Audio Based Object Recognition

A Major Qualifying Project (MQP) Report
Submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements
for the Degree of Bachelor of Science in

Electrical and Computer Engineering

By:

Zhuolin Liu, Michael Rothstein

Project Advisor:

Bashima Islam

Date: April 2023

Abstract

Audio-based shape recognition is often used in robotics and navigation. With the popularity of in-ear earphones and tiny robots and drones with low-power microphones, we see an opportunity where acoustic sensing can be used to monitor small enclosed spaces, e.g., ear canals, pipes, and machinery, for human, infrastructure, and machine health monitoring. This paper studies the opportunity of using acoustic sensing to identify and monitor various small hollow objects from the inside. We propose an algorithm that fuses signal processing and machine learning techniques to distinguish between three shapes and assess the health of the shapes.

The results show that under ideal conditions, audio can distinguish between three shapes with 90% accuracy, and distinguish between three shapes under non-ideal conditions(different heights, background noises, and setups) with 70% accuracy. We further determine shape deformation with 100% accuracy and identify the side of the deformation with 90% accuracy. Our results demonstrate that as individual features acoustic Loudness, power Spectrogram, mel Spectrogram, and spectral Entropy have great results. However the best combination of features is zero crossing rate, Acoustic Loudness, and mel Spectrogram. Additionally decreasing the number of microphones decreases the accuracy of the experiments. Furthermore, attempting localization techniques in small hollow areas to determine the height of the shape is difficult due to sound reflection and refractions. Localization through triangulation yields results with a high variance, while localization through a gradient boosted machine requires a large dataset.

Executive Summary

Wireless earbuds, widely used for entertainment purposes, have untapped potential in various applications, including health monitoring and ear authentication. This project investigates the feasibility of using microphone arrays for understanding the characteristics of small enclosed structures, with the long-term goal of making in-ear modeling more accessible and passive, potentially replacing or assisting devices that measure the inner ear's shape and detect potential ear ailments. Moreover, such acoustic-based monitoring of small enclosed structures opens a new direction of utilizing small robots[13] and insect drones[16] to monitor machines, infrastructures, and environment.

We conduct a series of experiments to empirically analyze the potential of using a microphone array to detect the shape, health, and depth of small enclosed objects. We test over 20 acoustic features and narrowed them down to 5 features that achieves the best results. We also observe that combining the audio features with the best individual performances does not necessarily produce the most accurate results across all experiments. The results from multiple features are in the figure below.

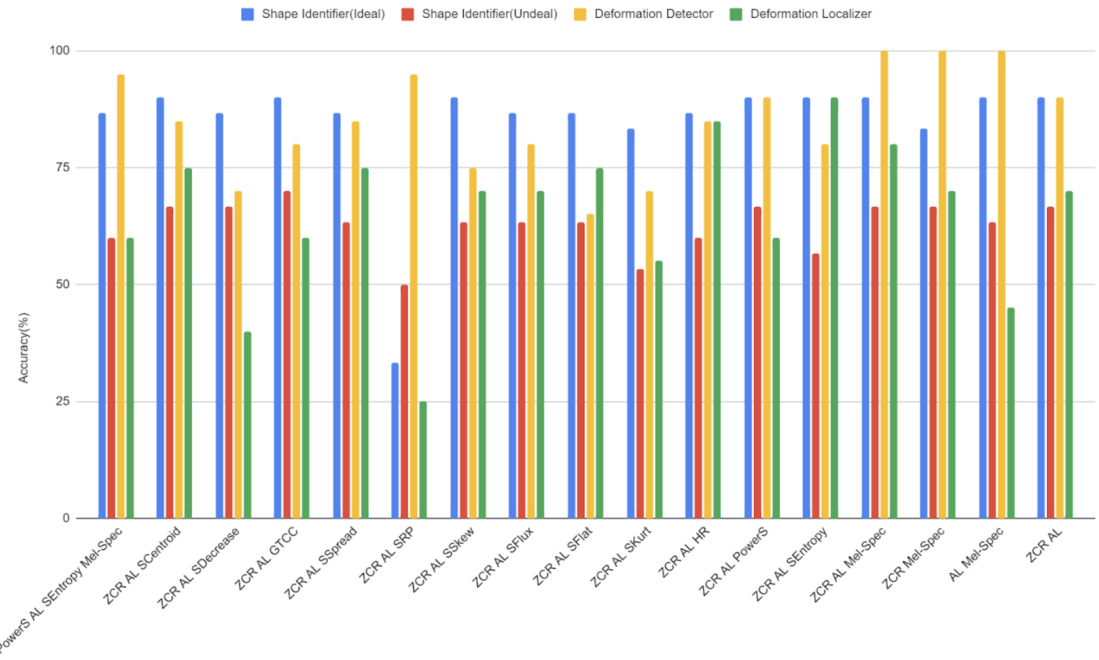

Figure 1: Results from feature combination

We also notice that the number of microphones significantly affects performance. In general, increasing the number of microphones increases the accuracy of the results, although the difference between using four and two microphones is not substantial. Analysis of the confusion matrices show that changing the number of microphones not only impacts the experiment's accuracy but also the manner in which errors occur. For instance, having more microphones reduces the number of ways errors occur.

Finally, we look into localizing the reflected acoustic signals to determine the shape parameters, e.g., the height of the cube, for 3D modeling. We implement two localization techniques, triangulation and Gradient Boosted Machine(GBM), and observe that neither method yields a satisfactory performance.

The impact of wireless earbuds on healthcare and industrial inspection is quite notable. This study highlights that these small devices have the capacity to change how we detect and prevent inner ear issues, as well as maintain the health of important structures in hard-to-reach areas. With the help of wireless connectivity and advanced sensors, the opportunities for improving our understanding of the world around

us are considerable. As we keep gathering data and refining this technology, it might become a useful tool for supporting human health and safety in both medicine and industry. The potential for audio sensors to influence the future is indeed impressive.

Acknowledgements

We would like to thank professor Bashima Islam and BASH LAB for their help, guidance, and assistance throughout the entirety of this project. Additionally we would like to thank Worcester Polytechnic Institute for providing us with the opportunity to complete this project.

Contents

List of Tables

List of Figures

1 Introduction

Wireless earbuds have become a staple in everyday life. Young adults used headphones(including earbuds) for an average of 1.54 hours per day in 2011[40]. Furthermore, in 2016, 77% of US adults used earbuds and headphones. This percentage is continually growing [30]. Currently, 87% of the people who use earbuds use them to listen to music, and 49% of people use them to watch movies or TV shows [30]. However, earbuds can be used for much more than entertainment.

Previous works show that the multiple modalities of sensors used in earbuds, e.g., multiple microphones, motion sensors, and wireless radios, present the potential of using earbuds in various applications. These multimodal measurements can be used for diverse health applications such as step counting [3], speaker recognition[2], and breathing detection[4]. Health applications are one of the most promising aspects of these earbuds due to their passive health monitoring capabilities. For instance, a temperature measurement in the ear can be used to determine if someone is ill, or a heartbeat sensor can be placed in the ear to accurately measure a heartbeat[3]. Although these applications are promising, they only demonstrate the beginning of what can be done. These measurements can be used in more complicated applications such as ear authentication purposes[27], in-ear fluid measurement [8], eating habit detection[1] and even ear-wax detection[8].

Inner ear modeling is often crucial to detect and assessing ear infections. An ear model can diagnose visual ear infections such as otitis externa, serous otitis media, and acute mastoiditis[39]. Current state-of-the-art techniques for ear modeling include expensive equipment and require experts to use them. The most user-friendly and portable solution is Otoscan [38] which includes a probe with tiny cameras to make a 3D ear model. We wonder if the earbuds have the potential to assist or even replace devices that measure the inner ear's shape and detect potential ear ailments.

Though audio-based 3D modeling is not new, they are often done in a large space for Room Impulse Response (RIR) detection, room reconstruction, or large object/shape reconstruction [41]. Few works have explored audio-based reconstruction for small enclosed spaces[11]. However, none of these works focus on shape deformation in small spaces. Ultrasonic and infrasonic sounds can be used for modeling. However, ultrasonic frequencies are above the range of human hearing (typically above 20 kHz), while infrasonic frequencies are below the range of human hearing (below 20 Hz). For this reason, it is crucial to focus on the audible range, generally between 20 Hz and 20 kHz, to ensure safety and minimize potential harm[19].

To satisfy the bigger goal of making in-ear modeling more accessible and passive, this project

explores the feasibility of using a microphone array to understand the characteristics of small enclosed structures. We study whether audible acoustic signals can detect an object's shape, health, and depth if the speaker and mic are placed inside the hollow object. We use a microphone array to get the audio response from different directions and model the responses to identify the object's properties. This technology can be applied to the human ear as well as factories, pipes, tunnels, and other structures to obtain health and maintenance information.

In this project, we show the potential of audio object classification by answering the following questions.

1. Can audible signal identify the shape of a small enclosed structure?
2. If variation in the structure parameter, e.g., height and materials, and various background noise is introduced, how does the shape identification get affected?
3. Can audible signal identify the shape correctly if the structure deforms? Can it differentiate between healthy and unhealthy (deformed) objects?
4. Is it possible to identify the location of the deformation?
5. Can intrinsic parameters of the object's shape, e.g., height, be estimated?

The findings indicate that in ideal conditions, audio can differentiate between three shapes with 90% accuracy, and under non-ideal conditions (varying heights, background noises, and setups), the accuracy is 70%. Moreover, shape deformation can be determined with 100% accuracy, and the side of the deformation can be identified with 90% accuracy. The results reveal that individual features such as acoustic loudness, power spectrogram, mel spectrogram, and spectral entropy yield excellent outcomes. However, the most effective combination of features includes zero crossing rate, acoustic loudness, and mel spectrogram. Reducing the number of microphones also decreases the accuracy of the experiments. Furthermore, applying localization techniques in small hollow spaces to ascertain the height of the shape is challenging due to sound reflection and refraction. Localization via triangulation produces results with high variance, while localization using a gradient boosted machine necessitates a large data set.

The literature review section discusses prior work on 3D model generation, 3D modeling for healthcare, and microphone array-based direction and distance estimations. In the system section, we introduce the hardware system we used to perform the experiment. This includes hardware selection, microphone selection, processor selection, speaker selection, and the programming environment.

In the Methodology section, many signal characteristics are explored. The primary four signal characteristics to achieve our goals are – frequency response (magnitude and phase), zero crossing rate, spectral centroid, and acoustic loudness. These characteristics are determined by analyzing the MEMS microphone signals in MATLAB. These characteristics and many more then become input parameters to a Discriminant Analysis machine learning classifier. Furthermore, localization techniques are used to learn about the shape's dimensions. These methods include triangulation and a Gradient Boosted Machine (GBM). Next, the dataset section outlines the data being used and the set ups for the different experiments.

In the results section, we mainly focused on 3 points, which are Limiting signal features, limiting the number of microphones, and localization. Within the initial two subcategories, four tests are conducted. The first test aims to differentiate among the three shapes (cube, triangular prism, and sphere) under perfect conditions (identical setup). The second test involves distinguishing between the shapes when positioned at varying heights and in an environment with diverse background noises. The third test seeks to differentiate between an obstructed and non-obstructed cube (covering shaped corners and faces with aluminum). The last test attempts to identify the location of an obstruction on a cube (left, right, back, or front). After the results, limitations and future areas of research are discussed.

2 Literature Review and Background Study

This section describes the existing works on 3D model generation using different sensors, audio-based localization, and earphone-based health applications.

2.1 3D model Generation

3D model generation creates a three-dimensional digital representation of an object, structure, or environment using computer software or hardware tools. 3D reconstruction is commonly used for mapping the environment using sensors like lidar, sonar, or cameras. The recent development of 3D reconstruction has moved the application to the human body and human-face reconstruction for Augmented Reality and Virtual Reality applications [44]. Furthermore, 3D reconstruction is commonly used in health applications to model body parts and diagnose diseases.

3D model generation is often done through a point cloud. A point cloud is a collection of discrete data points in space. Each data point has a location coordinate. Together, the data points can form a 3D model. An example of this process is shown in the following figure. There are many ways to gather this data

and create a 3D model[53].

