AVAL: Audio-Visual Active Locator
ECE-492/3 Senior Design Project
Spring 2014

Electrical and Computer Engineering Department
Volgenau School of Engineering
George Mason University
Fairfax, VA

Team members: Rony Alaghbar, Kelly Byrnes and Jacob Cohen
Faculty Supervisor: Dr. Kathleen E. Wage

Abstract:
Videoconferencing is a growing trend among businesses, as
the systems reduce need for travel, in turn reducing cost
and carbon footprints, and increasing productivity. There
has been steadily increasing growth in the industry.
However, the main drawbacks appear to be the large start-
up cost for equipment and installation, the need for
professional IT maintenance and setup, and overall lack in
system mobility. The proposed design alleviates the cost
while enhancing functionality of existing systems. This
project aims to design a compact, platform independent,
and affordable tabletop device for a terminal of a
videoconferencing system. The Audio-Visual Active Locator
(AVAL) takes input audio, and uses a microphone phased
array to locate the speaker’s position. This information is
used to position a camera toward the calculated location.
This functionality aims to enhance current conferencing
techniques, while embracing seamless integration into
existing systems.

1. Introduction
Videoconferencing is a widely used business practice for meetings and education. The General Services
Administration cited increased videoconferencing would reduce both travel expenses and pollutant emissions [2].
A single terminal of a room-based videoconferencing system requires audio and video input/output, processing,
and synchronization. Current systems typically utilize monitors as the hub for the videoconferencing terminal.
These monitors are either bulky portable systems or mounted within the room. The coupling of the terminal with
the large monitor increases the overall cost, and significantly hinders the mobility of the device. Not only do group
video systems require mounting or cumbersome relocation, but they also require personnel training and IT
support. Smaller, more transportable terminals are typically desktop-based and geared towards individual users. In
order to produce a product capable of mitigating these concerns, AVAL utilizes two, four-element microphone
arrays centered on the table to steer a servo-controlled camera towards the current speaker. Source localization
signal processing algorithms are used to find the person speaking.

2. Approach
A microphone array was designed specifically for speech signals between 400 and 1000 Hz. The array design
assumes point-to-point propagation. Equation 1 displays the necessary length of the array to assume point-to-
point propagation. $R$ represents the distance from the sensor, while $D$ represents the array length and $\lambda$ represents
the wavelength. This equation shows the minimum wavelength as the limiting factor, being about 34 cm for 1KHz.
Assuming the distance to the array is near 120 cm, the maximum array length to assume planar propagation is 45 cm long. Experimentation resulted in an array length of 51 cm as the best fit for finding the angle of incidence.

\[ R > \frac{2D^2}{\lambda} \]  

(1)

The aperture of the array, 51 cm, can be viewed as a rectangular spatial filter (Figure 1). Fourier transform of the spatial filter, also seen in Figure 1, is given with radians along the x-axis and gain along the y-axis. The length of the array is represented by \( L \) in the figures above, while \( \lambda/L \) dictates the beam-width of the array in radians. With a lower limit frequency of 400Hz, corresponding to a wavelength of 85cm, and an array length of 51 cm, the resulting beam-width is a maximum of \( \pi/2 \) radians. Discrete microphone points are placed along the entire length of the aperture. This discretization effectively turns the array into a spatial and temporal sampling problem.

There are many types of beamforming algorithms used today, but they all mimic filtering the signal in space and time. Beamforming a microphone array towards a given angle of incidence is analogous to filtering out every signal coming from alternate angles. Delay and Sum Beamforming, Figure 2, is susceptible to errors in time sampling. To synchronize accurately between the microphones, a high sampling frequency is needed along with high resolution to the applied delays to avoid destructive interference. This method also has no weightings associated with each sensor, which will make the signals coming from the furthest sensor just as powerful as the closest. This will mitigate the added noise in the signal from propagating longer to the furthest sensor. Filter and Sum beamforming adds the delays to each sensor using phase shifts associated with each filter. Since all the computations are done within the time domain, it is still susceptible to inaccuracies in phase delays.

AVAL implements source localization using frequency domain analysis and beamforming. Working in the frequency domain allows for analysis of broadband signals and improved accuracy of the calculation. Spatiotemporal Filtering (filtering in space and time) creates a Fourier Transform of the beamformed output. The output Fourier Transform is determined by weighting the individual transforms and conducting a matrix multiplied with a complex exponent, or phase shift, and finally summing these manipulated transforms. In order to localize on the speaker, AVAL has implemented a “fixed angle” approach. This has mitigated the need for dynamic filtering, which would require
amplification circuit

determined filters.

amplification circuit

microcontroller, as common microcontrollers are not powerful enough
to handle the required processing with the

PC and

Data Acquisition System (DAQ), the camera positioning unit with a Servo Controller and webcam, and
the microphone array consisting of eight electret condenser microphones. While the microphone array takes
the audio input, and the camera position system takes the video, the processing unit acts as a mediator between the
two. It determines the location of the speaker from the microphone input and forwards the information to the
camera positioning system. The processing unit also consolidates and outputs the singular audio-video feed.

