Speaker Localization and Speech Extraction with the EAR sensor.

Julien Bonnal[†‡], Sylvain Argentieri[*×], Patrick Danès[†‡] and Jérôme Manhès[†‡]
[†] Université de Toulouse ; UPS, INSA, INP, ISAE ; LAAS-CNRS : F-31077 Toulouse, France
[‡] CNRS ; LAAS ; 7, avenue du Colonel Roche, F-31077 Toulouse, France
[*] UPMC Univ Paris 06 ; UMR 7222 ; ISIR : F-75005, Paris, France
[×] CNRS ; UMR 7222 ; ISIR : F-75005, Paris, France
{jbonnal,danes,jmanhes}@laas.fr, sylvain.argentieri@upmc.fr

***Abstract*— This paper presents the Embedded Audition for Robotics (EAR) project internally developed at LAAS and its application to speaker localization and extraction. Hardware and software issues are first thoroughly depicted, concerning the development of an auditory sensor based on an array of microphones, a homemade dedicated acquisition chain and a FPGA based processing board. Then, the EAR sensor is assessed against various scenarios, in real noisy robotics environments. Localization results are presented when a speaker emits an utterance in the presence of a disturbing source. These validate the underlying theory and suggest further theoretical and experimental developments.**

I. INTRODUCTION

Robot Audition has constituted a fertile field of research for the last years. Many applications can be cited, e.g. visioauditive tracking, exteroceptive robot control or Human Robot Interaction. All these require the availability of auditory cues at high rate and thus imply large computational resources, while battery capacity and physical size are strictly limited on mobile platforms. Such heavy computational requirements lead to the exploitation of a high-end workstation, often also involved in other robotics functions, such as low-level control or vision-based procedures. Application-specific hardware is known to be one of the most effective solutions to these issues. Noticeably, this idea has been successfully applied in three-dimensional vision [1], while sound-specific hardware has remained fairly scarce in Robotics. For instance, [2] proposed the implementation of a sound source separation system on a dedicated Dynamically Reconfigurable Processor developed by NEC. Such a small, low power and flexible system is shown to be well suited to Robotics constraints, but remains hard to program.

Considering Robot Audition from the viewpoint of array processing requires the synchronous acquisition of the signals delivered by the microphones as well as their processing on hardware/software architectures endowed with real time performance. Some industrial off-the-shelf solutions do exist, yet they seldom fit the Robotics needs. For instance, their embeddability may be limited, their frequencies bandwidth may not cover the useful spectrum of a voice signal, or on the contrary their too high genericity may make them costly.

For these reasons, a specific hardware and software has been developed at LAAS-CNRS for Robot Audition. It constitutes the internal "Embedded Audition for Robotics" (EAR) project, which we plan to freely distribute under an open-source license. In the same vein, the Honda Research Institute and the Kyoto University have been developing a pure-software solution called HARK [3] based on a programming environment with modular architecture. The main difference between HARK and EAR relies on the hardware system which is deeply linked to the software for the EAR sensor, while HARK claims to be hardware-independent. So, this paper aims at presenting the prominent aspects of the EAR currently operational functionalities, as well as some ongoing developments. Hardware issues are first described in §II-A–II-B, from the array of transducers to the FPGA processing unit through the Data Acquisition (DAQ) board. VHDL and C softwares are then expounded in §II-C for the communication with a UNIX host, the fine tuning of the acquisition chain, and the data acquisition. Methods to the computation of acoustic localization cues constitute §III, and experimental results are reported in §IV. A conclusion ends the paper, which outlines ongoing extensions and higher-level auditory functions under development.

II. THE EAR SENSOR

The EAR hardware is to be coupled with appropriate microphones to provide reliable and high-quality measurements. This section begins by outlining its specifications. Next, its components are detailed, including the DAQ board internally developed at LAAS-CNRS and the computing unit relying on a commercial FPGA test board. The characteristics of the transducers constituting the microphone array are then given. Finally, the VHDL and C softwares enabling the control of the overall sensor are described.

