Sean – Sound Enhancing Autonomous Network

Annette Barboza, Ayanna Ivey, Brandon Kessler

Dept. of Electrical and Computer Engineering
University of Central Florida, Orlando, Florida, 32816-2450

***Abstract* — Sound Enhancing Autonomous Network (Sean) is a portable device that a user can employ to enhance the sound of another human's voice while maintaining a non-amplified steady state when the user is not engaged in conversation. Sean will use audio and visual data from the user's surroundings to amplify audio when there is a person speaking to the user. This is achieved by using a fully convolution neural network for face detection and a customized direction of arrival algorithm to determine where sources of sound are coming from. Combining these outputs allows Sean to make informed decisions on when to amplify audio.**

***Index Terms* — Microphone arrays, audio-visual systems, face recognition, interactive systems, logic circuits.**

I. Introduction

One goal of technological innovation is to make the lives of consumers more convenient in *almost* every conceivable way. [1] However, there have been very few attempts to make products for those who live with disabilities as accessible as products that are simply for convenience. In the case of hearing impairment, approximately 15% of American adults report having some trouble hearing.

Hearing loss presents itself in three different types: conductive hearing loss, sensorineural hearing loss, and a mix of both. In permanent cases, conductive hearing loss generally affects the overall loudness of a sound, while sensorineural can affect loudness and perception of tone [2]. For the majority of the people who live with any of these, their options are limited to potentially invasive procedures in an attempt to correct the hearing or the option of using hearing aids. Hearing-aids are one of the most widely available removable solutions to hearing loss. They help users by converting sound to digital signals, amplifying those signals, and passing the amplified signal back to the user as sound. However, a vast problem with the hearing aid is its amplification is not based on what the user is interested in hearing.

Sean (Sound Enhancing Autonomous Network) shows the viability and benefits of providing the user with a directed sound experience that amplifies sound only when a person is talking to the user. Sean shows that by integrating state of the art computer vision techniques for face detection with sound localization, a user can have amplified sound when it is highly required in a conversation without having to deal with irritating amplified background noise while in an idle state.

A. Market Analysis Research

In order to appropriately identify what are the current needs of hearing aid consumers, research on common issues and problems with standard hearing aids was conducted. Many users experience shortened lifespan, stunted amplification, incompetent automatic noise level adjustment, and inflated prices of hearing aids. The biggest issue found was the cocktail party problem where a user is distracted by several people talking in the background. [3][4][5] Other solutions and products that direct their efforts towards the same issue are the Earbeamer, Hear Active Listening, and the Widex Evoke. From the documentation that was provided on these solutions it was found that adding a visual aspect to our solution would improve the system's accuracy in human-to-human interaction while maintaining portability and minimizing latency when possible. [6] The budget for this project was set at $2000 which is in the range of prices for a standard set of hearing aids. The lifespan is improved in Sean's design since it will be utilizing a sufficient battery and not suffering from the moisture and wax build up that traditional hearing aids experience.

B. Objectives

The main goal for this project is to create a smart alternative to the common hearing aid. The system will make decisions on when to amplify and attenuate sound based on information provided from both the visual and audio domains. Effectiveness and efficiency are two main components that guided the decisions in framing the scope of the system. Sean must enhance a user's experience when it comes to the quality of the sound he or she receives. Keeping these main tenets in mind, the following core objectives relating to the function of Sean have been established:

1) Predict the human voice source the user wants to hear out of a range of potential sources and amplify that sound
2) Provide a mobile application interface in which the user can interact with Sean
3) Give a visual output of how Sean is making decisions
4) Create a comfortable wearable platform for Sean that can illustrate the potential of producing a portable system.
5) Improve the quality of life of hearing impaired people as well as providing an enhanced experience for average users
6) Be a user-friendly and intuitive product

