Acoustic Simultaneous Localization and Mapping for Autonomous Driving

Akul Madan and Lingxi Li

2024 International Conference on Cyber-Physical Social Intelligence (ICCSI) | 979-8-3503-7673-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCSI62669.2024.10799478

***Abstract*— The current technology employed for autonomous driving has provided us with satisfactory performance and results. However, some of the most commonly used sensors, such as LiDAR and cameras, are typically high-priced and often require tremendous amount of computational power to process the collected data. In many cases, the adverse weather and night-time conditions hinder the performance of vision-based sensors. In this paper, we propose an alternative approach for enhancing the current sensing methodologies by utilizing acoustic signatures of moving objects. This approach makes use of a microphone array to collect and process acoustic signatures captured for simultaneous localization and mapping (SLAM). Rather than using numerous sensors to gather information of the surrounding environment of the ego vehicle, this method investigates the benefits of considering the sound waves of different objects around the ego vehicle and offers a cost-effective approach for SLAM.**

I. INTRODUCTION

A. Motivation

With the rapid development of computational power and artificial intelligence technology, connected and automated vehicles (CAVs) have become a hot research area [1]–[6]. In autonomous driving, night-time conditions, dust, fog, snow, rain, or any other adverse conditions can hinder the functionality of various sensors. Vision-based sensors like cameras often struggle with recognizing objects within the ego vehicle's surroundings in dark conditions. Sudden changes in lighting conditions such as driving through an underpass or bright reflections makes camera-based sensors temporarily blind. Similarly, sensors that depend on the concept of reflection of light or sound waves like LiDAR, radar or ultrasonic are rendered useless in rainy, snowy, or foggy conditions due to the presence of numerous suspended particles around the sensors. Current technologies and algorithms require computers with high processing power: Capturing, processing, encoding, filtering, re-encoding, and syncing data with other sensors. All these tasks rely on the computational precision and any discrepancy in processing the crucial data could lead to fatal casualties [7]. Radar sensors are available as long-range and short-range, with short-range for tasks like detecting nearby vehicles and long-range for far objects. While radar sensors are generally cost-effective, they are prone to giving false results from reflections off of smaller objects like an empty soda can [8]. Many car manufacturers either use a single sensor or sensor fusion from various sensors to offer advanced driver assistance system (ADAS) features, for instance, lane departure warning, adaptive cruise control, and automatic emergency braking [9]. The price of an autonomous vehicle is highly associated with the number and type of sensors installed [10]. Typically, the cost of the industrial grade sensors like LiDAR, radar, or camera is high. There are some other alternatives that offer a combination of sensors, commonly known as MiDAR, but the cost is still extremely high [11].

In summary, there is an emerging need to investigate advanced sensing technology to make in-vehicle sensors cost-effective and do not produce complex data for processing In addition, a sensor that works during night-time, dusty, foggy or snowy conditions would be an optimal solution to the problems faced by some of the current sensors.

B. Methodology

In this paper, an alternative approach that uses a microphone array for autonomous driving is proposed. The main idea is to provide a cost-effective and computationally-efficient method, with sensing results comparable to other technologies being used [12]. The localization of the sound source is implemented using the concept of direction-of-arrival (DOA) of sound. The implementation method used for this purpose captures audio from the microphones and estimates the direction of the sound source by calculating the time of delay between signals. The governing concept behind this technique is that sound sources at larger distances lose their intensity due to their distance, and produce sounds with lower sound pressure level (SPL) [13]. This is true for both light and sound signals, and can be easily modeled by Inverse Square Law [13]. If a known sound source is at an unknown distance, the amplitude of sound I reaching a listener is lowered as the sound pressure level spreads over distance d [14], as shown in Equation (1).

$$I \propto 1/d^2 \tag{1}$$

The auditory characters can vary in terms of pitch, amplitude, reverberation, and timbre [15]. Although methods that make use of the concept of direct-to-reverberant-ratio (DRR) provide good results, those algorithms are computationally very expensive [16]. Hence, the goal of the algorithm is to provide distance approximations of moving vehicles by computing changes in the sound intensity levels in real-time. For example, as an ambulance approaches closer to a

979-8-3503-7673-9/24/$31.00 © 2024 IEEE

Akul Madan is with the Department of Electrical and Computer Engineering, Indiana University-Purdue University Indianapolis, 723 West Michigan Street, SL-160, Indianapolis, IN 46202, USA. Email: akul.madan@gmail.com.

