Sauron Security
Final Design Review Report

Jose LaSalle, Omid Meh, Walter Brown, Zachary Goodman

***Abstract*—Sauron is a security system that can be deployed in crowded areas to eavesdrop on individuals of interest. Sauron is an acoustic beamformer with a camera so that the operator can visually select targets. The beamformer is composed of a microphone array that records sound at different points. When the operator clicks on a target in the video, Sauron calculates the angle to the target and uses enhanced delay sum beamforming to extract what the target is saying.**

***Index Terms*—Acoustic, Source Isolation, Microphone Array, Delay Sum Beamforming, Compound Array.**

I. Introduction

SECURITY is a significant concern in public places, resulting in an increased interest in surveillance. Crowded places such as museums, markets, and airports are swarming with cameras. Sauron is a tool to further improve safety. Sauron allows security personnel to eavesdrop on individuals through the power of acoustic beamforming by simply identifying them in a video feed.

Sauron consists of a microphone array and camera that interface with a computer. An operator is able to hover their cursor over an individual in a crowded environment and the system plays what that individual is saying. This system can be adapted to be useful in almost any situation where a voice needs to be isolated. For example, an operator might record a lecture and click on students in the audience who are asking questions. Another use case would be video editing. A cameraman might record something and then want to eliminate a distraction in the background. Although Sauron is meant to improve safety, it has other applications as well.

Sauron is a threat to privacy. If enhanced, it could be deployed in a neighborhood to eavesdrop on conversations inside households and other private locations. The major obstacle in this task would be that potential targets would be at different distances from the array. Closer targets will be louder, meaning that delay sum beamforming would fail unless there were a large number of microphones, all of which would be sensitive enough to hear at a long range.

Sauron consists of a microphone array and camera that interface with a computer. An operator is able to hover their cursor over an individual in a crowded environment and the system plays what that individual is saying.

A. Established Solution

Squarehead Technology's new AudioScope is a device designed to listen in on players, coaches, and the like at sports events. This device performs acoustic beamforming with an array of around 300 microphones mounted on a disk on the ceiling to isolate locations selected by the operator [2].

Currently, airports have some of the most advanced surveillance. Video feeds are analyzed to identify individuals on watch lists, bags being left behind by their owners, people going the wrong way through checkpoints, and cars spending an abnormal amount of time in the parking lots [1]. However; audio is not as prevalent in airport security.

B. Use Case

A security guard with no knowledge of acoustic beamforming and very little training beyond the norm sits in a video surveillance room. One of the cameras is aimed at a line of people waiting to

J. LaSalle majors in Electrical Engineering and is a member of Commonwealth Honors College.

O. Meh majors in Electrical Engineering and in Computer Systems Engineering and is a member of Commonwealth Honors Collage.

W. Brown majors in Computer Systems Engineering and in Computer Science and is a member of Commonwealth Honors Collage.

Z. Goodman majors in Electrical Engineering.

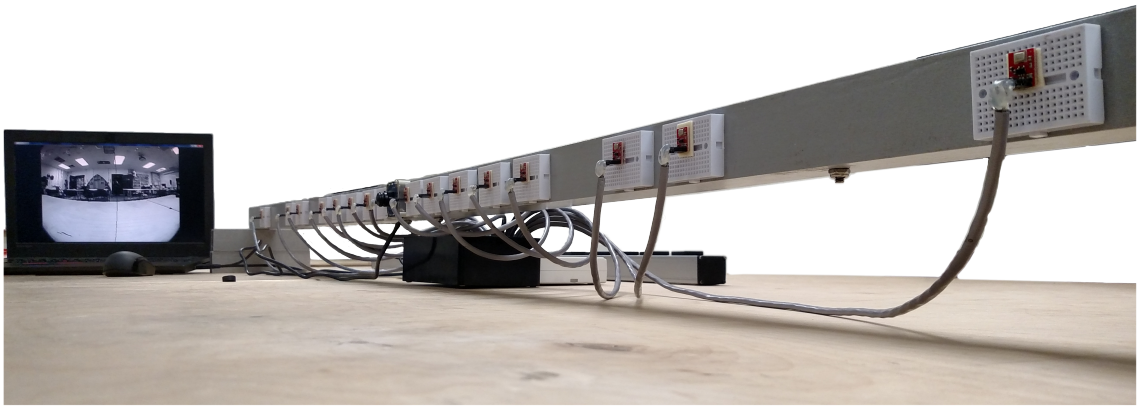


Fig. 1. Image of the physical array.

be screened at a checkpoint. Two individuals with suitcases are chatting near the back of the line. To be on the safe side, the guard hovers a cursor over the head of one of the speakers. The conversation can be heard through the guards headphones.