II. Design

Sean's design was guided by understanding the limitations of previous work and researching components that can together form a functional high-performing system. Audio quality and implementation time are important factors that were considered and led to the decision of the MATRIX Voice as the hardware for a Direction of Arrival (DOA) algorithm implemented to determine where in the field of view sources of sound are coming from. The SAMSON Go Mic was used for direct audio capture. Previous research indicated that adding a visual component in this application would help solve problems related to the complex issue of too much audio intention variation per environment; thus, a processor trade study was conducted to compare specifications and determine which would best meet the visual, audio, and communication needs of Sean. Ultimately, the NVIDIA Jetson TX2 Developer Kit was selected to be the main processor and integration platform. There was a consensus among existing products' results that a mobile app functioning as the user interface for the product was instrumental in allowing for user input, control of the device, and personalization of settings in a practical way. Sean communicates with an iOS app and acts as the mediator between Sean and the user's earbuds/headphones. The system includes a sound alarm temperature sensor that alerts the user when Sean is reaching a potentially harmful temperature.

The packaging originally designed to be 3D printed was instead formed with acrylic and laser cut to meet sizing needs and become a wearable device that allows users to pick their own earbuds/headphones according to what is most comfortable for them.

Sean will be able to be used either sitting on a flat surface near eye level or strapped to the user's chest to allow it to be a dynamic system.

A. Hardware Overview

Sean's hardware design takes advantage of many already developed sub-systems each optimized to perform specific tasks. Each component is strategically connected as a module in Sean to solve lower level problems. These lower level solutions are combined to produce a system capable of processing multi-domain data in parallel while maintaining a real-time output to the user. Sean's hardware diagram can be seen in Figure 1.

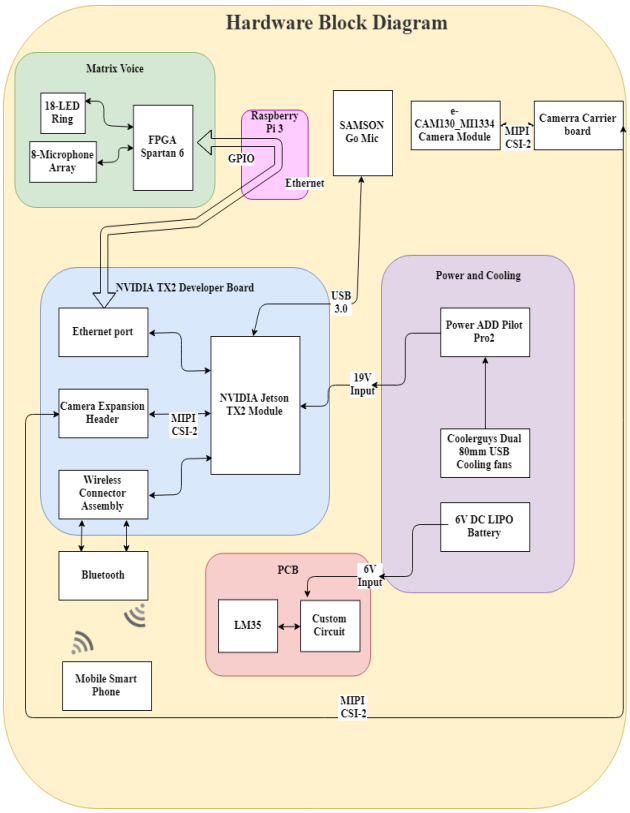

Fig. 1. Hardware block diagram of Sean. Shows how hardware is connected and interacts with the system.

The Matrix voice is a self-contained hardware component that contains the microphone array, an 18-LED ring, and an FPGA for audio processing. This component is connected to the Raspberry Pi 3 which sends information about the DOA of the detected source signals to the TX2 via ethernet. Sean's PCB, is a vital component to ensure the safe use of the system. When the outside of the acrylic packaging reaches 104 degrees Fahrenheit a beeper goes off to let the user know that the system is getting unsafe to wear. In the future the outputs of the temperature senor PCB would be sent to the mobile application so the user could have a visual indication as well.

The FPGA on board the MATRIX voice is responsible for performing the audio processing task of identifying separate signal sources in real-time that can then be combined with the visual results on the TX2 module. The Matrix Voice directly interfaces with the Raspberry Pi 3. Communication between the FPGA and the TX2 module takes place though ethernet connecting the TX2 and the Raspberry Pi 3.

