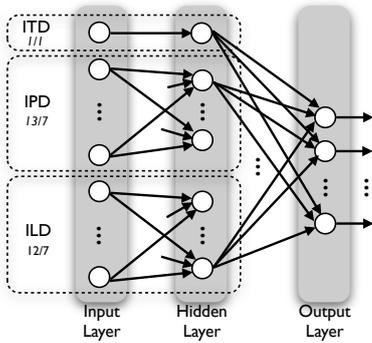
Fig. 4. Proposed neural network structure. The partial connections structure is outlined, with the *x/y* indication representing the input and hidden cells number respectively.
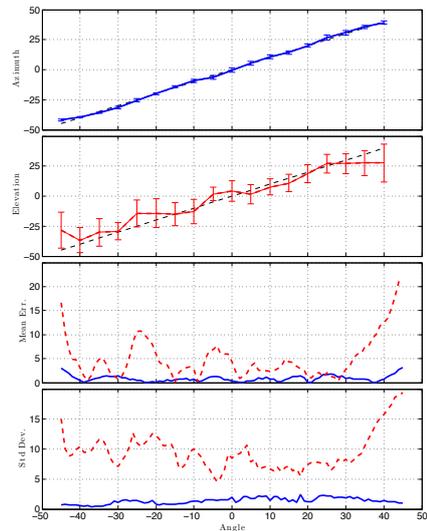

Fig. 5. Localization results. (First) Estimated azimuth as a function of the real azimuth. (Second) Estimated elevation as a function of the real elevation. (Third) Mean error between estimated and real azimuth (plain) and elevation (dashed). (Fourth) Mean standard deviation of the estimation for azimuth (plain) and elevation (dashed).

layer, as this number of hidden cells gave the best computational time-results performances, a 26-cell input layer and an output layer whose size will be adapted according to the scenario (see below). The partial connections structure is proposed here to cope with the heterogeneous input vector embedding ITD (s), IPD (rad) and ILD (dB). In such a structure, one hidden cell cannot be connected to two different kinds of inputs. As a consequence, only one hidden cell is connected to the ITD input, 7 are devoted to the 13 ILD inputs and 7 to the 12 IPDs (see Figure 4). On the contrary, the connections between the hidden and the output cells are complete, making them able to alter the output values properly. The iterative training of the neural network is based on a classical full gradient backpropagation algorithm, using the difference between the real network outputs (generated from the learning database) and the obtained ones to update the network weights. A cross-validation step is periodically performed to save the network characteristics producing the best results evaluated on a different database. The network training is stopped when it does not progress anymore, or when the learning error becomes smaller than a threshold.

3) Database and simulations: The CIPIC HRTF database provides the left and right impulse responses for various azimuths θ and elevations ϕ and a constant distance to the head. The azimuth is not uniformly sampled, but is more densely represented in the midsagittal plane, while the elevation ϕ is uniformly sampled, with $\phi(k) = -45 + 5.615k$ (in degrees) [2]. Spline interpolation is then used to obtain azimuth and/or elevation impulse responses that are not provided by the CIPIC database, resulting in an azimuth angle θ between $-80°$ and $80°$ with a $0.5°$-step, and an elevation angle ϕ ranging from $-45°$ to $45°$ with a $1°$-step. Consequently, 29211 known positions constitute the entire database, which is then split to form the learning (60%), cross-validation (20%) and test (20%) databases. Note that this high number is then reduced to about 8300 positions between $-45°$ and $45°$ with a $1°$-step for both θ and ϕ in order to guarantee the same learning conditions between azimuth and elevation.

