A BINAURAL SOUND SOURCE LOCALIZATION METHOD USING AUDITIVE CUES AND VISION

Karim Youssef, Sylvain Argentieri, and Jean-Luc Zarader

Abstract— **A fundamental task for a robotic audition system is sound source localization. This paper addresses the localization problem in a robotic humanoid context, providing a novel learning algorithm using binaural auditive cues to determine the sound source's position. Sound signals are extracted from a humanoid robot's ears. Binaural auditory cues are then computed to provide inputs for a neural network. The neural network uses pixel coordinates of a sound source in a camera image as outputs. This learning approach provides good localization performances as it reaches very small mean errors for azimuth and elevation angles estimates.**

Keywords — Binaural audition, Sound processing, Localization, vision.

I. INTRODUCTION

Robots and intelligent systems are becoming more and more reliable as partners in the humans' everyday life. Nowadays, it has become possible to envision machines in social interaction. One of the most important parts of social interaction is speech. Indeed, a sound signal holds various information: sound sources identities, their spatial locations and the contents of the emitted sounds. This brings to the fore multiple problems that a machine audition system has to deal with: Voice Activity Detection (VAD), speaker and speech recognition, and sound source separation and localization.

Sound source localization has been widely studied in the last few decades. Most of the previously built systems use microphone arrays together with techniques like beamforming [14]. But microphone-array based systems are often computationally expensive, which makes it important to use less complex methods. In this context, binaural audition has emerged as an interesting low-complexity and biologically-inspired sound processing domain. It is based on the use of the signals captured by only two microphones to reach human-like auditive capabilities. Indeed, binaural processing has been used in multiple applications, like sound source localization [12], [4]. Binaurality also provides good performances for speech enhancement [2], and voice activity detection [2]. We have also proposed in [15] a binaural speaker recognition system, which has been shown quite sensitive to the speaker position with respect to the robot's head.

Nevertheless, the impressive human auditive capabilities are not reached thanks to the two ears only. Vision plays a very important role in a scene analysis and some recent works hypothesize some visually guided auditory adaptation processes for seer people [6].

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

Studies that try to model the human head and to link the auditive cues to its geometry and to the sound source direction have been proposed before [8], [4] and [5]. Thus, an inversion the measured cues at a time instant, based on the built models permits to deduce the sound source position. But when such systems are used in experimental conditions, the models fail and do not comply with reality. On the contrary, learning approaches might be better adapted and more robust to such problems.

In this context, this paper presents a novel sound source localization system. It provides a new way of coupling vision and sound, in a learning based approach that provides effective localization capabilities. The system learns the relationship between the visualized positions of the sound source and the auditory cues extracted from two ears. So vision only provides a tool to represent the sound source's position in the scene, and the localization, expressed in pixel positions inside an image, relies on the information provided by the auditive cues. The approach is less complex and has a better spatial resolution than related works [11], [7] and [9]. For now, the work disregards the hypothesis of multiple sound sources, including noises and reverberations. The feasibility of the approach is tested and discussed in this paper, for a generalization to more realistic environments in following works. The proposed approach is based on a learning algorithm using a neural network. Contrarily to many studies that only address the azimuth estimation like [7], [11], this work aims at estimating both azimuth and elevation at the same time.

The paper is organized as follows: the azimuth and elevation estimation methods are presented in the next section. Simulation and experimental tests results are presented in Section III. The results are discussed in Section IV, Finally, a conclusion ends the paper.

II. AZIMUTH AND ELEVATION ESTIMATION

A. Azimuth estimation

For the azimuth estimation, the inputs of the aforementioned network are code vectors composed of Interaural Level Differences (ILD), Interaural Phase Differences (IPD) and Interaural Time Difference (ITD). These cues are precisely described in a first subsection. The neural network itself, together with the learning algorithm is depicted in a second subsection. Finally, the outputs of the networks are introduced in the last subsection.

1) Network inputs: auditory cues extraction: As it can be seen in Figure 1, signals from both robot ears are exploited to compute the interaural auditory cues. The human cochlear

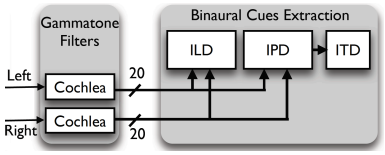

Fig. 1. Auditory cues extraction diagram.

filtering is artificially reproduced by a set of 20 gammatone filters defined in [10]. Their central frequencies $f_c(i)$ range from 100Hz to about half of the sampling frequency $f_s = 44100$Hz. This process leads to 20 signals per ear, the interaural cues being then extracted from these 20 signals through the following methodology.