The camera communicates with the TX2 through a MIPI CS2 connection. The camera carrier board is made specially for the TX2 so communication can occur very quickly.

Wireless communication occurs though the Wireless connector assembly. Here external wireless antennas are connected that enable wireless communication. A mobile smart phone will connect to our physical system. All of these hardware components will be communicating with the TX2. The TX2 module will be responsible for processing tasks that require high computing power. The GPU cores on the TX2 module will be responsible for the Computer Vision face detection algorithms. The hex-core CPU performs an analysis on the audio outputs along with outputs from the face detection results. This hardware layout is ideal for real-time processing.

B. Software Overview

In this section the algorithms and software that will be implemented on the audio and visual sides of the system will be discussed. The algorithms and techniques used to merge the information gathered from the audio and visual space into one are also defined.

In the initial proposal two preliminary audio processing filters were discussed. These were to remove microphone noise and the user's own voice. However, through initial testing it was shown that our new chosen microphone was not subject to noticeable noise, and our requirements were shifted to only focus on people talking to the user. In future development, a user-voice filter would be explored along with a hardware induced noise filter if smaller microphones proved to generate more noise.

[9] The DOA algorithm determines which microphone is closest to a source of sound and begins to store energy for that location. Energy is determined by the amplitude of the signal and duration of the signal in that location. Energy can be stored in 18 locations corresponding with our 18 microphones in the array. The longer and louder a signal persists, the more energy that location is allowed to build up. A closer person talking to the user will be more obvious allowing energy to build up quicker while further away sporadic sources will never build up enough energy to be meaningful. An energy threshold is set to determine when a signal produced enough energy for its location information to be sent to the TX2 for processing. Signals that build up more energy persist for longer amounts of time allowing someone to stop speaking and then continue without amplification being altered. This has a maximum cap of enough energy stored to allow information to be sent for up to 8 seconds. This behavior is represented by LED lights on the Matrix Voice in the position of each microphone. If a microphone is activated that light turns blue. Once enough energy is built up to send that information the light turns green. As the energy is depleted the light blinks until it eventually turns off. An example of this is shown in Figure 2.

Fig. 2. On the left a blue light represents a source signal being interrogated. On the right the source signal has built up enough energy for it's location information to be sent to the TX2 for processing.

[7][8] During video processing a Human Face Detection algorithm known as DetectNet is used to determine where faces are in the scene. This is a fully convolutional deep learning neural network that uses GoogLeNet as the base network. GoogleLeNet was retrained on the "Faces in the Wild" dataset to be repurposed for face detection. GoogleLeNet was as a good choice for our application as it uses Inception Modules that view 3 different images at the same time during training to become invariant to size, pose, and lighting conditions. This allows Sean to analyze the environment for humans that are potential sources of sound and compare results with the DOA algorithm to determine the best times for sound amplification/attenuation. The two spaces are merged by having the sensors aligned and placed in a common angle space. The results are compared in the confidence analysis portion of the software diagram.

TensorRT is a conversion tool provided by NVIDIA that allows for quick optimization of neural networks for streamlined GPU implementation. Before inference time this model is pushed through TensorRT optimization to increase the real-time capabilities of this network.

Fig. 3. Real-time human face detection capture of Brandon Kessler and participant on Sean. A false alarm can be viewed in the background. This shows how having 2 independent modules can increase the robustness of the system.

As the audio sub-system builds energy for each microphone Sean generate a confidence value for that location relating to the likelihood the sound we are processing is coming from a potential human source.

The visual sub-system will do the same. The values from the separate systems are then compared and if both pass their respective independent thresholds, Sean amplifies the source of sound. On the visual side, the faces being considered as sources are identified on the feed using boxes to show the user what sources are being considered. The Sean phone application will relay all this information to the user. Sean's software consists of the following components:

1) Start Up Process
2) User Input
3) App Calibration
4) Audio Capture
5) Direction of Arrival (ODAS)
6) Video Capture
7) Human Face Detection
8) Visual and Audio Space Alignment
9) Confidence Analysis
10) Final Visual Output
11) Final Audio Output

The first processing the video stream will go through is a frame capture process. The camera adapter board and carrier board allow the video input to be sent to the TX2 via a MIPI CSI-2 interface. Low-level drivers that are preloaded with the camera module enable streamlined communication between the camera and the TX2's CPU. These frames can be used through a Gstreamer interface that allows for real-time processing on the frames. Audio is also captured in this way allowing for real-time amplification and attenuation of the audio source.

