解 説

# The EAR Project

Julien Bonnal[*1], Sylvain Argentieri[*2], Patrick Danès[*1],
Jérome Manhès[*1], Philippe Souères[*1] *and* Marc Renaud[*1]
[*1]CNRS ; LAAS & Université de Toulouse; UPS, INSA, INP, ISAE; LAAS
[*2]UMPC Université Paris 06; UMR 7222; ISIR & CNRS; UMR 7222; ISIR

## 1. Introduction

Audition is often considered as the most important sense in humans, because of its fundamental role in learning, language and communication. However, its use in robotics is fairly recent in comparison to other exteroceptive sensing. Nevertheless, its complementarity to vision and its potentialities for Human-Robot Interaction have been widely acknowledged [1].

Besides binaural approaches, array processing constitutes a relevant way to endow robots with audition. The idea is to exploit the redundancy of the data sensed by an array of microphones so as to design robust and efficient functions. Most contributions in robotics have been rooted in the Computational Auditory Scene Analysis (CASA) framework. This in turn has given rise to unexpected original requirements. To cite few, any solution must be easily embeddable (geometry and energy consumption), perform in real time, handle wideband signals (e.g. the human voice), and cope with noise and reverberations.

In this context, LAAS-CNRS has developed an integrated auditory sensor named EAR ("Embedded Audition for Robotics"), on the basis of a linear array of eight microphones, a fully programmable acquisition board, a FPGA processing unit, and USB communication (**Fig. 1**). This paper describes its design and its exploitation in a robotics experiment. First, the codesign of its hardware and software is outlined in Section 2. Next, Section 3 highlights elements of source localization and source extraction by spatial filtering. The sensor effectiveness is demonstrated on experimental results in Section 4. A conclusion ends the paper.

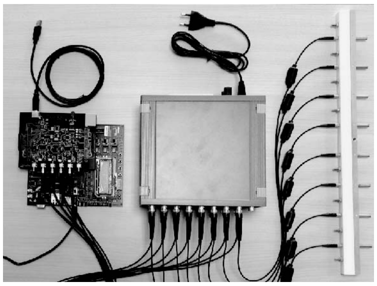


**Fig. 1** (from right to left) The EAR microphone array, conditioning system, FPGA and DAQ boards

## 2. The EAR sensor

The EAR sensor aims at delivering high-quality acoustic measurements and real time reliable auditory cues. Its hardware and software are hereafter outlined.

### 2.1 Hardware

The EAR hardware is connected to a uniform linear array of $N = 8 (1/4)''$ ICP® technology BSWA MP416 electret microphones, with 50 [mV/Pa] sensibility and ±3° phase match tolerance. The input signals are truncated to [300 [Hz]; 3 [kHz]] in order to keep utterances intelligible. So, the array interspace $d$ is defined as the Shannon spatial sampling period $d = \lambda_{3[\mathrm{kHz}]}/2 = 5.66$ [cm].

First, an embeddable conditioning system ensures the transducers supply, trans-impedance adaptation and amplification. Downstream, a dedicated homemade DAQ board applies identical analog filtering on the 8 channels prior to performing their synchronous 18-bits analog-to-digital conversion (ADC). The digital data are sent to a AVNET evaluation kit based on the Xilinx Virtex-4 FPGA, which includes multiple DSP cores and MAC blocks for massive parallel processing, and consumes less than 500 [mW]. Importantly, all the DAQ board gains and cut-off/sampling frequencies can be selected by software through the FPGA (**Fig. 2** (a)).

### 2.2 Software

The EAR software is composed of two parts. On the

原稿受付 2009 年 10 月 21 日
キーワード：Robot Audition, Auditory Sensor, Beamforming, Source Localization, Human-robot Interaction

[*1]7 avenue du colonel Roche, F-31077 Toulouse, France
[*2]Pyramide - T55/65, CC 173 - 4 Place Jussieu, F-75005 Paris, France