AVAL consists of three main components and a power system, as seen in Figure 3: the processing unit including
the PC and Data Acquisition System (DAQ), the camera positioning unit with a Servo Controller and webcam, and
the microphone array consisting of eight electret condenser microphones. While the microphone array takes
the audio input, and the camera position system takes the video, the processing unit acts as a mediator between the
two. It determines the location of the speaker from the microphone input and forwards the information to the
camera positioning system. The processing unit also consolidates and outputs the singular audio-video feed.

The processing unit introduces the main functionality of AVAL. It requires data acquisition from the microphones,
which is sent through an algorithm to produce a location for the speaker source. Then the location is sent to the
camera positioning system. These processes are implemented on a standard PC. A PC was chosen over a
microcontroller, as common microcontrollers are not powerful enough to handle the required processing with the
latency required to operate in real time. One camera is utilized to track the conversation in the room using the
location data from the PC. The microphones are connected directly to the pre-amplification circuit. The outputs
from the eight channels of pre-amplification circuit are connected directly to the DAQ. All other components are
connected via USB to the PC.

In order to properly mount various subcomponents, 3D printing was employed to create hardware specifically
designed to work with AVAL’s various subcomponents. The AVAL camera and servo mounts were designed in
OpenSCAD and 3D printed on a Makerbot Replicator 2X. The CAD renders can be seen in Figure 4.

changing the filters based on an input. AVAL’s fixed angle method utilizes multiple pre-determined filters.

2. System Design
AVAL operates in a standard office environment, is easy to implement, and is affordable. The system
autonomously tracks who is speaking with a camera using microphone arrays and interfaces with teleconferencing
software. To ensure ease of use, including installation and operation, AVAL implements a table-top design and
takes video and speech audio signals and outputs a cohesive audio-video stream of the current speaker.

AVAL consists of three main components and a power system, as seen in Figure 3: the processing unit including
the PC and Data Acquisition System (DAQ), the camera positioning unit with a Servo Controller and webcam, and
the microphone array consisting of eight electret condenser microphones. While the microphone array takes
the audio input, and the camera position system takes the video, the processing unit acts as a mediator between the
two. It determines the location of the speaker from the microphone input and forwards the information to the
camera positioning system. The processing unit also consolidates and outputs the singular audio-video feed.

The processing unit introduces the main functionality of AVAL. It requires data acquisition from the microphones,
which is sent through an algorithm to produce a location for the speaker source. Then the location is sent to the
camera positioning system. These processes are implemented on a standard PC. A PC was chosen over a
microcontroller, as common microcontrollers are not powerful enough to handle the required processing with the
latency required to operate in real time. One camera is utilized to track the conversation in the room using the
location data from the PC. The microphones are connected directly to the pre-amplification circuit. The outputs
from the eight channels of pre-amplification circuit are connected directly to the DAQ. All other components are
connected via USB to the PC.

In order to properly mount various subcomponents, 3D printing was employed to create hardware specifically
designed to work with AVAL’s various subcomponents. The AVAL camera and servo mounts were designed in
OpenSCAD and 3D printed on a Makerbot Replicator 2X. The CAD renders can be seen in Figure 4.

changing the filters based on an input. AVAL’s fixed angle method utilizes multiple pre-determined filters.
3. Microphone Array

The most important module of the AVAL is the microphone array. Different geometries of microphone arrays such as linear, annular, and planar have limits on the range of localization. In addition to designing the geometries of the microphone array, the spacing between each microphone was considered to limit aliasing of the audio signals.

The array spacing and dimensions were designed by considering spatial sampling, angular resolution of the array, and sound propagation to the array. Since sound signals are broadband signals, the array was designed for frequencies from 400Hz to 1KHz. According to the Nyquist Sampling Theorem, in order to sample effectively, the spacing between the two sensors needs to be less than half the minimum wavelength of the signal. This means that the microphones can be spaced a maximum of 17cm apart, half of our wavelength of 34 cm, when dealing with a 1 KHz source. The angular resolution of the array is also determined by the length of the aperture. As the length of the aperture increases, the beam-width decreases, as shown in Figure 5 below.

![Directivity Pattern of Each Array](image)

Figure 5: Length of Aperture vs Directivity

The last parameter considered was the limitation of Far Field Assumption. Since sound waves are originally spherical, our array must be far enough to assume that the waves are planar. Utilizing planar propagation allows for the assumption of point-to-point propagation, simplifying calculations and modeling in array signal processing. After experimenting with different spacing arrays of four microphones, data was taken and processed through a preliminary algorithm. Figure 6 shows the results of the spacing testing and its ability to output the correct angle. Data was taken for three different angles over four different trials. The RMS error of the outputted angle was calculated. The graph shows that the least amount of error occurred at 17cm, which determined the final dimensions of the microphone arrays.

