Speech Enhancement Techniques for Large Space Habitats Using Microphone Arrays

A Thesis

Presented to the

Graduate Faculty of the

University of Louisiana at Lafayette

In Partial Fulfillment of the

Requirements for the Degree

Master of Science

Peter Y. Achi

Fall 2018

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Speech Enhancement Techniques for Large Space Habitats Using Microphone Arrays

Peter Y. Achi

APPROVED:

Andi G. Petculescu, Chair Associate Professor of Physics	Natalia Sidorovskaia Head and Professor of Physics
Gabriela Petculescu Associate Professor of Physics	Mary Farmer-Kaiser Dean of the Graduate School

Acknowledgments

I would like to thank my advisor, Professor Andi G. Petculescu, for being a model of patience and perseverance, calmly showing me how to "get to the bottom of things." I would also like to thank Dr. Natalia Sidorovskaia and Dr. Gabriela Petculescu for their mentorship over the past 10 years, as well as for providing valuable feedback for this research.

We would like to thank Andy Romero and David Juge for welcoming us to the Audio Development Laboratory at NASA-JSC and helping us in the microphone calibration process.

This work was supported by the Louisiana Space Consortium (LaSPACE).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DOI	Direction of Arrival
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GCC	Generalized Cross-Correlation
ISS	International Space Station
MVDR	Minimum Variance Distorionless Response
PHAT	Phase Transform
PSD	Power Spectral Density
SNR	Signal-to-Noise Ratio
SOI	Signal of Interest
STFT	Short-time Fourier Transform
TDOA	Time Difference of Arrival
ULA	Uniform Linear Array

Chapter 1: INTRODUCTION

The purpose of this work is to perform speech enhancement in a noisy and reverberant environment, characteristic of large space habitats (e.g. spaceships or surface bases). The next two sections will describe typical noise conditions on a spacecraft, as well as set out short-term and long-term goals regarding communication enhancement, and methods of practical implementation.

1.1 Spacecraft Acoustic Environments for Long-duration Missions

The acoustic environment of manned habitats dedicated to long-duration space exploration (e.g. transit spacecraft, orbital stations, or surface bases) plays an important role in ensuring the success of missions. The current approach to spacecraft acoustics [1] is based on the "traditional" techniques and standardization in use for ground-based room acoustics (e.g. reverberation times, noise levels, impulse responses, etc.). If, however, the crew is to be efficient and psychologically healthy in space for long periods of time *and* far away from Earth, the scope of spacecraft acoustics needs to be broadened. For instance, from a habitability point of view, setting reverberation times to certain values as required by Earth-based fixed standard may not be the best solution. Instead, the astronauts may benefit from the ability to *tune* the reverberation level of their quarters as they see fit. This can be accomplished by adjusting the reflection coefficient of the walls by arrays of tunable Helmholtz resonators or metamaterials [2] [3]. From a logistics point of view, the operations aboard, say, a spaceship en route to Mars are likely to be overseen by a central computer. The crew members' day-to-day activities would be greatly eased if they could communicate with

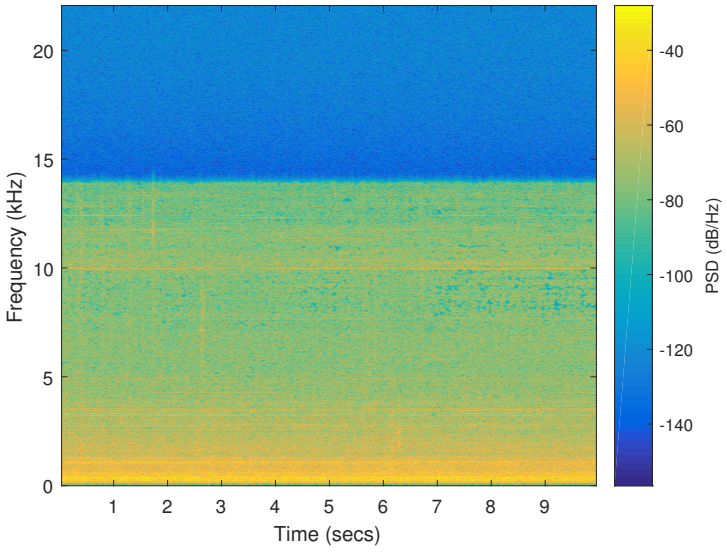
the computer and among themselves verbally, without the need of hands nor headsets.

