frontiers | Frontiers in Robotics and AI

TECHNOLOGY AND CODE
published: 11 May 2022
doi: 10.3389/frobt.2022.854444

ODAS: Open embeddeD Audition System

François Grondin, Dominic Létourneau, Cédric Godin, Jean-Samuel Lauzon, Jonathan Vincent, Simon Michaud, Samuel Faucher and François Michaud*

IntRoLab, Department of Electrical Engineering and Computer Engineering, Université de Sherbrooke, Sherbrooke, QC, Canada

Artificial audition aims at providing hearing capabilities to machines, computers and robots. Existing frameworks in robot audition offer interesting sound source localization, tracking and separation performance, although involve a significant amount of computations that limit their use on robots with embedded computing capabilities. This paper presents ODAS, the Open embeddeD Audition System framework, which includes strategies to reduce the computational load and perform robot audition tasks on low-cost embedded computing systems. It presents key features of ODAS, along with cases illustrating its uses in different robots and artificial audition applications.

Keywords: robot audition, sound source localization, microphone array, embedded computing, open source framework

OPEN ACCESS

Edited by:
Séverin Lemaignan,
Pal Robotics S.L., Spain

Reviewed by:
Akira Taniguchi,
Ritsumeikan University, Japan
Ha Manh Do,
University of Louisville, United States

****Correspondence:***
François Grondin
francois.grondin2@usherbrooke.ca

Specialty section:
This article was submitted to Computational Intelligence in Robotics, a section of the journal Frontiers in Robotics and AI

Received: *14 January 2022*
Accepted: *20 April 2022*
Published: *11 May 2022*

Citation:
Grondin F, Létourneau D, Godin C, Lauzon J-S Vincent J, Michaud S, Faucher S and Michaud F (2022) ODAS: Open embeddeD Audition System. Front. Robot. AI 9:854444. doi: 10.3389/frobt.2022.854444

1 INTRODUCTION

Similarly to artificial/computer vision, artificial/computer audition can be defined as the ability to provide hearing capabilities to machines, computers and robots. Vocal assistants on smart phones and smart speakers are now common, providing a vocal interface between people and devices Hoy (2018). As for artificial vision, there are still many problems to resolve for endowing robots with adequate hearing capabilities, such as ego and non-stationary noise cancellation, mobile and distant speech and sound understanding Ince et al. (2011), Deleforge and Kellermann (2015), Rascon et al. (2018), Schmidt et al. (2018), Shimada et al. (2019).

Open source software frameworks, such as OpenCV Culjak et al. (2012) for vision and ROS Quigley et al. (2009) for robotics, greatly contribute in making these research fields evolve and progress, allowing the research community to share and mutually benefit from collective efforts. In artificial audition, two main frameworks exist:

- HARK (Honda Research Institute Japan Audition for Robots with Kyoto University[1]) provides multiple modules for sound source localization and separation Nakadai et al. (2008), Nakadai et al. (2010), Nakadai et al. (2017b). This framework is mostly built over the FlowDesigner software Côté et al. (2004), and can also be interfaced with speech recognition tools such as Julius Lee and Kawahara (2009) and Kaldi Povey et al. (2011); Ravanelli et al. (2019). HARK implements sound source localization in 2-D using variants of the Multiple Signal Classification (MUSIC) algorithm Ishi et al. (2009); Nakamura et al. (2009), Nakamura et al. (2012). HARK also performs geometrically-constrained higher-order decorrelation-based source separation with adaptive step-size controlOkuno and Nakadai (2015). Though HARK supports numerous signal processing methods, it requires a significant amount of computing power (in part due to

[1]https://www.hark.jp/

the numerous eigenvalue decompositions required by MUSIC), which makes it less suitable for use on low-cost embedding hardware. For instance, when using HARK with a drone equipped with a microphone array to perform sound source localization, the raw audio streams needs to be transferred on the ground to three laptops for processing Nakadai et al. (2017a).

- ManyEars[2] is used with many robots to perform sound localization, tracking and separation Grondin et al. (2013). Sound source localization in 3-D relies on Steered-Response Power with phase Transform (SRP-PHAT), and tracking is done with particle filters Valin et al. (2007). ManyEars also implements the Geometric Sound Separation (GSS) algorithm to separate each target sound source Parra and Alvino (2002); Valin et al. (2004). This framework is coded in the C language to speed up computations, yet it remains challenging to run all algorithms simultaneously on low-cost embedding hardware such as a Digital Signal Processor (DSP) chip Briere et al. (2008).

Although both frameworks provide useful functionalities for robot audition tasks, they require a fair amount of computations. There is therefore a need for a new framework providing artificial audition capabilities in real-time and running on low-cost hardware. To this end, this paper presents ODAS[3] (Open embeddeD Audition System), improving on the ManyEars framework by using strategies to optimize processing and performance. The paper is organized as follows. **Section 2** presents ODAS' functionalities, followed by **Section 3** with configuration information of the ODAS library. **Section 4** describes the use of the framework in different applications.

2 OPEN EMBEDDED AUDITION SYSTEM

As for ManyEars Grondin et al. (2013), ODAS audio pipeline consists of a cascade of three main modules–localization, tracking and separation–plus a web interface for data vizualisation. The ODAS framework also uses multiple I/O interfaces to get access to raw audio data from the microphones, and to return the potential directions of arrival (DOAs) generated by the localization module, the tracked DOAs produced by the tracking module and the separated audio streams. ODAS is developed using the C programming language, and to maximize portability it only has one external dependency to the well-known third-party FFTW3 library (to perform efficient Fast Fourier Transform) Frigo and Johnson (2005).