Lingxi Li is with the Elmore Family School of Electrical and Computer Engineering, Purdue University, Indianapolis, IN 46202, USA. Email: lingxili@purdue.edu.

listener, its sound intensity increases in terms of amplitude, as a majority of the sound pressure levels are directed to the listener without attenuation [15].

II. Microphone Array Design

Generally, one single microphone would be enough for distance estimation based on the sound intensity levels. But in order to localize the direction of the sound, microphone pairs help in this regard [17]. To provide a higher degree of estimation in terms of DOA, multiple microphones can prove to be beneficial. With multiple microphones, each at equal intervals from the other, provides a better DOA tracking accuracy [18]. In order to ensure that the entire acoustic field is covered by the array, an eight microphone approach would be ideal under many conditions. This helps to localize a sound source by creating eight acoustic zones, each 45° wide. The sound source is then localized within one of the octets based on the sound intensity levels [19].

A. Array Design

Capturing the entire 360° field gives a better estimate of the fast changing surroundings of the host vehicle. This in turn allows the algorithm to capture sounds from different directions and process their intensity levels in real time. With an eight microphone array, there are enough sensors pointing in different directions, which can provide a vivid depiction of the surrounding sound sources. An eight microphone approach covers the acoustic field of the host vehicle by dividing them into eight distinct acoustic zones (Fig. 1). This helps separate various sound sources into different areas based on their sound intensity levels captured by the array. Higher number of microphones also provide a better estimate of the direction of sound, as the DOA algorithm has multiple data sources to compare the values with. This approach helps to increase the overall accuracy of the system.

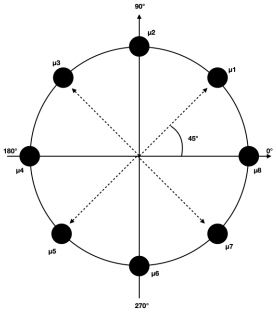

Fig. 1. Microphone array design.

B. Benefits Of The Design

An even placement provides the DOA algorithm with enough comprehensible data to estimate the direction of the sound source. Dividing the 360° field into eight sub-sectors enables the system to capture acoustic signatures at locations that are beyond the range of some other sensor types. By placing multiple sensors, each facing in a different direction provides the ability for blind-spot detection. The microphone array can easily pick up acoustic signatures of small children or pets around the vehicle. This is extremely beneficial for large commercial vehicles with a limited field of view around the vehicle. A similar application can also be extended to recognize certain types of sounds and provide higher priority to them, solely through the sound signatures [20]. This can include sirens of emergency vehicles or people walking.

III. Experimental Setup

A. Hardware Setup

The system design requires a microphone array for higher tracking and detection accuracy. For this reason, a microphone array was selected for testing and fine-tuning the algorithm. In order to provide an interface between the hardware and software, a portable computer was also used. The computer is also responsible for capturing, storing, and processing the data gathered by the sensor array. Therefore for the microphone array, a Microsoft Xbox One Kinect sensor was selected (Fig. 2). A speaker was also used to generate sounds of different vehicles and noises like rain or wind. The purpose of this hardware setup was to simulate different scenarios that would occur in a real-world situation.

Fig. 2. Test system setup.

