Improving HRI Capabilities for Serving Robots by Combining Visual and Aural Cues[1]

Hari Krishna Prannoy Namala

[a] *University of Maryland, College Park, 20740, Maryland, USA*

1. Introduction

Human body's sensing capabilities allow us to interact with the world and other humans. After millions of years of evolution, our senses have each become exemplary systems. The sensory systems of the human body have been an inspiration for developing sensory systems for robots. The inspirations from the human body have resulted in many breakthroughs in development of the sensory systems. Even with these breakthroughs, implementing human sensory systems is still an open research area.

Similar to humans, an autonomous robot interacting with humans on a regular basis needs to have well equipped sensor systems. Applications of robots with such capabilities are serving at events, waiting tables at restaurants, elderly care where robots can be helpful working alongside humans enabling humans to use the saved time to perform other tasks and interact with humans at levels in which a robot cannot.

The capabilities of humans to work in noisy environments has always been a topic of intrigue for autonomous robot researchers. Consider an example of waiters at restaurants, the noise levels at restaurants are high. Human waiters are able to navigate such environments without much difficulty most of the time. Achieving this level of autonomy is a really hard task in a robot. However, with the recent advancements in the sensory systems for robots, it is possible to integrate these senses to create a robot which can deal with lower levels of noise in the environment. The aim of this independent study is to use current state of the art sensory systems for robots and integrate

[1] This is the final report submission for ENPM808: Independent Study under Dr. Dinesh Manocha

them to create a sensory system which can interact better with humans.

This will be done by extending the previous work done on the COVID surveillance robot by Sathyamoorthy et al. (2021). In this work, the robot acts as a social distancing surveillance agent. When it sees a group of humans violating the social distancing requirement, the robot approaches to warn them. When not performing the social distance regulation, the robot acts as a serving robot. The robot randomly chooses a customer to attend to in its field of view to serve them items on the tray attached to its body. The aim is to improve on this aspect and interact better with humans by developing a system for the robot to respond to a customer's calls and attend to them. An identification system based on audio and visual input to help the robot estimate the direction of a customer's voice and identify the exact customer wanting the attention of the robot using body posture detection and existing visual tracking methods.

2. Related Work

The problem defined earlier is split into three components: Sound Source Localization, Wakeword Detection and human body posture detection. Due to the advancements in the fields of Natural Language processing, Autonomous Driving and Deep Learning, wakeword detection and human body posture detection have become mature topics with packages readily available to implement them. Some of the prominent techniques in sound source localization are discussed below.

There are a lot of approaches for sound source localization depending the kind of hardware and the approach used to detect the angle. One way is to use multiple microphones and inferred using various signal processing techniques. This approached was implemented in the work by Scola and Ortega (2010). In their work they approximate the location of a source using two microphones. Trigonometric techniques are used on the input from the microphones to triangulate the position of the source. Some other interesting approaches use sing microphone to localize the sound. Works from El Badawy and Dokmanić (2018) and from Saxena and Ng (2009) explore this technique. They try to construct a sturcture similar to that of the outer structure of the human ear. For estimating the direction they use machine learning techniques (Saxena and Ng, 2009) and matrix factorization techniques (El Badawy and Dokmanić, 2018). One of the more recent works from Adavanne et al. (2019a) and Adavanne et al. (2019b) use deep learning

to do various sound processing techniques like localization, event detection and so on. They use a convolutional recurrent neural network (CRNN) to give multiple outputs for single input file recorded from the microphones. Many of the techniques mentioned above are computationally demanding to train and use on an embedded system of a robot. So, a technique designed for embedded systems has been used in this study. This will be explained in more detail in the section 3.

3. Methods and Integration

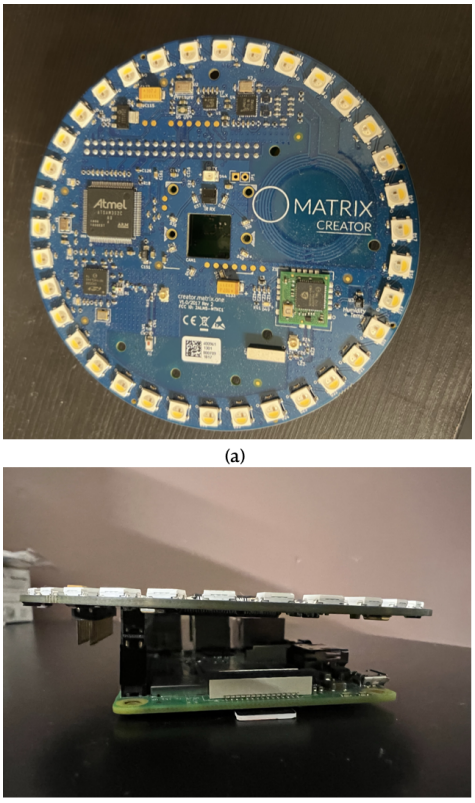
Sound Source Localization

Sound Source Localization has been an open research topic for a long time. As discussed above, there are several methods which have come up in this time. The technique of sound source localization used for this study is General Cross-Correlation with Phase Transform (GCC-PHAT). GCC-PHAT was first introduced by Joseph DiBisaeDiBiase (2000) in the year 2000. This work has been used in many implementations. In this work, we will be looking at Open Embedded Audition System (ODAS)Grondin et al. (2021). ODAS is a framework providing artificial audition capabilities in real-time while running on low-cost hardware. ODAS uses GCC-PHAT on multiple microphones arranged in a known configuration. These setups of multiple microphones is called a microphone array. The process of computing GCC-PHAT is optimised by using Inverse Fast Fourier Transforms (IFFT).