Space habitats like the International Space Station (ISS) contain sources of noise, like fans and pumps [1]. For an astronaut confined in the habitat twenty-four hours a day, an elevated amount of noise can cause auditory pain, headaches, ringing in the ear, and discomfort as well as psychological symptoms such as irritability, decreased productivity, and errors in judgment. The measures being taken to deal with noise are beyond the scope of this research. However, we do take into consideration the noise's deleterious effect on communication.

According to Goodman et al. [1] , "noise is one of the most important habitability issues in the ISS because it affects all operations and interferes with verbal communication." Goodman et al. recommend a reverberation time (the time it takes for the sound to decay by 60 dB, i.e. T_{60}) ranging from 0.2 to 0.5 seconds for quiet environments to preserve speech intelligibility. It is noteworthy that the reverberation time of Spacelab — a European orbital laboratory program — ranges from 0.566 seconds (for the 8 kHz octave band center frequency) to 1.907 seconds (for the 63 Hz octave band center frequency).

To have a better idea of the nature of the noise aboard some of these space habitats, we examine an actual noise sample from the ISS. Colonel Chris Hadfield [4] has published some sound samples on the internet containing the background noise that can be heard on the ISS. A spectrogram of this noise is displayed in Figure 1.1.

Figure 1.1. International Space Station Background Noise



1.2 Goals for Future

In the future, we would like to allow for quieter conditions to promote a healthy and efficient sound environment for the crew. Furthermore, places where astronauts live and work for an extended time should be designed with the possibility to adapt acoustic wall characteristics, such as the complex acoustic impedance, to the crew members' needs and/or moods.

1.3 Communication Goals

While the efforts to reduce noise aboard spacecrafts is underway, the presence of noise is still a reality that must be addressed, especially regarding communication. The main aim of this research is to allow hands-free and headset-free communication for astronauts in the presence of noise and interference. Let us suppose that an astronaut is moving about in a noisy part of the spacecraft. The goal is to allow him/her to

initiate a command to the computer using a cue word such as "Orion!". This would be feasible using dense microphone arrays along the walls of the chamber, only a small section of which would be activated near the astronaut to execute the beamforming for localization and speech enhancement.

1.4 Motivation for Arrays

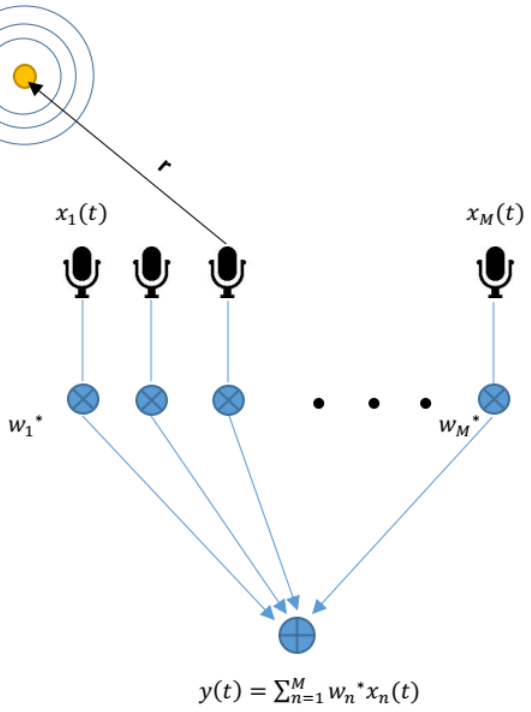
There are many reasons for using an array of microphones, rather than a single one. As suggested in the previous section, the abundance of microphones provides a greater freedom to the astronaut, such that they can be audible to the AI ship communication system regardless of their position. Aside from this practical reason, microphone arrays also have powerful advantages. As we will show in this thesis, by astutely combining the signals collected by the array, meaningful information that would not be available, if taken from a single microphone, may be extracted. For example, denoising speech using a single microphone is feasible but it introduces distortion. This problem is circumvented when using array processing [5]. Another benefit of array processing is the ability to localize the astronaut, which will be addressed in the following section.