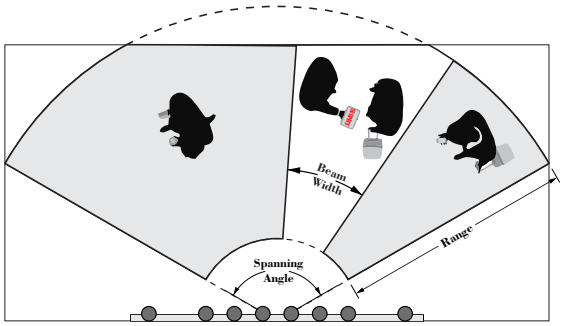


Fig. 2. Visual depiction of specifications.

C. Specifications

Table I lists the specifications of Sauron. The specifications for the targets distance, angle from the arrays center-line, and maximum beamwidth are about the same as the SDP 14 beamforming group had [3]. The SDP 16 group added context around the array, only increasing the performance of the array where absolutely needed. These old specifications are reasonable for a bottleneck like a corridor.

The beamwidth specification is for a -10dB bandwidth because -10dB will make a sound seem half as loud to a listener [4]. Tests done within the SDP 16 group showed that when one of two superimposed voices is amplified to 10dB above the other the amplified voice is easy to understand.

TABLE I
TABLE OF SPECIFICATIONS

Specification	Promised	Achieved
Range	1 to 3 meters	1 to 3 meters
Angle of Operation	-65° to 65°	-65° to 65°
Maximum -10dB beam width	40°	30°
Frequency Range	1kHz to 3.5kHz	500Hz to 5kHz
Real-time Delay	10s	5s
Error in angle selection	20°	10°

Experiments within the SDP 16 group found that higher frequencies were more important for determining what a person is saying than lower frequencies. These experiments involved taking sound clips of group members speaking and running them through a digital bandpass filter, expanding the passband until the message was clear. The specifications were changed to include this useful frequency range, as is reflected in Table I.

As security may need to quickly respond to a conversation, the operator must hear what the target said no longer than 10 seconds after they have said it. Reducing this delay is preferable even over hearing all that the target is saying. When the operator selects on a target, the actual angle that the system is focusing on must be within 20 ° of the intended target. More error than this and the beam will miss the target.

Figure 2 provides a visual depiction of these specifications.

II. Design

Figure 3 shows the layout of Sauron. Sauron uses a fisheye camera, a 16 microphone array, and

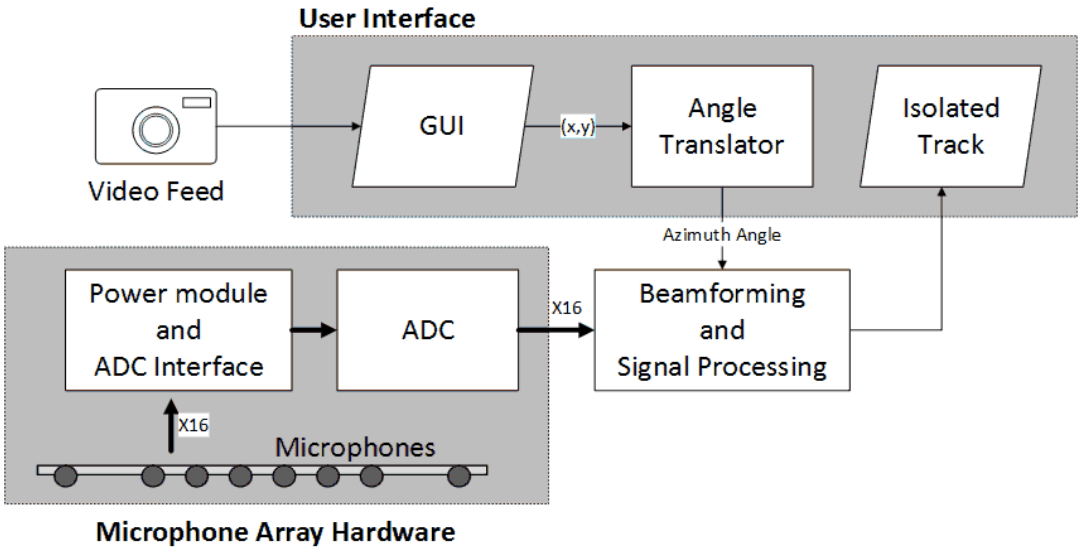


Fig. 3. System diagram for Sauron.

a computer. The video information is sent to the user interface so the operator can pick a target. The user interface maps the target location to an angle which is used by the audio processing portion of the program to perform beamforming on the microphone data. This yields an isolated sound that the user interface can hear.