To run ODAS, we used MATRIX Creator. MATRIX Creator is a microphone array with 8 microphones arranged in a circular configuration. The microphone array is attached to the GPIO pins of a Raspberry Pi 3B with Raspbian Buster OS. MATRIX, the manufacturer of the Creator, also has an outdated version of ODAS implementation using their hardware abstraction layer (HAL). A script combining the HAL and ODAS library is run to get the output of the Direction of Angle(DOA) for an input sound from the sound source. A snapshot of the program working is in figure 2. The finger is pointed by a human in the direction in which they are speaking and blue lights on the speaker denote the estimated direction of the voice as given by the script.

Wakeword Detection

Wakeword detection has been a key part in the field of speech recognition. The field has approached its maturity as there are lot of applications for the



(a)

(b)

Figure 1: (a) Matrix Creator(b) MATRIX Creator attached to Raspberry Pi

wakeword detection. The most prominent one are the voice assistants like Siri, Alexa and Google Assistant. There are also open source voice assistants like Rhasspy and Mycroft.

The frameworks mentioned above have many other capabilities including wakeword detection. Therefore, running the voice recognition frameworks requires a lot of computational power. Since the sensory system is run on an embedded system(Raspberry Pi 3B), we cannot afford to lose compute for functionalities we may not require. To solve this problem, we use Porcupine Wakeword Engine from PicoVoice.

Porcupine is a highly-accurate and lightweight wake word engine. It enables building always-listening voice-enabled applications. It is using deep neural networks trained in real-world environments while being compact and computationally-efficient. Porcupine also gives the developers the ability to

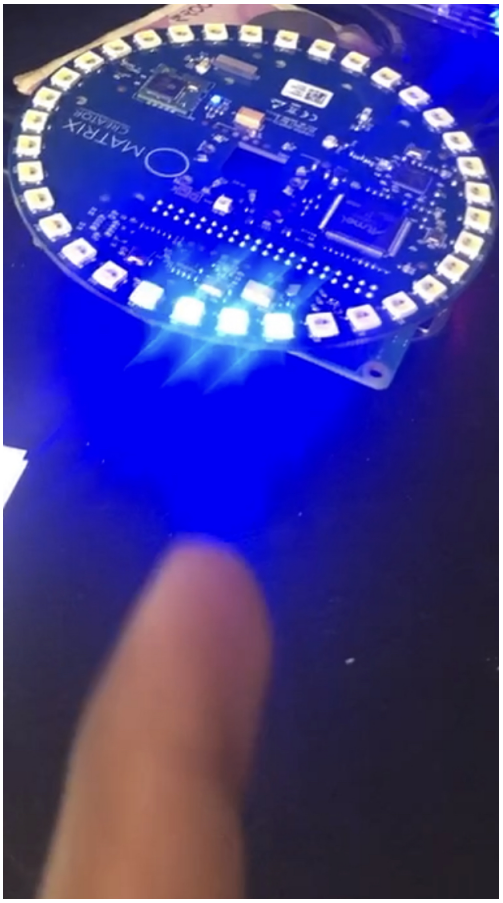


Figure 2: Localization Program working on MATRIX Creator

train custom wake words. The wakeword engine can be run on the Raspberry Pi alongside the sound source localization. The output from the wakeword detection will be a prompt whenever the word is detected.

Human Gesture Detection

The current microphone setup only outputs the angle for the direction of Arrival for the wakeword. There could be multiple people in the direction of the key word. It is necessary to have a gesture for the robot to be able to detect the human in its visible field of view when the robot is facing the direction of the voice. A simple hand raise gesture detection which can be integrated using the pre-trained models of the software development kit for Zed 2i, which is the onboard camera of the robot. The Zed SDK provides