Chapter 2: THEORY

2.1 General Time-domain Beamforming

Delay-sum beamforming, also known as conventional beamforming, is the simplest—and most intuitive—method of array processing. As far as the array geometry is concerned, the most straightforward arrangement is the uniform linear array (ULA), whose sensors are equally spaced. Let M be the number of sensors (here, microphones), d the sensor spacing, and $L = (M-1)d$ the effective length of the array. Let an acoustic signal originate at location vector $\mathbf{r}$ with respect to the reference sensor. The output of the general beamformer shown in Figure 2.1 is a weighed sum of the M inputs; obtaining the corresponding weight applied to each input signal forms the crux of array processing.

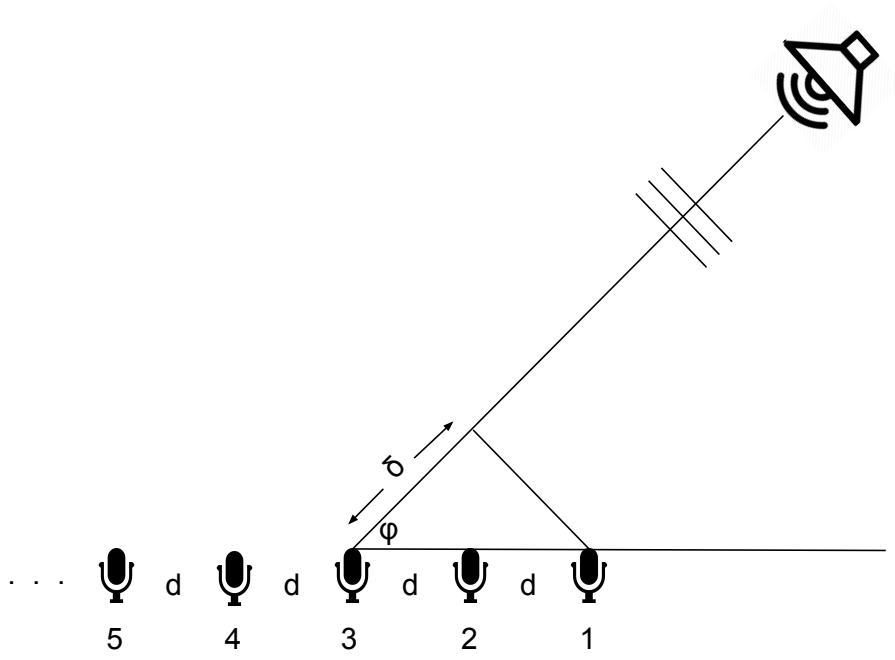
Figure 2.1. General beamforming



Van Veen and Buckley [6] define a beamformer as "a processor used in conjunction with an array of sensors to provide a versatile form of spatial filtering. The sensor array collects spatial samples of propagating waves, which are processed by the beamformer. The objective is to estimate the signal arriving from a desired direction in the presence of noise and interfering signals." Let $x_i(t)$ be the signal recorded by microphone i. Simply put, beamforming is the operation whereby we combine the microphone signals $x_i(t)_{\mathrm{i=1:M}}$ with the corresponding weights $w_{i\mathrm{i=1:M}}$ in such a way to produce the desired output. The purpose for this operation may be source localization, noise reduction, dereverberation, or a combination thereof.

If the distance r is in the far field of the array i.e. $r > \dfrac{2L^2}{\lambda}$, then one can assume that the wavefronts reaching the microphones are (quasi-)plane (Figure 2.2).

Figure 2.2. Delay-sum geometry



In this configuration, the wavefronts from the event will reach microphone 1

first, then microphone 2, and so on. The path difference δ_i for the i-th microphone with respect to microphone 1 is simply

$$\delta_i = (i-1)d\cos\phi. \tag{2.1}$$

Therefore, the time delay, τ_i of the acoustic event at microphone i is

$$\tau_i = \frac{\delta_i}{c} = \frac{(i-1)d\cos\phi}{c}, \tag{2.2}$$

where c is the speed of sound in the propagation medium (air). Accordingly, one can electronically steer the microphone array to a certain direction by (i) delaying each input signal, then (ii) adding all the signals—hence the name *delay-and-sum*. In other words, the output of this beamformer in the time domain, $y(t)$, will be

$$y(t) = \frac{1}{M}\sum_{i=1}^{M} x_i(t-\tau_i) \tag{2.3}$$

Delay-sum beamforming may be used to detect the direction of arrival (DOA) of a signal of interest (SOI). By steering the array to a range of angles and measuring the output power of the beamformer at each angle, the angle which maximizes this power represents the DOA. This method was used to detect the DOA of a SOI in an acoustically treated experiment which minimized reverberations.