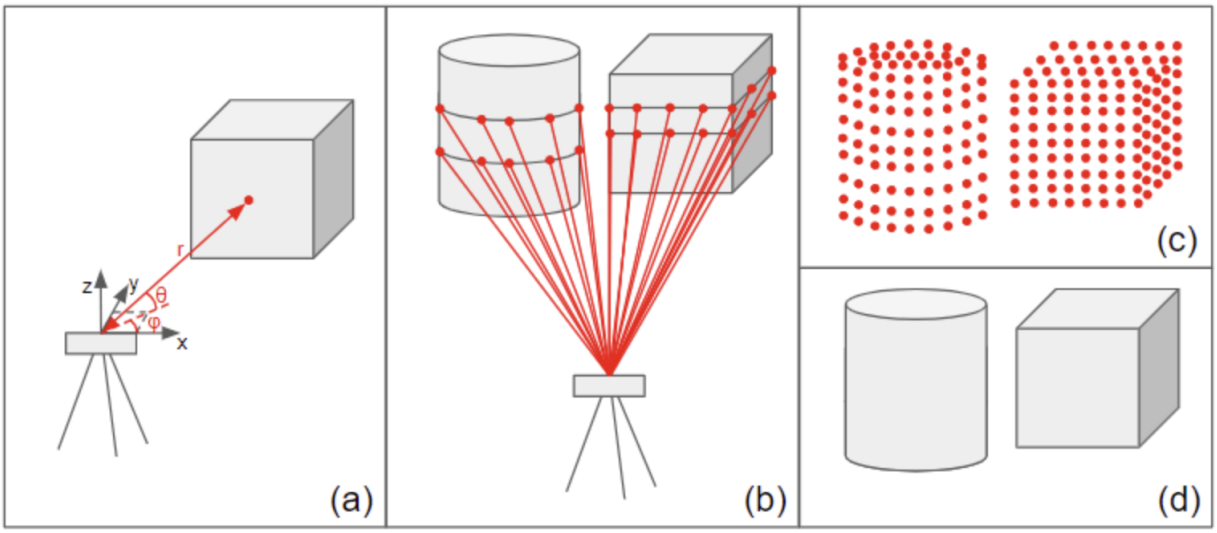

Figure 2: Point cloud from a laser scanner[53]

Camera-Based 3D modeling. Both single and stereo cameras have been used to make a 3D model of the environment. Multi-view stereo-based 3D reconstruction exploits multiple 2D views of a scene. However, this approach requires several processing steps, including feature matching, camera calibration, and triangulation, to estimate the 3D geometry of an object or scene[49]. One study proposes an algorithm that uses multi-view stereo to reconstruct a dense 3D scene[20]. Another study,Glimpse.3D, provides a multimodal solution for reconstructing sparse 3D scenes using body-worn stereo cameras[15]. Such 3D reconstruction methods have numerous applications, including robotics, computer graphics, virtual reality, and medical imaging [20]. However, these techniques require specialized equipment and may not be applicable in all industrial scenarios.

Lidar-Based 3D modeling. Lidar is another common tool for 3D reconstruction [42]. Lidars eject pulses of light, which get reflected by the object's surface, resulting in the depth information used to create models. Lidar utilizes a rotating system with multiple lasers shooting out light at a high rate. Schwarz et al [42]. use this depth information to create an accurate 3D model of the world. The downside of lidar is that it requires rotation to receive distance data from different angles. For this reason, lidar cannot be used in applications that prevent rotation.

Audio-Based 3D modeling. The sonarVisualizer F3D-S [24] is primarily used in water to create a 3D representation of ultrasound or sonar data. The product allows the user to see a 3D model of the sonar information while showing depth lines and different viewing angles. Though sonar is an excellent choice for underwater applications, they are highly inefficient in the air for two reasons – (1) the density of water is

significantly greater than the density of air, and (2) sound travels faster in water than in air. Besides, sonar can potentially hurt human ears, making it unideal for applications involving humans[6].

Thermal Camera-Based Modeling. Thermal cameras are often used to generate 3D models of various objects, especially living parts [45, 6] or active machinery. However, these cameras are required to be in very close proximity to the object, e.g., 95 cm [6]. Moreover, this approach cannot model objects without any thermal signature.

2.2 3D Model for Healthcare

3D models are increasingly used in healthcare for various applications, including surgical planning, medical training, and patient education.

Human Body-Part Modeling. Several previous works have developed 3D models of human body parts using a thermal camera. Despite being popular for anatomical modeling, thermal cameras are not a good solution for modeling inner-ear. These cameras are often larger and require rotational movement [9], which is impossible in this scenario.

Teeth Modeling. There are two traditional methods used for modeling teeth. The first method is a physical dental method. Physical models are tangible representations of a patient's teeth, usually made from materials like plaster or resin. These models can be created using traditional dental impressions or by 3D printing[54]. The second method is Digital 3D modeling. Digital 3D modeling uses a intraoral scanner to capture the shape and position of the teeth and generates a 3D digital model. This method is non-invasive, fast, and accurate[55]. A digital 3D model can also be created using cone-beam computed tomography. This imaging technique provides a detailed 3D images of teeth, bone, and soft tissues, which can be used to create accurate digital teeth models[56].

Inner-Ear Modeling. Generating 3D models of the inner ears of humans is extremely useful for diagnosing ear diseases. In the hospital, doctors generally model the inner ear by using a CT scan or an MRI. These scans are then used to create a physical model.[57]. If a 3D model is not necessary a doctor can use an otoscope to look into a patients ear. One of the smallest digital 3D ear scanners is Otoscan [38] which uses tracking cameras, ring lasers, and a trigger button to create a 3D ear model. The sophisticated procedure to insert the probe inside the user's ear demands use by trained operators only. The operator begins by setting the depth gauge and follows with a Canal ring, Pinna line, and Concha ring scan. This information collaboratively creates the 3d ear model [38]. Although the product can make an accurate 3D model, it requires an expert operator and is fairly expensive($12,500 capital investment and $1250 annual subscription). Therefore, this

is not a suitable solution outside the four walls of the hospital.

2.3 Microphone Array-based Direction and Distance Estimation

As rotating microphone and speaker is infeasible for a tiny enclosed structure like an in-ear application and surely is unfeasible for a sophisticated structure like a human ear, we explore the rich literature of microphone-array-based direction and distance estimation, which can be the base of generating a 3D model. We specifically look into the Direction of Arrival estimation, Beamforming, and Distance Estimation.

Direction of Arrival (DOA) Estimation. In essence, Direction of Arrival (DoA) refers to determining the spatial direction of an incoming signal or wavefront relative to a reference point or coordinate system. This information can be used to identify the position of the signal's source or to improve the reception quality of the signal by adjusting the receiver's orientation or beam pattern[58]. An illustration of DOA is in the following figure.

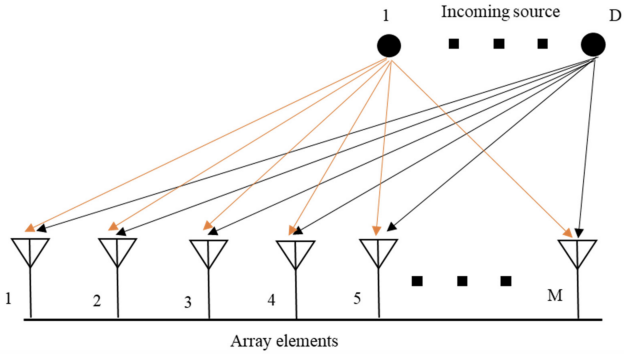

Figure 3: Direction of arrival illustration[59]

The most popular approach to estimate the direction of an arrived audio signal is MUltiple SIgnal Classification (MUSIC). The MUSIC algorithm estimates the Direction of Arrival (DOA) of audio signals using sensor arrays. It involves data collection, covariance matrix estimation, eigenvalue decomposition, and calculation of the MUSIC spectrum. By exploiting orthogonality between signal and noise subspaces and analyzing peaks in the MUSIC spectrum, the algorithm identifies the estimated DOAs of the audio sources[58].

However, traditional 2D MUSIC algorithms fix the azimuth or the elevation and then look for the other without taking source directions into account. To address this shortcoming, a fast 2D MUSIC algorithm is developed that only requires three rounds of search to do simultaneous azimuth and elevation searches (henceforth referred to as AESS) [23]. AESS first performs a circle search to identify the general

source directions. Based on these approximations, a subsequent search is started along numerous straight lines. Each source's 2D Direction of Arrival (DOA) is then calculated by searching several tiny concentric circles. AESS does not fix any azimuth and elevation parameters, unlike the 2D MUSIC algorithm. The spectrum search over the angular field of vision is avoided since each search's neighboring point has a distinct azimuth and elevation. Therefore, azimuth and elevation are simultaneously searched to ensure the search path is limited.

Previous work [7] on DOA estimation have shown good performance when the source is far from the microphone array (e.g., over 400m) but the performance degrades with lower SNR. Moreover, very few works have focused on enclosed space which has a lot of multipath. Thus, we see a need explore this unexplored space.

Beamforming. Beamforming,as shown in the figure below, is a signal processing technique used in sensor arrays (e.g., antennas or microphones) to direct the reception or transmission of signals in specific directions. By adjusting the weights and phases of the signals received or transmitted by each sensor, the array can create a focused beam that enhances the desired signal and suppresses interference or noise from other directions[60]. Researchers treat the Microphone array as a multiple-input multiple-output system and study

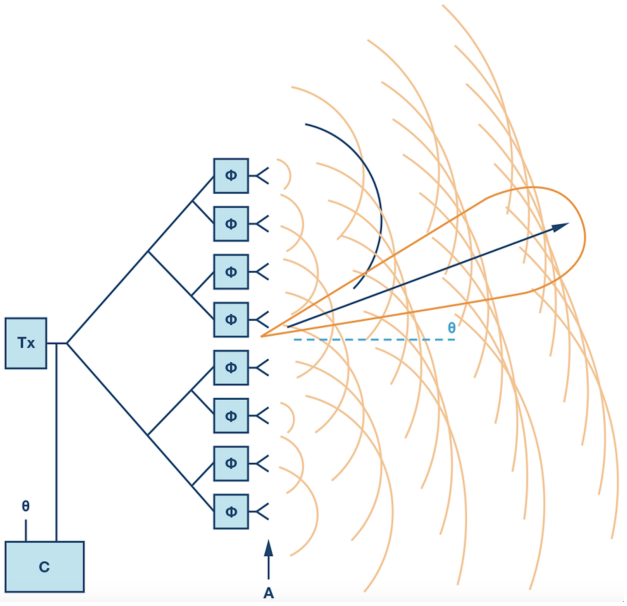

Figure 4: Beamforming illustration[61]

its signal-enhancement performance. Previous work [34] has developed a generic framework for evaluating the performance of beamforming algorithms. First, they create the impulse responses of acoustic MIMO channels. Next, the authors examine the performance limits of beamforming with respect to speech dereverberation and interference suppression. This involves looking at the constraints for the length of the

beamforming filter. Then they study the relationship between beamforming and the multiple-input/output inverse theorem (MINT). Finally, the paper addresses the fundamental connections between various classical beamforming approaches and outlines the conditions that must be met for those techniques to function from a channel condition standpoint. Current beamforming approaches face challenges such as computational complexity, calibration errors, mutual coupling, limited resolution, difficulties with wideband signals, sensitivity to environmental factors, computational latency, and source localization errors. These shortcomings can affect performance and robustness in various applications, and ongoing research aims to address and mitigate these issues.

Distance Estimation. Distance estimation refers to the process of determining the distance between two objects or points. In the context of signal processing and communication systems, it typically involves estimating the distance between a transmitter (source) and a receiver (sensor) based on the properties of the transmitted signal and the received signal[62]. Various audio features change depending on the distance between the sound source and the receiver providing the opportunity to estimate the distance between a microphone and a sound source. However, the way the features change depends on the setup of the microphone as well as the environment. To address this challenge machine learning approaches [50, 52] are used to find the distance between a microphone and a sound source. However, training a machine learning model requires collecting data which is often scarce. Therefore, few works use self-induced noise methods [51]. To illustrate, Calkins et al. [51] use the noise of a vehicle during normal operation as a base to estimate the distance of the vehicle from an unknown object. Even with these new approaches, the effects of the environment, microphone setup, and speaker setup remain a concern.

2.4 Earable bases Sensing

Sensing done through microphones and speakers has many applications that can be applied to humans for verification and health purposes. "Earable" is a term used to describe wearable devices that are designed to be worn on or around the ear, typically to track various physiological and environmental data such as heart rate, body temperature, blood pressure, the shape of the ear canal, and ambient noise levels.

Earable as Authenticator. Biometric authentication based on human ears has been mostly discussed in the context of image recognition. A promising approach to in-ear authentication is based on the unique acoustic characteristics of individual ear canals [27, 21].

EarEcho [27] is a wearable authentication device that users can easily use and conveniently access. EarEcho takes advantage of the unique shape of the human ear canal. A chirp signal is emitted from a

Product	**Size**	Creates a model	Modality	Accessibility	Price
Otoscan	Large	Yes	No	Light	$12,500
3d sonar visualizer	Large	Yes	Sound	No	$1k-10k
Mapping the world in 3d	Large	Yes	Light	No	$5-30/acre

Table 1: Existing Products

speaker and when a signal is received back, it is analyzed and classified using a Support Vector Machine classifier to authenticate the user. The shortcoming is that a low-frequency sinusoidal tone is used, which is painful to hear. Mahto et al. [21] propose a unique biometric authentication method that exploits the acoustic characteristics of human ears and achieves an EER of less than 1%. However, the audible sound signals used in this method are noticeable and may often interrupt user activities during authentication.