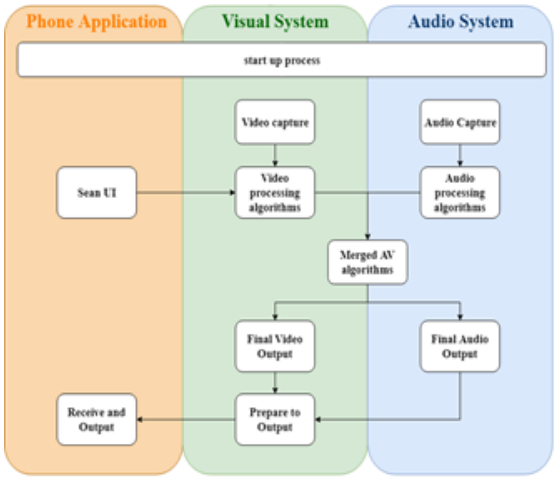

Fig. 4. Software diagram for Sean. Shows that the Phone Application, Visual System, and Audio System in interact with each other during operations.

C. iOS Mobile Application

This section will discuss the software components of the mobile applications in terms of functionality and ease of use. The user interface is a large part of any application or piece of software because if it is not set up correctly, the user will have a hard time setting up or using the software, rendering it ineffective. The user select mode takes some of the guesswork away from Sean because it essentially tells the system which settings to implement for a specific user which should allow for faster processing times. It also gives the user a sense of personalization for having their own voice profile. The environment selection section aids both the processor and microphone array because it tells these components what to expect from the surroundings of the user. Lastly there are a number of smaller features in addition the ones discussed about that will all be discussed in the following sections.

Upon clicking the icon for the mobile application on their device, the user will be taken to an introductory screen that displays the full name of the device as well as the logo. The logo is a graphic that depicts a simplified ear surrounded by small yellow circles, similar to what the microphone array looks like, with the name Sean to make it easily identifiable. This will give the application time to load and serve as a sort of preparation screen.

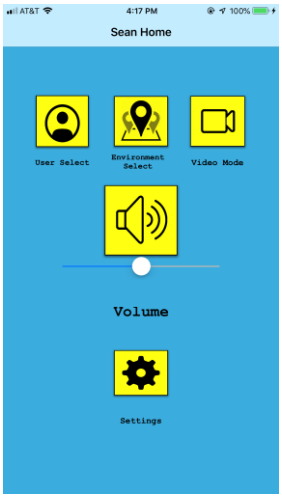

Fig. 5. Sample Home Screen for the Sean Mobile Application.

Following that the user will be taken to a home screen which includes a number of elements to interact with. On the top left there will be a button that takes the user to an area to create or select an existing user profile. This is one of the first things the user should interact with because it will allow for little to no set up time for future uses of the application. To the right of that is a button that allows the user to select the environment in which he or she will be

operating the device. To the right of that is the button that will the user to access Video Mode. The user does not have to use Video Mode while using Sean and can instead remain on the home screen if he or she chooses to do so. Near the center of the screen is a sound icon with a sliding bar underneath that will control the level of audio the user experiences. However, the user could also use the buttons on their device or headphones to control the volume as well. At the bottom of the screen will be a button that takes the user to the device settings page to ensure Sean and any secondary devices such as headphones are connected. Then the user will be prompted to return to the home screen and continue using the application.

Once the user clicks the user select icon, they will be taken to the main user select screen. This screen lists all the user voice profiles that are stored on this device. Each profile will stay on the device until the user decides to delete it. When the user selects a profile, it will be indicated on the screen then the user may return to the home page by using the home icon in the top left corner of the screen. At the bottom there is an option to create a new voice profile.