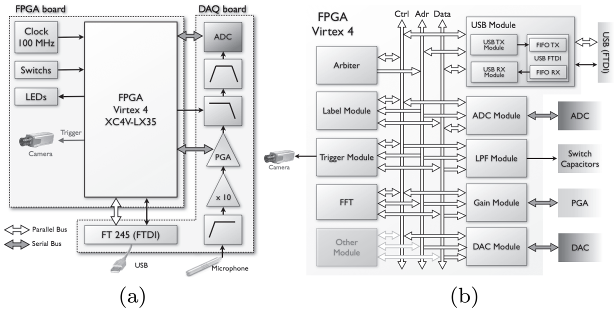


**Fig. 2** (a) Internal organization of the EAR sensor; (b) VHDL modular architecture

one hand, a VHDL homemade architecture is integrated on the FPGA (Fig. 2 (b)). Therein, modules communicate and exchange data thanks to standardized connexions with three internal buses. All the functions are scheduled by the `Arbiter` module. Four modules are related to the DAQ board (parametrization, reception of the sampled data, and digital-to-analog conversion of any signal handled by the FPGA). The `Label Module` performs data time stamping, and the `Trigger Module` generates calibrated digital signals for external triggering of video cameras. The `USB Module` is in charge of the construction and deconstruction/extraction of the USB frames. All the remaining modules are dedicated to the computation of specific cues. For example, `FFT` executes direct and inverse Fast Fourier Transforms. The hardcoding of source localization and spatial filtering as instances of `Other Module` is ongoing.

On the other hand, a C/C++ library enables the dynamic control of the sensor and data exchange from/with a UNIX host, as well as the implementation of auditory functions. It prototypes source localization and spatial filtering at the aim of their hardcoding. Thanks to multithread programming, a multitasking real time execution is guaranteed as far the number of concurrent activities remains low enough.

## 3. Source localization and extraction

This section introduces the theory underlying the low-level functions implemented on the EAR sensor, namely sound source localization and spatial filtering.

### 3.1 Beamforming based spatial filtering

By inserting digital (FIR) filters downstream the ADC stage and summing their outputs, the microphone array mimics a continuous antenna. The interest and key point of this approach, termed *beamforming* [2], lie in the computation of the filters coefficients which lead to a given directivity pattern. The wished spatial filtering typically consists in amplifying signals impinging on the array from a direction of arrival (DOA) of interest while attenuating other worthless DOAs.

Conventional beamforming points the array towards a selected azimuth $\theta_0$ by rephasing the waves impinging from $\theta_0$ prior to their summation. Though widely used, this strategy shows a poor resolution at low frequencies (**Fig. 3** (a)-top). Since much energy of human voice is located at these frequencies, the focusing of the array is likely to be very poor. Contrarily, the beamforming method [3] can lead to a pattern centered around $\theta_0$ with a nearly constant main lobe width over the range [300 [Hz]–3,000 [Hz]] (Fig. 3 (a)-bottom). Ref. [3] also analyzes the involvement of the theoretical beampattern in the global response of an EAR-like sensor.

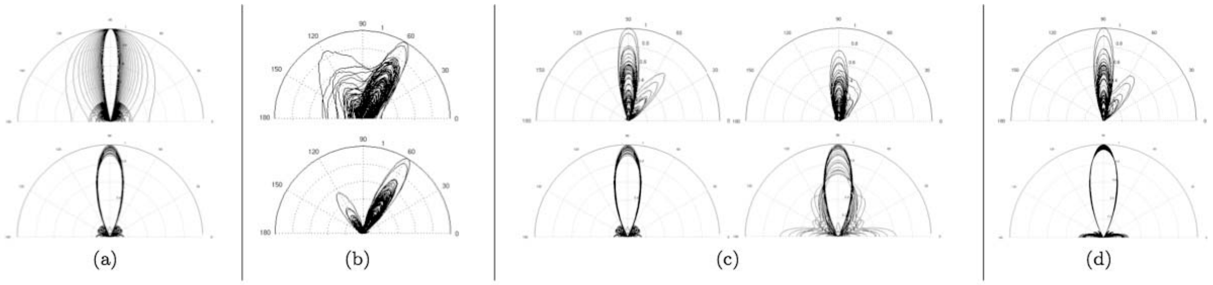
Two facts must be kept in mind. First, the FIR filters coefficients must be upper bounded so as to limit the array sensitivity to noise. Second, the pattern of a frequency-invariant beamformer synthesized under the farfield assumption gets distorted when a source gets closer (Fig. 3 (c)-bottom). The strategy [4], based on *modal analysis* and *convex optimization*, solves these problems at it enables the array azimuthal focusing onto a broadband source of known range while restricting the array white noise gain (Fig. 3 (d)-bottom).

From the above, azimuthal acoustic maps can be computed by scanning the environment with a bank of dedicated beamformers synthesized offline (one per scanned DOA), then by evaluating over a sliding temporal window the energy impinging from each azimuth. Fig. 3 shows simulated localization results using conventional and optimized beamforming, depending on the distance of the two emulated sources to the array.

### 3.2 Localization by beamspace MUSIC

The celebrated MUSIC (MUltiple SIgnal Classification) high-resolution method [5] enables the localization of $S$ narrowband independent sources with an array of $N > S$ microphones. From the generalized eigendecomposition of the array and noise covariance matrices pencil, the so-called signal and noise spaces are determined. The orthogonality of the latter with the array vector at the sources (range,azimuth)'s leads to a pseudo-spectrum, highly peaked at these locations.