B. Software Setup

A delay in processing the real-time data can lead to fatal errors [7]. MATLAB was selected as the ideal choice, as it provides toolboxes for processing different types of data. It also supports a wide array of hardware, along with toolboxes for each type of sensor. The data stored in the variables can also be exported to different file types for further analysis using various software tools. The DOA or localizing algorithm uses acoustic signatures of surrounding vehicles and computes an approximation of the direction of the sound source. The direction of the sound source is then displayed in the form of a vector pointing in the approximated direction of the sound source. The second main algorithm that is part of the software setup is the proximity modeling algorithm.

C. Benefits Of Test Setup

Using a microphone array for detecting acoustic signatures of objects, while using vision sensors for confirming the presence of those objects would provide better tracking results. In general, the test setup provides a real emulation of how the actual system setup would perform. This also provides enough flexibility testing and make changes to the system design by modifying the algorithms or hardware setup. Similarly, combination of sensors like RADAR with a microphone array can help localize still and moving objects [21] [22].

IV. Data Collection

The purpose of data collection is to test hypothetical situations that the system can en- counter in real-world applications. The data collected can be transformed into different forms, and the system's performance under different conditions can be identified [23].

A. Procedure

Sound sources were placed at different positions are the distance was computed by the algorithm, therefore the results were verified using distance measurement tools. In order to simulate windy conditions, a fan was placed close to the microphone array that creates the same effect as wind blowing into the microphones. The sensor also carries a power requirement of 12V DC for operation, which is indirectly supplied through the adapter. In order to simulate the scenarios, the three possible cases for the states of the host and the sound source were evaluated.

- Static host and source
- Static host and moving source
- Moving host and source

B. Static Host And Source

In this case the host and the source are both static. A situation like such can occur at a traffic light or in a parking lot. Similar applications of such scenarios can be used for blind-spot detection. Sounds generated by different sound sources are generally produced by the vibrations caused inside the vehicle's engine [24]. To simulate the presence of a static sound source, a monotonic sine wave was played using a wireless speaker. The speaker was placed at a fixed distance of two meters from the microphone array. This procedure was repeated with the monotonic tone being replaced by an engine sound, which simulates a static vehicle with the engine running. To test the system's performance with noise, an additional wireless speaker was introduced to the test condition. This speaker played the ambiance of a busy city with sounds of people walking, car horns, and various other acoustic elements. A vision-based sensor like a camera can detect still objects in the frame, while the acoustic array can detect moving objects that generate sounds [25]. The data collected for the case with a static sound source suggests that environmental noise can slightly affect the performance of the system, as it is unstructured data from random sources [26]. The types of noise vary in real-world conditions, the variations can be in terms of the amplitude of sounds of different objects [26]. The first type of sound source was using the monotonic sine wave. This type of sound generally mimics the sinusoidal vibrations produced by the movement of pistons in an engine [24]. In this scenario, the tracking results after noise removal are close to the results with the second type of sound source, with only a few variations in the blue line (Fig. 3).