From the home screen, the user will have the option to select the environment in which they will use Sean. There will be three preset options of a Low, Medium, or High Noise environment with examples listed to allow the user to make the appropriate selection. By default, Low Noise mode will be selected unless the user chooses otherwise. In order to provide additional customization for the user, a feature on the environment select screen is the option to choose the level of background noise the user wishes to hear. If the background noise is too distracting to a user they may set it to a minimum amount. If the user wishes for a natural audio experience, the user may choose a higher sound level.

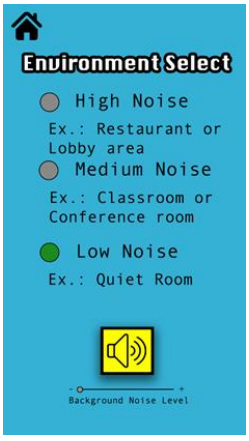

Fig. 6. Environment select for Sean. This app changes the difference between amplified and not amplified signals with respect to environment type selected.

From the home screen, the user will have the option to select video mode which changes the orientation of the application to a landscape mode. In this mode the user will be able to see the streaming video feed from the camera in Sean. There will be an option in the top corner to turn on or off subtitles for the audio that the user will hear. There will also be an "X" button will take the user back to the home screen when pressed.

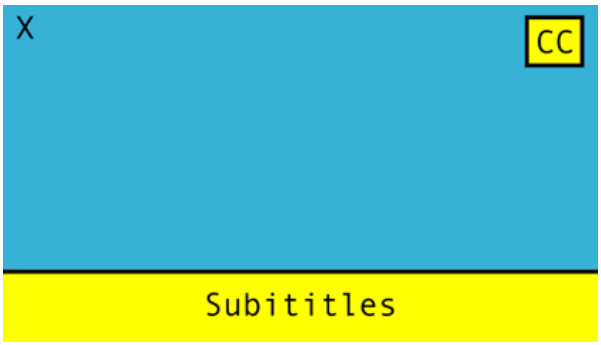

Fig. 7. Video streaming screen of Sean mobile application needs to be landscape view to watch live feed.

D. Printed Circuit Board

The printed circuit board of Sean is designed to be a sound alarm temperature sensor. When the back plate of the packaging—the part that is against the user's chest reaches 43 degrees Celsius-- which is just below the temperature that can start to cause first degree burns, the buzzer buzzes until temperatures drop below 43 degrees Celsius again. This allows time for the user to safely remove Sean from their chest to prevent harm to themselves. Safety is an important priority in all products but especially those that are intended to be wearable for people. The printed circuit board is one of the main measures towards safety in this project. The layout of the board is seen below in Figure 8.

The main components of the printed circuit boards are the LM35DZ temperature sensor and the LM358 op amp. The temperature sensor's output pin changes in voltage as the temperature changes and is compared to a reference voltage through the op amp that acts as a comparator. If the temperature sensor voltage is higher it toggles a 5V output at the output of the op amp sounding the alarm through a piezoresistive buzzer.

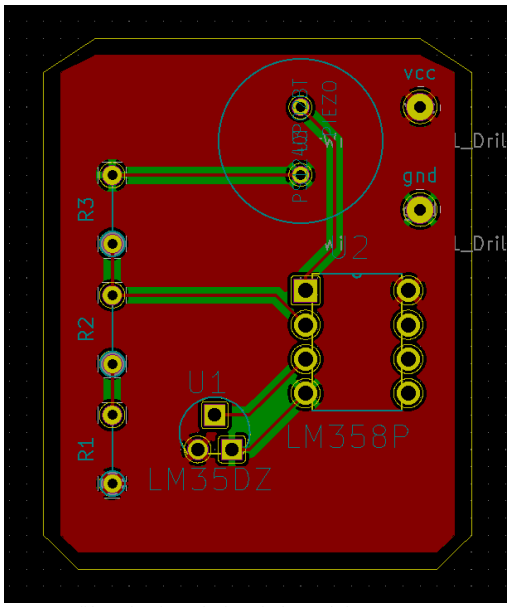

Fig. 8. CAD file of printed circuit board for Sean.

The red front copper layer is the voltage source layer and the green back copper layer is the ground plane.