2) Interaural Level Difference: The ILD is a frequency-dependent cue that reflects the difference in powers of the signals reaching the two ears. An ILD for each gammatone filter's frequency range can be extracted according to:

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

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 i^{th} gammatone response centered at frequency $f_c(i), i \in [1, 20]$.

3) Interaural Phase Difference: IPD refers to the difference in the phases of waves reaching the ears. It is obtained with:

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

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 i^{th} left and right gammatone filters respectively.

4) Interaural Time Difference: ITD reflects the difference between the lengths of the paths to be traveled by the sound wave before reaching the ears. It is computed by:

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

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 obtained by a least square operation performed on the IPD.

5) Code vector constitution: ILDs and IPDs are known to be not meaningful for low and high frequencies respectively. ILD is computed for frequencies higher than 1.5kHz, while IPD is taken into account for frequencies lower than 3kHz. These respective frequency intervals are considered, as the respective cues outside them don't carry much information. So, the final neural network's input code vectors are composed of 13 ILDs, 12 IPDs and a single ITD value, which makes a total input dimension of 26.

6) Network constitution and learning algorithm: The neural network used in this study is a feed-forward multi-layer perceptron (MLP) with one hidden layer composed of 15 cells. Since the input code vectors contain data of different types (amplitudes, phases and times), a regular complete connections neural network –i.e. a network where each hidden cell is connected to all input cells– is not physically adapted to these inputs. A hidden cell should not be connected to two inputs of two different types. Therefore, one hidden cell is dedicated to the ITD, 7 are dedicated to the ILDs and 6 to the IPDs. And the connections between the hidden cells and the outputs are kept unmodified.

The training of the neural network is performed with the full gradient backpropagation algorithm. Cross-validation steps are performed periodically, and the training is stopped when the performances of the network stop improving.

7) Network outputs: sound source representation: In this paper, the proposed sound source is a loudspeaker carrying three colored markers. A camera mounted on the robot's head takes movies of the moving sound source and an image processing system analyses the captured images and evaluates the line and column indices of each marker's center in the image. For the azimuth estimation, the outputs of the network are only the three column indices of the three markers.

B. elevation estimation

Tests performed with interaural cues on elevation estimation do not show satisfactory results. Indeed, interaural cues contain powerful information about the azimuth, and very weak information about the elevation [13].

The human pinna shape is at the origin of interferences with the waves directly entering the auditory canal, causing constructive and/or destructive reflections at specific frequencies depending of the sound source location. This phenomenon produces spectral peaks and notches which are supposed to be used by humans when evaluating the elevation of a sound source [3]. In this field, one can cite [13], where is presented a method using spectral cues for the elevation estimation with a robot having two logarithmic-shaped reflectors as pinnas. We propose here to compute for each source position the energies coming from the 2 cochlear filter-banks. These energies are expected to capture the aforementioned reflections translated by high and low energies in specific spectral areas, and thus to better the elevation estimation performances of the proposed approach. Only one regular neural network is now used to estimate the three line coordinates of the three markers in a simulated image. The input of this network is now made of 40 energy values corresponding to the 2×20 gammatone filters, and the outputs are the three markers' line coordinates.

III. Simulations and experiments

In order to evaluate the proposed approach, simulated and experimental databases have been elaborated. In both cases, the sound source emits a white discrete Gaussian noise (useful here as its spectrum spreads over a wide frequency

band). The cues are extracted on the basis of 1024-points time windows lasting 23ms with a sampling frequency of $f_s = 44.1$kHz.

In both cases, the learning of the neural networks is done with 60% of the total amount of data, the cross-validation uses 20% and the remaining 20% are used for testing. In the testing phase, the networks provide an estimation of the outputs (line and column indices of the three markers) based on the perceived auditory inputs. The mean Euclidean distance between the estimated outputs and the real ones is defined as the network's mean estimation error.

A. Simulations

This subsection presents an artificially generated database and the resulting localization performances. An artificial robot is placed in an environment where a sound source is moving. For each source position, the left and right ear signals are computed and a virtual camera placed on the robot's head detects the source position in an image.

1) Database generation: The simulated input database is generated from the noise signal coming from the source, convolved with impulse responses known as Head Related Impulse Responses - HRIRs for different spatial positions.This allows to obtain the left and right signals, with a sound source located in the HRIR specified position. The CIPIC database provides these left and right impulse responses for various azimuths and elevations [1].