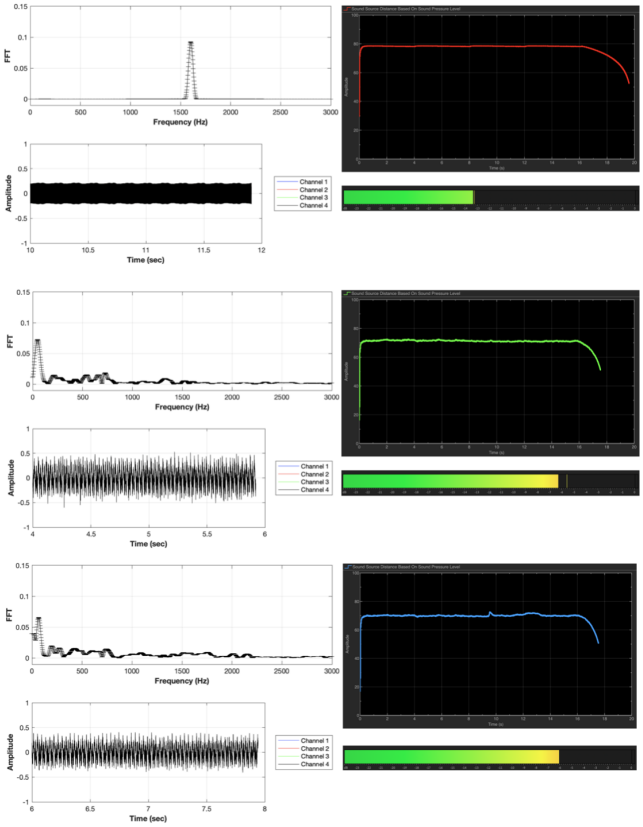

Fig. 3. Captured data for the first sound source.

C. Static Host And Moving Source

The second scenario is the case with a static host vehicle, with moving sound sources in its surroundings. An instance like such can occur when a host vehicle is at a red light and the vehicle in the opposite lane is passing by or the host vehicle is pulled over at the side of the road. In such cases, the system uses a combination of three to four adjacent microphones, in order to localize each sound source [17]. But for testing only a limited number of wireless speakers were available and only a four microphone array was used. Therefore, computing multiple sound sources with just four microphones would not provide accurate direction estimations. First, a monotonic sound source was used to play a fixed pitched sound. In order to simulate a moving sound source, the wireless speaker was being moved around the Kinect sensor. The initial starting distance was again used as two meters. This process was again repeated with a real-world sample of a car passing by, which provides a more

accurate representation of a moving vehicle. Another wireless speaker was used to play the sounds of a city landscape with sounds like traffic, car horns, and people walking. The speaker playing environmental noise was placed at a fixed position and was not moving like the other speaker. The results for the changing distance of the sound source are also modeled well by the system. With the second sound source, the approximations are still within a good range. But since the sound contains many harmonics when the source moves, the system takes a bit longer to process the data. In the third testing scenario with environmental noise added to the car engine sound, the distance approximations of the moving were similar to the first sound source (Fig. 4).

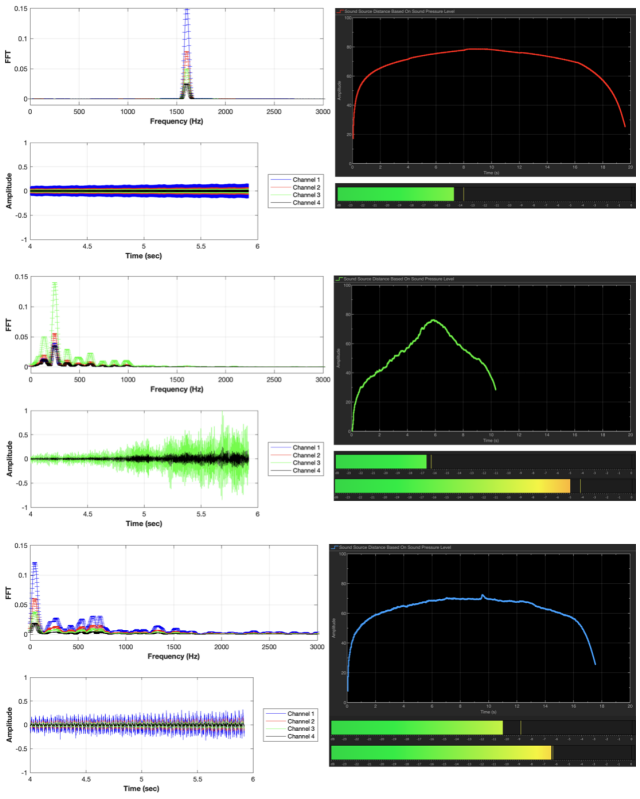

Fig. 4. Captured data for the second sound source.

