Visuo-acoustic Cues Integration in an Artificial Developing Agent

[1]Lorenzo Natale

[2]Giorgio Metta

[1]Giulio Sandini

[1]LIRA-Lab – DIST, University of Genoa, Genoa, Italy
[2]AI-Lab – Massachussets Institute of Technology, Cambridge MA, USA

Abstract

Sound localization has been widely studied with the aim of building artificial systems as well as understanding the mechanism underlying the same process in the biological systems. Nevertheless, in robotics the use of auditory cues has been rather limited. We present an artificial sound localization system and its implementation in an embodied, self-developing agent, capable of acquiring and enhancing its motor capabilities by interacting with the environment. The model includes the integration of visual and auditory cues and appropriate sensory to motor maps for generating meaningful actions.

Introduction

Robots able to move and interact in a real and possibly unknown environment must be able to take advantage of a wide range of stimuli. In the past, passive as well as active sensors have been employed for navigation; the recent development of humanoid robots has increased the importance of being able to exploit at best the available sensory information. This can be very important for at least two reasons. Firstly, humanoids robots are usually meant to interact with humans in a flexible way. In this case human-like senses can extend the ability of artificial beings to communicate with their human counterparts using a common language. Secondly, the intrinsic complexity of a natural environment and the absence

of a pre-specified task require the design of agents that can adapt to the external environment.

Acoustic cues have been widely studied with the aim of implementing artificial sound localization systems. In general, artificial heads with rubber pinnae were realized and the resulting head related transfer function (HRTF) explicitly employed to compute the position of a sound source using binaural and monaural cues [1, 2]. Though these systems turn out to be rather accurate, they cannot be easily used for real time applications. Less frequently, sound localization systems were embodied in fully operating robotic systems (in this sense, an example can be found in [3, 4]).

The goal of our work is to propose a biologically plausible functional model of the acquisition of visual, acoustic and multi-modal motor responses. ITD and ILD are used for estimating the position of auditory sources in space; acoustic and visual cues are then fused in a unified percept. The embodiment and the interaction of the system with the environment are used to test the correctness of the approach. The particular task considered is the control of the orienting behavior.

Our approach

The experimental setup consists of a five degrees of freedom robot head. The eyes can independently pan and are mounted on a common tilt axis. The neck itself can also pan and tilt. The robot acquires and processes images in a space variant format also known as log-polar [5], providing a high-resolution *fovea* and a low-resolution periphery. For practical reasons, the two microphones are mounted on the tilt of the eyes, surrounded by asymmetric ear lobes. Each joint has also a "proprioceptive" sense provided by high-resolution optical encoders.

Interaural level difference (ILD) and interaural time difference (ITD) are usually considered to be the most informative cues about the spatial location of a sound source. The different position of the ears makes both ITD and ILD to be related to the horizontal component (*azimuth*); the directional filtering of the head and the outer ear, at high frequency, can produce interaural level differences which vary along a constant azimuthal angle, that is with the elevation of the source (for a complete review see [6]).

Studies on birds [7, 8] as well as on humans [9] have shown an alignment of the auditory and visual maps even in the presence of alteration in the sensory sub-systems. It was shown also that vision guides the re-alignment of the maps [7, 10].
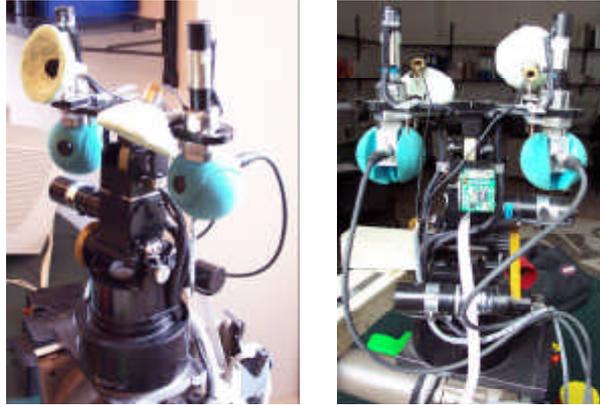
Figure 1 Two views of the robot head, frontal and rear view. Note the two cameras (surrounded by blue tennis balls) and the microphones arranged on top of them, with two plastic earlobes. The earlobes mimic the disparity of the external ears found in the barn owl.

Our approach mimics the sound localization apparatus of the barn owl. The horizontal position of the source is determined estimating the delay between the signals reaching the two microphones (ITD) by means of the generalized cross correlation algorithm [11]. As for the elevation, we built two artificial earlobes, which, mounted in an asymmetrical fashion on top of the head, reproduced the disparity of the external ear of the barn owl [12]. ILD is hence supposed to be proportional to the elevation of the source and is computed as the ratio between the average power spectra at the two channels. The retinal target position is extracted using a color segmentation algorithm. Given an initial cueing mechanism (motion detection), the robot selects an interesting brightly and uniformly colored region. Afterwards, color information is used to segment the target from the background [13].