A. Technical and functional specifications

The objective of the EAR project is to propose ingredients of an integrated acoustic sensor which can cope with the specific robotics constraints. The first major issue concerns *embeddability*, in term of size and energy consumption. As a solution, a specific hardware is proposed so as to synchronously acquire up to N audio channels and autonomously deliver auditory cues. This channels number N is of particular concern. It is selected between 2—for bio-inspired methods such as [4] or [5]—and 32—for the three-ring array from [6]. The EAR sensor is endowed with a maximum of $N = 8$ channels, as this number constitutes a good trade-off

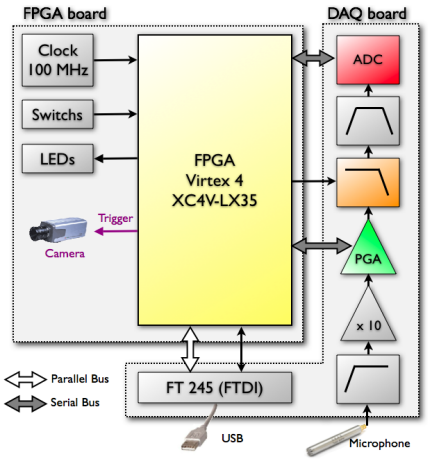

Fig. 1. Organization of the whole sensor. The DAQ board is connected to the FPGA board through dedicated extension connectors.

between the sensor embeddability and the redundancy of the acoustic information needed by the algorithms [7] [8].

The second fundamental concern is *real-time* performance. Indeed, acoustic primitives must be made available within a guaranteed short time interval in order to be exploited in low-level reflex functions such as auditive/visioauditive tracking. This leads to embed processing units with particular features. In the EAR project, a Xilinx FPGA is selected, for it includes multiple DSP cores and enables a massive parallelism. In addition, its power consumption is lower than 500mW.

The third constraint is directly induced by the *wideband* nature of audio signals, reaching up to 20 kHz for voice signals. In order to endow the $N = 8$ channels with an identical frequency response, the user must be able to parametrize the hardware through *hand-tunable parameters*, including *gains*, *cut-off frequency* of the various filters involved in the acquisition chain, or the signals *sampling frequency*.

The proposed solution is based on a hardware composed of two different boards and microphones described in the next subsection.

B. Hardware

1) The DAQ Board: The main role of the DAQ board is to perform the analog-to-digital conversion of its 8 inputs. Each channel entails six distinct stages, depicted on Figure 1. The first one is a passive 2^{nd}-order high-pass filter, presenting a cutoff frequency of 60 Hz, in order to eliminate the possible DC/low-frequency components of the microphone outputs. The filtered output is then sent to a first amplifier providing a gain G_1 ranging from 8.55 to 9.89. Next, a second digitally programmable amplifier follows, whose gain G_2 can be modified from the FPGA within $[-95.5\,\text{dB}; +31.5\,\text{dB}]$ by steps of 0.5 dB. The two next stages constitute the anti-aliasing filter, involving a 8^{th}-order low-pass elliptical switched capacitor filter followed by an active band-pass filter for clock noise removal. Noticeably, the anti-aliasing cut-off frequency value f_c can be easily adapted to the bandwidth of interest through the FPGA. Finally, the filtered signal is digitally converted through a high-speed 18 bits delta-sigma converter. Similarly, the sampling frequency f_s can be tuned in the range 2 kHz to 50 kHz.

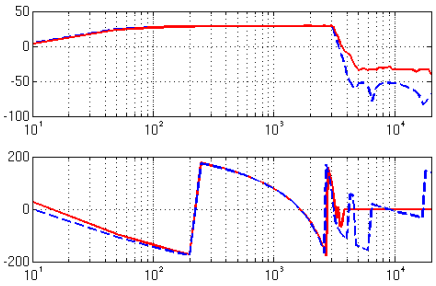

Fig. 2. Comparison between the experimental (plain) and the theoretical (dashed) frequency response of one channel of the DAQ board. The gain is in dB while the phase is expressed in degree.

