Auditory Perception for Interactive Robots: a Cognitive Framework to Include Motor Commands and Working Memory in the Process of Auditory Sound Localization

Omar Eldardeer*
omar.eldardeer@iit.it
Università di Genova
Istituto italiano di tecnologia
Genova, Italy

Alessandra Sciutti
alessandra.sciutti@iit.it
Istituto italiano di tecnologia
Genova, Italy

Matthew Tata
matthew.tata@uleth.ca
The University of Lethbridge
Lethbridge, Alberta, Canada

Francesco Rea
francesco.rea@iit.it
Istituto italiano di tecnologia
Genova, Italy

ABSTRACT

Auditory perception is fundamental for understanding the environment. It provides information about the surroundings, even beyond what visual perception provide, and it forms the basis for many human-human interactions. Accurate localization of sounds in the auditory scene is a fundamental aspect of auditory perception, however several challenges arise in using sound location in human-robot interactions. Drawing inspiration from the body of research showing that humans perform head motions to improve auditory localization, we propose here a cognitive framework to explore how head movements can improve localization abilities for humanoid robots. This Bayesian framework necessitates the role of a priori knowledge and auditory spatial working memory during auditory perception tasks.

KEYWORDS

audio localization, audio perception, HRI, motor actions, working memory, cognitive framework

1 INTRODUCTION

In typical human-human interactions, most of our sensorial inputs coregister to provide a complete understanding of the multisensory context in which we interact. In human-robot interaction, visual perception has been widely addressed, whereas other perceptual channels such as auditory perception are less well-studied. The auditory scene is dense with information about the people and their activities around a robot. This is true even for sound sources that are inaccessible to vision, such as around corners and off the field-of-view of the vision sensors. To make optimal use of this information, the spatial locations of auditory events must be correctly perceived and registered into an allocentric reference frame around the robot. Substantial research on this localization problem has yielded improving results for both binaural and microphone array systems ([2] [10]. An emerging theme in this research is the need to exploit the complex interaction between the robot's movements over time and the continuously updating memory of sound locations in the space around the robot. Binaural auditory systems, both biological and artificial, perform auditory scene analysis by computing either (or both) interaural time differences (ITD) and interaural level differences (ILD). However, such systems necessarily produce "phantom" images of sounds due to ambiguities in ITD and ILD computation [4]. One solution used by biological systems was proposed by Hans Wallach [12] [13], which involves integrating information across head rotations. This active-hearing approach has been successfully used (e.g. by [5] and [3] and [7]) to better resolve the sources in the scene. Optimal head movements thus not only reorient the hearing system to better perceive the sound sources, but also improve auditory localization.A key aspect of active-hearing approaches is that they use some variation on Bayesian memory (e.g. via Kalman-like or particle filters) and cannot exclusively rely on instantaneous evidence. Information about the pose of the robot as it changes in allocentric space over time must be integrated with egocentric instantaneous evidence about the auditory scene, to produce a posterior probability map of auditory objects. In fact, prior expectation of auditory information in the auditory scene is a crucial element in human hearing, especially for priors expectations extracted from vision. For example, the integration of visual evidence with auditory perception gives rise to the well-known ventriloquist illusion of sound localization [1] and the McGurk Effect in audiovisual speech perception [8].

We propose a cognitive framework that includes motor control and working memory modules to improve auditory localization in an audiovisual task. The model is bio-inspired [5] and has been used on the iCub humanoid robot [9] for sound localization [6] [11]. Our main goals are to present the cognitive framework, and to experimentally validate on the iCub humanoid the effect of different head motor actions on the performance sound localization. Here we report preliminary evidence that will inform future choices of motor behaviour models.

2 MODEL

Figure 1 shows the structure of our proposed framework. It consists of four main elements: 1) *Sensory input*: the two microphones located on the head of the robot; 2) *Memory*: a working memory element; 3) *Audio localization*; and 4) *Motor actions*. The localization algorithm is described elsewhere [6][5]. Briefly, the approach uses

a gammatone filterbank to spectrally decompose sound into narrow frequency bands, and a series of swept narrow-band beamformers that approximate the binaural temporal comparisons of ITD. Instantaneous egocentric auditory scenes are rotated in (*audio preprocessing*) module and used to update an allocentric Bayesian posterior map of probable sound sources in the 360 degrees around the robot in (*audio Bayesian processing*) module. Here we introduce the integration of a priori information about the auditory scene (*prior Knowledge integration*), biasing the system output toward known possible target locations. The system also solves the problem of sound onset detection, by computing the total power of the input sound and identifying the existence of a sound signal if the total power exceeds a predefined threshold and enable the on trigger. The off trigger is activated when the power is not exceeding the threshold.The working memory saves the priori information, action state (Not executed/In progress/executed), and the state of the sound (Not present/ present).The trigger sets and resets the state of the sound in the working memory. This trigger is connected also to the Action Linker, which executes a pre-defined motor policy based on the trigger and the state of the action from the working memory.

