MedCrave
Step into the World of Research

International Robotics & Automation Journal

Opinion

Open Access

CrossMark

# Robots that hear

## Abstract

The field of robot audition has blossomed into its own field throughout its 30 years of history. However, it is still not considered as an essential part of a robotic solution as other functionalities, such as navigation or manipulation. This paper presents some considerations of the overall current state of the field and proposes some ideas of how to push the community further.

**Keywords:** robot audition, service robots, rescue robots, human-robot interaction

Volume 2 Issue 4 - 2017

Caleb Rascon
CONACYT-University National Autonomy, Mexico

**Correspondence:** Caleb Rascon, CONACYT-University National Autonomy, Mexico, Email caleb.rascon@iimas.unam.mx

**Received:** May 13, 2017 | **Published:** June 06, 2017

**Abbreviations:** SSL, sound source localization; ASR, automatic speech recognition; SSS, sound source separation.

## Introduction

Since the introduction of the robot Squirt[1] from MIT in 1989, the concept of robot hearing has evolved to great lengths. Unfortunately, such concept for a long while converged into a local optimum: automatic speech recognition (ASR). Although it is an important topic of research, other hearing functionalities were left stagnant for a considerable amount of time. Functionalities such as sound source localization (SSL) and sound source separation (SSS) were rarely considered as part of a complete robotic solution until the start of this century. And here lies the topic of this writing: what is it that we aim for when we want to embed the ability to hear into a robot?

First off, the ability to hear is intrinsic to human nature, and a case can be aimed that it was as essential to our survival and evolution as any of the other four senses. However, what we mean by "human hearing" is rarely defined in terms that technologists can emulate. There are many reasons of why this is the case:

a) Biologists and medical experts rarely involve themselves with robotics (other than using it as a tool), and vice versa, while building academic bridges between two such divergent fields is non-trivial;

b) Breaking down an ability as complex as hearing is very challenging, especially since it appears that the whole process is holistic and technologists usually aim to tackle this issue by solving it modularly;

c) A research and development road map has not been established of where we want to take the emulation of such an ability, making the whole field feel sparse in terms of focus, thus the global interests of a robotics research group dictate the progress made into robot hearing; etc.

A cursory reading of the literature in robot audition can lead one to the works of well-known robot audition research groups/projects such as HARK,[2] ManyEars from IntRoLab,[3] BINAARH[4] and its consequent EAR project.[5] However, other than these efforts, the rest of the works are mainly carried out by robotics research groups that seem to see robot audition as an "additional" or "optional" part of their robotic solutions. This is not to say that they are in the wrong; it could be argued that the advancement of the robotics community with this philosophy has come a long way and, thus, it is evidence enough of them being right. Few can argue in favor of giving functionalities such as manipulation, vision and navigation a lower priority than audition, since these have put the community where it is today. Nevertheless, this brings us to a fundamental question: are the robots that we want to build aimed to be interacting with humans? If this is so, it is important to remember that it has been shown quite conclusively that speech is the main channel for human-human interaction. Thus, the author believes that it is time to bring audition in the same priority as the other aforementioned functionalities. Therefore, coming back to the original question of this writing: if we want robots to hear like humans do (to make human-robot interaction a natural endeavor), we need to define what is it that we want the robot to emulate from that ability. Up to this point, it seems that they are mainly four inter-connected abilities that are of interest to the robot audition community:

### Sound source localization

From the trivial point of view, knowing where the user when he/she is talking provides the information the robot needs to face him/her. It is astounding how just this is such a big step forward in making the interaction feel natural for the user, since it makes it appear that "the robot is putting attention" to the user. In addition, this information can be used internally to automatically label the user as "the person in my right" without requiring additional information from him/her. Furthermore, this information can be used to enhance the performance of other abilities of interest, such as:

### Sound source separation

Much of the successive processing stages, such as ASR, use models that were trained with the implicit assumption that one source (the user) is active (speaking) at a time. If the set of active sources only one of them is the user and the rest are noise sources (such as music or computer fans), sound source could be seen as a way of noise filtering. If in this set of active sources, more than one source are humans, their separated data could be fed to the ASR individually to provide multi-user speech recognition. Although it is rare that two users in a conversation speak over each other, in circumstances as a dinner party or restaurant (where a service robot is expected to interact in) several conversations could be carried out near each other. The author would like to point out his amazement in how the human hearing is able to make sense of these highly dynamic circumstances, and stands in humility of the challenge that this represents for us in the community to do the same.