The experimental frequency response of one channel, from the microphone input to the ADC input, is shown on Figure 2, for $G_1 = 8.65$, $G_2 = 10\,\text{dB}$ and $f_c = 3\,\text{kHz}$. Its characteristics on the bandwidth are the same as these of the theoretical response from the datasheets. Besides, stop-band rejection is about 55 dB, corresponding to less than a 10 mV voltage.

2) The FPGA Board: The above DAQ board is connected to a Xilinx Virtex 4 FPGA computing unit, which is integrated on an evaluation board from AVNET designed for tests or HDL developments. The selected Virtex 4 FPGA includes 192 fixed-point 500 MHz MAC (Multiplication and ACcumulation) blocks for massive parallel processing. Secondary connectors can be exploited in order to control external peripherals. Importantly, one of these is turned into the trigger of one or more cameras to be embedded on a robotics platform. This way, vision and audition can be accurately synchronized. Note that the image is not received nor processed by the EAR hardware, which just acts like a shutter. Finally, the overall connections of the FPGA board to the DAQ board through the aforementioned specific extension connectors are shown on Figure 1.

3) Microphones and preconditioning: So far, the EAR hardware has been connected with $N = 8$ $\frac{1}{4}$ inch-diameter and 60 mm-long BSWA MP416 microphones, arranged along a line with even $\frac{\lambda_{3\,\text{kHz}}}{2} = 5.66\,\text{cm}$ interspace. These exhibit a sensitivity of 50 mV/Pa and are phased matched with a $\pm 3°$ default tolerance. They are followed by a preamplifier for their supply, trans-impedance adaptation and amplification by a predefined factor of 1, 10 or 100. Noticeably, the compactness of this preconditioning unit and the possibility to supply it with a 24 V DC source eases its embedding.

C. Software

The EAR sensor integrates a VHDL software into the FPGA for the configuration of the DAQ board, as well as a C driver enabling the control of the sensor from a UNIX host. Each of them are expounded in the two next subsections.

1) FPGA Software: The program integrated into the FPGA has been entirely written in VHDL. Its functional

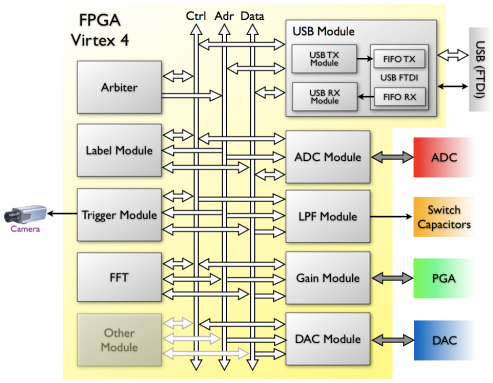

Fig. 3. FPGA software. The modularity of the design allows the easy integration of any `Other Module`.

block diagram is detailed on Figure 3. It relies on a modular design, each independent module being devoted either to control or to the computation of a specific feature, and linked to each other through three common intern parallel buses. The modularity of the design allows the easy integration of any new module, through a VHDL template providing the standard connections with the three buses. The nine basic modules of the VHDL architecture are made explicit on Figure 3. Four of them are dedicated to the control of the DAQ board and generate appropriate signals for the tuning of the cut-off frequency f_c, the gain G_2 and the sampling frequency f_s. The `ADC Module` also receives the 8 synchronously sampled microphone signals and makes them available to other modules. In addition, the FFT and/or the microphone signals can be sent through USB to a UNIX host thanks to the `USB Module`. Localization and Extraction functions come as separate instances of `Other module`.

2) C Library: The C `libfpga` library provides a list of functions enabling a UNIX host to dialog with the FPGA via USB communication. It must enable the dynamic configuration of the sensor as well as data exchange. USB communication is based on a FTDI chip whose role is to interface the FPGA with USB. On the top of the provided `libftdi` open-source library, `libfpga` implements a homemade communication protocol, and defines functions enabling the reception of data from the FPGA as well as the configuration of the acquisition chain.