Therefore, Xie et al.[47] attempt to overcome this problem by employing inaudible sound signals to achieve silent authentication so that repetitive and continuous authentication will not irritate or interrupt the user. The authors investigate novel biometric traits based on dental occlusion and discover that the bone-conducted sound of dental occlusion recorded in the canals of the ears contains distinctive characteristics of various bones and teeth. This inspired the authors to develop the new identification system TeethPass, which collects occlusal noises in binaural canals using headphones. To identify bone-conducted noises, they develop an event detection technique based on spectrum variance and double thresholds. After filtering out motion noises from the sounds using time-frequency analysis, three distinct user characteristics are derived: bone structure, occlusal position, and occlusal sound. Finally, the authors create a Siamese network based on incremental learning to build the classifier.

Earable for Monitoring Health. Sensors near the ear can be used to monitor many characteristics relating to health. This includes step counting, breath detection, heartbeat measurements, ear-wax detection, blood flow, and in-ear fluid monitoring[3]. As an example, the blood flow signal generated by a heartbeat can be captured using a flexible low-noise photoplethysmography (PPG) sensor that can be attached to the skin behind the ear[3], or the presence of liquid in the middle ear can be found using a speaker and microphone to generate and detect sound waves that penetrate the ear canal.[8].

While researching our earbud design, we encountered several products related to our goal. Each one with different advantages and disadvantages. Table1 demonstrates some broad details about existing products.

3 System Design

This section delves into the intricacies of the system design, touching upon the experiments, equipment setup, design choices, hardware comparisons, and the coding environment.

For acoustic monitoring, we use a speaker and microphone array where the speaker plays audible sounds inside the structure, e.g., ear. This sound is then reflected and refracted by different surfaces inside the structure, and the microphone array captures the original, reflected, and refracted signals. These reflected and refracted signals embed the characteristics of the structure, which can be used for 3D reconstruction or structure health monitoring. We envision that signal processing and machine learning algorithms have the potential to extract such information from the received signal at the microphone array. To evaluate our hypothesis, we develop a proof of concept focusing on small encloses structured of various shapes. The target objects (shapes and materials), data collection hardware, environment condition (i.e., background noise), and procedure are described in this section.

Small space acoustic monitoring can be difficult for several reasons. One of the main challenges is the presence of reflections and reverberation. In a small space, sound waves can bounce off surfaces, creating multiple reflections and reverberations that can be difficult to distinguish from the original sound. Furthermore, nearby sources of sound can interfere with the monitoring process, which can include sounds from adjacent rooms, people talking or moving nearby, and other sources that can make it difficult to distinguish the sounds of interest[63]. The traditional size of space in acoustic imaging can vary depending on the applications. However, acoustic imaging is typically used for imaging objects or structures that are larger than a few centimeters and smaller than several meters in size[64]. The limitations in resolution and frequency range make it challenging to accurately analyze objects that are smaller than a few centimeters in size. Additionally, hollow areas are prone to sound reflections making small hollow areas a distinct issue for acoustic imaging. As a results, alternative imaging methods, such as optical, are often used for study of smaller objects[64]. Through our evaluation, we aim to answer the following questions.

1. Can audible signal identify the shape of a small enclosed structure?
2. If variation in the structure parameter, e.g., height and materials, and various background noise is introduced, how does the shape identification get affected?
3. Can audible signal identify the shape correctly if the structure deforms? Can it differentiate between healthy and unhealthy (deformed) objects?

4. Is it possible to identify the location of the deformation?

5. Can intrinsic parameters of the object's shape, e.g., height, be estimated?

3.1 System Design Requirement

To determine and design the data collection platform for our application, we first identify the requirements the hardware needs to satisfy.

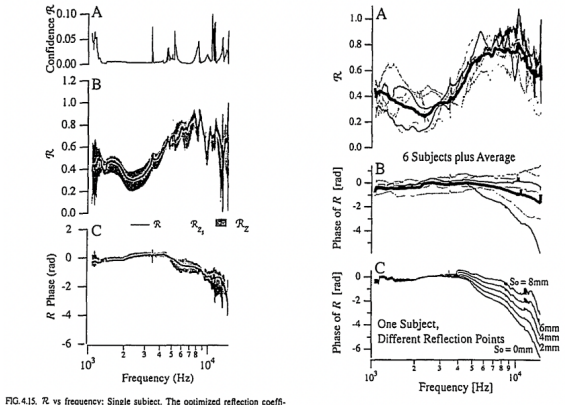

FIG.4.15. $\mathcal{R}$ vs frequency: Single subject. The optimized reflection coeffi-

Figure 5: Different Frequency and reflection points on a surface[12]

Q1. What should be the characteristics of the transmitted audible signal?

As our long-term goal is to use the proposed system for smaller structures like a human ear, we use this as the condition for determining the requirements. Although the ear can perceive frequencies between 20Hz and 20KHz, the spatial resolution of the ear canal when using a frequency below 9KHz can be very limited[12]. Thus, the microphone must have a frequency range including values greater than 9KHz to achieve good resolution of the human ear canal. As shown in Figure 5, an increase in frequency results in a smaller surface reflection. Thus, we use a 10 kHz square wave that can properly be reflected by surfaces with an area less than 8mm. Since the shapes being used in the experiments will have dimensions at least 10 times bigger than 8mm, 10KHz is more than sufficient. A 10 kHz square wave sound signal with a wavelength of approximately 3.4 cm can be properly reflected by surfaces with an area less than 8 mm. This is because the wave can diffract around surfaces that are smaller than its wavelength, allowing it to reflect off of surfaces that are much smaller than the wavelength of the wave itself.

Q2. What should be the minimum distance between the microphones in the microphone array?

For a Microphone array that has a sampling rate of 44.1kHz. The minimum distance required based on the equation below. The speed of sound is $340ms^{-1}$.

$$V = f \times \lambda \tag{1}$$

$$d = \frac{\lambda}{2} \tag{2}$$

$$V = 2 \times f \times d \tag{3}$$

$$d = \frac{V}{2f} \tag{4}$$

As a result, we can get a distance of 0.385cm

$$d = \frac{340ms^{-1}}{2 \times 44100} = 0.385cm \tag{5}$$

Q3. How many microphones are required at a minimum?

Though most embedded and tiny microphones are omnidirectional, a microphone array of multiple omnidirectional array can receive and process sounds from different directions using algorithms like beamforming. In general more microphones provides more accurate results. The minimum number of microphones needed to perform beamforming and obtain directional information is 2. However, a two-microphone array has limitations in terms of understanding the degree of arrival of sound. In order to solve this issue, microphone arrays with three or more microphones are typically required[65]. For these reasons, this project uses a four microphone array.

Q4. Which microcontroller can support the needed sampling rate and microphone array?

The choice of MCU is based on the specific requirements needed, including the amount of data, baud rate, number of microphones, and frequency range. The MCU must have a slot for a sim card in order to ensure space for all the data. Furthermore, a high baud rate is needed to provide a high quality and fidelity of the audio signal. The MCU must also be able to support four microphones that sample over 20KHz in order to support the 10KHz audio being played from the speaker. In this project, we use a 44.1KHz sampling because it meets the sampling requirements and is standard for WAV audio files.

Q5. What is the size of the system?

Sampling at 44.1KHz, a squared microphone array has to be at least 0.385cm by 0.385cm and must

have room for a central speaker. However, in this proof of concept phase, we are using a larger system. The system in this project(microcontroller, microphone array, and speaker) is 8.5cm by 6.5cm. This is much larger than the smallest possible system.

3.2 System Design Choices

We perform a value analysis to determine the system components that meet the system design requirements.

Microphone Selection. Table 2 compares a wide range of microphones to identify the optimal candidate. We extensively study the resources on this topic available on the internet. We mostly focus on Micro-electromechanical systems (MEMS) microphones. MEMS constitute the technology of microscopic devices, particularly the ones with moving parts, and are based on components of 1 and 100 micrometers in size. Thus, most MEMS devices range in size between 20μm to 1mm. A MEMS microphone is an electro-acoustic transducer with a sensor (MEMS) and an application-specific integrated circuit (ASIC) in a single package. The sensor converts variable sound pressure to capacitance variations that the ASIC transforms into analog or digital output.

Instead of choosing individual microphones and developing the board ourselves, we choose the Seeed studio ReSpeaker microphone array. Besides having the largest frequency range in a relatively small footprint, ReSpeaker has integrated data multichannel support using AC108, which has highly integrated quad-channel ADC with I2S/TDM output transition and can seamlessly integrate with many processing units. The Seeed studio Reaspeaker squared 4 mic array uses 4 analog microphones, the distance between each microphone is around 65mm. It features 12 GPIO pins, including 6 digital pins and 6 analog pins. The Reaspeaker connects to a host device via either a USB or I2S interface. While the UART protocol typically employs a specific baud rate for transmitting data bit by byte, both USB and I2S diverge from this approach. USB, a digital communication standard, facilitates high-speed data transfer, while I2S is a protocol specifically designed for the transmission of digital audio data. Utilizing packet-based data transfer, USB and I2S transfer data in chunks rather than bit by bit. As a result, The Reaspeaker does not have a specific baud rate.

Processor Selection. The integral component of this system is a microcontroller, the system's brain. The microcontroller must be compatible with the speaker, support the data transfer rate to collect data from the microphone array, and can support a high sampling rate. Table 3 compares several of the shelf microcontrollers. We choose Raspberry Pi 4 model B for developing this system for two reasons – (1) the processing

Microphone	Digital/Analog	Frequency(Hz)	Size of module	Number of mics	SNR(dbA)	Sensitivity	Mono/Stereo
I2S-SPH0645LM4H-B	Digital	64KHz	16.7x 12.7x 1.8 mm	1	65dB(A)	-26dBFS	Mono
SPH8878LR5H-1	Analog	80KHz	13.97x10.16mm	1	66dBV/Pa	-44dBV/Pa	Mono
ReSpeaker 4-Mic Array	Analog	96KHz	65mm x65mm x9mm	4	59dB(A)	-22dBFS	Mono and Stereo
ReSpeaker Mic Array - Far-field w/ 7 PDM Microphones	Digital	32KHz	148mm x100mm x32mm	7	61dB	-26dBFS	Mono and Stereo
UMA-8-SP USB mic array	Digital	48KHz	90 mm diameter 20mm height(with LED)	8	64 dB	-29dB	Stereo

Table 2: Microphone Comparison

performance and data transfer rate of Raspberry Pi are sufficient for this system, and (2) ReSpeaker is exclusively designed for Raspberry Pi that allows us to utilize the built-in software for synchronized multichannel data collection. Additionally, the architecture of the Raspberry Pi is ARM Cortex-A72, it has 26 GPIO pins, and has 8GB of RAM.

Microcontrollers	Architecture	Supports over 44.1kHz	GPIO Pins	DAC	ADC	RAM Size	Flash Size
Cortex M0	Armv6-M	Yes	20	YES	NO	92kb	256kb
Cortex M4	Armv7-M	Yes	25	Two(A0 and A1)	8 Analog Pin	192kb	512kb
Raspberry Pi	ARM Cortex-A72	Yes	26	No	No	8GB	microSD card Storage
Arduino Uno R3	ATMega328P	Yes	14	Analog input	Digital input	512B SRAM	16 KB ISP Flash
Bela		Yes	28				

Table 3: Processor Comparison

Figure 6: Speaker Picture

Speaker Selection. We choose a basic earbud speaker, Airpods Pro (shown in Figure 6), for three reasons – (1) it is small enough to fita small structure like human ear, (2)it is capable of playing audio at 15 KHz, which is higher than our desired 10 KHz, (3) has a good dynamic range, and (4) is widely available. Therefore, we can satisfy the size and dynamic range requirements.

System Overview. The Reaspeaker and Raspberry Pi are connected via I2S pins on the Pi. The connection between Reaspeaker and Raspberry pi is shown in Figure7 and listed below.