B. Learning results and discussion

Our main objective in this subsection is to verify if a neural network is able to learn the relationship between some acoustic cues and the corresponding source position. Consequently, the network output is the source position expressed in terms of azimuth $\hat{\theta}$ and elevation $\hat{\phi}$. Two different approaches are imaginable: one neural network with the two outputs $\hat{\theta}$ and $\hat{\phi}$, or two neural networks with only one output each. These two strategies have been evaluated, and the second one exhibits better results. They are hereafter presented. Note that these estimations are obtained by presenting feature vectors from the testing database. For each vector, the predicted position is compared with the real one, resulting in a mean and standard deviation error measurements. They are both depicted in Figure 5. The proposed system shows very good azimuth estimation capabilities. Indeed, the mean error is as low as $1.37°$, and the mean standard deviation is only $1.13°$. On the contrary, elevation estimation is quite poor, with a $8.08°$ mean error and a $7.48°$ mean standard deviation. This indicates that the neural network is not able to efficiently learn the elevation, which is not a surprise, as the proposed interaural features do not really capture this elevation information [7]. This limitation will be discussed in §V. But in comparison with related works, this system exhibits a higher resolution and looks less complex. For instance, one can cite [4] where mean errors of $2.5°$ and $11°$ (azimuth and elevation respectively) are obtained. In the same vein, the binaural system proposed in [11], also relying on ITDs and ILDs, exploits one GMM per source azimuth and frequency bin, thus requiring a high number of GMMs and long training durations. Moreover, the azimuth resolution is of 5 degrees. This resolution is also present in [14], where the localization system relies on Probability Density Functions (PDFs) based on normalized histograms,
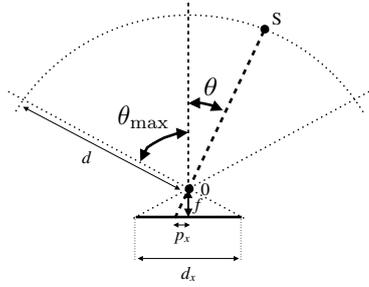
Fig. 6. Model of the proposed camera. S represents the sound source and 0 the head center.
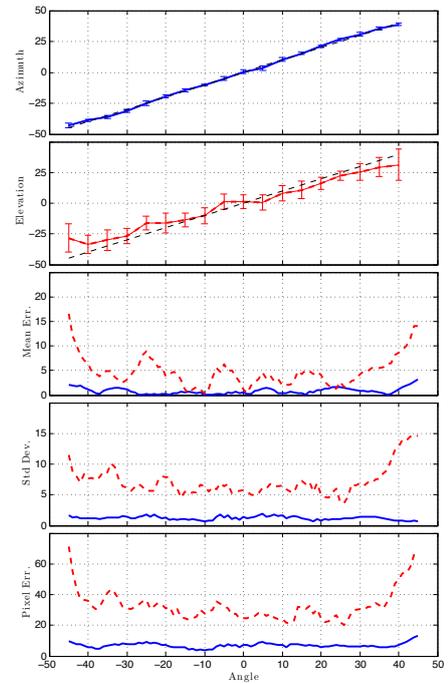

Fig. 7. Localization results. (First) Estimated azimuth (evaluated by inverting (4)) as a function of the real azimuth. (Second) Estimated elevation (evaluated by inverting (5)) as a function of the real elevation. (Third) Mean error between estimated and real azimuth (plain) and elevation (dashed). (Fourth) Mean standard deviation of the estimation for azimuth (plain) and elevation (dashed). (Fifth) Mean pixel error between estimated and real position of the three markers in the image for the x coordinate (plain) and the y coordinate (dashed).

with one PDF being computed for each interaural parameter, frequency and position. Another interesting result is shown in Figure 5. The azimuthal localization precision is at its best in front of the head. This result has also been reached in [16] where a parametric model computing ILDs and IPDs as a function of the azimuth is used.

These results demonstrate that a neural network is able to learn a HRTF, and is thus able to perform the sound source localization. The main problem of the proposed approach is the supervision of the learning step: one has to provide the expected source position. This can be of course achieved by some other ways. But does a robot really need to localize a sound source in terms of its angular position expressed in degrees? A human clearly does not compute such a position, and is more concerned with coincidence between two (or more!) modalities. This is exactly the idea of the forthcoming section, where the source position is learned with respect to its location in an image. Consequently, a sound localization error will now be expressed in terms of a pixel error. Of course, the link between this error and the (more classical) angular error will be investigated.

III. MULTIMODAL LEARNING OF THE SOURCE POSITION

In this section, an original multimodal sound source localization scheme is assessed. It uses both audition and vision. Classically, such a multimodal approach is conceived in such a way that one of the two modalities comes to validate or reinforce the localization obtained with the other [5], [10]. But unlike these studies using 3 or 4 microphones, our multimodal approach mimics the human perception using only both ears signals and vision mutual information to learn the relationship between them. This idea is first tested in simulation with a virtual camera placed at the center of the head. The details of the simulation parameters are described in a first subsection. Simulation results are discussed in a second subsection. The proposed method will then be evaluated with real data in section §IV.