III. FUNDAMENTALS OF VOICE EXTRACTION AND LOCALIZATION

Auditory functions have been built on the basis of the hardware and software described above. The lowest level routines relate to speaker localization and speech extraction. For each of them, the underlying theory is first outlined. Then, the experimental setup of the EAR sensor as well as the acquisition conditions are precisely depicted.

A. Beamforming for extraction and localization

Beamforming may be the straightest way to endow the EAR sensor with the ability to perform spatial filtering. By inserting suitable digital FIR filters downstream the transducers and summing their outputs, the microphone array can mimic a single continuous antenna, an operation termed *beamforming* [9]. So, signals impinging on the array from any direction of arrival (DOA) of interest can be amplified while other worthless DOAs are attenuated.

To make the array point towards the azimuth θ_0 of a speaker of interest—the value of θ_0 being computed beforehand by, say, a multisensor based tracking algorithm—the filters can be selected so as to rephase the waves impinging from θ_0 prior to their summation. This "conventional beamforming" strategy, though commonly used, shows a significantly poor resolution at low frequencies. Since the energy contents of human voice is mainly located at these frequencies, the focalization of the array is likely to be very poor. This is why another array pattern synthesis was envisioned to obtain directivity patterns centered on a given azimuth θ_0 with a nearly constant main lobe width over the frequency range $[300\,\text{Hz} - 3000\,\text{Hz}]$ [10]. Broadband frequency-invariant beamforming not only can lead to a better focusing on a human speaker, but also can preserve some important features of the extracted signal, which may be valuable for post-processing issues.

Two important facts must be kept in mind when synthesizing such beamformers. On the one hand, the FIR filters coefficients magnitudes must be limited in order to keep the array response to noise within acceptable limits. On the other hand, the pattern of a frequency-invariant beamformer synthesized under the farfield assumption gets distorted when a source comes closer to the array. Importantly, the strategy proposed in [11], based on convex optimization and on the modal decomposition of beampatterns, theoretically enables the azimuthal focalization onto one broadband sound source under the knowledge of its range while limiting the array white noise gain.

The application of beamforming to the computation of horizontal (i.e. azimuthal) acoustic energy maps is straightforward. First, the environment is scanned through the successive "electronic polarization" of the array towards a set of spatial directions by means of a bank of dedicated beamformers synthesized offline (one per scanned DOA). Then, for each hypothesized azimuth, the impinging acoustic energy is evaluated over a sliding temporal window.

B. Localization by beamspace MUSIC

An extension of the celebrated high-resolution MUSIC (MUltiple SIgnal Classification) method [12] to the azimuth and range localization of wideband sources was recently developed by [13] on the basis of [14]. It consists in combining the outputs from dedicated beamformers—namely the spherical harmonics of increasing order—onto separate narrow frequency bins. As the results proposed in [11] perfectly fit its array pattern synthesis needs, the union of both algorithms has been assessed in [15] on simulated data, and compared with the single use of MUSIC in robotics so far [16]. The new strategy turns to be better suited to robotics thanks to its higher performance, its relatively

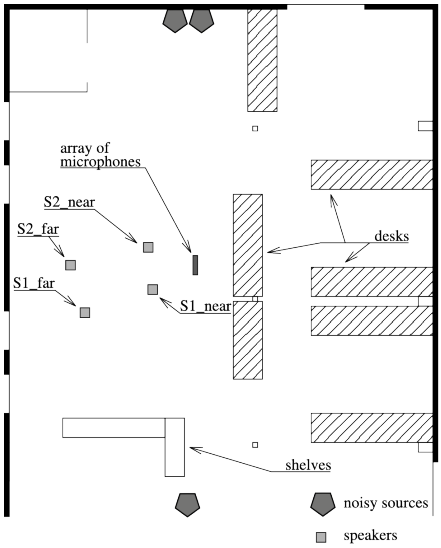

Fig. 4. Map of the LAAS robotics hall. Two sets of experiments have been conducted, the sources `S1` and `S2` being either at 3.3 m far from the array midpoint (`S1_far`@117°,`S2_far`@90°) or at 1.2 m (`S1_near`@120°,`S2_near`@68°).