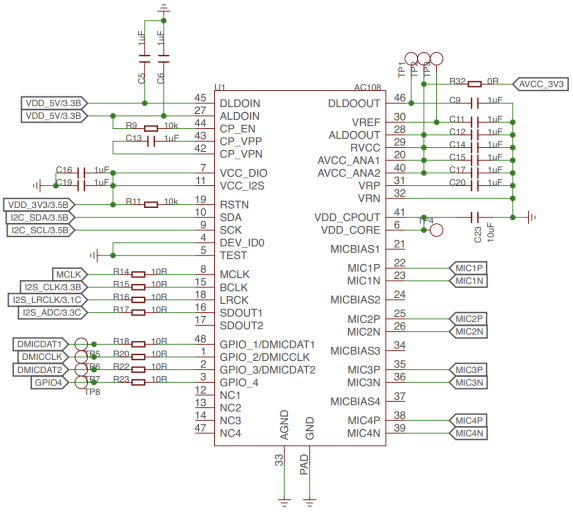

Figure 7: Circuit Diagram of Reaspeaker

- I2S Data: GPIO 20 (Pin 38) on the Raspberry Pi
- I2S BCLK (Bit Clock): GPIO 18 (Pin 12) on the Raspberry Pi
- I2S LRCLK (Left/Right Clock): GPIO 19 (Pin 35) on the Raspberry Pi
- I2S MCLK (Master Clock): GPIO 12 (Pin 32) on the Raspberry Pi
- LED Data (for the programmable RGB LED ring): GPIO 5 (Pin 29) on the Raspberry Pi

During the installation of the most recent Raspberry Pi OS, we face several challenges, such as black screens while recording and unforeseen interruptions while configuring the required environments. To overcome these issues, we chose to utilize a 2021 version of the Raspberry Pi OS. Additionally, we employ Audacity to capture audio data for subsequent analysis. Moreover, we have incorporate the ODAS (Open embeddeD Audition System) Studio, a tool commonly employed for sound source localization, tracking, and separation tasks. However, we discover that its application is significantly constrained within confined spaces, such as the 80mm x 80mm 3D-printed structure we utilize in our case Figure8.

Programming Environment. The programming environment for our project utilizes Python and an Integrated Development Environment (IDE) on the Raspberry Pi. We have installed the 32-bit Raspberry Pi OS, which includes the Python IDE, onto an SD card. The Raspberry Pi is powered through a USB-C port, and the display output is facilitated by a Micro HDMI port. To record the audio reflections from each shape on the Raspberry Pi, we employ Audacity, a versatile audio recording and editing software. The audio data captured by the four microphones is then uploaded to MATLAB, a powerful computational software,

for processing and analysis of the reflection signal data. This comprehensive approach enables us to study the audio properties of the shapes in a systematic and accurate manner.

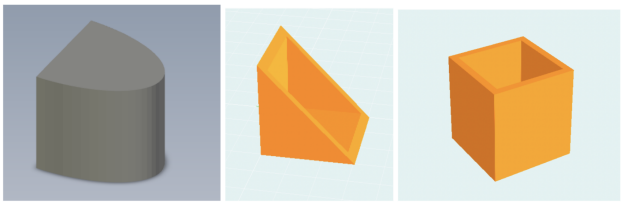

Figure 8: 3D Printed Objects

4 Methodology

This section analyzes the potential of different acoustic features and signal characteristics for (i) detecting different shapes, (ii) identifying any deformation, and (iii) localizing the deformed sides. A block diagram is shown in the following figure.

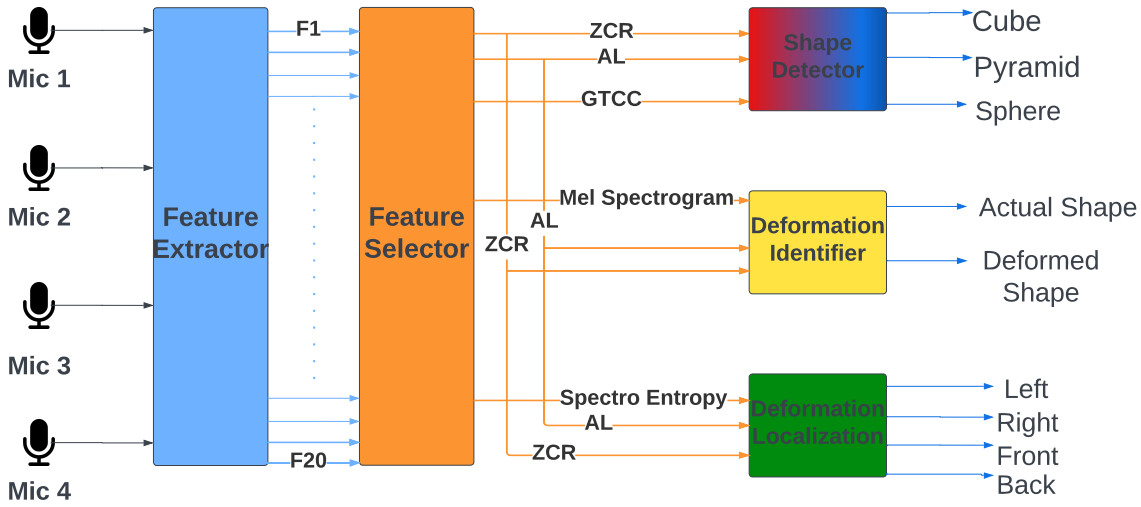

Figure 9: Block Diagram

4.1 Acoustic Features Calculation

We analyze 20 different acoustic features. The Significant 15 features are – (i) frequency response, (ii) zero crossing rate, (iii) acoustic loudness, (iv) Mel spectrogram, (v) spectral centroid, (vi) spectral entropy, (vii) spectral decrease, (viii) spectral spread, (ix) spectral roll-off point, (x)spectral skewness, (xi) spectral flux, (xii) spectral flatness, (xiii) spectral kurtosis, (xiv) gamma-tone cepstral coefficient, (xv) harmonic ratio, and (xvi) power spectrogram. Following is the description of different features.

Frequency Response (FR). Frequency response is a system characteristic that describes how it reacts to different input frequencies. In the context of audio, signal processing, frequency response is used to assess the performance and behavior of a sound across a range of frequencies[66]. The frequency response aims to observe frequency changes due to different reflections and refractions and gain insight into which audio features can best differentiate between small hollow areas. The first step in finding the frequency response is to collect time-domain audio signals from the four microphone channels. Next, various signal processing techniques are

performed to extract the frequency response. One of the methods is the Fast Fourier Transform(FFT)[67]. FFT offers significant advantages regarding computational efficiency, applicability, and frequency resolution. However, it also has limitations, such as reduced time-frequency resolution for non-stationary signals, sensitivity to noise, and issues related to leakage and spectral smearing[67]. Another technique used to capture the frequency response is the mel Spectrogram. Mel Spectrograms offer several advantages over the FFT, including better alignment with human perception, improved frequency discrimination, and noise reduction. However, also comes with some disadvantages, such as reduced frequency and time resolution and sensitivity to windowing and other parameters[69].

Zero Crossing Rate (ZCR). The zero crossing rate (ZCR) is a measure of the number of times that the amplitude of a signal changes from positive to negative (or vice versa) over a given period [46]. The ZCR is calculated by counting the number of times the signal crosses the zero amplitude line within a fixed window of time and then normalizing the count by the length of the window. ZCR is a standard tool to differentiate between audio signals. The ZCR is defined by the following equation below[46].

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} |sgn[x_i(n)] - sgn[x_i(n-1)]| \tag{6}$$

Acoustic Loudness (AL). Acoustic loudness measures perceived sound pressure [33]. Acoustic loudness is measured in sones, a useful tool for distinguishing signal sound pressure differences.

Mel Spectrogram (Mel-Spec). Mel Spectogrm is a way of finding the frequency response of a signal on the mel-scale. Notably, the mel-scale closely mimics human perception, providing an accurate representation of the frequencies humans perceive[71].

Spectral Centroid (SCentroid). The spectral centroid [36] measures the center of mass of the frequency components of a signal. We calculate it by taking the weighted mean of the frequencies present in the signal, with the weighting applied based on the amplitude of each frequency. The spectral centroid can represent the "brightness" or "tonality" of a signal. The following equation defines the spectral centroid.

$$C_i = \frac{\sum_{k=1}^{Wf_L} kX_i(k)}{\sum_{k=1}^{Wf_L} X_i(k)} \tag{7}$$

Spectral Entropy (SEntropy). Spectral Entropy is a measure of a signals power distribution[72].

Spectral Decrease (SDecrease).

Spectral Decrease is a method of measuring the decrease of a spectrum while highlighting the slope

of the lower frequencies present in the signal[73].

Spectral Spread (SSpread). Spectral Spread denotes the average deviation of the rate-map from the signals centroid[74].

Spectral Roll-Off Point (SRP). The Spectral Roll-Off Point of a signal is the point at which 85% of the power is at lower frequencies. This measures the right-skewedness of the power spectrum[74].

Spectral Skewness (SSkew).

Spectral Skweness is a measure of the location of the concentration of spectral energy with respect to the spectral center[74].

Spectral Flux (SFlux). Spectral Flux measures the rate at which the power spectrum changes[70].

Spectral Flatness (SFlat). Spectral Flatness quantifies how close a sound is to being a pure tone or noise[68].

Spectral Kurtosis (SKurt). Spectral Kurtosis indicates the non stationary or non-Gaussian behavior of a signal in the frequency domain[75].

Gamma Tone Cepstral Coefficients (GTCC). GTCC are a set of features extracted from audio that describe the spectral shape of a signal. GTCC is commonly used for audio recognition tasks[76].

Harmonic Ratio (HR). Harmonic Ratio is a ratio used to distinguish periodic audio and non periodic audio. This is done by comparing the harmonic portion of audio to the total energy of the audio[74].

Power Spectrum (PowerS). The power Spectrum indicates the magnitude of the different frequencies that make up the entire signal[77].

4.2 Acoustic Features Selection.

By analyzing both the frequency and time responses of the recorded audio signals, we identify signal characteristics that appear to be dependent on the shape of the objects. After pinpointing these shape-related characteristics, we extract specific audio features that quantify these attributes, enabling a more accurate assessment of the relationship between the signal and the object's shape. Once these individual audio features are found they are tested to see which ones produce the most accurate results. The ones that individually work best are combined to attempt to increase the accuracy of the results. However, combining the best individual features does not always increase the accuracy and therefore random combinations are also tried.

4.3 Shape Identifier

A shape identifier is developed using a discriminant machine learning classifier to determine the shape of an object based on extracted audio features. Discriminant Analysis classifiers are a family of machine learning algorithms. The main goal of this method is to find combinations of features that best separate the groups, maximizing the differences between them while minimizing the variance within each group. There are two primary types of Discriminant Analysis Classification: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). QDA does not assume equal covariance matrices for the groups. Instead, it allows for different covariance matrices and fits a quadratic decision boundary between the groups. We use a QDA which can be more flexible than LDA but may require more data to provide accurate estimates of the covariance matrices. They are particularly useful for classification tasks and dimensionality reduction. The classifiers input is individual audio features and combinations of audio features, and the output is the predicted shape(cube, pyramid, or sphere) Figure8. In this case the best feature combination under ideal and unideal conditions is Zero Crossing Rate, Acoustic Loudness, and GTCC.

4.4 Deformation Detector

A deformation detection system has been designed using the same discriminative machine learning classifier as the Shape Identifier, which discerns whether a shape is healthy or obstructed by analyzing extracted audio features. The classifier processes individual audio features and their combinations as input and generates an output indicating whether the shape is healthy or obstructed, as illustrated in Figure10. Notably, when using the Mel Spectrogram as a standalone input feature, the classifier achieved remarkably high accuracy.

4.5 Deformation Localizer

A deformation localizer is developed in the same manner as the Deformation Detector to identify which side of a shape (left, right, back, or front) is obstructed based on extracted audio features. The classifier's input consists of individual audio features and various combinations thereof, while the output predicts the deformation's location (left, right, back, or front), as depicted in Figure11. In this scenario, the optimal feature combination includes Zero Crossing Rate, Acoustic Loudness, and Spectral Entropy.

4.6 Shape Dimension Estimator

Utilizing height measurements alongside the shape identifier can facilitate the creation of a 3D model. Once the shape is identified using the estimator, height information can be combined with other dimensions to generate a 3D representation of the shape. For instance, if the estimator determines the shape to be a cube, the height can be used to define its proportions. This approach can be applied to various shapes, creating detailed three-dimensional representations.

A shape dimension estimator is developed using two different approaches to identify the height of the object based on extracted audio features. The first approach is triangulation and the second approach is a Gradient Boosted Machine (GBM). Triangulation is performed by combining time difference of arrival (TDOA) and direction of arrival (DOA) to estimate the height of the shapes. Through the process of triangulation, we notice that the cube and sphere display the most accurate results, while the pyramid demonstrates a considerably larger variation. This inconsistency can be ascribed to the fluctuations present in the pyramid's top surface, which add greater complexity and influence the model's capacity to generate accurate predictions for this particular shape, as illustrated in Figure19. A Gradient Boosting Machine (GBM) is also employed to estimate the height of an object by leveraging extracted audio features. The model's input is acoustic Loudness, while the output predicts the numerical value corresponding to the object's height. By utilizing the powerful ensemble learning technique of GBM, the model effectively captures complex relationships between acoustic Loudness and the target height, as depicted in. In this project the accuracy of the GBM is limited due to a small data set.

We perform sound localization through the use of time of arrival, direction of arrival, triangulation, and a Gradient Boosted Machine. Together these elements provide information on the location of the sound source(ideally the sound source is the sound reflected from the top of the 3D shape). The goal of this technique is to obtain information on the height of the shape. For the simple shapes used in this proof of concept, just knowing the height can be used to create a 3D plot of the shape, however many more parameters are needed to plot a 3D ear.