A. Microphone Array

The purpose of the microphone array is to record sound from different locations. It sends this information to the audio processing software described in section II-D.

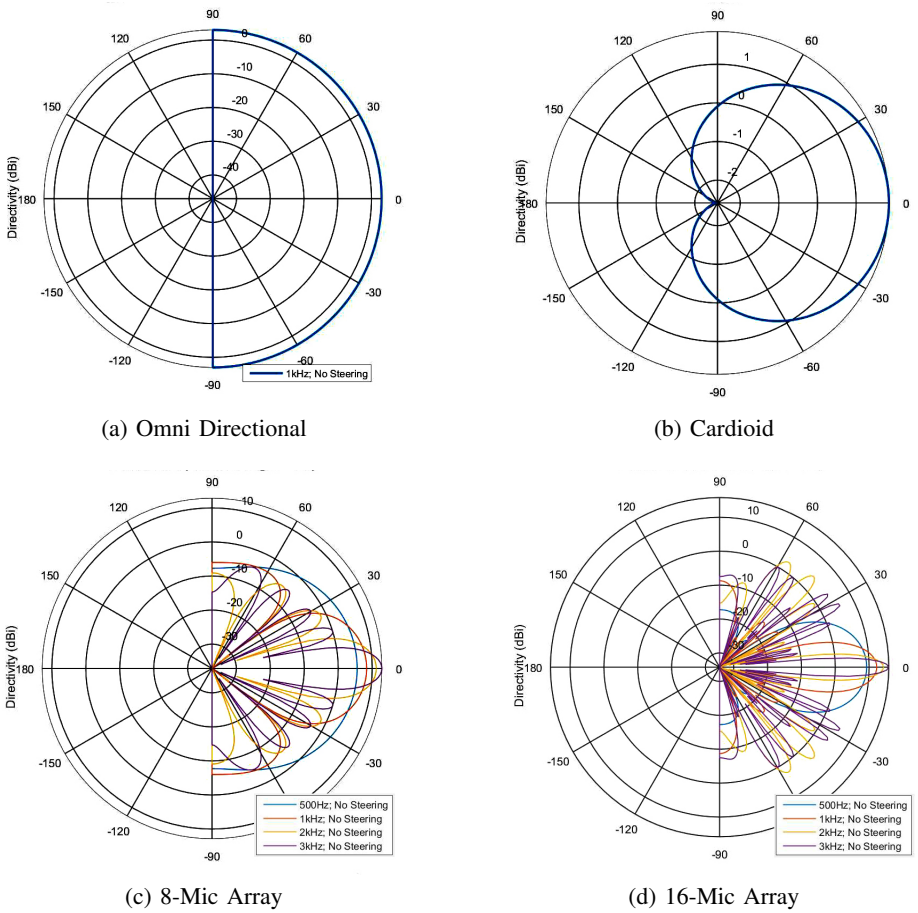
Our array needs to produce high-quality sound across our desired frequency range with a relatively constant beamwidth.

Beamforming involves processing multiple microphone outputs to create a directional pickup pattern. It is important that the microphone only picks up sound from one direction and attenuates the sound that is off the main axis. Beamforming capabilities are determined by the geometry of the microphone array, the polar pattern of the microphones, and the speed of sound (which could be more accurately determined using a temperature sensor).

Information about the geometry of the microphone array and speed of sound are used to determine the time delays used in the beamforming algorithm. The array geometry also influences what frequencies the array operates at. Smaller microphone spacing is optimal for high frequencies while larger spacing is superior at lower frequencies.

1) Microphones: Microphone (or microphone array) directionality describes the pattern in which the microphones sensitivity changes with respect to changes in the position of the sound source. An omnidirectional pattern is equally sensitive to sound coming from all directions regardless of the orientation of the microphone. A cardioid polar pattern means that there is minimum signal attenuation when the signal arrives from the front of the microphone (0° azimuth), and maximum signal attenuation when the signals arrive from the back of the microphone (180° azimuth), referred to as the null. Figure 4b shows a 2-axis polar plot of the omnidirectional and cardioid microphone responses. This plot looks the same regardless of whether the microphones port is oriented in the x-y, x-z, or y-z plane [5].