D. Moving Host And Source

In the third test scenario, both the host and the source are moving. This situation is most commonly encountered while driving on the road. First the monotonic sound source is tested. For this test scenario, both the microphone array and the wireless speaker are moving. For this scenario, the initial starting distance was not needed as the positions of both the microphone array and the sound source keep changing over time. This setting was repeated again but this time with the car engine sounds, which provides a better representation of the sounds on a road. Then finally, the engine sound was tested with environmental sounds like traffic noise, which was played through a different wireless speaker. The speaker playing the noise was kept at a fixed position but the volume of the noise was altered over time. In this case, the proximity estimation was good but was not as accurate as with the first two cases. The line in figure is not as smooth in the third segment due to the loss of data during processing (Fig. 5). This is because the number of microphones available in the array is limited to four.

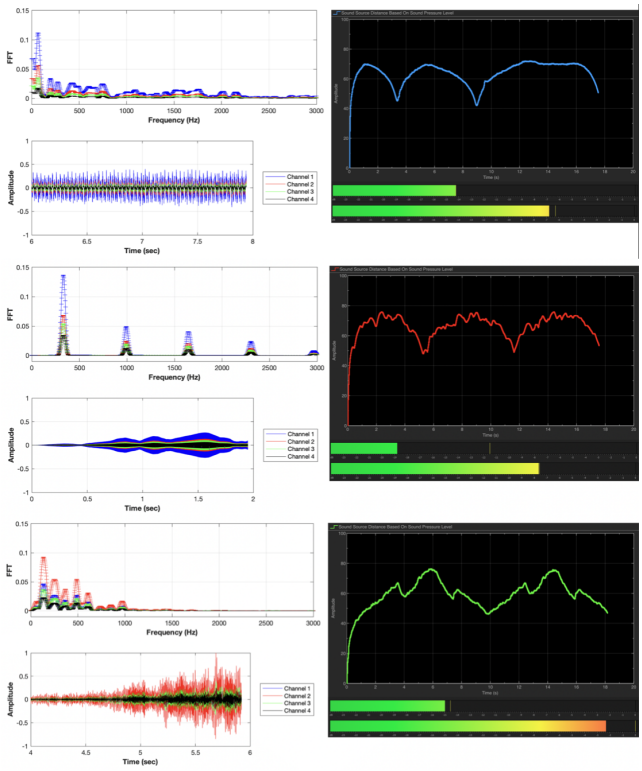

Fig. 5. Captured data for the third sound source.

V. Results

In general, if an object is farther from a listener, then there are more reverberations present in the sound once it reaches the listener [16] [27]. Whereas, sounds closer to the listener have a smaller number of reverberations [16]. This is due to the fact that as distance increases, the sound gets reflected off of multiple surfaces [28]. The transients of a sound source that has reverberations present are generally longer in length [27].

A. Accuracy Comparison

The verification process for the direction of the sound source is as follows. A printed protractor was used to measure the angle at which the wireless speaker was placed (Fig. 6). In order to accurately get the angle of the speaker, a string was used to get the direction from the protractor to the source. This setup helps to verify the results provided by the DOA algorithm. Multiple DSP techniques such as sampling, noise reduction or frequency domain conversion is performed on the captured data, in order to extract information. The core concept of the noise reduction process is to minimize the noise in the raw data. The overall estimation accuracy of the direction of the sound source is 94.58% (Table I).

TABLE I
DOA ALGORITHM ACCURACY COMPARISON

Test Angle	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
90°	95°	100°	90°	90°	85°
80°	80°	75°	80°	70°	80°
70°	60°	70°	65°	70°	65°
60°	60°	70°	60°	60°	65°
50°	40°	55°	50°	50°	50°
40°	40°	35°	45°	40°	50°
30°	30°	30°	35°	30°	30°
20°	20°	10°	20°	20°	20°
10°	5°	10°	5°	10°	10°
0°	-5°	0°	-5°	0°	0°
-10°	-5°	-15°	-10°	-10°	-10°
-20°	-20°	-20°	-15°	-10°	-15°
-30°	-25°	-30°	-25°	-30°	-35°
-40°	-40°	-45°	-45°	-40°	-40°
-50°	-50°	-50°	-50°	-60°	-55°
-60°	-55°	-55°	-60°	-60°	-60°
-70°	-70°	-65°	-70°	-70°	-70°
-80°	-70°	-85°	-80°	-75°	-80°
-90°	-90°	-85°	-90°	-95°	-80°