A. Proposed methodology

For the moment, the interaural cues database remains exactly the same. The only difference is the neural network itself, which will now have to learn the source position expressed in pixels. To perform such an operation, a simple camera (see Figure 6) is simulated at the head center. In order to obtain a basic visual representation of the position, three punctual markers, whose positions depend on the azimuth θ and elevation ϕ, are placed around the simulated sound source. These three markers are then projected in the image plane to obtain their horizontal and vertical positions pix_x and pix_y in pixels, with:

$$\text{pix}_x = \text{round}\left(\frac{X_{res}}{2} + X_{res}\frac{f}{d_x}\tan(\theta)\right), \quad (4)$$

$$\text{pix}_y = \text{round}\left(\frac{Y_{res}}{2} + Y_{res}\frac{f}{d_y}\tan(\phi)\right), \quad (5)$$

where $X_{res} = 640$ and $Y_{res} = 480$ represent the image resolution in pixels, f is the focal length of the simulated camera, and d_x, d_y are respectively the horizontal and vertical sizes of the image plane. These three last parameters have been adjusted in order to adapt the field of view to the maximal azimuth $\theta_{\max}$ and elevation $\phi_{\max}$ provided by the CIPIC HRTF database.

B. Learning results

So, a sound source position is now represented by three points in a simulated image. We propose here to use three identical neural networks, each of them learning the position of one point in the image on the basis of the aforementioned binaural cues database. During the test step, the three networks will provide an estimation of the three points coordinates in the image. The estimation error can now be evaluated not only in terms of pixel errors, but also in degrees by inverting (4) and (5). Localization results are reported in
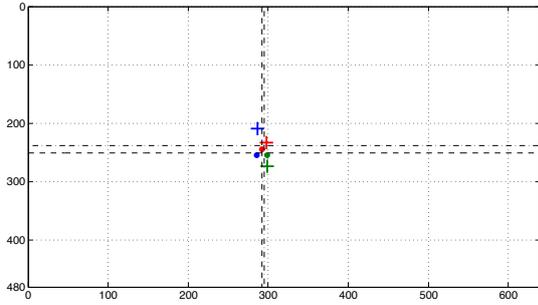
Fig. 8. Comparison between actual and estimated patches positions. The three actual patches are represented by three circular markers, while the three estimated positions are symbolized by three crosses. Dashed lines (resp. dotted lines) represent the mean actual horizontal and vertical positions (resp. estimated horizontal and vertical positions).

Figure 7. Azimuth estimation remains very accurate when working with images. Indeed, the mean error is only $0.82°$, for a $1.22°$ mean standard deviation. In comparison with the previous section, the mean error is even smaller, while the standard deviation remains quite constant. The same applies for the elevation estimation, which is still less accurate with a $8.4°$ mean error and a $7.15°$ mean standard deviation. These results clearly demonstrate that the proposed system is able to make an efficient projection of the auditive perception into the visual one. But for the moment, only the auditive part of the algorithm has been discussed. Figure 7 (bottom) shows the mean error in pixels, obtained by computing the difference between the actual pixel positions and their estimated values provided by the neural networks. It clearly appears that the vertical pixel position estimation –directly related to the elevation– is imperfect, with a mean error reaching up to 70 pixels. On the contrary, the horizontal pixel position estimation exhibits a small mean error of only 7 pixels. This fact is illustrated in Figure 8, where the three actual patches are represented by three circular markers, while the three estimated positions computed by the three neural networks are symbolized by three crosses. Clearly, a good vertical positioning of the 3 estimated points can not be guaranteed. Finally, note that the angular and pixel errors are highly correlated, allowing the evaluation of the localization precision without the ground true azimuth and elevation values. Importantly, and from an experimental point of view, it could prevent the precise measurement of these two angles.