Figure 1 illustrates the audio pipeline and the I/O interfaces, each running in a separate thread to fully exploit processors with multiple cores. Raw audio can be provided by a pre-recorded multi-channel RAW audio file, or obtained directly from a chosen sound card connected to microphones for real-time processing. The Sound Souce Localization (SSL) module generates a fixed number of potential DOAs, which are fed to the Sound Source Tracking (SST) module. SST identifies tracked sources, and these DOAs are used by the Sound Source Separation (SSS) module to perform beamforming on each target sound source. DOAs can also be sent in JSON format to a terminal, to a file or to a TCP/IP socket. The user can also define fixed target DOA(s) if the direction(s) of the sound source(s) is/are known in advance and no localization and no tracking is required. The beamformed segments can be written in RAW audio files, or also sent via a socket.

The ODAS Studio Web Interface, shown in **Figure 2**, can be used to visualize the potential and tracked DOAs, and also to get the beamformed audio streams. This interface can run on a separate computer connected to ODAS via sockets. The interface makes it possible to visualize the potential DOAs in three dimensions on a unit sphere with a color code that stands for their respective power, and in scatter plots of azimuths and elevations as a function of time. The tracked sources are also displayed in the azimuth/elevation plots, as continuous lines with a unique color per tracked source.

ODAS relies on many strategies to reduce the computational load for the SSL, SST and SSS modules, described as follows.

2.1 Sound Source Localization

ODAS exploits the microphone array geometry to perform localization, defined at start-up in a configuration file. In addition to the position, the orientation of each microphone also provides useful information when microphones lie in a closed array configuration (e.g., when installed on a robot head or torso). While microphones are usually omnidirectional, they can be partially hidden by some surfaces, which make their orientation relevant. Localization relies on the Generalized Cross-Correlation with phase Transform method (GCC-PHAT), computed for each pair of microphones. ODAS uses the inverse Fast Fourier Transform (IFFT) to compute the cross-correlation efficiently from the signals in the frequency domain. When dealing with small arrays, ODAS can also interpolate the cross-correlation signal to improve localization accuracy and to cope with the TDOA discretization artifact introduced by the IFFT. While some approaches rely on the Head-Related Transfer Function (HRTF) to deal with closed array Nakadai et al. (2003), ODAS proposes a simpler model that provides accurate localization and reduce the computational load. In fact, the framework exploits the directivity of microphones to only compute GCC-PHAT between pairs of microphones that can be simultaneously excited by the direct propagation path of a sound source. To illustrate this, **Figure 3** shows an 8-microphone closed array, for which it is assumed that all microphones point outside with a field of view of 180°. Because microphones on opposite sides cannot capture simultaneously the sound wave coming from a source around the array, their cross-correlation can be ignored. Consequently, ODAS computes GCC-PHAT for the pairs of microphones connected with green lines only. With such a closed array configuration, the pairs connected with red lines are ignored as there is no potential direct path for sound waves. Therefore,

[2]https://github.com/introlab/manyears

[3]http://odas.io

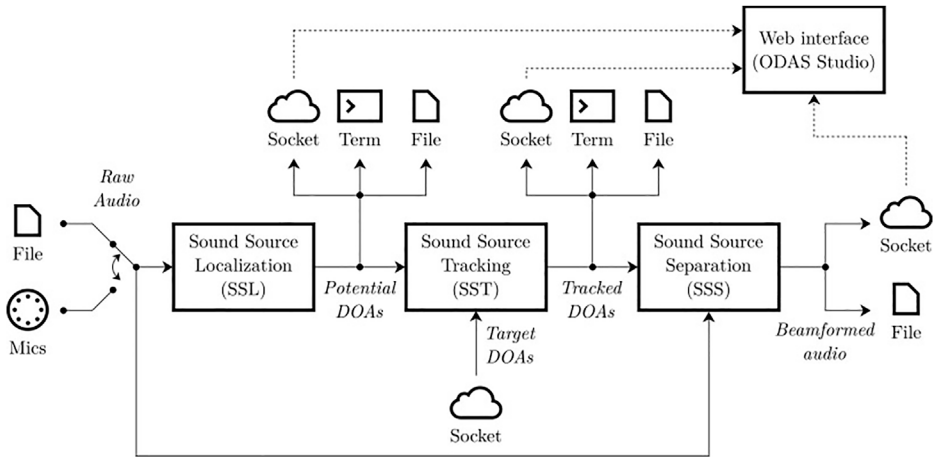

FIGURE 1 | ODAS processing pipeline.