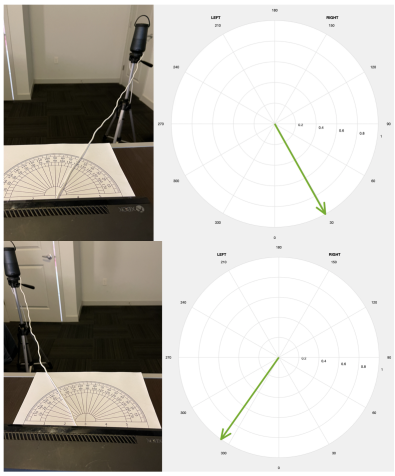

Fig. 6. Direction of arrival test setup.

The first test scenario was with the static sound source. In this case, sound sources like a monotonic sine wave, or less complex sounds are easier to detect and process. When additional environmental noise is introduced to the condition, the data gets more complex, and more processing is needed for extracting specific information. Slight deviations in accuracy are observed when environmental noise is added to the scenario. This occurs due to the fact that the noise goes through a reduction process, which can alter the original source sound in some way. The second scenario was with the static host and a moving source. This scenario was simulated by moving the sound source around, in order to create an illusion of a moving vehicle. Similar to the first scenario, the monotonic sine wave, and the car sound produce accurate results. Sometimes if the object is moved too fast between the microphones in the array, then the tracking and the results are not quite accurate. With the environmental noise induced into the scenario, the results are still observed to be close to the actual measurements. The third and final scenario tested was with the moving host and source. In order to simulate this scenario, both the microphone array and the sound source were moving to replicate a moving vehicle. A test situation like such exposes the microphones to wind while in motion. The monotonic sine wave results were the most accurate in this scenario. Then the results for the engine sound were close to the actual motion of the source. After inducing environmental noise in the scenario, the results were close to the actual measurements but were not as accurate as they were with other test conditions. Some additional improvements can be made on the software end by using neural networks for better separation of noise from the sound source [29].

B. Blind Spot Detection

Blind-spot regions or dead zones around a vehicle could be areas that are beyond the range of a sensor, or are below the level of detection [7]. Spots like these could be behind the vehicle, or by the sides where the field of view is limited. Especially in large commercial vehicles, blind-spot detection can prevent some major fatalities. A microphone array can pick up acoustic signatures of activities such as pedestrians, small vehicles, small children, or animals. This can be used for regular vehicles where the field of view is limited. This can also be used as a combination with some other sensors in order to detect still objects, as they do not produce any sound. Other applications of blind-spot detection can also be applied to large warehouses, where workers are moving around large commercial machines. Overall, an acoustic sensor has some benefits in terms of object detection over other sensor types, through detection of their acoustic signatures. Mel-Frequency spectrum can provide an accurate representation of the way different sounds at specific distances are perceived by human ears [19] [30]. Acoustic event recognition or sound-activated decision-making can help avoid certain obstacles like firetrucks [26] [31]. Generating a risk evaluation matrix for different sounds can help prioritize tasks like emergency vehicles.

C. Economic Benefits

As microphone array only considers the data that reaches the sensor in terms of changes in acoustic intensities, thus producing data that is easy to process. A lot of redundant information present within the data about the surroundings is eliminated in the processing stages. This creates a system that is economic to operate. The data sets produced are not too complex and hence do not require extensive processing power. Overall, the cost of operation is relatively low compared to some other sensor types like LiDAR. Cost efficiency is one of the major benefits of a system like such. Starting from the cost of hardware components, to the software elements. The system also proves to be economic in the long run, with regard to the cost of replacing hardware components.