The cardioid polar pattern offers beamforming capabilities by creating a beam where the signal is attenuated except for where the beam is steered, while an omnidirectional polar pattern has no attenuation in any direction relative to the microphone. A cardioid polar pattern with a wide angle of operation and narrow beamwidth is desired from our beamforming array in order to focus our beam on a single individual and



(a) Omni Directional (b) Cardioid

(c) 8-Mic Array (d) 16-Mic Array

Fig. 4. Polar plots that depict directivity and beamwidth for (a) omnidirectional microphones (b) cardioid microphones (c) our 8-microphone array and (d) our 16-microphone array. As one can see, there are significant improvements in both the directivity and beamwidth when the number of omnidirectional microphones are increased.

operate in the largest area possible. We use omnidirectional MEMS microphones for our array to create cardioid polar patterns across our operating frequency range. MEMS stands for Micro-Electro-Mechanical Systems, which include microsensors and microactuators that act as transducer elements that convert acoustic pressure waves into electrical signals [6]. MEMS microphones enable improvements in sound quality for multiple-microphone applications. Microphone arrays can take advantage of the small form factor, sensitivity matching, and frequency response of a MEMS design for beamforming to help isolate a sound in a specific location [7].

High Input sound quality is the result of high sensitivity microphones, a uniform output level across our operating frequency, and low noise.

Microphone sensitivity is defined as the ratio of the analog output voltage to the input pressure. The standard reference input signal for microphone sensitivity measurements is a 1 kHz sine wave at 94 dB sound pressure level (SPL), or 1 pascal (Pa) pressure. Microphone sensitivity is determined using the reference input signal. As microphone sensitivity increases, the output level for a fixed acoustic input increases. Microphone sensitivity measured in decibels (dB) is a negative value, meaning that higher sensitivity is a smaller absolute value [8]. The sensitivity of the microphone array is higher than that of each individual array because their outputs are summed.

- Cardioid
  - -54dBV sensitivity
  - 50-15kHz frequency range
- Electret
  - -44dBV sensitivity
  - 20-20kHz frequency range
- MEMS

- -38dBV sensitivity
- 100-15kHz frequency range

The frequency response of a microphone describes its output level across the frequency spectrum. The high and low frequency limits are the points at which the microphone response is 3 dB below the reference output level (normalized to 0 dB) at 1 kHz. Figure 5 shows the frequency response of the ADMP510 omnidirectional MEMS microphone [5].

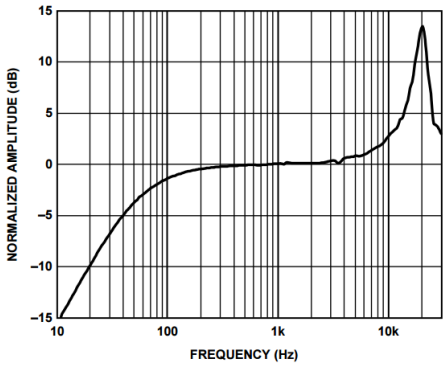


Fig. 5. Frequency response of ADMP510 MEMS microphone.

When building the microphone array, knowing the microphones frequency response enables us to choose microphones based on what frequency range we want to cover. In our desired operating range (1kHz - 3.5kHz), we can see that MEMs microphones have a flat, linear frequency response, meaning we do not have to attenuate or amplify our signals differently at different frequencies to achieve a uniform output across the frequency spectrum.

Figure 6 shows the design of the MEMS microphone modules. These are commercial modules provided by sparkfun that we purchased as a way to test our methods. Consequently, these modules proved to be of excellent quality and allowed us to meet our specifications. Therefore, we decided to use these products in our project.

Low noise is essential for high quality audio. Following the microphones, op amps are available with significantly lower noise than the microphones themselves, making the microphones the limiting factor regarding the noise of the overall design. The cable connections must be shielded and/or filtered to prevent the wires from picking up electromagnetic interference (EMI) or RF noise.

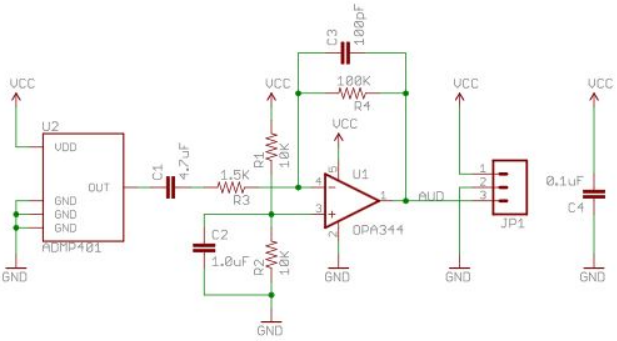


Fig. 6. Circuit schematic for MEMS microphone board.

