Robotics and Autonomous Systems 96 (2017) 184–210

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Localization of sound sources in robotics: A review

Caleb Rascon *, Ivan Meza

Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas, Universidad Nacional Autonoma de Mexico, Circuito Escolar S/N, Mexico 04510, Mexico

HIGHLIGHTS

- A highly detailed survey of sound source localization (SSL) used over robotic platforms.
- Classification of SSL techniques and description of the SSL problem.
- Description of the diverse facets of the SSL problem.
- Survey of the evaluation methodologies used to measure SSL performance in robotics.
- Discussion of current SSL challenges and research questions.

ARTICLE INFO

Article history:
Received 18 August 2016
Received in revised form 24 June 2017
Accepted 21 July 2017
Available online 5 August 2017

Keywords:
Robot audition
Sound source localization
Direction-of-arrival
Distance estimation
Tracking

ABSTRACT

Sound source localization (SSL) in a robotic platform has been essential in the overall scheme of robot audition. It allows a robot to locate a sound source by sound alone. It has an important impact on other robot audition modules, such as source separation, and it enriches human–robot interaction by complementing the robot's perceptual capabilities. The main objective of this review is to thoroughly map the current state of the SSL field for the reader and provide a starting point to SSL in robotics. To this effect, we present: the evolution and historical context of SSL in robotics; an extensive review and classification of SSL techniques and popular tracking methodologies; different facets of SSL as well as its state-of-the-art; evaluation methodologies used for SSL; and a set of challenges and research motivations.

1. Introduction

The goal of sound source localization (SSL) is to automatically estimate the position of sound sources. In robotics, this functionality is useful in several situations, for instance: to locate a human speaker in a waiter-type task, in a rescue scenario with no visual contact, or to map an unknown acoustic environment. Its performance is of paramount influence to the rest of a robot audition system since its estimations are frequently used in subsequent processing stages such as sound source separation, sound source classification and automatic speech recognition.

There are two components of a source position that can be estimated as part of SSL (in polar coordinates):

- Direction-of-arrival estimation (which can be in 1 or 2 dimensions)
- Distance estimation.

SSL in real-life scenarios needs to take into account that more than one sound source might be active in the environment. Therefore it is also necessary to estimate the position of multiple simultaneous sound sources. In addition, both the robot and the sound source are mobile, so it is important to track its position through time.

SSL has been substantially pushed forward by the robotics community by refining traditional techniques such as: single direction-of-arrival (DOA) estimation, learning-based approaches (such as neural network and manifold learning), beamforming-based approaches, subspace methods, source clustering through time and tracking techniques such as Kalman filters and particle filtering. While implementing these techniques onto robotics platforms, several facets relevant to SSL in robots have been made evident including: number and type of microphones used, number and mobility of sources, robustness against noise and reverberation, type of array geometry to be employed, type of robotic platforms to build upon, etc.

As it is shown in this review, the SSL field in robotics is quite mature, proof of which are the recent surveys in this topic. For instance, [1,2] present a survey on binaural robot audition, [3] offers a general survey of SSL in Chinese, [4] presents some SSL

* Corresponding author.
E-mail addresses: caleb.rascon@iimas.unam.mx (C. Rascon), ivanvladimir@turing.iimas.unam.mx (I. Meza).

http://dx.doi.org/10.1016/j.robot.2017.07.011

works based on binaural techniques and multiple-microphone arrays, and [5] presents an overview of the robot audition field as a whole. The aim of this work is to review the literature of SSL implemented over any type of robot, such as service, rescue, swarm, industrial, etc. We also review efforts that are targeted for an implementation in a robotic platform, even if they were not actually implemented in one. In addition, we review resources for SSL training or evaluation, including some that were not collected from a robotic perspective but could be applied to a robotic task. Finally, we incorporate research that uses only one microphone for SSL that, although not applied in a robotic platform, we believe has an interesting potential for the SSL robotic field.

In this work we present: the evolution of the field (Section 2); a definition of the SSL problem (Section 3); a classification of techniques used in SSL within the context of robot audition (Section 4); an overview of popular tracking techniques used for SSL (Section 5); several facets that describe the areas that SSL techniques are tackling (Section 6); a review of different evaluation methods that are currently being used for measuring the performance of SSL techniques (Section 7); and an insight on potentially interesting challenges for the community (Section 8). Finally, we highlight several motivations for future research questions in the robot audition community (Section 9).