IV. EXPERIMENTS

In this section, real binaural signals acquired with a dummy head are exploited, together with the image flow originating from a camera placed on top of the head, within the preceding framework. The experimental setup and the binaural/visual database is outlined in the first subsection. The resulting signals are then exploited to perform the offline learning of the neural networks. The resulting localization estimations are then discussed in the second subsection.
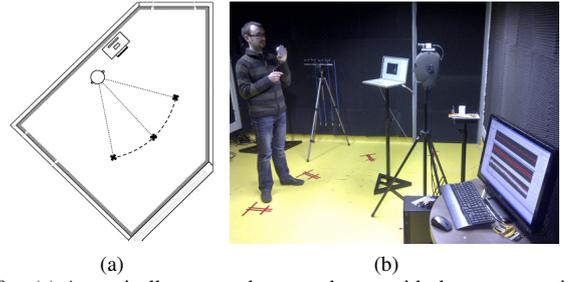

Fig. 9. (a) Acoustically-prepared room scheme, with the representation of the camera field of view (approx. $60°$ of azimuth). (b) Experimental setup during an acquisition.

A. Experimental setup

All the forthcoming experiments take place in an acoustically prepared room (see Figure 9). A KU100 dummy head from Neumann, with two microphone capsules built into two ears imitating the human pinnae, is used to record binaural signals. The two microphone outputs are synchronously acquired by a National Instruments PCI acquisition card through 24 bits delta-sigma converters operating at the sampling frequency $f_s = 44100$Hz. A camera is placed on top of the head, delivering 44 images per second with a resolution of 640×480. This frame rate corresponds to time windows lasting about 23ms. Note that the camera is not calibrated, as no projection in the $3D$ space is required. Next, a small portable loudspeaker with a frequency response ranging from 200Hz to 16kHz is used to emit a white noise. 3 colored patches are sticked in front of it so as to ease the vision algorithm. It only consists in finding the loudspeaker form, and then in detecting the three center points of the patches previously identified, see Figure 10. The three points coordinates are next extracted and used during the learning step as the actual coordinates to be associated with the auditive perception. Finally, a person presents the loudspeaker in front of the camera during an experiment, and moves in the left, right, or vertical directions at an almost constant distance to the head. This constitutes the binaural audio-visual database used in the following (about 7000 examples).

B. Localization results

In the following, only the pixel coordinates of the three loudspeaker markers are taken into account. The estimation
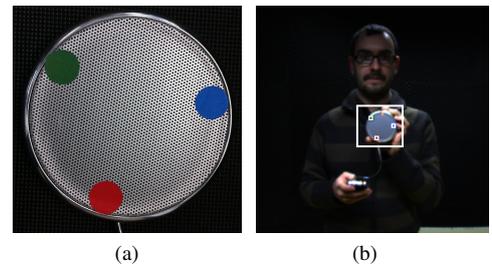

Fig. 10. (a) Used loudspeaker, with the three colored markers. (b) Image actually viewed by the camera, with the three markers correctly detected (cropped view of the real image).
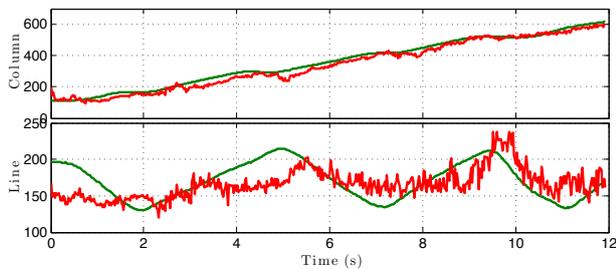
Fig. 11. Experimental results. Horizontal and vertical actual (resp. estimated) pixel positions as a function of time in green (resp. in red).

of the pixel coordinates is highly correlated with the angle values. The approach is first tested with a sound source moving up and down while passing from the left to the right side of the image. This allows the evaluation of the azimuth and elevation estimation performances together. Results are reported in Figure 11. A small error on the horizontal axis is obtained, with an about 20-pixel mean error associated to a 10-pixel standard deviation, whereas a much higher error exists in the case of the vertical axis, with an about 40-pixel mean error associated to a 50-pixel standard deviation. So, experimental conclusions are quite identical to those obtained during simulations: the azimuth is well determined while the elevation's estimation is much less effective.