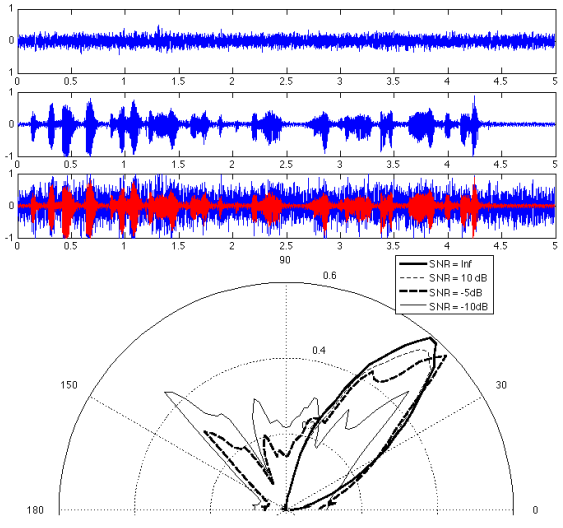

Fig. 5. Top: pattern of the sensed noise signal in the room, simulated voice signal and simulated perceived signal on one microphone with $SNR = -5\,\mathrm{dB}$, as a function of time. Bottom: energy map for various SNRs on one snapshot.

low computational cost, and because multipath propagation in reverberant environments can theoretically be handled. Interestingly, the localization comes as a sequence of two one-dimensional searches: first, the azimuths $\theta_1^*, \theta_2^*, \ldots$ maximizing the obtained pseudospectrum $h(r,.)$ for $r = +\infty$ are looked for; then, for each θ_i^*, the range r_i^* comes as the argument r maximizing $h(.,\theta_i^*)$. References [16] [17] constitute a significant theoretical improvement over azimuthal energy maps, in that no prior knowledge on the distance to the sources is needed. Its implementation on an FPGA is however more involved, as generalized eigendecompositions must be hardcoded.

One sharp issue is that the number of sources is assumed given beforehand. To make MUSIC fully functional, it has been coupled with an online detector of this number. The algorithm implements its minimum Akaike information criterion estimate (MAICE), along the lines of [14] [18].

C. Experimental setup

The extraction and localization capabilities of the EAR sensor have been evaluated in a real noisy environment. Experiments have been conducted in the LAAS robotics hall, within a $8\,\mathrm{m} \times 5\,\mathrm{m}$ open space. This hall is about 6 m-high. About 20 computers are running continuously, and an impressive air-conditioning system keeps on humming, so that this environment can be defensibly qualified as very noisy. No specific sound absorbing material has been laid on the walls. On the contrary, materials of various types have been used, including large windows.

The 40 cm-width linear array of $N = 8$ microphones described in §II-B.3 has been mounted on a 1.5 m-high tripod, and positioned as shown on Figure 4. The characteristics of the EAR sensor have been set to $G_2 = -10\,\mathrm{dB}$, $f_c = 3\,\mathrm{kHz}$ and $f_s = 15024\,\mathrm{Hz}$.

IV. Experimental results

A. Inluence of noise on the localization and extraction

The first conducted experiment consists in measuring the acoustics of the hall in the absence of any speaker. The resulting signal, shown on Figure 5–top for microphone 1, covers about 20% of the DAQ full-scale input dynamic range. Such a high noise level may significantly hinder the localization process. It also highlights the need of spatial filtering to extract the signal-of-interest out of the environment. In our previous work [11] [15], the environmental noise had not been explicitly taken into account. So, we hereafter assess the robustness of our method against noise by simulations.

All the source azimuths are measured with respect to endfire. In the following, the real hall noise sensed by the microphones is mixed with the simulated propagation of a speech signal uttered from the azimuth $\theta = 45°$ and distance $r = 2\,\mathrm{m}$ to the array. The utterance is then amplified or attenuated to obtain various signal-to-noise ratios (SNRs) on the microphones. The voice signal as well as one microphone signal with a $-5\,\mathrm{dB}$ SNR are shown on Figure 5. For each SNR, a series of acoustic energy maps has been computed from the outputs of 91 optimized beamformers over successive sliding snapshots, these beamformers being polarized from $\theta_0 = 0°$ to $\theta_0 = 180°$ with a 2° step.