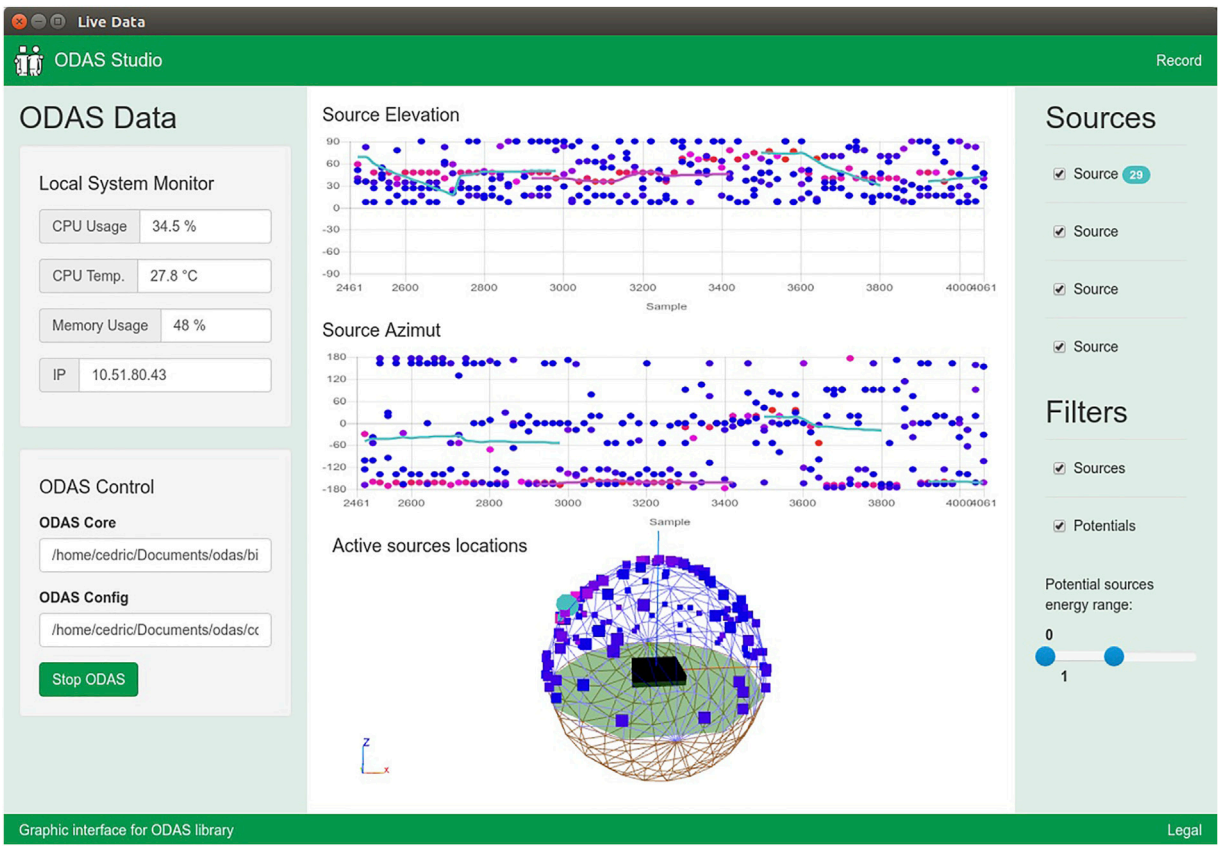

FIGURE 2 | ODAS Studio Web Interface. Colored dots represent potential DOA with power levels, and solid lines illustrate the tracked sound source trajectories over time.

ODAS computes the cross-correlation between 20 pairs of microphones out of the 28 possible pairs. This simple strategy reduces the cross-correlation computational load by 29% (i.e., (28–20)/28). With open array configurations, ODAS uses all pairs of microphones because sound waves reach all microphones in such cases.

ManyEars computes the Steered-Response Power with phase Transform (SRP-PHAT) for all DOAs that lie on a unit sphere

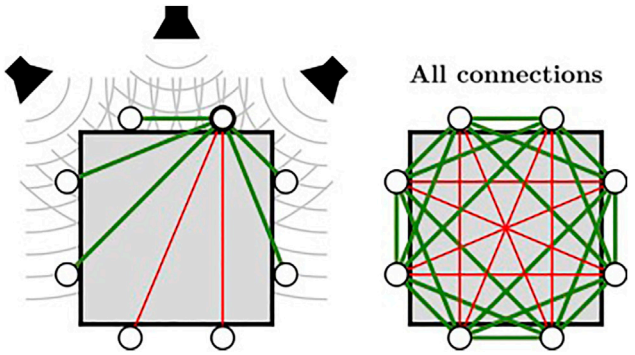

FIGURE 3 | ODAS strategy exploiting microphone directivity to compute GCC-PHAT using relevant pairs of microphones in a closed array configuration.

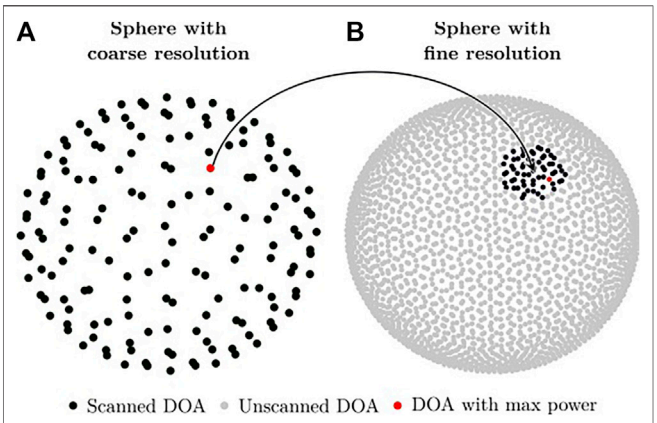

FIGURE 4 | Illustration of the two unit sphere search, first with coarse resolution **(A)**, and then more precise search with finer resolution **(B)**.

discretized with 2562 points. For each DOA, ManyEars computes the SRP-PHAT power by summing the value of the cross-correlation associated to the corresponding time difference of arrival (TDOA) obtained with GCC-PHAT for each pair of microphones, and returns the DOA associated to the highest power. Because there might be more than one active sound source at a time, the corresponding TDOAs are zeroed, and scanning is performed again to retrieve the next DOA with the highest power. These successive scans are usually repeated to generate up to four potential DOAs. However, scanning each point on the unit sphere involves numerous memory accesses that slow down processing. To speed it up, ODAS uses instead two unit spheres: 1) one with a coarse resolution (made of 162 discrete points) and 2) one with a finer resolution (made of 2562 discrete points). ODAS first scans all DOAs in the coarse sphere, finds the one associated to the maximum power, and then refines the search on a small region around this DOA on the fine sphere Grondin and Michaud (2019). **Figure 4** illustrates this process, which reduces considerably the number of memory accesses while providing a similar DOA estimation accuracy. For instance, when running ODAS on a Raspberry Pi 3, this strategy reduces

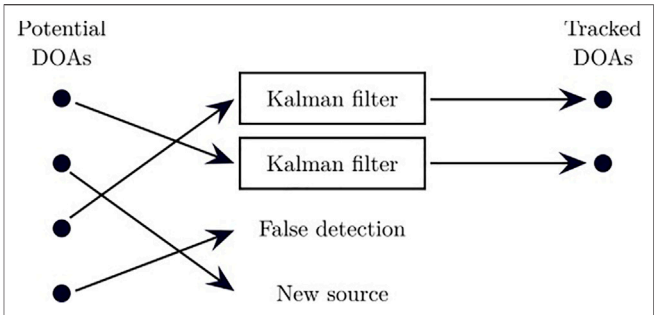

FIGURE 5 | Tracking sound sources with Kalman filters: each DOA is associated to a previously tracked source, a false detection or a new source.