V. Discussion and ongoing work

We have presented here a multimodal approach of the sound source localization problem. The proposed method exhibits promising results. Therefore, we will be working on the following points in the forthcoming months. First, the ability of the system to precisely estimate the elevation is relatively poor, which is related to the nature of the used binaural cues. We will try to add additional cues to the feature vector, like the spectral cues helping to perform vertical localization [7], [18]. Another ongoing work deals with the ability of the neural network to generalize its learning to human speech, substituting the loudspeaker with a person talking to the robot, in various acoustic constraints, including reverberations and additive noisy and unwanted sounds. Finally, an active scheme, relying on the head movement, will be required in order to cope with non-visible sound sources to the camera. A first approach could consist in learning the relationship between the movement and the perception, like in [13] or [9].

VI. Conclusion

A multimodal sound localization system, based on binaural acoustic cues and visual information has been presented. It is based on a partially-connected neural network that learns how a sound source, represented by three markers, is projected into an image. The system exhibits very good azimuth estimation, allowing to determine where the sound source is in the image. Vertical localization is less precise, as the used binaural cues do not efficiently capture the elevation information. A discussion on ongoing works is finally proposed, where short-term guidelines are outlined so as to improve and generalize the approach.

References

[1] S. M. Abel and C. Tikuisis. Sound localization with monocular vision. *Applied Acoustics*, 66(8):932–944, August 2005.

[2] V. Algazi, R. Duda, R. Morrisson, and D. Thompson. The cipic hrtf database. *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to audio and Acoustics*, pages pp. 99–102, 2001.

[3] B. Breteau, S. Argentieri, J.-L. Zarader, Z. Wang, and K. Youssef. Binaural speaker recognition for humanoid robots. *IEEE International Conference on Robotics and Biomimetics, ROBIO*, 2010.

[4] H. Finger, P. Ruvolo, S.-C. Liu, and J. R. Movellan. Approaches and databases for online calibration of binaural sound localization for robotic heads. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[5] S. S. Ge, A. Loh, and F. Guan. 3d sound localization based on audio and video. *International Conference on Control and Automation*, 2003.

[6] A. A. Handzel and P. S. Krishnaprasad. Biomimetic sound-source localization. *IEEE Sensors Journal*, 2:607–616, 2002.

[7] W. M. Hartmann. How we localize sound. *Physics Today*, 52(11):24–29, 1999.

[8] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda. Sound localization for humanoid robots - building audio-motor maps based on the hrtf. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2006.

[9] A. Laflaquière, S. Argentieri, B. Gas, and E. Castillo-Castenada. Space dimension perception from the multimodal sensorimotor flow of a naive robotic agent. In *IEEE International Conference on Intelligent Robots and Systems*, pages 1520 – 1525, 2010.

[10] B.-g. Lee, J. Choi, D. Kim, and M. Kim. Sound source localization in reverberant environment using visual information. In *IEEE International Conference on Intellligent Robots and Systems*, pages 3542 – 3547, October 2010.

[11] T. May, S. van de Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on Audio, Speech and Language Processing*, 19(1), 2011.

[12] K. Nakadai, H. Okuno, and H. Kitano. Epipolar geometry based sound localization and extraction for humanoid audition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 1395–1401, Nov. 2001.

[13] H. Nakashima and T. Mukai. 3d sound source localization system based on learning of binaural hearing. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3534 – 3539, January 2005.

[14] J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *Journal of the Acoustic Society of America*, 119(1), 2006.

[15] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. In *International Symposium on Hearing, Auditory physiology and perception*, pages 429–446, 1992.

[16] M. Raspaud, H. Viste, and G. Evangelista. Binaural source localization by joint estimation of ild and itd. *IEEE Transactions on Audio, Speech and Language Processing*, 18(1), 2010.

[17] T. Rodemann, G. Ince, F. Joublin, and C. Goerick. Using binaural and spectral cues for azimuth and elevation localization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, September 2008.

[18] T. Shimoda, T. Nakashima, M. Kumon, R. Kohzawa, I. Mizumoto, and Z. Iwai. Spectral cues for robust sound localization with pinnae. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[19] K. Youssef, S. Argentieri, and J.-L. Zarader. From monaural to binaural speaker recognition for humanoid robots. In *IEEE-RAS International Conference on Humanoid Robots*, pages 580 – 586, Dec. 2010.

[20] M. Zwiers, A. Van Opstal, and J. Cruysberg. A spatial hearing deficit in early-blind humans. *Journal of Neuroscience*, 21:RC142(1-5), 2001.