Figure 5 shows energy maps computed on a common snapshot for $SNR \in \{\mathrm{Inf\ (no\ noise)}, 10\,\mathrm{dB}, -5\,\mathrm{dB}, -10\,\mathrm{dB}\}$. In the case $RSB = \mathrm{Inf}$ or $RSB = 10\,\mathrm{dB}$, they exhibit a main lobe pointing towards the emulated speaker direction. In such a favorable case, the localization is clearly accurate and leads to the same type of results published in the above references. When lowering the utterance level until a $-5\,\mathrm{dB}$ RSB, the azimuth of the speaker can still be deduced from the computed energy map though with a slight error. Additionally, some grating lobes appear. Decreasing the *RSB*

Fig. 6. The EAR sensor (top left) — The two experiments (right & bottom).

to -10 dB leads to a very disturbed energy map, where no clear dominant main lobe can be exploited for localization. In fact, a kind of *noise signature* appears in the energy map, related to the intrinsic noise properties, e.g. its frequency contents, the positions of noise sources, etc. So, in such an unfavorable context, the localization of low-level sources is no longer possible, these being buried in the ambient noise.

B. Assessment of acoustic localization against ground truth

This section presents the evaluation of the localization algorithms on real data sensed within the noisy robotics hall described above. Two loudspeakers `S1` and `S2` respectively emit a voice signal and a 1 kHz pure tone. These loudspeakers and the microphone array are fitted with markers reflecting infra-red light. From the localization of such markers in 3D space by a commercial Human Motion Capture system, "ground truth" values of the sources ranges and azimuths w.r.t. the array are deduced. Two sets of experiments have been considered, with either (`S1_far`@(3.3 m,117°);`S2_far`@(3.3 m,90°)) or (`S1_near`@(1.2 m,120°);`S2_near`@(1.2 m,68°)), see Figures 4 and 6.

Acoustic maps and MUSIC pseudo-spectra are hereafter presented on polar plots. Importantly, the way the two representations must be understood is totally different. On the one hand, the distance between the origin and a point of an acoustic energy map is related to the magnitude of the impinging energy at its azimuth. On the other hand, MUSIC pseudo-spectra are functions—in dB—both of the range and azimuth. So, their iso-levels are presented in contour plots, the "hot" values tending to the peaks.

Due to the powerful ambient noise, `S1_far` and `S2_far` can be efficiently localized only if they are loud enough. Figure 7 shows three MUSIC pseudo-spectra, namely, the signature of the environment when `S1` and `S2` are mute, the localization of `S1_far` (human voice) when `S2` is inactive, and the localization of `S2_far` (pure 1 kHz sinusoid) when `S1` is inactive. Noticeably, the low RSBs may make the number of detected sources meaningless.

Experiments conducted when the sources get closer to the array give better results, in spite of the ambient noise. Figure 8 shows MUSIC pseudo-spectra and acoustic maps computed at two distinct instants. The two plots on the left correspond to a snapshot when both `S1_near` and `S2_near` are active. As the voice uttered by `S1_near` is louder, `S2_near` does not appear in the acoustic map, while it is clearly detected and localized by MUSIC. On the right, only the `S2_near` (1 kHz pure tone) is active. In both cases, the peaks of the pseudo-spectra lead to accurate estimates of the source azimuths, but the range is not faithful. This can be explained by oversimplifying assumptions made within the MUSIC algorithm, e.g. by the mismatch between the assumed statistics of the ambient noise and their true values. However, the detected number of sources is more meaningful than in the above case. As for the acoustic maps, notice that they have been computed through broadband beamformers synthesized along the lines of [11] so as to get a frequency invariant beampattern at a range of 1 m. This way, a well-oriented lobe can be got—though small—which points towards `S2_near`.