### Sound source identification

This is another way to refer to speaker identification. However, there is an interesting challenge that comes with this ability. There are tasks (such as a waiter or store attendant) in which the robot does not have information from the user with which it can train its models

to identify him/her. Thus, it does not only require identifying the user without re-training, but it needs to do so with using a very small amount of information (since such interactions are short).

### Emotion/mood estimation

Speech carries a vast amount of non-verbal information, of which it seems that the mood or emotion of the user is the one with the most use for service robots (other than identifying its source). Annoyance, sarcasm, panic, etc. have an important use of robotic tasks ranging from elderly care to rescue situations. It is important to mention that this list is not intended to be exhaustive; however, the author does feel that it encompasses that overall tendency of the robot audition community. The author would like to propose other items to add to the aforementioned list that are worthwhile to consider in future efforts and that haven't been seen tackled as much:

### Room response estimation

Reverberation is one of the most ubiquitous acoustic features of an environment, and unfortunately, considerably decreases the performance of several popular SSL and SSS techniques. Estimating and modeling its effect on the captured data could benefit them. However, carrying this out in a generalized is a definite challenge, since the room response changes not only between environments, but also by changing the position of the robot inside a single environment.

### Environment characteristics estimation

Having estimated the room response, there are several spatial features that can be estimated such as its size, the materials of its walls, even the position of the robot.

### Other types of "users"

It is implicitly assumed that the users of interest are human. However, other type of additive tasks involve other type of users, such as: Bioacoustics, in which the "users" are animals being identified and counted for environmental census; locating fire arm discharges in domestic conflicts or military encounters; monitoring drone activity; etc. However, the reader may have realized that, by doing this list, the author has fallen into one the pits mentioned in the beginning of this writing: it is being assumed that human hearing can be broken down into modules, while it is in fact a holistic process. Unfortunately, to develop a technological approach that carries robot audition in a holistic fashion, an unfathomably enormous training corpus must be recollected to train a holistic model that encompasses: number of sources, their location, their separation, their identification, their mood, the characteristics of the environment, etc. Since this model would be aimed to be applied in a generalized fashion, the corpus should be recollected in a very wide range of acoustic circumstances. And even if this is carried out, the outputs of such a model still rely on the modular nature that the aforementioned list bares. The point being made here is that, as far as the author can envision, modularity cannot be escaped, and we need the support from bio-medical experts to obtain, at least, a semi-modular model of human hearing to further push our efforts.

Finally, although the field feels sparse, progress has been made in focusing itself. Several of the important robotic-themed academic conferences (IROS, ICRA, RO-MAN, etc.) have sessions and/or workshops dedicated to robot audition. And there are small research groups around the globe that have branched out from robotics that are focused in emulated hearing in non-human entities. The author would like to add a final item to the list:

### Other types of "listening"

Throughout this writing (and most of the history of the field), human hearing has been the gold standard with which most of the techniques compare themselves to (or, at the very least, strive to emulate). However, it is important to consider that in some of the aforementioned points, such as SSL, the community is close to surpassing "human levels" of performance. Therefore, it may be appropriate to start putting objectives that go beyond that of emulating a human.

## Conclusion

Robot audition has definitely come a long way since its emergence from the robotics community. Although it is still not considered as an essential part of a typical robotic solution, several ways to push the community further into essentialness is in the horizon. Defining the specifics of what it is the community wants to emulate from human hearing, bringing into the fold bio-medical experts, and pushing further than "human levels" are a good step forward in that regard.

## Acknowledgments

The author would like to thank the editor(s) and anonymous reviewers for their constructive comments which helped to improve the present paper. This paper is jointly supported by the Natural Science Foundation of China (Grant 61175028, Grant 61374161), and the Key Project of Natural Science Foundation of Shanghai (16JC1401100).

## Conflict of interest

Author declares that there is none of the conflicts.

## References

1. Flynn AM, Brooks RA, Wells WM, et al. Squirt: The prototypical mobile robot for autonomous graduate students. *Artificial Intelligence Lab*. 1989.
2. Nakadai K, Takahashi T, Okuno HG, et al. Design and implementation of robot audition system 'HARK' open source software for listening to three simultaneous speakers. *Advanced Robotics*. 2010;24(5–6):739–761.
3. Grondin F, LÃl'tourneau D, Ferland F, et al. The Many Ears open framework. *Autonomous Robots*. 2013;34(3):217–232.
4. Portello A, Danès P, Argentieri S. Acoustic models and kalman filtering strategies for active binaural sound localization, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2011. pp. 137–142.
5. Bonnal J, Argentieri S, Danès P, et al. The EAR Project. *Journal of the Robotics Society of Japan*. 2010;28(1):10–13.