By using an array of high sensitivity microphones, low noise preamplifier circuitry, and shielded transmission wires, we achieve high quality audio input into our computer interface for frequencies based on the array geometry.

2) Array Organization: The geometry and the number of elements in the array directly affect the performance. In general, given a fixed linear array, as the frequency increases, the beam width decreases. To understand this, look at Figure 7. The signal is parallel to the mic array which is gives us the maximum delay. The phases for the 500Hz signal arrive at the microphones in the array at [0 36 73 110] degrees, which are close and difficult to distinguish in terms of coherency. However, as the frequency increases the phases for the 1500Hz signal arrive at [0 110 146 330] degrees. For higher frequencies, the maximum phase difference becomes larger and during the analysis it will be easier to distinguish how incoherent signals due to large phase differences in the received signal.

In other words, the larger phase differences allow us (as long as we are in the same cycle) to determine the direction of the source more clearly, thus giving the microphone array higher directivity. Notice that the directivity is different for different frequencies, as for different frequencies we have different range of arrival phase difference.

Smaller microphone spacing is better for high frequencies and larger microphone spacing is desirable for lower frequencies. To achieve the best result for all frequency bands, we use a compound microphone array, which is the superposition of multiple arrays with different microphone spacings. Bandpassing the signal to the proper frequency range for each array and subarray, perform-

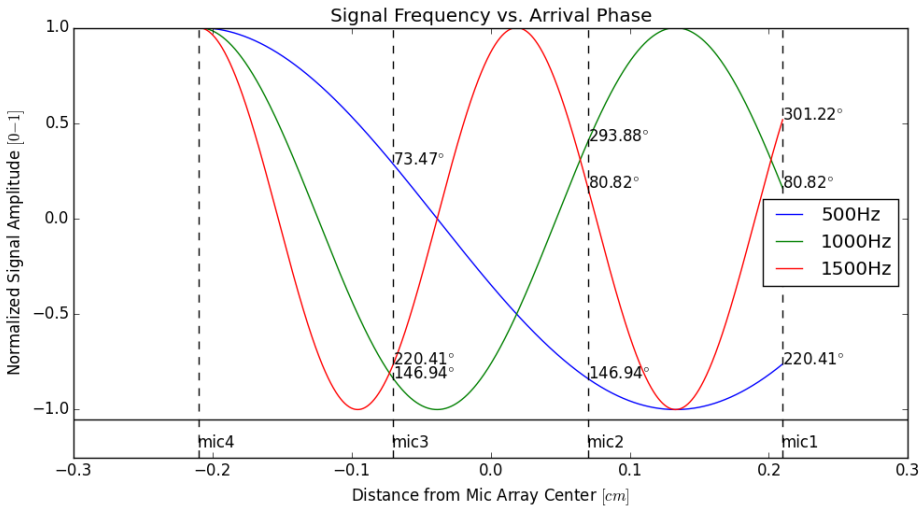


Fig. 7. Wave phases over time for different frequencies.

ing the delay-sum for the specific band, and finally, summing the results of the different bands to obtain result with maximum beam precision for multiple frequency bands. Using equal microphone spacing prevents the array to create a precise beam for a wider range of frequencies. Figure 8 depicts the layout of our compound array.

3) Analog to Digital Converter: National Instrument's USB-6210 [9] is the A/D used for this project to handle the microphones. This A/D can sample above the needed Nyquist rate of 7 kHz.

This A/D supports 16 microphones. [10] describes how to physically attach the A/D to its inputs. The SENSE line was left floating.

Before connecting the A/D to the laptop, the CD-ROM that came with the A/D was used to install the appropriate drivers. During the installation of the A/D drivers, daqlib was installed in Simulink. A block can be added called "daqlib/Analog Input" that allows access to the readings of the A/D. For our setup, we needed to set the Analog Input block to use a "referenced" signal as there was a common ground across all devices connected to the A/D.

To demonstrate the real-time functionality of the A/D with the rest of the array, a "Sinks/Scope" was attached to the end of each output line of the Analog Input block. This allowed us to watch what the A/D detects.

B. Camera

The purpose of this block is to produce a video that the operator can reference to choose a target to listen to. This produces visual data that is displayed by the user interface described in section II-B.

We interfaced a USB fisheye camera with Simulink to give a wide field of view on which to beamform. The camera's functionality was tested by attaching it to a computer, running the Simulink script, and observing video streaming.