CPU usage for performing localization using a single core for an 8-microphone array by almost a factor of 3 (from a single core usage of 38% down to 14%) Grondin and Michaud (2019). Note that when all microphones lie in the same plane in 2-D, ODAS scans only a half unit sphere, which also reduces the number of computations.

2.2 Sound Source Tracking

Sound sources are non-stationary and sporadically active over time. Tracking therefore provides a continuous trajectory in time for the DOA of each sound source, and can cope with short silence periods. This module also detects newly active sound sources, and forgets sources that are inactive for a long period of time. Sound source localization provides one or many potential DOAs, and the tracking maps each observation either to a previously tracked source, to a new source, or to a false detection. To deal with static and moving sound sources, ManyEars also relies on a particle filter to model the dynamics of each source Grondin et al. (2013). The particles of each filter are associated to three possible states: 1) static position, 2) moving with a constant velocity, 3) accelerating. This approach however involves a significant amount of computations, as the filter is usually made of 1,000 particles, and each of them needs to be individually updated. Briere et al. (2008) proposed to reduce the number of particles by half to bring down the computational load. Experiments however demonstrated that this impacts accuracy when tracking two moving sound sources. Instead ODAS uses Kalman filter for each tracked source, as illustrated by **Figure 5**. Results demonstrate similar tracking performance, with a significant reduction in computational load on a Raspberry Pi 3 device (e.g., by a factor of 30, from a single core usage of 24% down to 0.8% when tracking one source, and by a factor of 14, from a single core usage of 98% down to 7% when tracking four sources Grondin and Michaud (2019).

2.3 Sound Source Separation

ODAS supports two sound source separation methods: 1) delay-and-sum beamforming, and 2) geometric sound source separation. These methods are similar to the former methods implemented in ManyEars Grondin et al. (2013), with the exception that ODAS also considers the orientation of each microphone. Because ODAS estimates the DOA of each sound

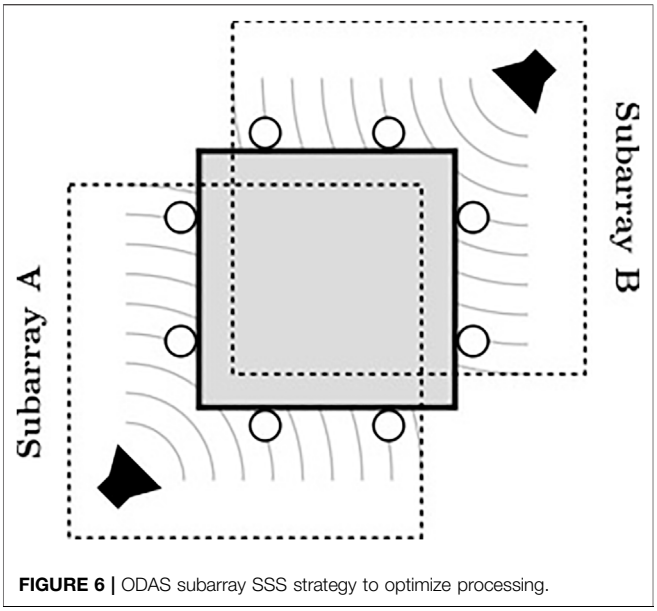

FIGURE 6 | ODAS subarray SSS strategy to optimize processing.

source, it can select only the microphones oriented in the direction of the target source for beamforming. For a closed array configuration, this improves separation performance [e.g., when using a delay-and-sum beamformer, this can result in a SNR increase of 1 dB when compared to using all microphones Grondin (2017)], while it reduces the amount of computations. **Figure 6** presents an example with two sound sources around a closed array, where ODAS performs beamforming with the microphones on the left and bottom to retrieve the signal from the source on the left, and performs beamforming with the microphones on the right and top to retrieve the signal from the source on the right. Subarray A only uses four out of the eight microphone, and subarray B uses the other four microphones. This reduces the amount of computations and also improves separation performance.

ODAS also implements the post-filtering approach formerly introduced in ManyEars Grondin et al. (2013). Post-filtering aims to improve speech intelligibility by masking time-frequency components dominated by noise and/or competing sound source Valin et al. (2004).

3 CONFIGURING THE OPEN EMBEDDED AUDITION SYSTEM LIBRARY

ODAS relies on a custom configuration file that holds all parameters to instantiate the modules in the processing pipeline, with some parameters determined by the microphone array hardware. There are some useful guidelines to follow when the microphone array geometry can be customized: 1) the microphones should span all x, y and z dimensions to localize sound sources in the full elevation and azimuth ranges. When microphones only span two dimensions, the localization is limited to a half sphere; 2) the microphones should be a few tens of centimeters apart. Increasing the microphone array aperture provides better discrimination in low-frequencies, which is well-suited for speech; 3) using a directivity model for closed arrays with microphones installed on rigid surface further reduces the computational load and improves accuracy. The structure of each file obeys the configuration format, which is compact and easy to read[4]. The configuration file is divided in many sections:

3.1 Raw Input

This section indicates the format of the RAW audio samples provided by the sound card or the pre-recorded file. It includes the sample rate, the number of bits per sample (assuming signed numbers), the number of channels and the buffer size (in samples) for reading the audio.