E. Packaging

The packaging design for Sean was modeled in SOLIDWORKS® as seen in Figure 9. Acrylic was chosen as the material for the packaging as it was easily accessible and machines that could easily cut it were available. Due to the nature of the material of the packaging, acrylic, the packaging was made in six separate parts and glued together with acrylic welding glue with the exception of the front piece which was attached with Velcro to allow for access to Sean's components. Of the six pieces of the packaging, two are the right and left sides that have vents to allow for heat dissipation, another two are to top and bottom sides, on the back is the base that contains the mounting holes for the TX2 and the battery.

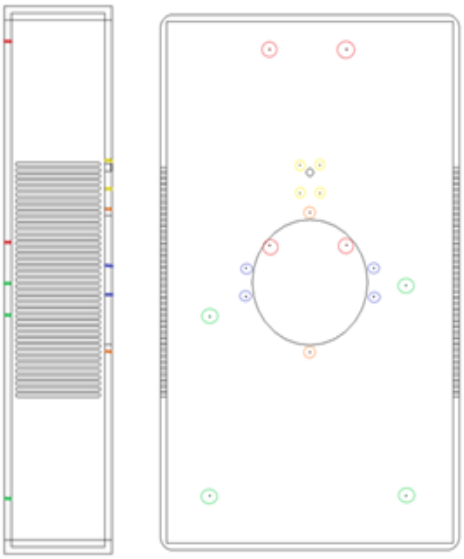
Fig. 9. Solidworks drawing of Sean packaging. This figure includes mounting hole diagram.

[11] To make Sean's packaging wearable, it was attached to a harness with an extra strength Velcro that has the ability of holding up to 15 lbs. The whole device weighs 10.6 lbs and thus comfortable is held by the Velcro. This allows the user to safely walk around with Sean attached to their chest.

Table 1 -- Dimensions of Sean Packaging

Component	Part Name	Length (mm)	Width (mm)	Height (mm)	Thickness (mm)
Rectangular casing	Housing for processor, PCB, and battery	394.46	240	76.20	5

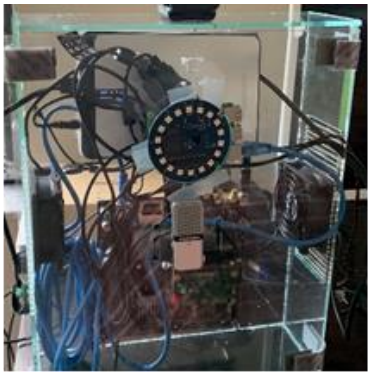
Fig. 10. Sean packaging with all components housed and mounted.

III. Results

The results in this section are going to be based on the success and functionality of the device in comparison to a standard hearing aid. These results and feedback come from a participant (not to be disclosed) who has suffered a loss of 60% of their hearing. The following cases in Table 2 below were tested; these cases were created to outline and test the limitations of the Sean while also proving the worth of having a visual component.

Table 2 – Test Cases for Sean

Case	Desired Results
(1) User is alone, no other people, no unexpected noise.	"Natural" and clear environment noise.
(2) User is with one other person (not talking) and noise in the background.	"Natural" and clear environment noise.
(3) User is with one other person who is in front of the camera talking.	Target's voice is amplified and clear.
(4) Case 2 with another person in the background not talking.	Target's voice is amplified and clear.
(5) User is with two people having their own conversation in the background.	Target's voice is amplified and clear. Background is attenuated but still clear and present.

Of these cases, the results with the participant were mostly positive; the participant picked a number on a scale one to five (i.e. one is definitely worse than the hearing aid, three is the same as the hearing aid, five is definitely better than the hearing aid). Although this system is not a completely accurate way of measuring any success or failure it will and does tell us that adding this visual component is beneficial in this practice. Below in Table 3 are the results from the participant. Along with these number values, the participant said these values could potentially change for better or worse in a nonidealized environment. The next step in the testing process would be migrating from a low noise and optimally lit environment for testing.