VI. Conclusion

In this paper, a novel idea that utilizes an array of acoustic sensors for SLAM was presented. The overall performance and accuracy of the system were tested and were comparable to the actual measurements. One major advantage that the system has over other sensor setups is the low cost. The datasets captured by the sensor array are easy to process, as they are not too complex and contain acoustic data stored in four channels. The simplicity of the data means that the system does not have a high operational cost in terms of computational power needed. This helps the system capture and process data, in order to display the results in near real-time. Another advantage the system has is that it can function in adverse conditions. Conditions like heavy snow, heavy rain, or foggy environment do not impact the system's accuracy. Similarly, dark or low light conditions also do not affect the system's functionality. Since conditions like rain and wind have white noise-like characteristics, this can be removed during the processing stage by applying a high pass filter to the data. This makes it a viable low-cost alternative. The system's design is relatively simple and only depends on the positioning of the microphones for better accuracy. This provides enough flexibility for modifications or for sensor fusion. Sensor fusion with a vision-based sensor can help the system reach even higher detection accuracy.

In the future, we will investigate how to improve the accuracy of the proposed system. It is also interesting to study how to incorporate the proposed system with other sensors to achieve better results.

References

[1] J. Yin, D. Shen, X. Du, and L. Li, "Distributed stochastic model predictive control with Taguchi's robustness for vehicle platooning," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 9, pp. 15967-15979, September 2022.

[2] S.-S. Han, D. Cao, L. Li, S. E. Li, N.-N. Zheng, and F.-Y. Wang, "From software-defined vehicles to self-driving vehicles: A report on CPSS-based parallel driving," IEEE Intelligent Transportation Systems Magazine, vol. 11, no. 1, pp. 6–14, Spring 2019.

[3] S. Li, J. Yan, and, L. Li, "Automated guided vehicle: The direction of intelligent logistics," in Proc. 2018 IEEE International Conference on Service Operations and Logistics, pp. 250-255, Furama RiverFront, Singapore, July 2018.

[4] L. Li and F.-Y. Wang, "The automated lane-changing model of intelligent vehicle highway systems," in Proc. IEEE International Conference on Intelligent Transportation Systems, pp. 216-218, Singapore, September 2002.

[5] Y. Liu, B. Tian, Y. Lv, L. Li, and F.-Y. Wang, "Point cloud classification using content-based transformer via clustering in feature space," IEEE/CAA Journal of Automatica Sinica, vol. 11, no. 1, pp. 231–239, 2023.

[6] D. Cao, X. Wang, L. Li, C. Lv, X. Na, Y. Xing, X. Li, Y. Li, Y. Chen, and F.-Y. Wang, "Future directions of intelligent vehicles: Potentials, possibilities, and perspectives," IEEE Transactions on Intelligent Vehicles, vol. 7, no. 1, pp. 7–10, 2022.

[7] J. Wang, L. Zhang, Y. Huang, and J. Zhao, "Safety of autonomous vehicles," Journal of advanced transportation, vol. 2020, 2020.

[8] M. Skolnik. (Nov. 2020). "Factors affecting radar performance." Last date accessed: 05-13-2021, [Online]. Available: https://cleantechnica.com/2021/03/12/lidar-may- be-harmful-to-people-cameras/.

[9] B. A. Jumaa, A. M. Abdulhassan, and A. M. Abdulhassan, "Advanced driver assis- tance system (adas): A review of systems and technologies," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 8, no. 6, 2019.

[10] L. Tang, Y. Shi, Q. He, A. W. Sadek, and C. Qiao, "Performance test of au- tonomous vehicle lidar sensors under different weather conditions," Transportation research record, vol. 2674, no. 1, pp. 319–329, 2020.

[11] NASA. (Aug. 2018). "Midar - active multispectral imaging." Last date accessed: 08-29-2021, [Online]. Available: https://www.nasa.gov/ames/las/midar.