![RMS Error of Spacing Testing](image)

Figure 6: RMS Error of Spacing Testing

The audio signals obtained through the microphone array must now be converted into data the device can use. It is imperative that these audio signals are sampled synchronously. If not, all calculations will be incorrect and output audio may be distorted or mistimed with the video. Before amplification, all extraneous signals beyond 2KHz (high frequency noise) and below 60Hz (DC components) are eliminated using a high pass and low pass passive filter. These anti-alias filters condition the signal before amplification. Before the audio signals are...
processed, they must be amplified. Since the audio signals from the microphones are weak (micro-volts), a preamp amplifier connected to the microphones strengthens the signal (up to 10 Volts peak-to-peak) for further processing [3]. Once the analog signal is ready for processing, it is converted to a set of digital data. For this process a Data Acquisition System (DAQ) is used. Many DAQs use a “round robin” approach when sampling multiple channels. However, “round robin” does not guarantee synchronous data acquisition, which is key to the required beamforming processing. An 8-channel ADC samples the signal and quantifies the values to send to the location calculator. The DAQ used to prepare the signal must support the number of channels required. Since maximum frequency of speech signals are around 3400Hz, a minimum sampling rate of 6800Hz is needed across each channel of input. However, as conversational frequencies lie under 2KHz, a smaller sampling frequency may suffice. We are using analog, omnidirectional electret condenser microphones as inputs to our system.

Once the sampled audio signals are stored, the signals are sent to the Location Calculator module of the device. The location calculation is processed within the PC, as speed and signal processing power are vital for signal tracking. The processor implements spatiotemporal filters for source localization. This decreases the need for extremely high sampling rates required for time-domain processing. The spatiotemporal method is utilized for the location calculation and also for our optional audio enhancement.

One camera mounted on the device is able to rotate 360° in the azimuth direction. Motors with precision in position, such as servos, are used to actuate the mounts of the webcams. The camera control device was designed to be quiet, as to not distort the microphone input. During operation, the camera always points to the speaker while they are talking if nobody else has spoken yet.

4. Audio Processing Algorithms
The scan function or source localization function acts to find a persistent sound source at a new location (Figure 7). First, the processing unit takes a short time Fourier transform of the data coming in from the DAQ, about 150ms. Then, a spatial transform is applied to the frequencies under 2KHz. The resulting two-dimensional array consists of a wavelength vs. frequency plot. Peaks within the plot above 95% of the maximum value are chosen and the plot is summed across at the discretized angles that the system is sampling (50, 70, 90, 110, 130 degrees). The largest sum across the oblique lines corresponding to each angle determines the position of the speaker.

![Figure 7: Scan Algorithm](image)

Figure 8 shows a spectrogram of a speech signal recorded. Most of the spectral content within a voice falls under 2KHz. The spectral content does not change drastically as words change with time. Additionally, there are three major noise bands that are consistent with time; this is due to the room acoustics and equipment running in the
room. Therefore, in our algorithm, only frequencies under 2KHz were processed in further stages. Figure 8 also shows a 2048 point temporal FFT of a 150ms speech signal with the frequencies over 2KHz cutoff. The FFT consists of three dominant peaks where most of the spectral content is contained. Every other frequency that arises in the FFT is at least 10dB less than the three dominant peaks.

Figure 8: Spectrogram of speech signal (left), and FFT of speech signal (right)

Figure 9 shows the result of a spatial FFT taken across the temporal FFT data. This process results in a wave number vs. frequency plot that has characteristics which relate to the source angle. Below is the result of applying a peak algorithm to the previous dataset. The final dataset consists of 1’s at each peak and 0’s otherwise. If the data is taken while a person is speaking from an angle, a series of peaks results along the linear relationship. The output angle is determined by summing across the five discretized angles and selecting the largest sum as the speaker’s angle.

Figure 9: Spatial FFT of Temporal FFT (left) and peak algorithm (right)

5. Experimental Evaluation and Validation

Individually, each component, including the arrays, DAQ, and the servo, were found to fall well within operational requirements. The first experiment focused on location accuracy. Three speakers spoke at angles varying from 50° to 130° and 230° to 310° with intervals of 10° on each side of the array. This data is represented in Figure 10. The
gold X’s signify AVAL’s predefined angles and the green X’s signify the midpoints between two predefined angles. Examining the chart above it becomes clear that all predefined angles are precise within 2-3° of the theoretical value. This falls well within the 15° tolerance defined in the requirements. Moving to the midpoints it should be noted that the camera is always pointed towards one of AVAL’s predefined angles. An anomaly can be seen at 120°, where the system will default to broadside (90°). When the system cannot determine an accurate speaker location (110° and 130° in this example), the ambient noise in the room will win out, which defaults the system to broadside. Overall, the system test proved successful as the speaker was always detected at an AVAL predetermined angle.

The second test consisted of measuring time to location. In each test the camera started at broadside of the opposite array. Each speaker spoke at one of AVAL’s predefined angles and the time to locate the speaker was measured. Observing the data below it can be seen that AVAL has an average response time of roughly 2.25 seconds, well within the 4 seconds stated in the requirements. The algorithm response time is under a quarter second; AVAL takes its time locating the speaker in order to ensure smooth, necessary transitions between speakers. With these experiments completed, it can be said that the AVAL system is functional well within its predefined requirements.

References