To obtain the outputs, a camera model has been simulated. It projects the three markers in the image to obtain their horizontal and vertical positions pix_x and pix_y in pixels, based on the loudspeaker's center's given position. In the testing phase, the network provides estimations of the line and column indices, $\widehat{\text{pix}}_y$ and $\widehat{\text{pix}}_x$ respectively. An inverse of the camera model gives then the corresponding estimated angles $\widehat{\phi}$ and $\widehat{\theta}$, the estimation errors being then defined as $\epsilon_\phi = |\phi - \widehat{\phi}|$ and $\epsilon_\theta = |\theta - \widehat{\theta}|$ respectively.

2) Localization results: During the learning and testing steps, the database is restricted to angles between $-45°$ and $45°$ with a $1°$ step for both azimuth an elevation for a total number of 8281 examples. This allows to have the same resolution for both angles and to efficiently compare their relative results. Recall that after the training step, the neural network is able to produce an estimation of the line and column indices $\widehat{\text{pix}}_y$ and $\widehat{\text{pix}}_x$, which are then expressed in terms of the two angles $\widehat{\phi}$ and $\widehat{\theta}$. The resulting estimations are shown in Figure 2. As expected, they show a high accuracy in the azimuth and elevation estimation, having mean errors of only $0.82°$ and $2.06°$, and mean standard deviations of $1.22°$ and $1.69°$ respectively.

B. Experiments

This subsection presents experimental results obtained with real binaural signals recorded by using a dummy head and images provided by a camera mounted on top of it.

1) Database: The experimental database has been recorded in an acoustically prepared room A KU100 dummy head from Neumann is employed. It has two microphone

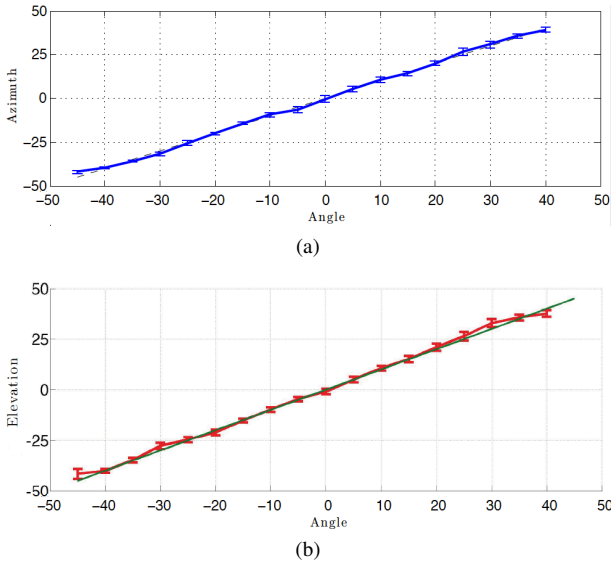

Fig. 2. Simulation: estimation results, predicted angles as a function of the real angles. (a) azimuth angle, (b) elevation angle.

capsules built inside two human-like ears, thus reproducing the effects of the human head and outer ears on a sound signal, before reaching the inner ear. The two microphone outputs are synchronously acquired by a National Instruments PCI acquisition board through 24 bits delta-sigma converters operating at a sampling frequency $f_s = 44.1$kHz. A camera from Baumer is placed on top of the head, and provides 44 photos per second with a 640*480 resolution. This frame rate is selected to easily synchronize each frame with an approximately 23ms sound frame. A small portable round loudspeaker with a frequency response ranging from 200Hz to 16kHz is used to emit a white Gaussian noise. 3 colored patches are sticked in front of it, an image processing algorithm gives then the coordinates of the centers of the patches. During a recording, a person holds the loudspeaker emitting the noise in the camera field of view, and moves it in different directions (left, right, up, or down).

2) Localization results: When working with this experimental setup, the exact relative angular location of each marker with respect to the head is unknown. Indeed, the approach uses the camera since this relative position is directly reflected by the corresponding pixel coordinates in the image plane. So, the experiments will be assessed by comparing the actual pixel coordinates pix_y and pix_x to the prediction $\widehat{\text{pix}}_y$ and $\widehat{\text{pix}}_x$ produced by the neural networks. Having three points, the mean real and estimated coordinates are compared, so as to estimate the mean real and estimated loudspeaker centers. Note that the results presented in this section are obtained on a 12s-recording during which the sound source moves in the image. Estimations are reported in Figure 3. They show that the predicted pixel coordinates follow the real ones quite well, while the column estimation results are better than the line estimation results, which is inline with the simulation results showing a better estimation in the azimuth case.