Once an object has been located, the spatial percept has to be converted into the motor command required to saccade toward the target. This is carried out by using maps linking together visual and auditory information with motor commands suitable to drive the robot head. In the barn owl the acquisition of these maps is mainly driven by vision [10], but other possibilities have been discussed [14]. Because the eyes can move with respect to the head not only are vision and sound represented with respect to different reference frames, but a coordinate conversion has to take into account the actual position of the eyes in order to generate suitable motor commands. In our implementation, we simplified the problem considering only a given eye-head orientation (i.e. eyes almost centered with respect of the head) and only horizontal movements (i.e. the azimuth). As the learning of the multi-modal map progresses (driven by vision), the robots estimates

3

also the relationship between visual and auditory information, so that, after a certain amount of training, the same map can be addressed even in absence of vision

A second experiment is reported, showing the acquisition of the motor map needed to control the neck by means of sound alone. In this second case, sound localization is performed in both azimuth and elevation.

Conclusions and future work

Vision plays a major role during the life of biological systems. Nevertheless, sound turns out to be useful in situations where the visual system is inadequate – like in absence of light or when something occurs outside the subject's field of view. We employed binaural cues like ITD and ILD to control the orienting behavior of a humanoid robot. The acquisition of the appropriated motor maps can be driven by sound alone or by vision. Given an initial set of learning rules, the whole learning process is carried out autonomously by the robot just by interacting with the environment. The integration between visual and auditory cues can enhance the adaptive capabilities of the robot. As it was previously argued, plasticity of the sensory representation in the biological systems is useful to keep optimal performance during development. Although artificial systems do not physically grow, adaptive capabilities might help to cope with alterations in the motor and sensory subsystems or to avoid manually tuning the internal parameters [3].

Finally, future work will be directed toward a more sophisticated exploitation of the information in different frequency bands, in order to enhance the sound localization sub-system.

Bibliography

1.	Duda, R.O.a.C., W. *Combined Monaural and Binaural Localization of Sound Sources.* in *29th Asilomar Conference on Signals, Systems, and Computers.* 1995.

2.	Henderson, N., *Estimating Azimuth from Speech in a Natural Auditory Environment.* 1996, Department of Electrical Engineering, San Jose State University.

3.	Rucci, M., G.M. Edelman, and J. Wray, *Adaptation of Orienting Behavior: From the Barn Owl to a Robotic System.* IEEE Transactions on Robotics and Automation, 1999. **15**(1, February 1999): p. 96-110.

4.	Rucci, M. and J. Wray, *Binaural cross-correlation and auditory localization in the barn owl: a theoretical study.* Neural Networks, 1999. **12**: p. 31-42.

5.	Sandini, G. and V. Tagliasco, *An Anthropomorphic Retina-like Structure for Scene Analysis.* Computer Vision, Graphics and Image Processing, 1980. **14**(3): p. 365-372.

6.	Blauert, J., *Spatial Hearing: the Psychophysics of Human Sound Localization.* Second Printing, 1999 ed. 1983, Cambridge: M.I.T. Press. 494.

7.	Brainard, M.S., Knudsen, E. I., *Experience-dependent Plasticity in the Inferior Colliculus: A Site for Visual Calibration of the Neural Representation of Auditory Space in the Barn Owl.* The Journal of Neuroscience, 1993. **13**(11): p. 4589-4608.

8.	Knudsen, E.I., *Early auditory experience aligns the auditory map of space in the optic tectum of the barn owl.* Science, 1983. **222**: p. 939-942.

9.	Hofman, P.M., Van Riswick, J.G.A.,Van Opstal A.J., *Relearning sound localization with new ears.* Nature Neuroscience, 1998. **1**(5): p. 417-421.

10.	Knudsen, E.I., Knudsen, P. K., *Vision Guides the Adjustment of Auditory Localization in Young Barn Owls.* Science, 1985. **230**: p. 545-548.

11.	Knapp, C.H., and Carter, G.C., *The Generalized Correlation Method for Estimation of Time Delay.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 1976. **24**: p. 320-327.

12.	Knudsen, E.I., *The Hearing of the Barn Owl.* Scientific American, 1981. **245**: p. 82-91.

13.   Metta, G., *Babyrobot: A Study on Sensori-motor Development*, in *Dipartimento di Informatica, Sistemistica e Telematica.* 1999, University of Genoa (PhD Thesis): Genova, Italy.

14.   Knudsen, E.I.a.M., J., *Vision-independent Adjustment of unit Tuning to Sound Localization Cues in Response to Monaural Occlusion in Developing Owl Optic Tectum.* The Journal of Neuroscience, 1992. **12**(9): p. 3485-3493.