V. CONCLUSION

A sound-specific hardware and software dedicated to robot audition has been presented. It enables the simultaneous acquisition of up to 8 microphone outputs and can be fully controlled via USB by an UNIX host. All this work is planned to be released under an open-source license so as to propose to the Robotics community an integrated easy-to-use audio system. The operation of the sensor has been experimentally assessed. Two speaker localization strategies previously evaluated in simulation were successfully tested. The first one is based on broadband frequency-invariant beamformers. In comparison with conventional beamforming, this method conducts to sharper azimuthal energy maps thanks to the better spatial filtering of low frequencies. The second approach is based on a broadband beamspace MUSIC strategy. Though this point hasn't been stressed in the paper, it enables source azimuth and range estimation at a much lesser computational cost compared to the popular broadband elementspace MUSIC algorithm [16].

Current investigations are twofold. On the one hand, the trade-off offered by the convex optimization based beamforming method [11] between the focalization of the array and the limitation of its white noise gain has been studied in more detail, and its influence on the consequent acoustic energy maps has been assessed. Ongoing work also concerns a deeper analysis of the sensitivity and tuning of the MUSIC method, so as to get more trustable pseudo-spectra. This includes deeper insights into the detection of the number of sources [14] [18], as well as theoretical analyses of the effect of errors in the array vectors and/or in the noise statistical description [19], [20].

Up to this point, the EAR sensor provides facilities to the acquisition of acoustic signals, the focusing of the array, and the computation of source localization primitives, at a rate up to 15 Hz. The optimized wideband array pattern synthesis

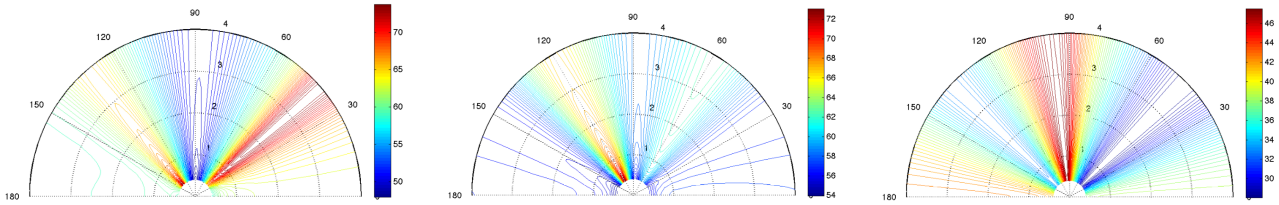

Fig. 7. (left) MUSIC signature of the environment; (middle & right) localization of `S1_far`@(3.3 m,117°) and `S2_far`@(3.3 m,90°), respectively. Note that the algorithm can distinguish between the peak corresponding to `S1_far` (middle) and a peak at $\sim 130°$ due to the noise (left).

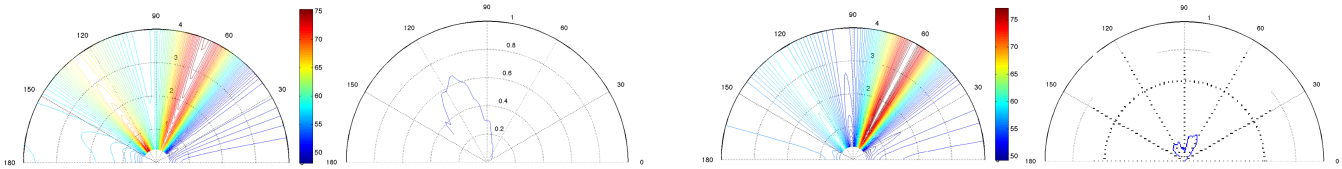

Fig. 8. MUSIC pseudo-spectra and Optimized beamformers based acoustic energy maps in the nearfield: (left) both `S1_near` (voice@(1.2 m,120°)) and `S2_near` (1 kHz tone@(1.2 m,68°)) are active; (right) only `S2_near` (1 kHz tone) is active. In the right (zoomed) acoustic map, notice that the main lobe, though small, is well-oriented.

may constitute a sound basis to Geometrical Source Separation (GSS). To cope with robotics environments, which are intrinsically variable, evolutive, and subject to noise and reverberation, several higher-level auditory functions will be developed on the basis of the EAR sensor. These will include spatiotemporal Speaker Tracking, Voice Activity Detection and Speaker Recognition. Extending audio recognition to other specific sounds of the environment will be envisioned as well, as it raises new perspectives in Human Robot cooperative tasks.