C. User Interface

The purpose of this block is to let the user easily interact with the system.

This is a graphical user interface that takes video information from the camera described in section II-B and displays it. The user is able to hover his or her cursor on a target in the video to listen to it. A curve is drawn on the display to show the user the region he or she is listening to. This is necessary because Sauron listens to an angle, not a precise spot. This block then calculates the angle from the center-line of the array described in section II-A to the target. This value is sent as an input to the audio processing software described in section II-D. The audio processing software calculates and provides the audio coming from the selected point so that the user interface can play it to the user.

The interface was written in Simulink.

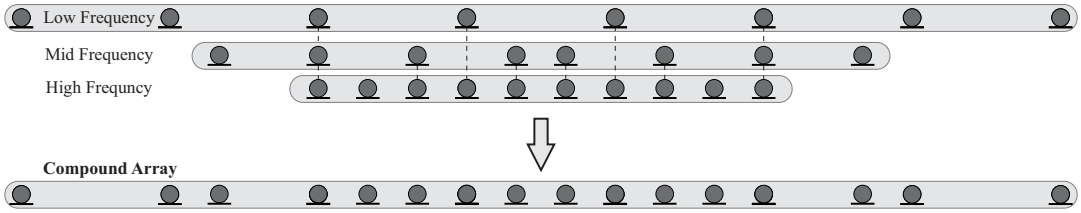
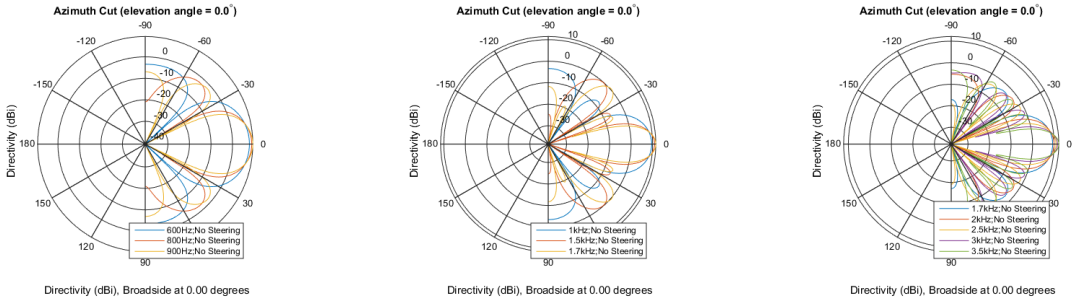


Fig. 8. Drawing of our compound array design. The low frequency array has a spacing of 21cm. The array for middle frequencies has a spacing of 14cm, except for he middle two microphones which are 7cm apart. The highest frequency array has a spacing of 7cm.



(a) Low Frequency Subarray (b) Mid Frequency Subarray (c) High Frequency Subarray

Fig. 9. Simulation results for sub-arrays within the system. 9a is the low frequency array in the Fig 8 with four elements at $21cm$ spacing. 9b is the middle frequency array in the Fig 8 with six elements at $[14cm\ 14cm\ 7cm\ 14cm\ 14cm]$ spacing. 9c is the high frequency array in the Fig 8 with six elements at $7cm$ spacing. 9a is tuned for [600Hz, 1kHz], 9b is tuned for [1kHz, 1.7kHz], and 9b is tuned for [1.7kHz, 3.5kHz]

This block was tested by having a human user observe the system respond to selecting an individual on the video feed and and hearing the audio.

D. Audio Processing Software

The purpose of this block is to isolate the target's voice. It is given the angle to the target by the user interface described in section II-C. It gets the necessary audio data from the microphone array described in section II-A. This block gives the isolated voice of the target to the user interface.

1) Delay-Sum Beamforming: Beamforming, also known as Spatial Filtering, is a signal processing method used with sensor arrays allowing directional reception or transmission of the signal. For this project we are interested in directional reception of the human voice. Since the speech is a broadband signal, we decided to use a delay-sum beamforming with a linear array which allows us to process a wideband signal and relatively low computational complexity.

Figure 10 is an illustration of a simple microphone array composed of three microphones and a summing module. As shown, when the signal is produced at the -45° it arrives at the left, middle, then right microphones in order, and when the signal is produced at the +45° angle it arrives at the right, middle, then left microphones in order. In both cases, when all three signals are summed the signals will be off by some time delay and will not constructively add up. However, if the signal is produced perpendicular to the array, it arrives at the three microphones at the same time resulting in a constructive signal sum. This microphone array is called a non-steered (focused on 0° azimuth) 3-element linear microphone array.