Table 3 – Test Results

Case	Results and Feedback
1	3, Sound is the same as hearing aid in ideal environment.
2	3, Sound is the same as hearing aid in ideal environment.
3	4, Voice comes through clear and accurate timing of amplification. Again, consider ideal environment.
4	4, Voice comes through clear and accurate timing of amplification. Inconsistent recognition and amplification.
5	4, Nothing was amplified as we did not care about that separate conversation.

IV. Conclusions and Recommendations for Future Work

The results section supports the assertion made early on that adding a visual component to hearing aid systems would improves the intended user's experience. Hearing impaired persons that use standard hearing aids often complain of excessive noise amplification and not enough of a focus on solving the cocktail party problem. Sean accurately detects a face and is able to map it to a source detected on the microphone array. The whole signal is then amplified and will continue to do so until they have finished talking. The algorithms have been optimized for close (1-2m) human-to-human interaction. The sensitivity for the source detection of the microphone array is low to prevent random noises from being amplified in the case that the Human Face detection accidentally identifies a face in the same quadrant as the sound that was mapped.

Unfortunately, time was the most valuable resource for this project. Even with the aggressive schedule the group tried to maintain, a lack of time ended up resulting in more limitations. The original limitations of Sean were that the user only wants to talk to one person at a time; the limitations are now that the user cannot speak and the user wants to hear anyone talking in their field of view that is speaking loud enough.

With time and more familiarity with C++ the possibility to separate different source signals and amplify them according the confidence tables originally developed is achievable. The hardware for this project comfortably performs the algorithms with minimal program crashing from the Raspberry Pi. The audio's latency could potentially be improved with designing a custom framework or using one better suited than Gstreamer.

Acknowledgement

The authors wish to acknowledge Lockheed Martin for the generous sponsorship and the ECE department of the University of Central Florida for the support provided throughout the research and development of Sean. The

authors also wish to thank Jonathan Tucker for being an advisor for the duration of two semesters of Senior Design.

REFERENCES

[1] Quick Statistics About Hearing. (2018, October 05). Retrieved from https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing

[2] Hearing aids: How to choose the right one. (2018, May 22). Retrieved from https://www.mayoclinic.org/diseases-conditions/hearing-loss/in-depth/hearing-aids/art-20044116.

[3] Gupta, N., Lucia, A., Dunn, N., & Puzella, M. (2017). Earbeamer: A Parallel Beamforming Hearing Aid System. Retrieved March 28, 2019, from

[4] Hear Active Listening - Change The Way You Hear The World. (n.d.). Retrieved from https://www.kickstarter.com/projects/dopplerlabs/here-active-listening-change-the-way-you-hear-the

[5] Elgan, M. (2015, August 17). New earbuds give you super-hearing. Retrieved from https://www.computerworld.com/article/2971267/new-earbuds-give-you-super-hearing.html

[6] Gandel, C. (2016, October 03). Hearing Aid Price, How To Keep Costs Down. Retrieved March 28, 2019, from https://www.aarp.org/health/conditions-treatments/info-2016/hearing-aid-costs-prices-cs.html

[7] Vidanapathirana, M., & Vidanapathirana, M. (2018, March 24). Real-time Human Detection in Computer Vision - Part 1. Retrieved from https://medium.com/@madhawavidanapathirana/https-medium-com-madhawavidanapathirana-real-time-human-detection-in-computer-vision-part-1-2acb851f4e55

[8] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition (p./pp. 1--9), .

[9] Source Localization Using Generalized Cross Correlation. (n.d.). Retrieved March 28, 2019, from https://www.mathworks.com/help/phased/examples/source-localization-using-generalized-cross-correlation.html

[10] Entertainment, F. (n.d.). Wi-Fi vs Bluetooth: Is There an Impact on Sound Quality? Retrieved from https://www.fusionentertainment.com/pulse/wi-fi-vs-bluetooth-is-there-an-impact-on-sound-quality

[11] Lamkin, P. (2016, February 17). Wearable Tech Market To Be Worth $34 Billion By 2020. Retrieved from https://www.forbes.com/sites/paullamkin/2016/02/17/wearable-tech-market-to-be-worth-34-billion-by-2020/#174bb3b3cb55