Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 87–112

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A survey on sound source localization in robotics: From binaural to array processing methods[☆,☆☆]

S. Argentieri[a,b,*], P. Danès[c,d], P. Souères[c,e]

[a] *Sorbonne Universités, UPMC Univ. Paris 06, UMR 7222, ISIR, F-75005 Paris, France*
[b] *CNRS, UMR 7222, ISIR, F-75005 Paris, France*
[c] *CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France*
[d] *Univ. de Toulouse, UPS, LAAS, F-31400 Toulouse, France*
[e] *Univ. de Toulouse, LAAS, F-31400 Toulouse, France*

Received 3 June 2014; received in revised form 13 February 2015; accepted 10 March 2015
Available online 20 March 2015

Abstract

This paper attempts to provide a state-of-the-art of sound source localization in robotics. Noticeably, this context raises original constraints—e.g. embeddability, real time, broadband environments, noise and reverberation—which are seldom simultaneously taken into account in acoustics or signal processing. A comprehensive review is proposed of recent robotics achievements, be they binaural or rooted in array processing techniques. The connections are highlighted with the underlying theory as well as with elements of physiology and neurology of human hearing.

Keywords: Robot audition; Source localization; Binaural audition; Array processing

1. Introduction

“Blindness separate us from things but deafness from people” said Helen Keller, a famous American author who was the first deafblind person to obtain a Bachelor in Arts, in 1904 (Kohlrausch et al., 2013). Indeed, hearing is a prominent sense for communication and socialization. In contrast to vision, our perception of sound is nearly omnidirectional and independent of the lighting conditions. Similarly, we are able to process sounds issued from a nearby room without any visual information on their origin. But human capabilities are not limited to sound *localization*. We can also *extract*, within a group of speakers talking simultaneously, the utterance emitted by the person we wish to focus on. Known as the term *Cocktail Party Effect* (Haykin and Chen, 2005), this separation capacity enables us to process efficiently and selectively the whole acoustic data coming from our daily environment. Sensitive to the slightest tone and level variations of an audio message, we have developed a faculty to *recognize* its origin (ringtone, voice of a colleague,

[☆] This works has been conducted within the European FP7 TWO!EARS project under grant agreement n° 618075 and the BINAAHR project funded by ANR (France) and JST (Japan) under Contract n° ANR-09-BLAN- 0370-02.

[☆☆] This paper has been recommended for acceptance by R.K. Moore.

[*] Corresponding author. Tel.: +33 144276355.
E-mail addresses: sylvain.argentieri@upmc.fr (S. Argentieri), patrick.danes@laas.fr (P. Danès), philippe.soueres@laas.fr (P. Souères).

http://dx.doi.org/10.1016/j.csl.2015.03.003
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

etc.) and to *interpret* its contents. All these properties of localization, extraction, recognition and interpretation allow us to operate in dynamic environments, where it would be difficult to do without auditory information. All the above impressive human capabilities have stimulated developments in the area of *Robot Audition*. Likewise, the recent research topic of human–robot interaction (HRI) may have constituted an additional motivation to investigate this new field, with the aim to artificially reproduce the aforementioned localization, extraction, recognition and interpretation capabilities. Nevertheless, contrarily to computer vision, robot audition has been identified as a scientific topic of its own only since about 15 years (Nakadai et al., 2000). Since then, numerous works have been proposed by a growing community, with contributions ranging from sound source localization and separations in realistic reverberant conditions to speech or speaker recognition in the presence of noise. But as outlined in Argentieri et al. (2013), the robotics context raises several unexpected constraints, seldomly taken into account in signal processing or acoustics. Among them, one can mention:

Geometry constraint: Though the aim is to design an artificial auditory system endowed with performances inspired by human hearing, there is no need to restrict the study to a biomimetic sensor endowed with just two microphones. Indeed, bringing redundant information delivered by multiple transducers can improve the analysis and its robustness to noise. Straight connections thus appear with the field of array processing. Yet, the robotics context imposes an *embeddability* constraint. While array processing can consider large arrays of microphones—e.g. several meters long, robotics implies a trade-off between the size of the whole sensor and its performances, so that it can be mounted on a mobile platform, be it humanoid or not.

Real time constraint: Many existing methods to sound analysis rely on heavy computations. For instance, a processing time extending over several tens of seconds is admitted to perform the acoustic analysis of a passenger compartment. Contrarily, localization primitives involved in low-level reflex robotics functions—e.g. sensor-based control or auditive/visioauditive tracking—must be made available within a guaranteed short time interval. So the algorithms computational complexity is a fundamental concern. This may imply the design of dedicated devices or computational architectures.

Frequency constraint: Most sound signals valuable to robotics are *broadband*, i.e. spread over a wide bandwidth w.r.t. their central frequency. This is the case of voice signals, which show significant energy on the bandwidth [300–3300 Hz] used for telephony. Consequently, narrowband approaches developed elsewhere do not straightly apply in such broadband contexts. Noticeably, this may imply a higher computational demand.

Environmental constraint: Robotics environments are fundamentally dynamic and unpredictable. Contrarily to acoustically fully controlled areas, unexpected noise and reverberations are likely to occur, which depend on the room characteristics—dimensions, walls, type of the building materials, etc.—and may singularly deteriorate the analysis performance. The robot itself participates to these perturbations, because of its self-induced noise, e.g. from fans, motors, and other moving parts. A challenge is to endow embedded sound analysis systems with robustness and/or adaptivity capabilities, able to cope with barge-in situations where both the robot and a human are possibly both speaking together.

Generally, most of embedded auditory systems in robotics follow the following classical bottom-up framework: as a first step, the sensed signals are analyzed to estimate sound sources positions; next the locations are used to separate sound of interests from the sensed mixture in order to provide clean noise or speech signals; finally, sound/speaker/speech recognition systems are fed with these extracted signals. Of course, other alternatives have also been proposed (Otsuka et al., 2012), but this approach remains by far the most used framework in robot audition. Nevertheless, it exhibits the importance of sound localization in the overall analysis process. It has been indeed the most widely covered topic in the community, and a lot of efforts have been made to provide efficient sound localization algorithms suited to the robotics context. Moreover, independently of any high-level interpretation of the acoustical scene, having access to the low-level source localization information itself is mandatory for any applications related to HRI. Indeed, a natural intuitive HRI might heavily depends on how responsive a robot will be to acoustical information. Among them, source localization allows the robot to quickly react to an auditory stimulus by turning the head towards the source (turn-to reflex), or even to focus its other sensors in the direction of interest (e.g. by moving a camera field of view towards a speaker). Since, in our opinion, Robot Audition has reached an undeniable level of scientific maturity, we feel that the time has come to summarize and organize the main publications of the literature. This paper then attempts to review the most notable contributions specific to sound source localization. Another intent is to underline their connections with theoretical foundations of the field, including with basics of human physiology and neuroscience.