3.2 Mapping

The mapping selects which channels are used as inputs to ODAS. In fact, some sound cards have additional channels (e.g., for playback) and it is therefore convenient to extract only the meaningful channels. Moreover, this option allows a user to ignore some microphones if desired to reduce computational load.

3.3 General

This section provides general parameters that are used by all the modules in ODAS' pipeline. It includes the short-time Fourier Transform (STFT) frame size and hop length (since all processing is performed in the frequency domain). It also provides a processing sample rate, which can differ from the initial sample rate from RAW audio in the sound card (ODAS can resample the RAW signal to match the processing sample rate). The speed of sound is also provided, along with some uncertainty parameter, to cope with different speeds of sound. All microphone positions are also defined, along with their orientation. It is also possible to incorporate position uncertainty to make localization more robust to measurement errors when dealing with microphone arrays of arbitrary shape.

3.4 Stationary Noise Estimation

ODAS estimates the background noise using the minima controlled recursive averaging (MCRA) method Cohen and Berdugo (2002), to make localization more robust and increase post-processing performances. This section provides the MCRA parameters to be used by ODAS.

3.5 Localization

The localization section provides parameters to fine tune the SSL module. These parameters are usually the same for all setups, except for the interpolation rate which can be increased when dealing with small microphone arrays to cope with the discretization artifact introduced by the IFFT when computing GCC-PHAT.

[4]http://hyperrealm.github.io/libconfig/

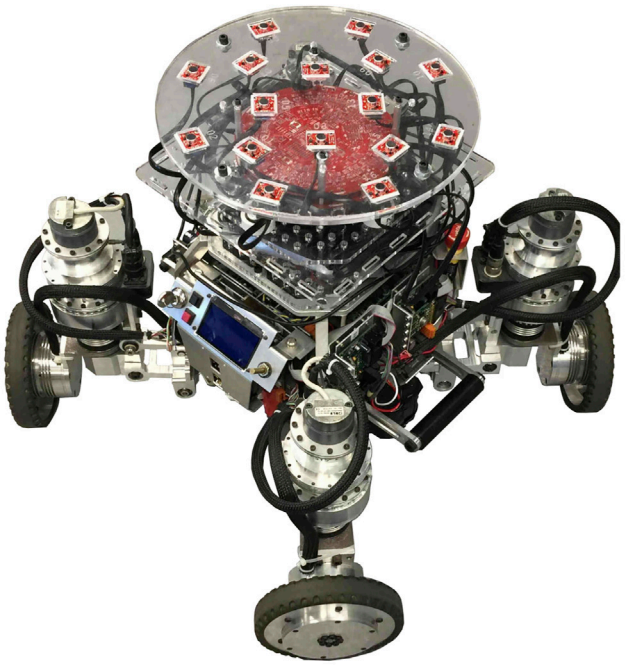

FIGURE 7 | Azimut-3 (open array configuration, 16 microphones).

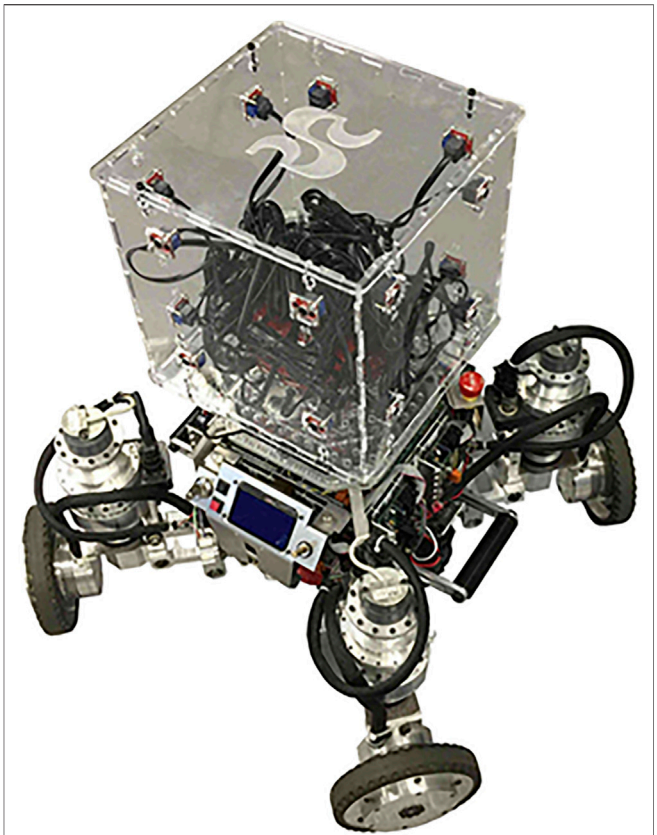

FIGURE 8 | Azimut-3 (closed array configuration, 16 microphones).

3.6 Tracking

ODAS can support tracking with either the former particle filter method, or the current Kalman filter approach. Most parameters in this section relate to the methods introduced in Grondin and Michaud (2019). It is worth mentioning that ODAS represents the power distribution for a DOA generated by the SSL module as a Gaussian Mixture Model (GMM). Another GMM also models the power of diffuse noise when all sound sources are inactive. It is therefore possible to measure both distributions experimentally using histograms and then fit them with GMMs.

3.7 Separation

This section of the configuration file defines ODAS which separation method to use (Delay-and-sum or Geometric Source Separation). It also provides parameters to perform post-filtering, and information regarding the audio format of the separated and post-filtered sound sources.

A configuration file can be provided for each commercial microphone array or each robot with a unique microphone array geometry, and to use ODAS for processing.