Time Difference of Arrival.

Time Difference of Arrival, is a method of estimating distance. The distance is calculated by measuring the difference in time it takes for a signal to reach different microphones. A common way the time difference is found is through the use of cross correlation functions. The delay corresponding to the peak value of the cross correlation function is used as an estimate for the delay between functions.[18]

Direction of Arrival Estimate and triangulation.

The Direction of arrival estimates(DOA) denote the direction from which sound waves arrive at the microphone. The result of direction of arrival estimates is the direction the sound waves are relative to the microphone array[48]. The estimations are done by utilizing the phase information obtained by the microphone separation inside the array[7]. For every two microphones one direction of arrival is found using the equation shown below. In the equation is the direction of arrival v is the speed of sound, is the time difference of arrival, d is the distance between the two microphones being used, R is the average distance between the sound source and microphones, and is the wavelength. When multiple directions are found, trigonometry is used to calculate the location of the sound source.[18]

$$\varphi = cos^{-1} * (\frac{v * \tau}{d}, forR > \lambda) \tag{8}$$

Many of these localization techniques rely on a clear sound source and are performed in large spaces with minimal sound reflections and refractions. These conditions do not exist in our applications and therefore are susceptible to inaccurate results.

Gradient Boosted Machine.

Another localization technique used is a Gradient Boosted Machine(GBM). GBM is a machine learning technique that creates a strong learner by combining weak learners. It belongs to the category of boosting algorithms that correct errors made by previous models through sequentially adding new models. It is effective and robust for classification problems[22]. The training data used in the GBM was the acoustic loudness when sound was played at different distances from the microphone array. The testing data was the acoustic loudness measured when sound was being played inside of the shapes. Similar to the other localization technique(DOA) this method was implemented in an unideal scenario. For one the training data was collected without a shape over the microphones, additionally there was background noise while data was being collected.

5 Dataset

This section outlines a study that examines the impact of small hollow areas on audio features using 3D-printed objects of various shapes. We choose three shapes with diverse top surfaces to test the system's ability to identify shapes, and deformations, and predict the location of the deformation. The section also explores the robustness of the shape differentiation algorithms by altering the base height of the objects.

We collect a total of 210 samples (14 samples for each shape and each task). Each sample is 8 seconds long and consists of 4 WAV files, one for each microphone. The data collected is divided into training and testing sets, and the accuracy of the machine learning models is evaluated through 5 cross-validations.

Details about different environmental noise, room size, and other sources are to be provided in the "Noisy Environment" section. The study collects an unspecified number of samples for each class and divided the data into training (80%) and testing sets. The machine learning models are trained on the training set and tested using the test dataset. The accuracy is reported using cross-validation.

5.1 Target Object Design

We design a group of 3D objects with Solidwork and printed them using a lulzbot TAZ 6 3D printer. These objects are made of ABS plastic and are approximately 3.5 x 3.5 x 3.5 inches. We select three distinct hollow shapes for our study: a cube, a pyramid, and a sphere (Figure8). The reasoning behind choosing these shapes is the diversity of their top surfaces. The cube has a flat top, the pyramid has a pointed top, and the sphere has a rounded top. This variation allows us to investigate the effects of different surface properties on the audio reflections. Additionally, these objects are easy to reproduce, making them practical choices for consistent experimentation.

The deformations are done by blocking a certain part of the cube with tin foil. An example of this is shown in Figure10. In addition to generally deforming the shape, specific sides(left, right, back, and front) of the cube are deformed. This is shown in Figure11. The generally obstructed shapes are used to see if our system can distinguish between a healthy and non healthy cube, and the side obstructed shapes are used to see if our system can predict the location of deformation.

Figure 10: Obstructed cube

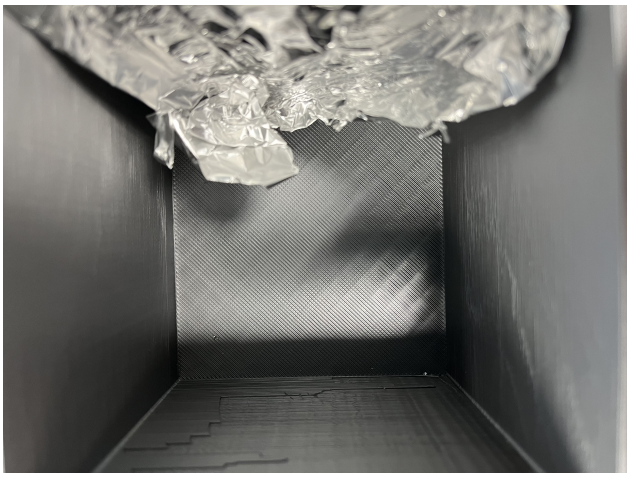

Figure 11: Side Obstructed cube

Since the base of all three shapes is square, multiple 3D printed extension of the base are made to alter the height of the shapes (Figure12). This is used to vary the height by up to 60mm. The varied heights are used to test the robustness of the shape differentiation algorithms.

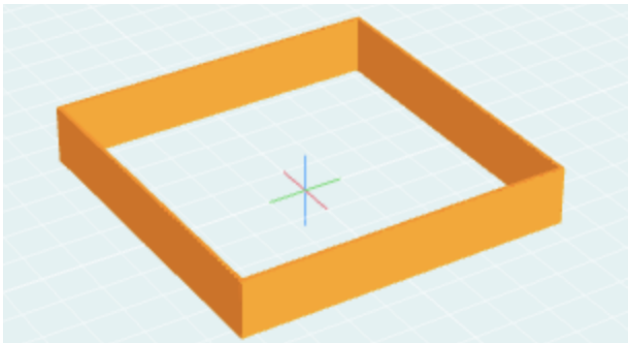

Figure 12: 3D Printed elevation

5.2 Audio Signal

As the dimensions of the shapes involved in the experiments are at least 10 times larger than 8mm, a 10kHz frequency is deemed adequate. A square wave sound signal of 10kHz has a wavelength of approximately 3.4 cm, which can effectively reflect off surfaces with an area smaller than 8 mm. This is possible because the wave can diffract around surfaces that are smaller than its own wavelength, enabling it to reflect off surfaces considerably smaller than the wave's actual wavelength.

5.3 Noisy Environment

The presence of background noise and variations in shape size can considerably affect the recorded audio signals, resulting in a marked decrease in the accuracy of the model's predictions. These external factors introduce further complexities and challenges for the classifier, complicating the accurate identification of patterns and relationships between the audio features and the target outcomes. This experiment was conducted in an uncontrolled environment where people were moving and various noise sources were present, adding to the background noise. These sources include computer fans, air conditioners, human voices, and buzzing sounds from electronic devices. The dynamic nature of the environment, combined with these noise sources, increases the difficulty in isolating the desired signal for accurate classification, posing additional challenges for the model in discerning the relevant information.

5.4 Data Distribution

Around 210 total audio files are collected and divided into 4 classes(each with around 50 audio files) to be used in different experiments. We split the data into two sets – (i) training and (ii) testing. We randomly choose 80% samples for the training set on which we train the machine learning models and we test the machine learning model with the test dataset. We perform 5-fold cross-validation to report the accuracy.

6 Results

This section provides an in-depth analysis of the features and algorithms shown in the previous section.

6.1 Performance of Shape Identification

First, we analyze the performance of the shape detection module that distinguishes between small hollow areas of three different shapes – cube, triangular prism, and sphere. We consider two different scenarios for this evaluation – *ideal condition* and *unideal condition*. Under the ideal condition, the experimental setup and the environment remains the same for both the training and testing scenario. In the unideal condition, we introduce noises as described in the previous section and change the height of the test objects. Moreover, the environment of training and testing are different to test the robustness of the proposed system.

Figure 13 shows that when using a single feature Power Spectrum (Power) and Acoustic Loudness

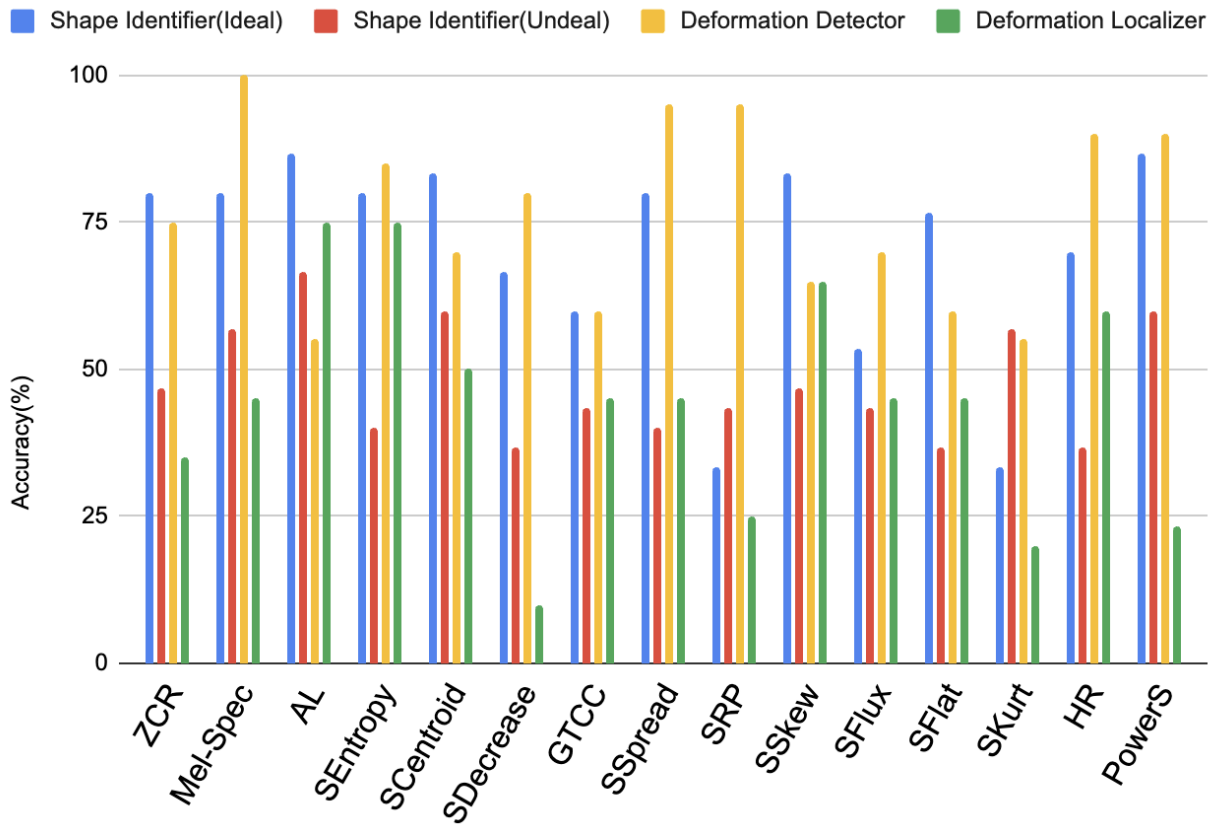

Figure 13: Classification accuracy of shape identification, deformation detection, and deformation localization with individual of acoustic features.

(AL) achieve the highest accuracy of 86.6% in identifying shapes. These features are effective because variations in shape can alter the reverb and sound reflections, which have a direct impact on Acoustic Loudness and Power Spectrum. However, Spectral Kurtosis and Spectral Roll-Off Point achieve lower performance (33.33%) due to the signals having comparable power spectrum distributions, making it challenging to differentiate between them.

In the unideal condition, Acoustic Loudness achieves the highest accuracy of 66.67%, which is 20% less than that of the ideal condition using the same feature. However, Power Spectrum's performance has more accuracy drop (26.67%), indicating the lack of robustness of the feature. Interestingly, the Spectral Roll-Off Point and Spectral Kurtosis have higher accuracy in an unideal scenario compared to an ideal scenario. But as both results are below 50%, we discard these features. Besides these two features, all features experience a drop in accuracy in the unideal scenario.