[12] A. Madan, "Acoustic Simultaneous Localization And Mapping (SLAM)," MS Thesis, Purdue University, December 2021.

[13] N. Voudoukis and S. Oikonomidis, "Inverse square law for light and radiation: A unifying educational approach," European Journal of Engineering and Technology Research, vol. 2, no. 11, pp. 23–27, 2017.

[14] M. Yiwere and E. J. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," Int. J. Appl. Eng. Res, vol. 12, no. 22, pp. 12 384–12 389, 2017.

[15] J. O'Reilly, S. Cirstea, M. Cirstea, and J. Zhang, "A novel development of acous- tic slam," in 2019 International Aegean Conference on Electrical Machines and Power Electronics (ACEMP) & 2019 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM), IEEE, 2019, pp. 525–531.

[16] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," IEEE Transac- tions on Audio, Speech, and Language Processing, vol. 18, no. 7, pp. 1793–1805, 2010.

[17] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," Robotics and Autonomous Systems, vol. 96, pp. 184–210, 2017.

[18] I.-J. Jung and J.-G. Ih, "Distance estimation of a sound source using the multiple intensity vectors," The Journal of the Acoustical Society of America, vol. 148, no. 1, EL105–EL111, 2020.

[19] G. Tzanetakis, G. Essl, and P. Cook, "Audio analysis using the discrete wavelet transform," in Proc. conf. in acoustics and music theory applications, Citeseer, vol. 66, 2001.

[20] C. Couvreur, "Environmental sound recognition: A statistical approach," Doctorat en sciences appliqu ees, Facult e Polytechnique de Mons, Mons, Belgium, 1997.

[21] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: A survey from 2010 to 2016," IPSJ Transactions on Computer Vision and Applications, vol. 9, no. 1, pp. 1–11, 2017.

[22] C. Evers and P. A. Naylor, "Optimized self-localization for slam in dynamic scenes using probability hypothesis density filters," IEEE Transactions on Signal Process- ing, vol. 66, no. 4, pp. 863–878, 2017.

[23] J. Foote, "An overview of audio information retrieval," Multimedia systems, vol. 7, no. 1, pp. 2–10, 1999.

[24] H. Wu, M. Siegel, and P. Khosla, "Vehicle sound signature recognition by frequency vector principal component analysis," in IMTC/98 Conference Proceedings. IEEE Instrumentation and Measurement Technology Conference. Where Instrumentation is Going (Cat. No. 98CH36222), IEEE, vol. 1, 1998, pp. 429–434.

[25] Z. Zeng, J. Tu, M. Liu, T. S. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition," IEEE Transactions on multimedia, vol. 9, no. 2, pp. 424–428, 2007.

[26] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," IEEE Transactions on Audio, Speech, and Lan- guage Processing, vol. 17, no. 6, pp. 1142–1158, 2009.

[27] G. Chen and Y. Xu, "A sound source localization device based on rectangular pyramid structure for mobile robot," Journal of Sensors, vol. 2019, 2019.

[28] E. Georganti, T. May, S. Van De Par, and J. Mourjopoulos, "Sound source distance estimation in rooms based on statistical properties of binaural signals," IEEE transactions on audio, speech, and language processing, vol. 21, no. 8, pp. 1727– 1741, 2013.

[29] K. Ashraf, B. Elizalde, F. Iandola, M. Moskewicz, J. Bernd, G. Friedland, and K. Keutzer, "Audio-based multimedia event detection with dnns and sparse sam- pling," in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015, pp. 611–614.

[30] M. S. Puckette, M. S. P. Ucsd, T. Apel, et al., "Real-time audio analysis tools for pd and msp," University of California San Diego, 1998.

[31] P. Dhakal, P. Damacharla, A. Y. Javaid, and V. Devabhaktuni, "A near real- time automatic speaker recognition architecture for voice-based user interface," Machine Learning and Knowledge Extraction, vol. 1, no. 1, pp. 504–520, 2019.