Figure 4 shows a comparison between a predicted tra-

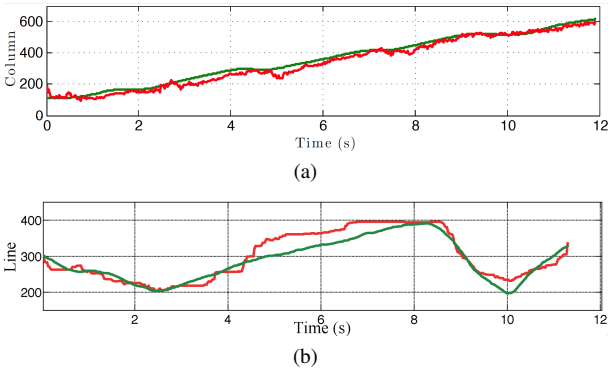

Fig. 3. Experiments: estimation results, predicted dimensions as a function of time. (a) columns, (b) lines.

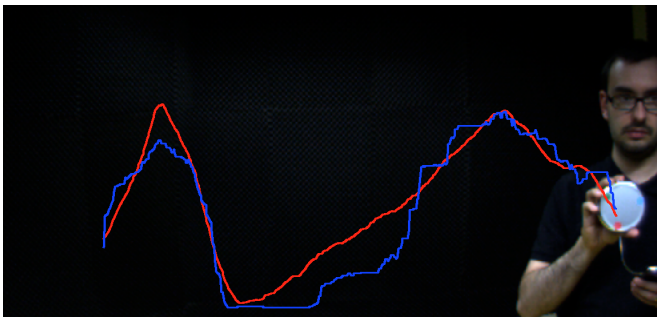

Fig. 4. Experiments: a truncated view of an image taken by the camera. It shows a predicted trajectory made by the sound source (blue) and the corresponding real trajectory (red).

jectory and a real one. It can be seen that the system follows the target source quite accurately. The observed differences between the two trajectories are mainly caused by the line coordinates whereas the column coordinates are better estimated.

IV. Discussion

The tests made on the simulated and the experimental databases lead to the same conclusions: the system is able to efficiently estimate the position of the sound source, with better performances in the azimuth estimation than in the elevation estimation. Compared to related works, this system has a higher resolution and is less complex. For example, one can cite [11] where a parametric model computing ILDs and IPDs as a function of the azimuth is used, and these cues are inverted to deduce the azimuth. But the estimation errors are higher than those obtained in our study, and the resolution is weaker (also 5 degrees in azimuth). Also in [7] and [9], the binaural systems rely on probabilistic approaches needing large databases and computational capabilities, and have position resolutions of 5°.

V. Conclusion

A sound source localization system has been presented. It deals with the localization problematic in a new learning fashion using cues extracted from both human-like ears of a humanoid robot and visual indices from a camera placed on its head. While the interaural cues provided very satisfactory results for azimuth estimation, output energies from a set of cochlear filters allowed to efficiently determine the source's elevation. The described works provided an efficient tool in adequate acoustic conditions, current works are aiming at generalizing the tests to more complex situations. Noises and reverberations are to be taken into account, with human voice sound signals. In such a case, the image processing stage will then consist in a face detection system to provide the training data's speaker pixel position.

Acknowledgment

This work was conducted within the French/Japan BINAAHR (BINaural Active Audition for Humanoid Robots) project under Contract n°ANR-09-BLAN-0370-02 funded by the French National Research Agency.

References

[1] V.R. Algazi, R.O. Duda, R.P. Morrisson, and D.M. 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.

[2] R. Brueckmann, A. Scheidig, and H.-M. Gross. Adaptive noise reduction and voice activity detection for improved verbal human-robot interaction using binaural data. *IEEE International Conference on Robotics and Automation*, April 2007.

[3] J. Garas. Adaptive 3d sound systems. *Kluwer*, 2000.

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

[5] Marco Jeub, Matthias Dörbecker, and Peter Vary. A semi-analytical model for the binaural coherence of noise fields. *IEEE Signal Processing Letters*, 18(3), March 2011.

[6] J. Lewald and S. Getzmann. Horizontal and vertical effects of eye-position on sound localization. *Hearing Research*, 213(1-2):99–106, Mar 2006.

[7] Tobias May, Steven van de Par, and Armin 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.

[8] Kazuhiro Nakadai, Daisuke Matsuura, Hiroshi G. Okuno, and Hiroaki Kitano. Applying scattering theory to robot audition system: Robust sound source localization and extraction. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003.

[9] Johannes Nix and Volker 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.

[10] R.D. 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.

[11] Martin Raspaud, Harald Viste, and Gianpaolo Evangelista. Binaural source localization by joint estimation of ild and itd. *IEEE Transactions on Audio, Speech and Language Processing*, 18(1), 2010.

[12] 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.

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

[14] J.-M. Valin, F. Michaud, and J. Rouat. Robust 3d localization and tracking of sound sources using beamforming and particle filtering. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP Proceedings*, 2006.

[15] 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.