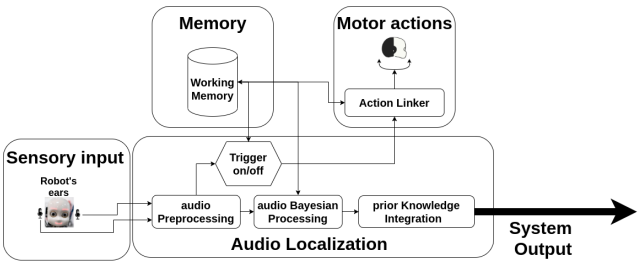

Figure 1: System diagram

3 EXPERIMENT AND RESULTS

On each trial, a complex tone (1KHz with 3 harmonics) was presented for 10 seconds from one of four identical boxes located horizontally on a table in front of the robot (-23° , -8°, 8°, 23° with respect to the midline). The robot's task was to deterimine which of these boxes produced the target sound. Three movement conditions were considered: no head movement, rotation of the midline toward the direction of the target, and rotation away from the target. Rotation was at 5 degrees/seconds for 2 seconds. The robot completed 24 trials for sounds from each box. From the allocentric posterior map described above, we extracted the azimuth of maximum probability as it changed over time and expressed this as percentage of time points at which the max probability corresponded to the target sound location. The results in Fig. 2 show that the accuracy depends on both time and direction of head rotation: In all conditions accuracy improves quickly from chance but stabilized at 100% only when the head rotated toward the target. Other aspects of the system were stable. The system detected the sound almost instantaneously and the motor actions were performed correctly.

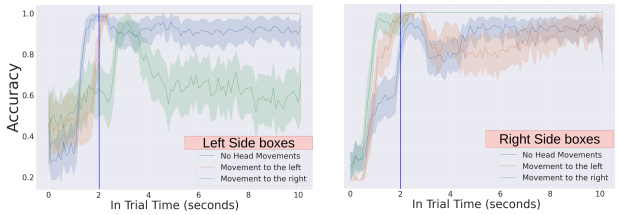

Figure 2: Average accuracy & std in the three motion conditions. Blue lines indicate the end time of the movement.

4 DISCUSSION

We proposed a framework to empower the robot with motor actions with the target of enhancing the audio perception of the robot. Demonstrating our concept practically, for example, the model can be applied in industrial applications that use an active auditory signal to promote interaction between the industrial robot and the human operator. Our work will directly improve the interaction with the human by better accurate audio localization for the sound signal in the working environment as the human localize. We tested our implementation on the iCub robot and explored the effect of specific head movements: the azimuthal rotation of the robot head (yaw) in two directions, relative to static pose. Overall the frameworks worked well. The audio power based trigger successfully captured the starting of the sound signal, both signal and action state was tracked in the working memory component correctly, and both the action and reset to the home position were executed in the correct timing. Regarding the performance, the experiment showed encouraging results about the use of prior information and motor actions : the system converged from chance accuracy (25%) to near perfect accuracy (100%) within about 2 seconds, particularly for the optimal motion strategy. Interestingly, during the initial seconds of the trial, the solution tended to be less stable, and the static pose also improved considerably from chance. This means that time - and not head movements - is sufficient for a Bayesian system to achieve relatively good localization in this task. However, beyond about two seconds, head rotation toward, but not away from the target was necessary to achieve stable 100% accuracy. Head rotations away from the target were not as useful, suggesting that other factors such the speed of the motor action and the initial orientation of head before sound onset might also be consequential.

This study provides insight about how auditory Artificial Intelligence might provide relevant details about the world and human partners in the HRI context. Given that complex human-robot interactions might present conflicting requirements for the robot to orient its microphones both toward a target (for optimal localization) and away from a target (in response to an instruction or to accomplish some other behaviour) it will be interesting to explore the integration of various motor behaviours into active auditory perception. More importantly the evidence that motor commands improve auditory perception in humans give us confidence that exploiting proactive behaviour for robots will help improving the next generation of perceptual skills in interactive scenarios.

REFERENCES

[1] David Alais and David Burr. 2004. The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology* 14, 3 (2004), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

[2] Sylvain Argentieri, Patrick Danes, and Philippe Souères. 2015. A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech & Language* 34, 1 (2015), 87–112.

[3] C. Baumann, C. Rogers, and F. Massen. 2015. Dynamic binaural sound localization based on variations of interaural time delays and system rotations. *Journal of the Acoustical Society of America* 138, 2 (2015), 635–650. https://doi.org/10.1121/1.4923448

[4] E. Blauert. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization.* The MIT Press.

[5] D. A. Hambrook, M. Ilievski, M. Mosadeghzad, and M. S. Tata. 2017. A Bayesian computational basis for auditory selective attention using head rotation and the interaural time-difference cue. *PLoS One* 12, 10 (2017). https://doi.org/10.1371/journal.pone.0186104

[6] Austin Kothig, Marko Ilievski, Lukas Grasse, Francesco Rea, and Matthew Tata. 2019. A Bayesian System for Noise-Robust Binaural Sound Localisation for Humanoid Robots. In *2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE).* IEEE, 1–7.

[7] M. Kumon and S. Uozumi. 2011. Binaural Localization for a Mobile Sound Source. *Journal of Biomechanical Science and Engineering* 6, 1 (2011), 26 – 39.

[8] H. McGurk and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (1976), 746–8. https://doi.org/10.1038/264746a0

[9] Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco Nori. 2008. The iCub humanoid robot: an open platform for research in embodied cognition.. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems.* https://doi.org/10.1145/1774674.1774683

[10] Caleb Rascon and Ivan Meza. 2017. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems* 96 (2017), 184–210. https://doi.org/10.1016/j.robot.2017.07.011

[11] Francesco Rea, Austin Kothig, Lukas Grasse, and Matthew Tata. 0. Speech Envelope Dynamics for Noise-Robust Auditory Scene Analysis in Robotics. *International Journal of Humanoid Robotics* 0, 0 (0), 2050023. https://doi.org/10.1142/S0219843620500231 arXiv:https://doi.org/10.1142/S0219843620500231

[12] H. Wallach. 1939. On Sound Localization. *Journal of the Acoustical Society of America* 10 (1939), 270–274.

[13] H. Wallach. 1940. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology* 27, 4 (1940), 339 – 368.