As illustrated in Figure 11, this concept can be further expanded to steer the array beam to an arbitrary direction. A delay block is added to each signal before the summer which further delays the signal. The added delay is to reverse the expected time delays for the signal coming from the desired direction. For instance, in Figure 11, we desire to listen to the target wavefront (top speaker), this we mathematically calculate the expected time delay for the signal to arrive at each microphone. Next, the received signals are shifted back in time (in the steering unit) to look as if they were all received at the same time by mics. At the summing stage,

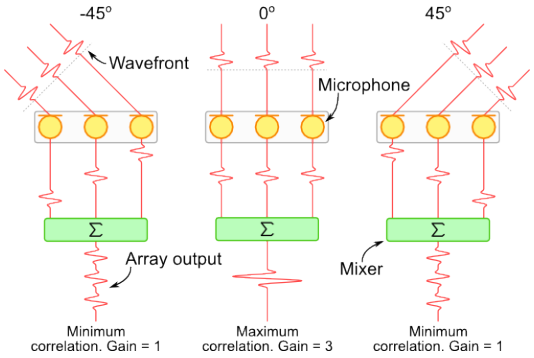


Fig. 10. Simple microphone array with sounds coming from the direction of -45°, 0° , and 45°. Reprinted with permission from [14].

this will result in the constructive interference for the signals coming from the target direction and destructive- or incoherent- interference for the signals coming from other directions.

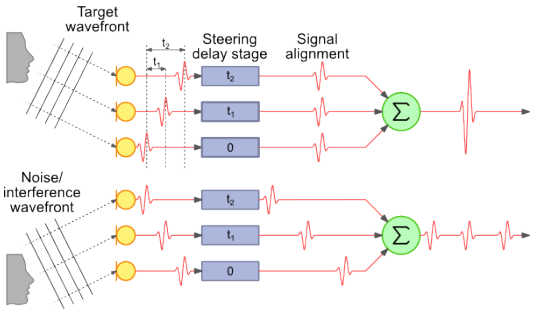


Fig. 11. Illustration of delay-sum beamforming. Reprinted with permission from [14].

III. Project Management

Our team has shown a lot of vitality and perseverance since the beginning of this project, and through that we continue to learn how to work together efficiently and effectively. With communication and personal accountability as our mode of operation, coupled with frequent meetings and clearly delegated tasks, we were able to accomplish all of our MDR goals despite a late start. We achieved our goal of demonstrating voice isolation between two speakers by establishing four specific sub-goals that were tailored to each team member's area of strength. Analysis of the hardware for the mic array was headed by Zach, as amplifier design and use of electronic elements are in his field of study as an electrical engineer. Walter was responsible for interfacing the hardware into Simulink and building a software block for calibrating the array. Omid took on the beam-forming algorithm given his CSE background, and Jose was responsible for noise reduction as this was involved in his REU. As an execution of our plans unfolded, an overlap of our knowledge bases lead to a very integrated experience of one helping the other, resulting in a very rewarding experience so far.

IV. Conclusion

Project Sauron proceeded as planned after MDR. Table I details our desired and accomplished specifications. These deliverables demonstrated that our group could interface with an array and that our group could isolate voices. Our final product has resolved the physical boundaries that the group feared would stop them.

For CDR, our group was able to demonstrate that a user can hover over a point in a fisheye video feed and Sauron will isolate the audio at that point. A new 16-microphone array was built to support the tight beamwidth called for by the specifications. A fisheye camera was implemented as promised to provide a visual of the environment for aiding in security applications. There is a successful mapping between the video and the target angle.

The major challenge of this project was implementing a realtime system. This challenge mutually arises with the 5 second sampling buffer required to acquire enough samples from all 16 microphones. Another consideration that presented a challenge for this project was an elegant arrangement of the microphones that provides an ease of use for the operator. Figure 1 displays the setup, which fits a form factor that would be easily deployable on an airport terminal wall.

A. Future Work

There are extensive options for improving this project:

- Post Processing: Implementing post processing would allow the user to perform beamforming on media previously recorded through the beamformer, a valuable application for video forensics.
- Multi-Dimensional Array: Developing a multi-dimensional microphone array to

improve the beamformer's directivity. There is potential for a distributed mesh array that can span a large space.