Next, we combine multiple of these acoustic features to improve the identifiers' performance. Figure 14 shows that in an ideal scenario the highest accuracy is 90% achieved with the following 8 feature

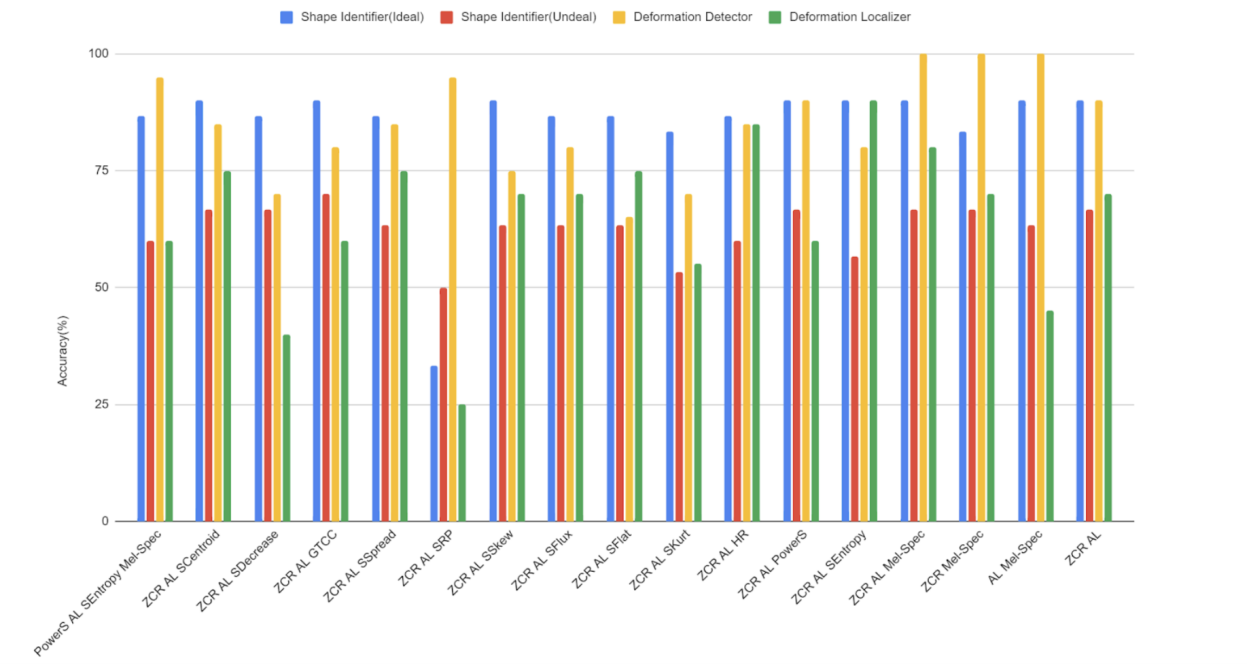

Figure 14: Classification accuracy of shape identification, deformation detection, and deformation localization with combinations of acoustic features.

combinations: [ZCR, AL,SCentroid], [ZCR, AL, GTCC], [ZCR, AL, SSkew], [ZCR, AL, PowerS], [ZCR, AL, SEntropy], [ZCR, AL, Mel-Spec], [AL, Mel-Spec], [ZCR, AL]. However, under unideal conditions, the highest accuracy is 70%, achieved with ZCR, AL, and GTCC. This combination of features works best under ideal and unideal conditions, making it the overall best for shape identification. Combining features increases the accuracy of both the ideal shape identification and the nonideal shape identification by approximately 3%.

6.2 Performance of Deformation Detection

In Figure 13, Mel Spectrogram, Spectral Spread, Spectral Roll-Off Point, Harmonic Ratio have more than 80% accuracy in differentiating between a perfect and deformed shape. However, Mel Spectrogram achieves 100% accuracy because it accurately captures the 20Hz-20KHz changes caused by the shape deformities. Notice that though Acoustic Loudness achieves the overall best performance in shape identification, it performs poorly (55%) when detecting deformation.

Figure 14 shows that 3 combinations of features achieve 100% accuracy: [ZCR, AL, Mel-Spec], [AL, Mel-Spec], [ZCR, Mel-Spec].

All three of these combinations contain Mel-Spec, which achieves 100% accuracy on its own. The

worst feature combination for deformation detection is ZCR, AL, SFlat achieving 65% accuracy. This is 5% higher than SFlats individual deformation detection accuracy.

6.3 Performance of Deformation Localization

We look into the performance of the deformation localizer which detects whether the deformation happened on the left, right, back, or front of a cube. Figure 13 shows that Spectral Entropy and Acoustic Loudness achieve the highest accuracy (75%) for this four-class classification while Spectral Decrease achieves only 10% accuracy, implying that deformation does not have a large impact on the slope of the lower frequencies in the audio signals.

The best combination of features(ZCR, AL, and SEntropy) 14 increases the accuracy to 90%. This is 15% higher than the individual accuracy's of SEntropy and AL. ZCR, AL, and Harmonic Ratio obtain a slightly lower accuracy (85%) which is interesting given that Harmonic Ratios individual accuracy is only 60%. ZCR, AL, and SEntropy result in a higher accuracy than the other combinations and therefore should be used for deformation localization. Although this combination of features performs best for deformation localization and ideal shape identification, it does not perform best for non ideal shape identification and deformation detection. This implies that ZCR, AL, and SEntropy are more sensitive to side obstructions and ideal shape changes than general obstructions and non ideal shape changes.

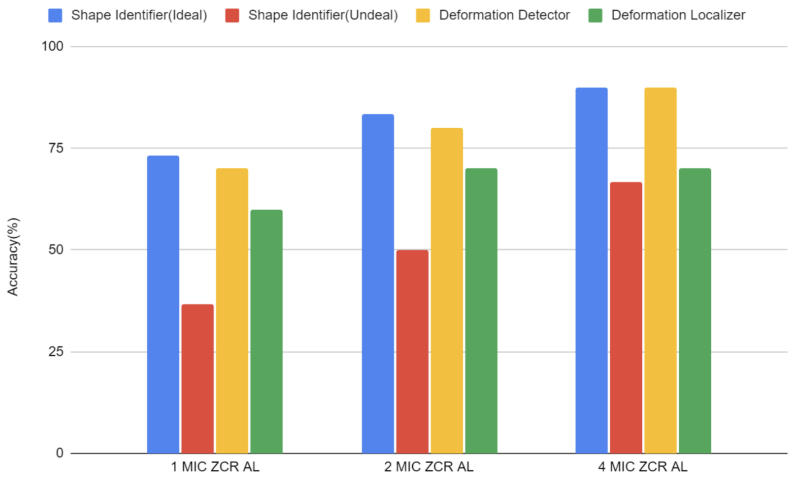

Figure 15: Effect of the number of microphones of performance.

6.4 Effect of Number of Microphones

To understand the relationship between the number of microphones and the performance, we analyze the performance of shape identification, deformation detection, and deformation localization using a combination feature – Zero Crossing Rate (ZCR) and Acoustic Loudness (AL). We consider this combination of features because - It shows a stable results for both individual feature and multiple features in accurately identifying the shape, detecting deformation, and localizing deformation. Figure 15 shows that performance drops with the number of microphones. While deformation localization remains relatively unaffected by the number of microphones used, shape identification experiences the most significant performance decrease in non ideal scenarios. Interestingly, our results also show that the number of microphones has a varying impact on different experiments. For instance, we observed a 10% increase in performance between one to two microphones, but no significant improvement between two to four microphones. These findings suggest that the optimal number of microphones required may depend on the specific application and use case. In some scenarios, two microphones may be sufficient to achieve optimal performance, while in others, a greater number may be required.

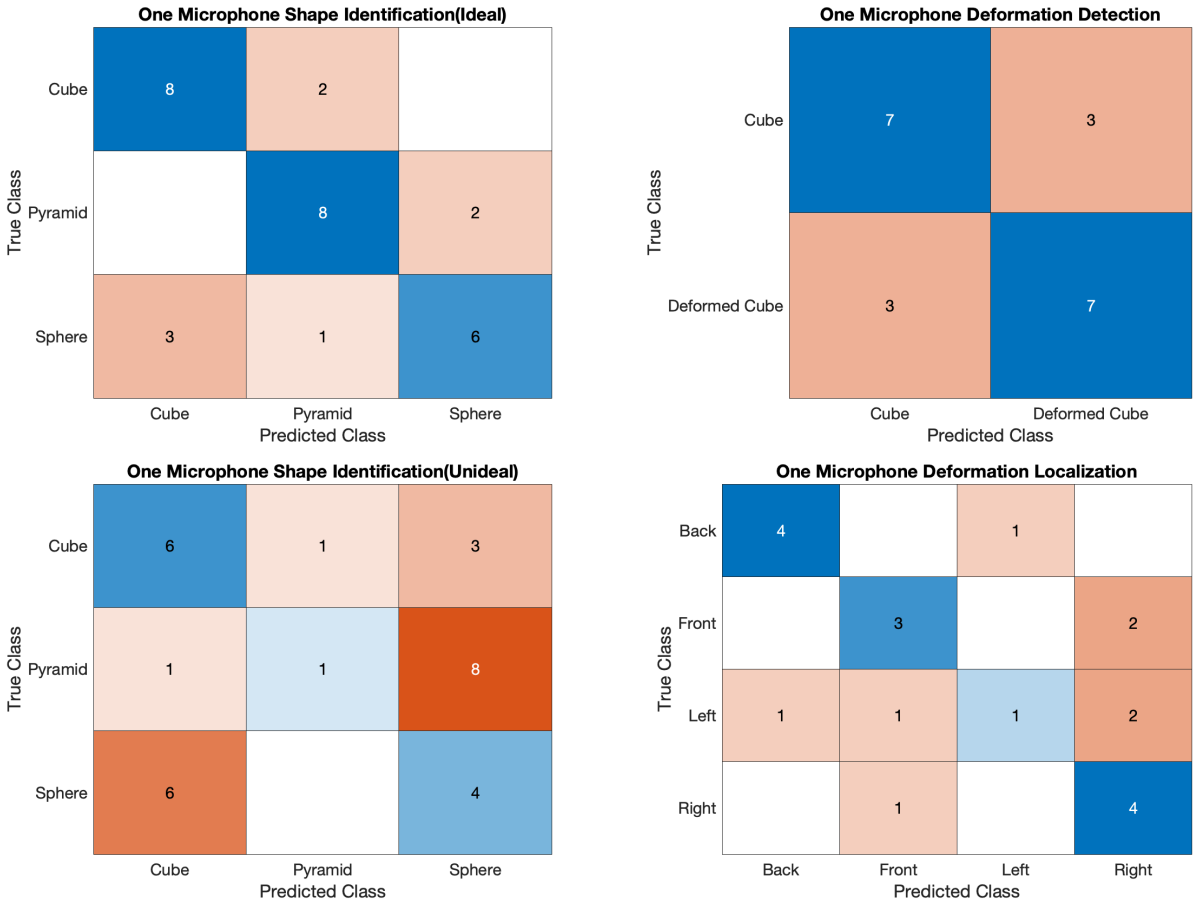

Figure 16: Confusion matrices for a single microphone setup using two features for shape identification in an ideal scenario (Top Left), shape identification in an unideal scenario(Bottom Left), deformation detection (Top Right), and deformation localization (Bottom Right).

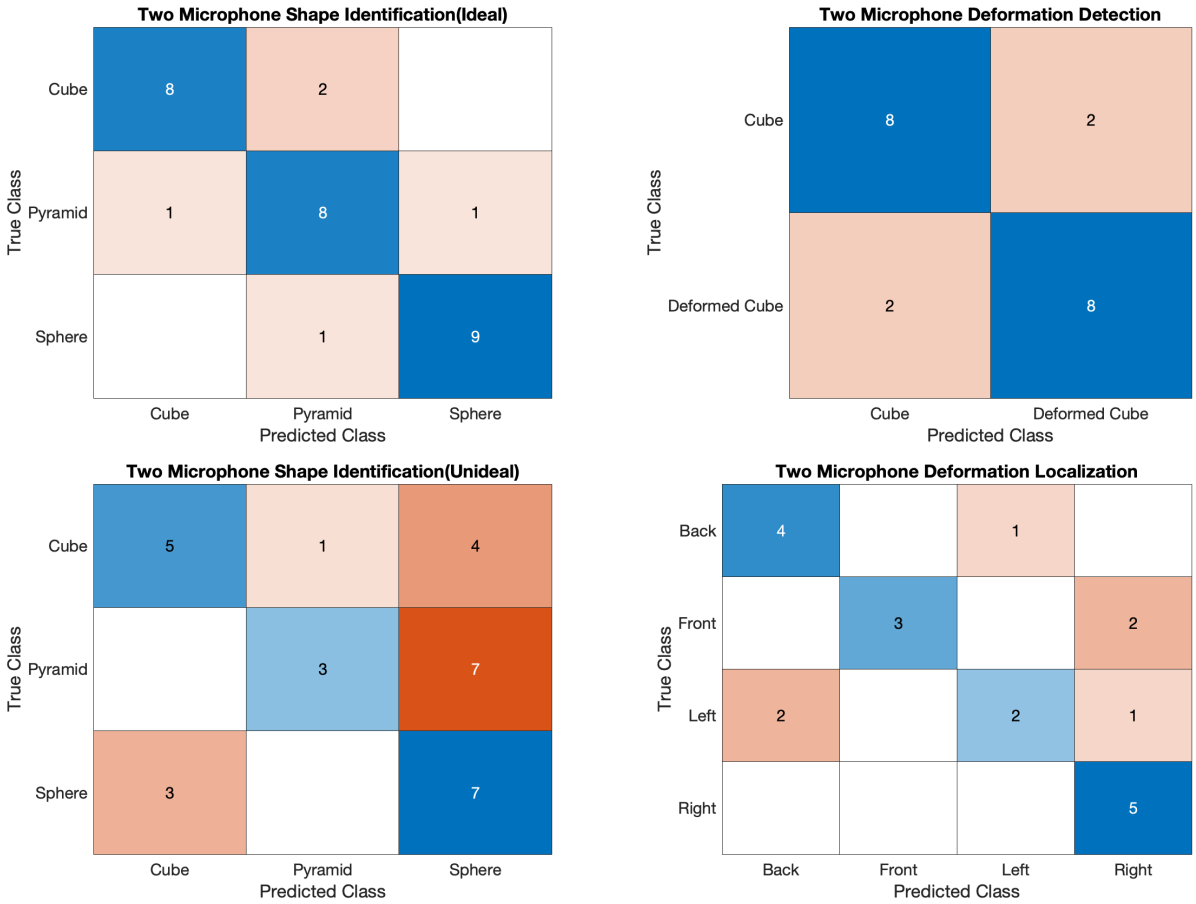

Figure 17: Confusion matrices for a two-microphone setup using two features for shape identification in an ideal scenario (Top Left), shape identification in an unideal scenario(Bottom Left), deformation detection (Top Right), and deformation localization (Bottom Right).