Broadband extensions follow two lines. The popular method [6] averages narrowband pseudospectra computed separately on $B$ frequency "bins". This can be computationally expensive, as $B$ generalized eigendecompositions of $N \times N$ complex Hermitian symmetric



(a) Conventional (top) *vs* optimized frequency-invariant (bottom) beampatterns for $\theta_0 = 90°$. – *One curve per frequency.* – Farfield assumption.
(b) Consequent acoustic energy maps. – *Each curve is a map computed over a time snapshot.* – The sources true azimuths are 60° and 120°.
(c) Acoustic energy maps obtained through optimized frequency-invariant farfield beamformers, considering two sources at azimuths 50° and 90°, when these sources are at infinite distance (top-left) *vs* when they are at $r = 0.8$ m to the array (top-right). – A frequency-invariant beamformer synthesized in the farfield (bottom-left) shows a distorted pattern in the nearfield (bottom-right), and thus leads to a degraded map (top-right).
(d) Typical optimized frequency-invariant beamformer in the nearfield, for $r = 0.8$ m (bottom). – Consequent acoustic energy map (top).

**Fig. 3** Beampatterns and consequent acoustic maps, computed from Ref. [4]

pencils are required to compute the broadband pseudo-spectrum. *Coherent* alternatives can be proposed, along the suggestion of Ref. [7]. A $\mathbb{C}^{T\times N}$-valued *focalization* matrix function, $T > S$, is first introduced, whose product by the array vector is frequency invariant. So-called *array and noise focalized covariance matrices* are then obtained by averaging on the $B$ bins the covariance matrices of the signals built by premultiplying the array and noise frequency contents by the focalization matrices. The generalized eigendecomposition of this focalized covariance matrices pencil leads to a pseudo-spectrum. This method is theoretically sound even if the sources are correlated, so that it can handle multipath propagation in reverberant environments. Its complexity is highly reduced, as a single generalized eigendecomposition of a $T \times T$ complex Hermitian symmetric pencil is required for each hypothesized source range. On this basis, Ref. [8] defines the aligned vector signals as the outputs from $T$ dedicated frequency-invariant beamformers, namely the spherical harmonics of increasing order. As the beamforming method [4] fits the needs of this approach, their union was assessed in Ref. [9].

One sharp issue is the required prior knowledge of the sources number. An online detector of this number has been implemented as its minimum Akaike information criterion estimate (MAICE), along [7] [10]. Importantly, such a theoretically grounded detection could not be coupled with a noncoherent broadband MUSIC.

## 4. Experimental results

Experimental localization results are hereafter sketched, in the framework of object localization for collaborative human-robot manipulation. The microphone array is fixed to a mast and oriented downwards, so as to privilege sounds emanating from a table. An object noise is mimicked by a mobile phone playing soundtracks. A calibration chart, laid on the table and parallel to the array, enables the knowledge of the true object azimuth. The distance to the array is about 70 [ cm] and the environment is silent.

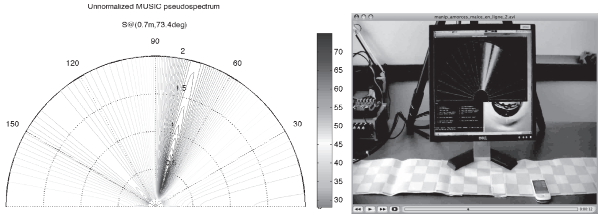


**Fig. 4** (left) Sample MUSIC pseudospectrum (in [dB]); (right) The EAR sensor in action

Localization is performed using the MUSIC method outlined in section 3. 2. As four beamspaces are used, up to three sources can be simultaneously localized. Each pseudo-spectrum is computed on 1,024 samples at 7,971 [ Hz]. **Fig. 4** shows a pseudo-spectrum computed offline for a source azimuth of 73° (left), together with a photograph of real time operation (right). The level sets of the pseudo-spectra are reported on polar plots (*vs* azimuth and range), with peaks in dark. In the worst case, the estimated azimuth error keeps inferior to ±10°. The pseudo-spectra get inconsistent during the soundtrack pauses if MUSIC is not aware that there is no source in the environment. Detecting the number of sources as the MAICE solves this problem. Offline computations of farfield pseudo-spectra with Ref. [9] show a time saving factor exceeding 15 w.r.t. Ref. [6].