VI. ACKNOWLEDGMENTS

This work was partially conducted within the EU Project CommRob (`www.commrob.eu`) under Contract FP6-IST-045441, and within the AMORCES project funded by the French National Research Agency.

REFERENCES

[1] J.-L. Boizard, M. Devy, P. Fillatreau, J.-Y. Fourniols, P. Lacroix, N. Nasreddine, F.-X. Bernard, and T. Sentenac, "Real-time stereovision by an integrated sensor," in *6th IFAC Symposium on Intelligent Autonomous Vehicles*, September 2007.

[2] S. Kuratori, N. Suzuki, K. Nakadai, H. Okuno, and H. Amano, "Implementation of active direction-pass filter on dynamically reconfigurable processor," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Aug 2005, pp. 515–520.

[3] K. Nakadai, H. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition hark and its evaluation," in *IEEE-RAS International Conference on Humanoid Robots*, December 2008, pp. 561–566.

[4] J. C. Murray, S. Wermter, and H. R. Erwin, "Bioinspired auditory sound localisation for improving the signal to noise ratio of socially interactive robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2006.

[5] R. Brueckmann, A. Scheidig, and H.-M. Gross, "Adaptive noise reduction and voice activity detection for improved verbal human-robot interaction using binaural data," in *IEEE International Conference on Robotics and Automation*, April 2007, pp. 782–1787.

[6] Y. Tamai, Y. Sazaki, S. Kagami, and H. Mizoguchi, "Three ring microphone array for 3d sound localization and separation for mobile robot audition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Aug. 2005, pp. 903–908.

[7] J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H. Okuno, "Robust recognition of simultaneous speech by a mobile robot," *IEEE Transactions on Robotics*, vol. 4, pp. 742 – 752, August 2007.

[8] J.-S. Choi, M. Kim, and H.-D. Kim, "Probabilistic speaker localization in noisy environments by audio-visual integration," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2006.

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

[10] S. Argentieri, P. Danès, P. Souères, and P. Lacroix, "An experimental testbed for sound source localization with mobile robots using optimized wideband beamformers," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Aug. 2005.

[11] S. Argentieri, P. Danès, and P. Souères, "Modal analysis based beamforming for nearfield or farfield speaker localization in robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2006, pp. 866–871.

[12] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," in *RADC Spectrum Estimation Workshop*, Oct. 1979.

[13] D. B. Ward and T. D. Abhayapala, "Range and bearing estimation of wideband sources using an orthogonal beamspace processing structure," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, May 2004, pp. 109–112.

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

[15] S. Argentieri and P. Danès, "Convex optimization and modal analysis for beamforming in robotics: Theoretical and implementation issues," in *European Signal Processing Conference*, September 2007.

[16] F. Asano, H. Asoh, and T. Matsui, "Sound source localization and signal separation for office robot jijo-2," in *IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Aug. 1999, pp. 243–248.

[17] S. Argentieri and P. Danès, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2007.

[18] H. L. Van Trees, *Optimum Array Processing*, ser. Detection, Estimation, and Modulation Theory. John Wiley & Sons, Inc., 2002, vol. IV.

[19] A. Paulraj and T. Kailath, "Eigenstructure methods for direction of arrival estimation in the presence of unknown noise fields," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 13–20, 1986.

[20] A. L. Swindlehurst and T. Kailath, "A performance analysis of subspace-based methods in the presence of model errors, Part I: The MUSIC algorithm," *IEEE Transactions on Signal Processing*, vol. 40, no. 7, pp. 1758–1774, 1993.