4 APPLICATIONS

ODAS' optimized processing strategies makes it possible to perform all processing on low-cost hardware, such as a Raspberry Pi 3 board. **Figures 7–10** present some of the robots using the ODAS framework for sound source localization, tracking and separation. ODAS is used with the Azimut-3 robot, with two different microphone array configurations Grondin and Michaud (2019): 16 microphones lying on a single plane, or with all microphones installed around a cubic shape on top of the robot. For both setups, the sound card 16SoundsUSB[5] performs signal acquisitions and then ODAS performs SSL, SST and SSS. **Figure 9** illustrates the 16 microphones configuration on the SecurBot robots Michaud et al. (2020). ODAS is also used with the Beam robot Laniel et al. (2017), placing eight microphones on the same plane and using the 8SoundsUSB sound card[6]. ODAS exploits the directivity of microphones for setups with the Azimut-3 robot (closed array configuration) and SecurBot, which reduces the amount of computations. On the other hand, ODAS searches for DOAs on a half sphere for the Azimut-3 robot (open array configuration) and the Beam robot, as all microphones lie on the same 2-D plane. The proposed framework is also used with the T-Top robot Maheux et al. (2022), which base is equipped with 16 microphones. The Open-Tera Panchea et al. (2022) microservice architecture solution for telepresence robots also uses the ODAS framework.

ODAS is also used for drone localization using three static 8-microphone arrays disposed on the ground at the vertex of an equilateral triangle, with edges of 10 m Lauzon et al. (2017). DOA search is performed on a half-sphere over the microphone array

[5]https://github.com/introlab/16SoundsUSB
[6]https://sourceforge.net/p/eightsoundsusb/wiki/Main_Page/

FIGURE 9 | SecurBot (16-microphone configuration on top and sides).

plane. Each microphone array is connected to a Raspberry Pi 3 through the 8SoundsUSB sound card, which runs ODAS and returns potential DOAs to a central node. The central node then performs triangulation to estimate the 3D position of the flying drone.

ODAS is also extensively used with smaller commercial microphone arrays, for sound source localization or as a preprocessing step prior to speech recognition and other recognition tasks. These are usually small circular microphone arrays, where the number of microphones varies between 4 and 8. ODAS is referenced on Seeed Studio ReSpeaker USB 4-Mic Array official website as a compatible framework with their hardware[7]. The Matrix Creator board has numerous sensors, including eight microphones on its perimeters, and has online tutorials showing how to use ODAS with the array[8]. ODAS was also validated with the miniDSP UMA-8 microphone array[9], and the XMOS xCore 7-microphone array[10]. For all circular arrays, ODAS searches for DOAs on a half-sphere, and also interpolates the cross-correlation results to improve accuracy since microphones are only a few centimeters apart.

Configuration files with the exact positions of all microphones for each device are available online with the source code.

[7]https://respeaker.io/4_mic_array/

[8]https://www.youtube.com/watch?v=6ZkZYmLA4xw

[9]https://www.minidsp.com/aboutus/newsletter/\\listid-1/mailid-68-minidsp-newsletter-an-\\exciting-new-chapter

[10]https://www.youtube.com/watch?v=n7y2rLAnd5I

FIGURE 10 | Beam (8-microphone on a circular support).

5 CONCLUSION

This paper introduces ODAS, the Open embeddeD Audition System framework, explaining its strategies for real-time and embedded processing, and demonstrates how it can be used for various applications, including robot audition, drone localization and voice assistants. ODAS' strategies to reduce computational load, consist of: 1) partial cross-correlation computations using the microphone directivity model, 2) DOA search on coarse and fine unit spheres, 3) search on a half sphere when all microphones lie on the same 2-D plane, 4) tracking active sound sources with Kalman filters and 5) beamforming with subarrays using simple microphone directivity models. In addition to use cases found in the literature, ODAS source code has been accessed more than 55,000 times, which suggests that there is a need for a framework for robot audition that can run on embedded computing systems.

ODAS can also be part of the solution for edge-computing for voice recognition to avoid cloud computing and preserve privacy.

While ODAS performs well in quiet or moderately noisy environments, its robustness should be improved for more challenging environments (noisy conditions, strong reflections, important ego-noise, etc.). In future work, additional functionalities will be added to ODAS, including new algorithms that rely on deep learning based methods, as machine learning has became a powerful tool when combined with digital signal processing for sound source localization Chakrabarty and Habets (2019), speech enhancement Valin (2018); Valin et al. (2020) and sound source classification Ford et al. (2019); Grondin et al. (2019). Additional beamforming methods could also be implemented, including Minimum Variance Distortionless Response (MVDR) beamformer Habets et al. (2009), and generalized eigenvalue (GEV) beamforming Heymann et al. (2015); Grondin et al. (2020), as these approaches are particularly suited for preprocessing before automatic speech recognition. ODAS would also benefit from ego-noise suppression algorithms to mitigate the impact of motor noise while doing sound source localization, tracking and separation.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/introlab/odas.

AUTHOR CONTRIBUTIONS

FG designed and wrote the ODAS framework C library. DL assisted with maintenance and integration of the library on all robotics platforms. CG designed the ODAS studio web interface. J-SL and JV debugged the ODAS library. SM and SF worked on integrating the framework in ROS. FM supervised and led the team.

FUNDING

This work was supported by FRQNT—Fonds recherche Québec Nature et Technologie.