## 5. Conclusion and Prospects

This paper presented the EAR project, which aims at an embeddable integrated auditory sensor fitting the requirements of robotics. In the mean-term, localization and spatial filtering will be hardcoded on the FPGA in addition to acquisition, so that the UNIX host is dedicated to higher-level functions (voice activity and audio patterns detection, sound recognition, tracking, etc.).

Other experiments have been conducted within a noisy uncontrolled environment, and confronted to "ground truth" obtained through a human motion capture system, see Ref. [11]. Ongoing work concerns (i) a thorough statistical evaluation of the sensor, (ii) the hardcoding of beamforming through an "overlap-and-save" fast convolution technique, (iii) the hardcoding of MUSIC based on Jacobi algorithm and CORDIC, (iv) the theoretical analysis of MUSIC sensitivity to modeling errors, and (v) the development of a next-generation FPGA board.

**Acknowledgements** The authors acknowledge the support from French ANR AMORCES (Algorithms and MOdels for a Collaborative Robot, Eloquent and Social - *Cooperative Perception/Action Loops* for joint human-robot objects manipulation) and EU FP6-STREP CommRob (Advanced Behavior and High-level Multimodal Communication with and Among Robots - *Human Motion Interpretation* from intelligent trolleys deployed in shopping malls).

## References

[1] H.G. Okuno and K. Nakadai: "Computational auditory scene analysis and its application to robot audition," IEEE Wkshop on Hands-free Speech Communication and Microphone Arrays, pp.124–127, 2008.

[2] B.D. Van Veen and K.M. Buckley: "Beamforming: a versatile approach to spatial filtering," IEEE ASSP Mag., vol.5, no.2, pp.4–24, 1988.

[3] S. Argentieri, P. Danès, P. Souères and P. Lacroix: "An experimental testbed for sound source localization with mobile robots using optimized wideband beamformers," IEEE/RSJ Int. Conf. on Intell. Robots and Systems, pp.2536–2541, 2005.

[4] S. Argentieri, P Danès and P. Souères: "Modal analysis based beamforming for nearfield or farfield speaker localization in robotics," IEEE/RSJ Int. Conf. on Intell. Robots and Systems, pp.866–871, 2006.

[5] R.O. Schmidt: "Multiple emitter location and signal parameter estimation," RADC Spect. Estim. Wkshop, pp.276–280, 1979.

[6] F. Asano, H. Asoh and T. Matsui: "Sound source localization and signal separation for office robot Jijo-2," IEEE/SICE/RSJ Int. Conf. on Multisensor Fusion and Integration for Intell. Systems, pp.243–248, 1999.

[7] H. Wang and M. Kaveh: "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol.33, pp.823–831, 1985.

[8] D.B. Ward and T.D. Abhayapala: "Range and bearing estimation of wideband sources using an orthogonal beamspace processing structure," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp.109–112, 2004.

[9] S. Argentieri and P. Danès: "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," IEEE/RSJ Int. Conf. on Intell. Robots and Systems, pp.2009–2014, 2007.

[10] H.L. Van Trees: Optimum Array Processing, volume IV of Detection, Estimation, and Modulation Theory. John Wiley & Sons, Inc., 2002.

[11] J. Bonnal, S. Argentieri, P. Danès and J. Manhès: "Speaker localization and speech extraction with the EAR sensor," IEEE/RSJ Int. Conf. on Intell. Robots and Systems, 2009.

**Julien Bonnal**

Julien Bonnal graduated in Computer Science and Signal Processing. He is preparing a Ph.D thesis on robot audition at LAAS-CNRS, Toulouse, France.

**Patrick Danès**

Patrick Danès, Ph.D, is an Associate Professor at Univ. Paul Sabatier and LAAS-CNRS, Toulouse, France. His interests are robot audition, tracking, and visual servoing.

**Philippe Souères**

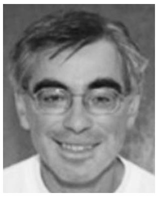
Philippe Souères, Ph.D, is a Director of Research at LAAS-CNRS, Toulouse, France. His interests include robot control and the link between robotics and neurosciences.

**Sylvain Argentieri**

Sylvain Argentieri, Ph.D, graduated in Electronics and Robotics. He is an Associate Professor at Univ. Pierre et Marie Curie and ISIR, Paris, France. His interests concern robot audition.

**Jérome Manhès**

Jérome Manhès graduated in Electrical Engineering and Computer Science. He is a research engineer in electronics and robotics at LAAS-CNRS, Toulouse, France.



**Marc Renaud**

Marc Renaud, Ph.D, is a Professor in Mechatronics at INSA and LAAS-CNRS, Toulouse, France. His interests are modeling and control of mobile manipulators and robot audition.