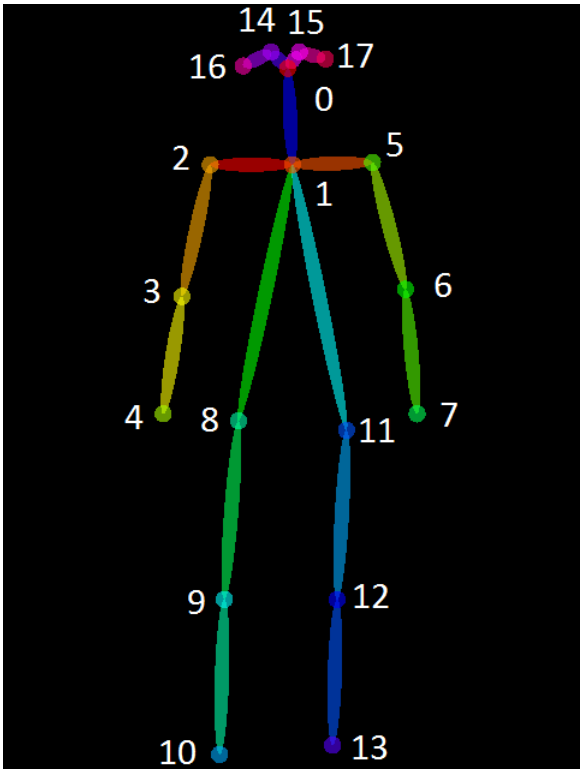


Figure 3: Caption

models for human body tracking using 18 points (as shown in figure 3) which represent critical features and joints of the human body.

Integration

The system integration framework used is ROS. ROS has a basic communications version which can be used on a Raspberry Pi. The process flowchart is shown in figure 4. The sound localization program with ODAS is run continuously to give DOA of any sounds the microphone is hearing. Once the wakeword is detected, the output from the localization program will be recorded. Once the angle is recorded, it is used as a reference by the navigation stack of the robot to turn towards the output angle. After turning, the robot starts the body tracking module. The body tracking module is well integrated with ROS using the Zed ROS SDK. The Zed ROS SDK has launch files based on the Zed Camera. Once the appropriate roslaunch file is run, the camera constantly looks for human bodies and overlays the

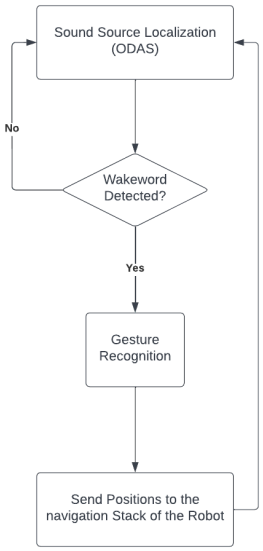


Figure 4: Sensor System Process Flowchart

18 point model. This 18 point model is a rostopic. A simple ros subscriber has been used to read the 18 points output by the camera model. Using this data, a simple if-else logic is used to recognize a hand raise by tracking the points which represent the hand joints.

4. Conclusion

In this study, we explored the capabilities of current state of the art for sensor systems in autonomous robots. Based on the analysis, an approach is proposed for a light weight audio and video based human tracking system for a service robots. The individual components of the system were working perfectly. A standalone sensor system was designed which was tested by placing the system on a static mount. The integrated system works as expected. Due to time constraints, the integrated system was not coupled with the navigation stack of the robot. In future, the robot implementation can be performed to test the limits of the sensory system on a moving base. To have a better performance in noisy environments, speech separation can introduced in the system before the localization part. Lastly, other body sign detection, such as hand sign detection, can be added in addition to the currently implemented basic hand raise detection.

References

Adavanne, S., Politis, A., Nikunen, J., Virtanen, T., 2019a. Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks. IEEE Journal of Selected Topics in Signal Processing 13, 34–48. URL: https://ieeexplore.ieee.org/document/8567942/, doi:10.1109/JSTSP.2018.2885636.

Adavanne, S., Politis, A., Virtanen, T., 2019b. Localization, Detection and Tracking of Multiple Moving Sound Sources with a Convolutional Recurrent Neural Network, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), New York University. pp. 20–24. URL: http://hdl.handle.net/2451/60768, doi:10.33682/xb0q-a335.

DiBiase, J.H., 2000. A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays. Ph.D. thesis. Proquest Information and Learning.

El Badawy, D., Dokmanić, I., 2018. Direction of arrival with one microphone, a few legos, and non-negative matrix factorization. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26, 2436–2446. doi:10.1109/TASLP.2018.2867081.

Grondin, F., Létourneau, D., Godin, C., Lauzon, J.S., Vincent, J., Michaud, S., Faucher, S., Michaud, F., 2021. Odas: Open embedded audition system. URL: https://arxiv.org/abs/2103.03954, doi:10.48550/ARXIV.2103.03954.

Sathyamoorthy, A.J., Patel, U., Paul, M., Savle, Y., Manocha, D., 2021. Covid surveillance robot: Monitoring social distancing constraints in indoor scenarios. PLOS ONE 16, 1–20. URL: https://doi.org/10.1371/journal.pone.0259713, doi:10.1371/journal.pone.0259713.

Saxena, A., Ng, A.Y., 2009. Learning sound location from a single microphone, in: 2009 IEEE International Conference on Robotics and Automation, pp. 1737–1742. doi:10.1109/ROBOT.2009.5152861.

Scola, C.F., Ortega, M.D.B., 2010. Direction of arrival estimation – A two microphones approach. Ph.D. thesis. Blekinge Institute of Technology.