REFERENCES

Brière, S., Valin, J.-M., Michaud, F., and Létourneau, D. (2008). "Embedded Auditory System for Small Mobile Robots," in Proceedings of the IEEE International Conference on Robotics and Automation, Pasadena, CA, May 19–23, 2008, 3463–3468. doi:10.1109/robot.2008.4543740

Chakrabarty, S., and Habets, E. A. P. (2019). Multi-Speaker DOA Estimation Using Deep Convolutional Networks Trained with Noise Signals. *IEEE J. Sel. Top. Signal Process.* 13, 8–21. doi:10.1109/jstsp.2019.2901664

Cohen, I., and Berdugo, B. (2002). Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement. *IEEE Signal Process. Lett.* 9, 12–15. doi:10.1109/97.988717

Côté, C., Létourneau, D., Michaud, F., Valin, J.-M., Brosseau, Y., Raievsky, C., et al. (2004). "Code Reusability Tools for Programming Mobile Robots," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan, September 28–October 2, 2004, 1820–1825.

Culjak, I., Abram, D., Pribanic, T., Dzapo, H., and Cifrek, M. (2012). "A Brief Introduction to OpenCV," in Proceedings of the International Convention on Information, Communication and Electronic Technology, Opatija, Croatia, May 21–25, 2012, 1725–1730.

Deleforge, A., and Kellermann, W. (2015). "Phase-Optimized K-SVD for Signal Extraction from Underdetermined Multichannel Sparse Mixtures," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, South Brisbane, QLD, April 19–24, 2015, 355–359. doi:10.1109/icassp.2015.7177990

Ford, L., Tang, H., Grondin, F., and Glass, J. R. (2019). "A Deep Residual Network for Large-Scale Acoustic Scene Analysis," in Proceedings of Interspeech, Graz, Austria, September 15–19, 2019, 2568–2572. doi:10.21437/interspeech.2019-2731

Frigo, M., and Johnson, S. G. (2005). The Design and Implementation of FFTW3. *Proc. IEEE* 93, 216–231. doi:10.1109/jproc.2004.840301

Grondin, F., Glass, J., Sobieraj, I., and Plumbley, M. D. (2019). "Sound Event Localization and Detection Using CRNN on Pairs of Microphones," in Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, New York City, NY, October 25–26, 2019.

Grondin, F., Lauzon, J.-S., Vincent, J., and Michaud, F. (2020). "GEV Beamforming Supported by DOA-Based Masks Generated on Pairs of Microphones," in Proceedings of Interspeech, Shanghai, China, October 25–29, 2020, 3341–3345. doi:10.21437/interspeech.2020-2687

Grondin, F., Létourneau, D., Ferland, F., Rousseau, V., and Michaud, F. (2013). The ManyEars Open Framework. *Auton. Robot.* 34, 217–232. doi:10.1007/s10514-012-9316-x

Grondin, F., and Michaud, F. (2019). Lightweight and Optimized Sound Source Localization and Tracking Methods for Open and Closed Microphone Array Configurations. *Robotics Aut. Syst.* 113, 63–80. doi:10.1016/j.robot.2019.01.002

Grondin, F. (2017). Système d'audition artificielle embarqué optimisé pour robot mobile muni d'une matrice de microphones. Ph.D. thesis. Sherbrooke, QC: Université de Sherbrooke.

Habets, E. A. P., Benesty, J., Cohen, I., Gannot, S., and Dmochowski, J. (2009). New Insights into the MVDR Beamformer in Room Acoustics. *IEEE Trans. Audio, Speech, Lang. Process.* 18, 158–170. doi:10.1109/TASL.2009.2024731

Heymann, J., Drude, L., Chinaev, A., and Haeb-Umbach, R. (2015). "BLSTM Supported GEV Beamformer Front-End for the 3rd CHiME Challenge," in Proceedings IEEE Automatic Speech Recognition and Understanding Workshop, Scottsdale, AZ, December 13–17, 2000, 444–451. doi:10.1109/asru.2015.7404829

Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Med. Ref. Serv. Q.* 37, 81–88. doi:10.1080/02763869.2018.1404391

Ince, G., Nakamura, K., Asano, F., Nakajima, H., and Nakadai, K. (2011). "Assessment of General Applicability of Ego Noise Estimation," in Proceedings of the IEEE International Conference on Robotics and Automation, Sanghai, China, May 9–13, 2011, 3517–3522. doi:10.1109/icra.2011.5979578

Ishi, C. T., Chatot, O., Ishiguro, H., and Hagita, N. (2009). "Evaluation of a MUSIC-Based Real-Time Sound Localization of Multiple Sound Sources in Real Noisy Environments," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, October 10–15, 2009, 2027–2032. doi:10.1109/iros.2009.5354309

Laniel, S., Letourneau, D., Labbe, M., Grondin, F., Polgar, J., and Michaud, F. (2017). Adding Navigation, Artificial Audition and Vital Sign Monitoring Capabilities to a Telepresence Mobile Robot for Remote Home Care Applications. *IEEE Int. Conf. Rehabil. Robot.* 2017, 809–811. doi:10.1109/ICORR.2017.8009347

Lauzon, J.-S., Grondin, F., Létourneau, D., Desbiens, A. L., and Michaud, F. (2017). "Localization of RW-UAVs Using Particle Filtering over Distributed Microphone Arrays," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Vancouver, BC, September 24–28, 2017, 2479–2484. doi:10.1109/iros.2017.8206065

Lee, A., and Kawahara, T. (2009). "Recent Development of Open-Source Speech Recognition Engine Julius," in Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference, Sapporo, Japan, October 4–7, 2009, 131–137.

Maheux, M.-A., Caya, C., Létourneau, D., and Michaud, F. (2022). "T-Top, a SAR Experimental Platform," in Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction, 904–908.