2. The evolution of SSL

The surge of SSL in robotics is relatively new. To our knowledge, it started in 1989 with the robot *Squirt*, which was the first robot to have a SSL module [6,7]. *Squirt* was a tiny robot with two competing behaviors: hiding in a dark place and locating a sound source. The idea of using SSL as a behavior to drive interaction in a robot was later explored by Brook's own research team and it culminated with a SSL system for the *Cog* robot [8–11]. In the meantime, several Japanese researchers started to investigate the potential of SSL in a robot as well. In 1993, Takanashi et al. explored an anthropomorphic auditory system for a robot [12,13] (as described by [10]). This research was followed by notable advances in the field: Chiye robot [14], RWIB12-based robot [15–18], Jijo-2 [19,20], Robita [21] and Hadalay [22]. This first generation of robots tackled difficult scenarios such as human–robot interaction, integrating a complete auditory system (source separation feeding speech recognition), active localization, dealing with mobile sources and capture systems, and by exploring different methodologies for robust SSL.

At the turn of the 20th century, the binaural sub-field of robot audition started to become an important research effort, including SSL. Although robots from the first generation were technically binaural (e.g., Squirt, COG, Chiye, Hadalay), it is with the arrival of the SIG robot [23] that the field of binaural robot audition started to generate interest. SIG was built to promote audition as a basic skill for robots and was presented as an experimental platform for the RoboCup Humanoid Challenge 2000 [24]. This resulted in SIG becoming popular for researching robot perception. Binaural robot audition has been followed by other research teams and progress in the field has been constant [25–36].

During the 2000s, an important rift occurred in terms of the research motivations in the robot audition field, specifically in SSL techniques. Binaural audition cemented itself by the motivation to imitate nature: using only two ears/microphones. On the other hand, there was the motivation to increase performance (detailed in Section 4.3), which pushed for the use of more microphones. This opened the door for source localization techniques that use a high amount of sensors (such as MUSIC and beamformers) to carry out SSL in a robot. Subsequently, the facets of the SSL problem were broadened, which yielded a wide variety of solutions from the robot audition community.

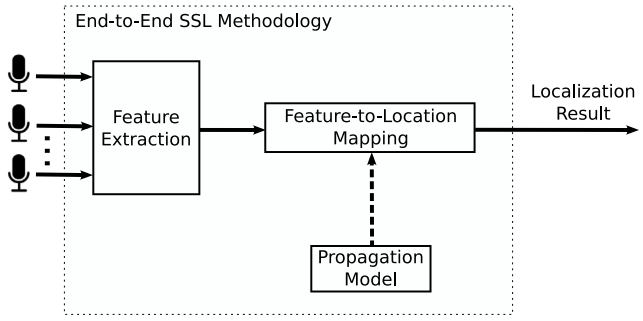

Fig. 1. The complete data pipeline of an end-to-end SSL methodology.

Throughout its history, a central goal for robots with a SSL system has been to support interaction with humans. In the first generations, an important contribution was to face the user, since it indicates that the robot is paying attention. One of the first robots to carry out this attention-based interaction was the Chiye robot [14] which has made its way into recent products such as the Paro robot [37]. Further on, SSL has been used in more complex settings in which other skills intertwine together to reach a specific goal, such as: playing the Marco-Polo game, acting as a waiter, taking assistance and finding its user when it visually lost him/her [38]; logging and detecting the origin of certain sounds while interacting with a caregiver [39]; playing a reduced version of hide and seek in which hand detection and SSL are used to guide the game [40]; providing visual clues from the sound sources as a complement of a telepresence scenario [41]; and directing a trivia-style game [42]. Given the evolution of SSL in robots, we are certain that the complexity of the scenarios will keep growing. In fact, we foresee that the challenges to come will definitely be more demanding (see Section 8 for further discussion).

3. Definition of the sound source localization problem

Sound source localization (SSL) tackles the issue of estimating the position of a source via audio data alone. This generally involves several stages of data processing. Its pipeline is summarized in Fig. 1.

Since this pipeline receives the data directly from the microphones and provides a SSL estimation, we consider a methodology that carries this out as *end-to-end*. Features are first extracted from the input signals. Then, a feature-to-location mapping is carried out, which usually relies on a sound propagation model. These three phases are referenced as such in the explanation of each methodology and their relevant variations in Section 4.

In this section a brief overview of these three phases is presented for ease of reference in the later detailed explanations.

3.1. Propagation models

The sound propagation model is proposed depending on: the positioning of the microphones, as there may be an object between them; the robotic application, as the user may be very close or far away from the microphone array; and the room characteristics, as they define how sound is reflected from the environment. In addition, the propagation model generally dictates the type of features to be used.

The most popular propagation model used is the free-field/far-field model, which assumes the following:

- *Free field:* The sound that is originated from each source reaches each microphone via a single, direct path. This