- Detection of Arrival and Tracking: Create detection of arrival and tracking algorithms to control the beamformer. These functionalities would enable the beamformer to operate without a user and could isolate audio in areas making noise.
- Temperature Sensor: Add a temperature sensor to more accurately determine the speed of sound in the beamforming environment.
- Wireless Array: Creating an wireless array that interacts with other wireless arrays and/or the main computer. This could be useful for implementing a system that is easy to setup and fit to any environment.
- Stand Alone System: Implementing a stand alone system, where the beamforming occurs on an FPGA instead of laptop would allow you reduce the delay in our beamforming system by performing signal processing in parallel.
- HoloLens, Virtual Reality, Speech-to-Text: Integrating beamforming with Microsoft HoloLens and/or Virtual Reality. This could be applied towards applications for audio impaired individuals, allowing people with spatial hearing difficulties, making it hard for them to locate the source of the audio they are hearing. This could help them focus their hearing where they want, and not be distracted by other noises. By implementing speech to text functionality, you can provide transcripts of targeted individuals even in a crowded environment. This could be used to develop a solution to language barrier issues by showing the transcript of a persons speech in the users native language.

B. Acknowledgments

We would like to thank Professor Hollot and Professor Moritz for their feedback and guidance in establishing realistic goals. We would also like to send a big thanks to Professor Wolf who took the time to meet with us each week and helped us stay on track and organized. An additional thanks for Alumni John Shattuck for coming back to UMass to meet with us as we evolve his old project.

REFERENCES

[1] Airports [Online]. Available: https://www.videosurveillance.com/airports.asp [Accessed Web. 18 Jan. 2016.]

[2] Catherine de Lange Audio zoom picks out lone voice in the crowd [Online]. Available: https://www.newscientist.com/article/dn19541-audio-zoom-picks-out-lone-voice-in-the-crowd/ [Accessed Web. 21 Jan. 2016.]

[3] J. A. Danis, et al. *The Acoustic Beamformer*. Available: http://www.ecs.umass.edu/ece/sdp/sdp14/team15/assets/Team15FinalMDRReport.pdf [Accessed Web. 18 Jan. 2016]

[4] University of WisconsinMadison, About Decibels (dB) [Online]. Available: http://trace.wisc.edu/docs/2004-About-dB/ [Accessed Web. 24 Jan. 2016.]

[5] InvenSense, Microphone Specifications Explained [Online]. Available: http://43zrtwysvxb2gf29r5o0athu.wpengine.netdna-cdn.com/wp-content/uploads/2015/02/MICROPHONE-SPECIFICATIONS-EXPLAINED.pdf [Accessed Web. 2 Dec. 2015.]

[6] InvenSense, Analog and Digital MEMS Microphone Design Considerations [Online]. Available: http://43zrtwysvxb2gf29r5o0athu.wpengine.netdna-cdn.com/wp-content/uploads/2015/02/Analog-and-Digital-MEMS-Microphone-Design-Considerations.pdf [Accessed Web. 2 Dec. 2015.]

[7] Digi-Key, MEMS Technology for Microphones in Audio Applications [Online]. Available: http://www.digikey.com/en/articles/techzone/2012/aug/mems-technology-for-microphones-in-audio-applications [Accessed Web. 2 Dec. 2015.]

[8] Analog Devices, Understanding Microphone Sensitivity [Online]. Available: http://www.analog.com/library/analogDialogue/archives/46-05/understanding_microphone_sensitivity.pdf [Accessed Web. 2 Dec. 2015.]

[9] National Instruments, "NI USB-610 - National Instruments" [Online]. Available: http://sine.ni.com/nips/cds/view/p/lang/en/nid/203223 [Accessed Web. 3 May 2013.]

[10] National Instruments, "Bus-Powered M Series Multifunction DAQ for USB - 16-Bit, up to 400 kS/s, up to 32 Analog Inputs, Isolation" [Online]. Available: http://www.ni.com/datasheet/pdf/en/ds-9 [Accessed Web. 3 May 2013.]

[11] Mathworks, Database Toolbox [Online]. Available: http://www.mathworks.com/products/database/ [Accessed Web. 23 Jan. 2016.]

[12] Mathworks, Acquire Images from Webcams [Online]. Available: http://www.mathworks.com/help/supportpkg/usbwebcams/ug/acquire-images-from-webcams.html [Accessed Web. 24 Jan. 2016.]

[13] National Instruments, Least Mean Square (LMS) Adaptive Filter [Online]. Available: http://www.ni.com/example/31220/en/ [Accessed Web. 18 Jan. 2016.]

[14] A. Greensted, Delay Sum Beamforming, The lab book Pages An online collection of electronics information, 01-Oct-2012. [Online]. Available: http://www.labbookpages.co.uk/audio/beamforming/delaySum.html. [Accessed: 24-Jan-2016].