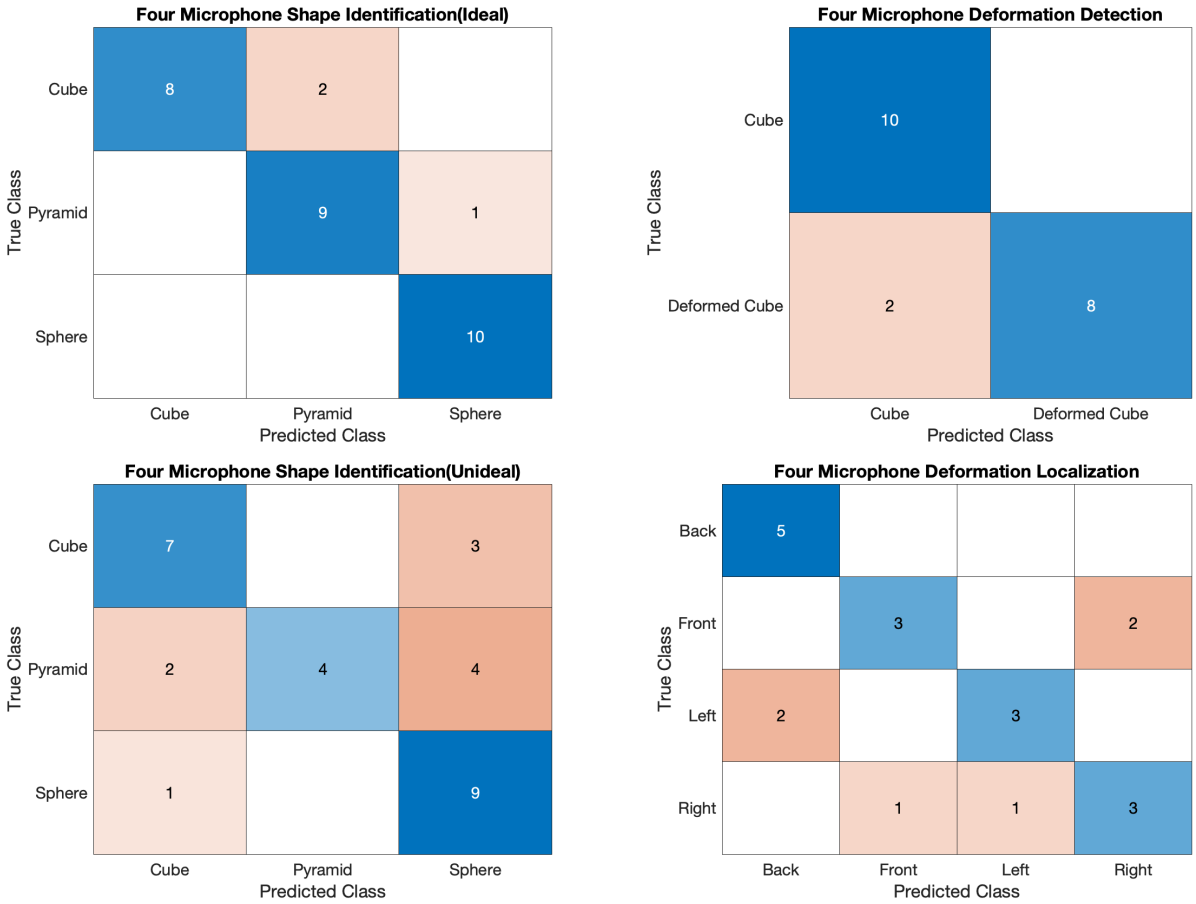

Figure 18: Confusion matrices for a four-microphone setup using two features for shape identification in an ideal scenario (Top Left), shape identification in an unideal scenario(Bottom Left), deformation detection (Top Right), and deformation localization (Bottom Right).

Figures 16,17 and 18 show the confusion matrices for all the experiments with one, two, and four microphones, respectively. From these confusion matrices, we see that for non ideal shape identification, less than four microphones confuse Pyramid with Sphere. This is the main contributor to performance degradation in this scenario. Moreover, we observe that the number of microphones impacts not only the accuracy of the experiment but also how the errors occurred. Notably, the number of ways an error occurred decreased with increased microphones. For instance, shape identification in an ideal scenario confuses Sphere and Cube with one microphone, but two microphones can eliminate this confusion.

6.5 Shape Dimension Estimator

First, we use the triangulation method with a 10 sample lag limit to determine the height of different shapes. Figure 19 shows that triangulation-based height estimation is inconsistent, and the estimation error varies from -53.72mm to 12.89mm, 26.0326389mm to 738.95mm, and -40.11mm to 190.34mm for the cube, pyramid, and sphere, respectively. Note that the actual height of each shape is approximately 85mm. We limit the maximum sample delays between the microphones to minimize this error. Additionally, Figure 19 shows that the average estimation error for the height of the cube is only -9.84mm, while for the sphere, it increases to 48.96mm. The pyramid's height estimation is highly inaccurate. This error is due to the special structure of the top surface.

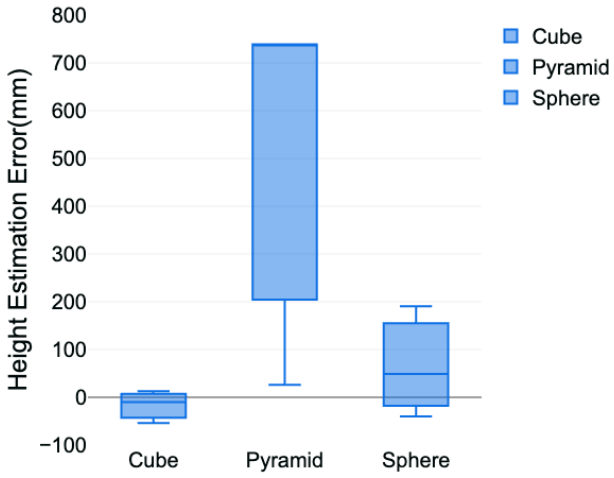

Figure 19: Performance of the Shape Dimension (Height) Estimator with Triangulation and Sample Delay Limitation.

However, when the number of lags is not limited to ten and triangulation is performed the results are imaginary4. This is because finding the delay between microphones through a correlation function produces wildly varying and often illogical results causing the trigonometry function used to calculate distance to output imaginary heights. These results indicate that triangulation based on the delay between microphones is not ideal for this application.

Next, we perform a gradient-boosted machine (GBM)technique that utilizes acoustic Loudness, to predict the height of the shape. Due to a limited amount of data, this approach could not distinguish between heights that were not specifically imputed as data. Thus, the GBM always predicted the height of

Shapes	**Height(mm)**
Cube_1	28.7182333373835 + 0.0487262010197422i
Cube_2	0 + 0.0365403917080402i
Cube_3	0 + 0.0399341621598897i
Pyramid_1	0 + 0.0399999657677704i
Pyramid_2	0 + 0.0360547465672955i
Pyramid_3	0 + 0.0354565634580843i
Sphere_1	0 + 0.0399995682894515i
Sphere_2	0 + 0.0368586158570601i
Sphere_3	0 + 0.0399990640022761i

Table 4: Performance of the Shape Dimension (Height) Estimator with Triangulation only

.

the shapes to be 0.063mm. However, we envision that with more data we can achieve better performance with this approach because the acoustic loudness recorded by the microphones varied based on the height of the shape. For instance, an 80mm box resulted in an acoustic Loudness of MIC1=37.1410dB, MIC2= 38.6732dB, MIC3= 36.3973dB, MIC4= 38.94930dB and a lower box(40mm) resulted in an acoustic Loudness of (MIC1=40.6866dB, MIC2= 41.5314dB, MIC3= 40.8353dB, MIC4= 39.1811dB).

7 Limitations and Future Work

This section highlights the shortcomings and challenges of the study and presents potential avenues for future research and improvements.

Limited Characteristics of Objects. This paper mostly focused on three simple shapes – a cube, a pyramid, and a sphere. However, whether the proposed technique is suitable for more complex shapes is yet to be explored. Moreover, the objects are 3D printed with ABS plastic and we have used aluminum foils for creating the deformation. In our future work, we want to explore more complicated shapes and different materials including wood, steel, and cardboard. This will test the robustness of the proposed algorithm for complex shapes and materials.

Limited Environment. The experiment utilized a single room with specific dimensions. The room contains a computer server and an HVAC system, which introduces robust noises, proving beneficial. However, further experiments and analysis are required to understand the algorithm's performance in more challenging situations like larger rooms, outdoor settings, or rooms with more people. In the future, we aim to conduct these tests and develop the proposed algorithm to be environment invariant.

Dataset Size. This project involves 210 samples, which is a relatively small number. To evaluate the model's generalizability, it is important to gather more data in future studies. By creating a more robust dataset that encompasses various shapes, materials, and environments, as previously mentioned, the model's applicability can be further assessed and potentially enhanced.

Limited Acoustic Sound. In this study, we focus solely on using a 10 kHz square wave sound signal. This frequency can be unpleasant and potentially harmful to the ear. Moving forward, we aim to investigate alternative sound types and frequencies. Additionally, we are interested in exploring the potential of incorporating music into our approach.

Comparatively Larger Microphone Array. The size of the ReSpeaker is currently not suitable for in-ear research, necessitating the development of a smaller microphone array to investigate much smaller enclosed shapes, which would better mimic in-ear studies. This is the primary motivation for creating a more compact microphone array.

Realtime System. In the current system, there is a two-part process. Data is collected on a Raspberry Pi, sent to the computer, and subsequently processed. In the future, the goal is to first eliminate the manual sending. Then, the aim is to shift the computation to the device itself, creating a fully integrated system.

Shape Dimension Estimation. As shown in Section 6.5, we have not achieved satisfactory performance in estimating the dimension (height) of the shapes. By obtaining more data, we can enhance the training of the GBM model, potentially leading to improved performance. The additional data enables the model to better understand and identify patterns, relationships, and trends within the dataset. As a result, the model is more likely to generalize well to new, unseen data, offering more accurate and reliable predictions. The increased volume of data strengthens the model's foundation, allowing it to adapt and evolve to handle complex tasks and challenges more effectively. We also plan to explore deep neural networks (DNN) for this estimation as DNN has shown great performance in many real-world applications. As more data collection may require large annotation efforts we plan to utilize semi-supervised learning techniques to improve our performance.

In summary, while the current study provides the foundation for future research in this area, there is still much work to be done to fully understand the potential of this method for shape identification and deformation detection.

8 Conclusion

Through this project it has been shown that audio can be used to differentiate between small hollow shapes. Although the algorithms had fairly accurate results, they are far from perfect. In order to improve the results, more audio features could be explored, different machine-learning algorithms could be implemented, more data could be collected, and improved localization techniques could be applied.

Not only can this technology be applied to earbuds for ear modeling and health purposes, but it can also be applied to other applications such as aiding in the maintenance of factories(including pipelines and machinery that is hard for humans to reach) and providing information on the health of structures(such as bridges and tunnels). For instance, since the shapes used in the experiments are geometric they are similar to the shape of pipes and other parts used in factories. Thus, playing sounds into these parts can give an indication of the health of the factory parts and help diagnose problems. Similarly, playing sounds into structures such as tunnels can be used to determine if the structure is eroding or broken. This project demonstrates the potential of audio technology to improve various industries beyond ear modeling and health, making it a promising area for future research and development.

References

[1] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, (New York, NY, USA), pp. 1029–1040, Association for Computing Machinery, 2015.

[2] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[3] J. A. Patterson, D. C. McIlwraith, and G.-Z. Yang, "A flexible, low noise reflective PPG sensor platform for ear-worn heart rate monitoring," in *2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks*, pp. 286–291, 2009.