Michaud, S., Faucher, S., Grondin, F., Lauzon, J.-S., Labbé, M., Létourneau, D., et al. (2020). "3D Localization of a Sound Source Using Mobile Microphone Arrays Referenced by SLAM," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, October 25–29, 2020, 10402–10407. doi:10.1109/iros45743.2020.9341098

Nakadai, K., Kumon, M., Okuno, H. G., Hoshiba, K., Wakabayashi, M., Washizaki, K., et al. (2017a). "Development of Microphone-Array-Embedded UAV for Search and Rescue Task," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Vancouver, BC, September 24–28, 2017, 5985–5990. doi:10.1109/iros.2017.8206494

Nakadai, K., Matsuura, D., Okuno, H. G., and Kitano, H. (2003). "Applying Scattering Theory to Robot Audition System: Robust Sound Source Localization and Extraction," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, October 27–31, 2003, 2, 1147–1152.

Nakadai, K., Okuno, H. G., Nakajima, H., Hasegawa, Y., and Tsujino, H. (2008). "An Open Source Software System for Robot Audition HARK and its Evaluation," in Proceedings of the IEEE-RAS International Conference on Humanoid Robots, Daejeon, South Korea, December 1–3, 2008, 561–566. doi:10.1109/ichr.2008.4756031

Nakadai, K., Okuno, H. G., Okuno, H. G., and Mizumoto, T. (2017b). Development, Deployment and Applications of Robot Audition Open Source Software HARK. *J. Robot. Mechatron.* 29, 16–25. doi:10.20965/jrm.2017.p0016

Nakadai, K., Takahashi, T., Okuno, H. G., Nakajima, H., Hasegawa, Y., and Tsujino, H. (2010). Design and Implementation of Robot Audition System 'HARK' - Open Source Software for Listening to Three Simultaneous Speakers. *Adv. Robot.* 24, 739–761. doi:10.1163/016918610x493561

Nakamura, K., Nakadai, K., Asano, F., Hasegawa, Y., and Tsujino, H. (2009). "Intelligent Sound Source Localization for Dynamic Environments," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, October 10–15, 2009, 664–669. doi:10.1109/iros.2009.5354419

Nakamura, K., Nakadai, K., and Ince, G. (2012). "Real-Time Super-Resolution Sound Source Localization for Robots," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal, October 7–12, 2012, 694–699. doi:10.1109/iros.2012.6385494

Okuno, H. G., and Nakadai, K. (2015). "Robot Audition: Its Rise and Perspectives," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, South Brisbane, QLD, April 19–24, 2015, 5610–5614. doi:10.1109/icassp.2015.7179045

Panchea, A. M., Létourneau, D., Brière, S., Hamel, M., Maheux, M. A., Godin, C., et al. (2022). Opentera: A Microservice Architecture Solution for Rapid Prototyping of Robotic Solutions to COVID-19 Challenges in Care Facilities. *Health Technol. (Berl)* 12, 583–596. doi:10.1007/s12553-021-00636-5

Parra, L. C., and Alvino, C. V. (2002). Geometric Source Separation: Merging Convolutive Source Separation with Geometric Beamforming. *IEEE Trans. Speech Audio Process.* 10, 352–362. doi:10.1109/tsa.2002.803443

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). "The Kaldi Speech Recognition Toolkit," in Proceedings of the IEEE Automatic Speech Recognition and Understanding, Waikoloa, Hawaii, December 11–15, 2011.

Quigley, M., Conley, K., Gerkey, B. P., Faust, J., Foote, T., Leibs, J., et al. (2009). "ROS: An Open-Source Robot Operating System," in ICRA Workshop on Open Source Software, Kobe, Japan, May 12–17, 2009, 1–6.

Rascon, C., Meza, I. V., Millan-Gonzalez, A., Velez, I., Fuentes, G., Mendoza, D., et al. (2018). Acoustic Interactions for Robot Audition: A Corpus of Real Auditory Scenes. *J. Acoust. Soc. Am.* 144, EL399–EL403. doi:10.1121/1.5078769

Ravanelli, M., Parcollet, T., and Bengio, Y. (2019). "The PyTorch-Kaldi Speech Recognition Toolkit," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Brighton, UK, May 12–17, 2019, 6465–6469. doi:10.1109/icassp.2019.8683713

Schmidt, A., Löllmann, H. W., and Kellermann, W. (2018). "A Novel Ego-Noise Suppression Algorithm for Acoustic Signal Enhancement in Autonomous Systems," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Calgary, AB, April 15–20, 2018, 6583–6587. doi:10.1109/icassp.2018.8462211

Shimada, K., Bando, Y., Mimura, M., Itoyama, K., Yoshii, K., and Kawahara, T. (2019). Unsupervised Speech Enhancement Based on Multichannel Nmf-Informed Beamforming for Noise-Robust Automatic Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 27, 960–971. doi:10.1109/taslp.2019.2907015

Valin, J.-M. (2018). "A Hybrid Dsp/deep Learning Approach to Real-Time Full-Band Speech Enhancement," in Proceedings of the IEEE Workshop on Multimedia Signal Processing, Vancouver, BC, August 29–31, 2018, 1–5. doi:10.1109/mmsp.2018.8547084

Valin, J.-M., Isik, U., Phansalkar, N., Giri, R., Helwani, K., and Krishnaswamy, A. (2020). "A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech," in Proceedings of Interspeech, Shanghai, China, October 25–29, 2020, 2482–2486. doi:10.21437/interspeech.2020-2730

Valin, J.-M., Michaud, F., and Rouat, J. (2007). Robust Localization and Tracking of Simultaneous Moving Sound Sources Using Beamforming and Particle Filtering. *Robotics Aut. Syst.* 55, 216–228. doi:10.1016/j.robot.2006.08.004

Valin, J.-M., Rouat, J., and Michaud, F. (2004). "Enhanced Robot Audition Based on Microphone Array Source Separation with Post-filter," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan, September 28–October 2, 2004, 3, 2123–2128.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.