[4] J. Lazaro, E. Gil, R. Bailón, A. Mincholé, and P. Laguna, "Deriving respiration from photoplethysmographic pulse width," *Medical & Biological Engineering & Computing*, vol. vol. 51, pp. pp.233–242, Feb. 2013.

[5] I. Vogel, J. Brug, E. Hosli, C. van der Ploeg, and H. Raat, "MP3 players and hearing loss: Adolescents' perceptions of loud music and hearing conservation," *The Journal of pediatrics*, vol. 152, pp. 400–4, Apr. 2008.

[6] R. Bayareh Mancilla, B. P. Ta, C. Daul, J. Gutirrez Martnez, L. Leija Salas, D. Wolf, and A. Vera Herndez, "Anatomical 3D Modeling Using IR Sensors and Radiometric Processing Based on Structure from Motion: Towards a Tool for the Diabetic Foot Diagnosis," *Sensors*, vol. 21, p. 3918, Jan. 2021.

[7] Â. M. C. R. Borzino, J. A. Apolinário Jr., and M. L. R. de Campos, "Consistent DOA Estimation of Heavily Noisy Gunshot Signals Using a Microphone Array," *IET Radar, Sonar & Navigation*, vol. 10, no. 9, pp. 1519–1527, 2016.

[8] J. Chan, S. Raju, R. Nandakumar, R. Bly, and S. Gollakota, "Detecting middle ear fluid using smartphones," *Science Translational Medicine*, vol. 11, p. eaav1102, May 2019.

[9] C.-Y. Chen, C.-H. Yeh, B. R. Chang, and J.-M. Pan, "3D Reconstruction from IR Thermal Images and Reprojective Evaluations," *Mathematical Problems in Engineering*, vol. 2015, p. e520534, Aug. 2015.

[10] X. Chen, H. Sun, and H. Zhang, "A New Method of Simultaneous Localization and Mapping for Mobile Robots Using Acoustic Landmarks," *Applied Sciences*, vol. 9, p. 1352, Jan. 2019.

[11] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and Martin Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12186–12191, 2013.

[12] B. L. Farmer-Fedor and R. D. Rabbitt, "Acoustic Intensity, Impedance and Reflection Coefficient in the Human Ear Canal," *The Journal of the Acoustical Society of America*, vol. 112, pp. 600–620, Aug. 2002.

[13] D. Floreano and R. J. Wood, "Science, technology and the future of small autonomous drones," *Nature*, vol. 521, pp. 460–466, May 2015.

[14] I. C. o. M. Systems, Applications, N. . . t. . Services Bretton Woods, A. International Conference on Mobile Systems, S. C. Author, and A. Campbell, "I am a smartphone and i can tell my user's walking direction," in *MobiSys '14 : Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services : June 16-19, 2014, Bretton Woods, NH, USA*, MobiSys'14: The 12th Annual International Conference on Mobile Systems, Applications, and Services, ([Place of publication not identified]), ACM.

[15] B. Islam, M. T. Islam, and S. Nirjon, "Glimpse.3D: A motion-triggered stereo body camera for 3D experience capture and preview," in *Proceedings of the 17th ACM/IEEE International Conference on Information Processing in Sensor Networks*, IPSN '18, (Porto, Portugal), pp. 176–187, IEEE Press, Apr. 2018.

[16] V. Iyer, R. Nandakumar, A. Wang, S. Fuller, and S. Gollakota, *Living IoT: A Flying Wireless Platform on Live Insects*. Dec. 2018.

[17] C. Junu Jahana, \. M. Sinith, and \. P. Lalu, "Direction of Arrival Estimation Using Microphone Array," in *2021 Fourth International Conference on Microelectronics, Signals & Systems (ICMSS)*, pp. 1–6, 2021.

[18] D. Kowerko, "3D Indoor Audio Localization of Moving Objects,"

[19] T. G. Leighton, "What is ultrasound?," *Progress in Biophysics and Molecular Biology*, vol. 93, pp. 3–83, Jan. 2007.

[20] J. Mahmud and J.-M. Frahm, *VPFusion: Joint 3D Volume and Pixel-Aligned Feature Fusion for Single and Multi-view 3D Reconstruction*. Mar. 2022.

[21] S. Mahto, T. Arakawa, and T. Koshinak, "Ear Acoustic Biometrics Using Inaudible Signals and Its Application to Continuous User Authentication," in *2018 26th European Signal Processing Conference (EUSIPCO)*, (Rome), pp. 1407–1411, IEEE, Sept. 2018.

[22] A. Natekin and A. Knoll, "Gradient Boosting Machines, a Tutorial," *Frontiers in Neurorobotics*, vol. 7, p. 21, Dec. 2013.

[23] W. Nie, K. Xu, D. Feng, C. Q. Wu, A. Hou, and X. Yin, "A Fast Algorithm for 2D DOA Estimation Using an Omnidirectional Sensor Array," *Sensors*, vol. 17, p. 515, Mar. 2017.

[24] "3D Sonar Visualizer™ F3D-S — FURUNO."

[25] "Direction Of Arrival Estimation Using Microphone Array — IEEE Conference Publication — IEEE Xplore."

[26] "Direction-of-Arrival Estimation - an Overview — ScienceDirect Topics."

[27] "EarEcho: Using Ear Canal Echo for Wearable Authentication: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies: Vol 3, No 3."

[28] "The Effect of Sonar on Human Hearing — IntechOpen."

[29] "Fever Temperatures: Accuracy and Comparison – Health Information Library — PeaceHealth."

[30] "Headphones Usage in the US 2017 — Statista."

[31] "How Does Sound in Air Differ from Sound in Water? – Discovery of Sound in the Sea."

[32] "An Introduction to Beamforming with MEMS Microphones."

[33] "Loudness — Acoustics — University of Salford."

[34] "On Microphone-Array Beamforming From a MIMO Acoustic Signal Processing Perspective — IEEE Journals & Magazine — IEEE Xplore."

[35] "Optical Heart Rate Measurement at the Earbud — Analog Devices."

[36] "Spectral Centroid - an Overview — ScienceDirect Topics."

[37] "Survey: Headphones Penetration Up 20% Since 2016.."

[38] J. Phillips, "Otoscan - 3D Ear Scanning: The Future Is Now!."

[39] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomedical Engineering*, vol. 2, pp. 158–164, Mar. 2018.

[40] C. Portnuff, B. Fligor, and K. Arehart, "Teenage Use of Portable Listening Devices: A Hazard to Hearing?," *Journal of the American Academy of Audiology*, vol. 22, pp. 663–77, Nov. 2011.

[41] S. Purushwalkam, S. V. A. Gari, V. K. Ithapu, C. Schissler, P. Robinson, A. Gupta, and K. Grauman, "Audio-Visual Floorplan Reconstruction," Dec. 2020.

[42] B. Schwarz, "Mapping the World in 3D," *Nature Photonics*, vol. 4, pp. 429–430, July 2010.

[43] D. o. H. . H. Services, "Ear Infections."

[44] S. Sharma and V. Kumar, "3D Face Reconstruction in Deep Learning Era: A Survey," *Archives of Computational Methods in Engineering*, vol. 29, pp. 3475–3507, Aug. 2022.

[45] S. Suriyal, C. Druzgalski, and K. Gautam, "Mobile Assisted Diabetic Retinopathy Detection Using Deep Neural Network," in *2018 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, (Porto), pp. 1–4, IEEE, Mar. 2018.

[46] A. A. Torres-García, O. Mendoza-Montoya, M. Molinas, J. M. Antelis, L. A. Moctezuma, and T. Hernández-Del-Toro, "Pre-Processing and Feature Extraction," in *Biosignal Processing and Classification Using Computational Learning and Intelligence*, pp. 59–91, Elsevier, 2022.

[47] Y. Xie, F. Li, Y. Wu, H. Chen, Z. Zhao, and Y. Wang, "TeethPass: Dental Occlusion-based User Authentication via In-ear Acoustic Sensing," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pp. 1789–1798, May 2022.

[48] Z. Yang, J. Li, P. Stoica, and L. Xie, "Chapter 11 - Sparse Methods for Direction-of-Arrival Estimation," in *Academic Press Library in Signal Processing, Volume 7* (R. Chellappa and S. Theodoridis, eds.), pp. 509–581, Academic Press, Jan. 2018.

[49] T. Price, J. L. Schönberger, Z. Wei, M. Pollefeys, and J.-M. Frahm, "Augmenting crowd-sourced 3d reconstructions using semantic detections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[50] T. Gburrek, J. Schmalenstroeer, and R. Haeb-Umbach, "On source-microphone distance estimation using convolutional recurrent neural networks," in *Speech Communication; 14th ITG Conference*, pp. 1–5, 2021.

[51] L. Calkins, J. Lingevitch, J. Coffin, L. McGuire, J. Geder, M. Kelly, M. M. Zavlanos, D. Sofge, and D. M. Lofaro, "Distance estimation using self-induced noise of an aerial vehicle," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2807–2813, 2021.

[52] D. de Godoy, B. Islam, S. Xia, M. T. Islam, R. Chandrasekaran, Y.-C. Chen, S. Nirjon, P. R. Kinget, and X. Jiang, "Paws: A wearable acoustic system for pedestrian safety," in *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pp. 237–248, 2018.

[53] Bricsys, "What is Point Cloud in CAD - Bricsys Blog."

[54] A. Moldi, V. Gala, S. Puranik, S. Karan, S. Deshpande, and N. Neela, "Survey of impression materials and techniques in fixed partial dentures among the practitioners in india," *ISRN Dentistry*, vol. 2013, pp. 1–5, 2013.

[55] F. Mangano, A. Gandolfi, G. Luongo, and S. Logozzo, "Intraoral scanners in dentistry: a review of the current literature," *BMC Oral Health*, vol. 17, 12 2017.

[56] W. C. Scarfe and A. G. Farman, "What is cone-beam ct and how does it work?," *Dental Clinics of North America*, vol. 52, pp. 707–730, 10 2008.

[57] C. Morgan, C. Khatri, S. A. Hanna, H. Ashrafian, and K. M. Sarraf, "Use of three-dimensional printing in preoperative planning in orthopaedic trauma surgery: A systematic review and meta-analysis," *World Journal of Orthopedics*, vol. 11, pp. 57–67, 01 2019.

[58] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, pp. 276–280, 03 1986.

[59] P. K. Eranti and B. D. Barkana, "An overview of direction-of-arrival estimation methods using adaptive directional time-frequency distributions," *Electronics*, vol. 11, p. 1321, 04 2022.

[60] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, pp. 4–24, 04 1988.

[61] K. Benson, "Phased array beamforming ics simplify antenna design — analog devices," 01 2019.

[62] T. S. Rappaport, *Wireless Communications*. Prentice Hall, 2002.

[63] S. Wang, J. Lin, T. Wang, X. Chen, and P. Huang, "Recent advances in photoacoustic imaging for deep-tissue biomedical applications," *Theranostics*, vol. 6, pp. 2394–2413, 2016.

[64] L. E. Kinsler and E. Al, *Fundamentals of acoustics*. John Wiley & Sons, 2000.

[65] B. Rafaely, "¡i¿microphone array signal processing¡/i¿microphone array signal processingjacobbenestyjingdongchenyitenghuangspringer, berlin, 2008. 240 pp. price $119 (hardcover). isbn: 3540786112," *The Journal of the Acoustical Society of America*, vol. 125, pp. 4097–4098, 06 2009.

[66] M. Leach and W. M. Leach, *Introduction to Electroacoustics and Audio Amplifier Design*. Kendall Hunt Publishing Company, 2003.

[67] O. Gazi, *Understanding digital signal processing*. Springer, 2018.

[68] S. G. Mallat and G. Peyer, *A Wavelet Tour of Signal Processing, 3rd Edition*. Academic Press, 2009.

[69] M. Muller and S. O. Service, *Fundamentals of Music Processing : Audio, Analysis, Algorithms, Applications*. Springer International Publishing, 2015.

[70] I. Fujinaga, "Machine recognition of timbre using steady-state tone of acoustic musical instruments," 04 1999.

[71] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357–366, 08 1980.

[72] H. Azami and J. Escudero, "Amplitude- and fluctuation-based dispersion entropy," *Entropy*, vol. 20, p. 210, 03 2018.

[73] O. Lartillot, "A matlab toolbox for musical feature extraction from audio," *... of the 10th International Conference on ...*, 01 2007.

[74] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 293–302, 07 2002.

[75] J. Antoni, "The spectral kurtosis: a useful tool for characterising non-stationary signals," *Mechanical Systems and Signal Processing*, vol. 20, pp. 282–307, 02 2006.

[76] R. D. Patterson, M. Unoki, and T. Irino, "Extending the domain of center frequencies for the compressive gammachirp auditory filter," *The Journal of the Acoustical Society of America*, vol. 114, pp. 1529–1542, 09 2003.

[77] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70–73, 06 1967.