Since a time delay in the time domain corresponds to a phase shift in the frequency domain, we can express this beamformer in the frequency domain as:

$$y(f) = \frac{1}{M}\sum_{i=1}^{M} x_i(f)e^{j(-2\pi f\tau_i)}. \tag{2.4}$$

Care must be taken when implementing this form of the beamformer, as the output of the beamformer is frequency-dependent. If the SOI is not a monochromatic (single-frequency) signal, multiple narrow-band filters must be used in the pre-processing stage. Since speech is a broadband signal, and is more efficiently processed in the time-frequency domain (see Section 2.4), the delay-sum beamformer proves unsuitable for our purposes. Furthermore, delay-sum beamforming breaks down in the presence of reverberation, as spurious directions are detected due to the multiple reverberations. To overcome this problem, the GCC-PHAT [7] (generalized cross-correlation phase transform) algorithm was used for localization due to its robustness against reverberation. The details of this algorithm are elucidated in Section 2.5.

2.2 Array Parameters and Aliasing

According to McCowan [8], a microphone array can be treated as a sampled version of a continuous linear aperture. In general, the response of an aperture depends on the frequency of an incident signal and the direction of arrival of that signal; it is described by a function known as the array directivity pattern. For a ULA of M microphones with a spacing of d meters, the directivity pattern is:

$$D(f,\phi) = \sum_{i=1}^{M} w_i(f) e^{j\frac{2\pi f}{c}(i-1)d\cos\phi}. \tag{2.5}$$

Therefore, there are three variable parameters in designing a microphone array: the number of microphones, M, the microphone separation distance, d, and the frequency of the incident signal, f. To avoid temporal aliasing, the Nyquist criterion

must be observed, that is, the sampling frequency must be at least twice the highest frequency involved. A spatial extension of this criterion allows us to avoid spatial aliasing. Specifically, d must be less than half of the minimum wavelength of the incident signal:

$$d < \frac{\lambda_{min}}{2}. \tag{2.6}$$

Using Equation 2.5, we can simulate the microphone array which will be used in the experiment, with twenty-seven microphones equally spaced by 6.67 cm, and the frequency ranging from 100 Hz to 8 kHz. A top view of the normalized directivity pattern is shown in Figure 2.3 and a 3-D version is shown in Figure 2.4 for a clearer demonstration of spatial aliasing as a function of incident frequency.

Figure 2.3. Directivity pattern ($M = 27$, $d = 6.67cm$)

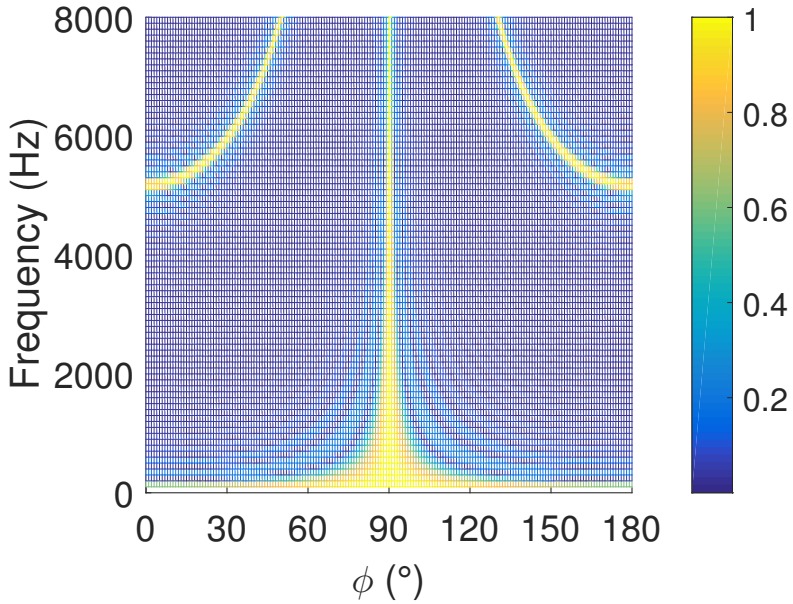
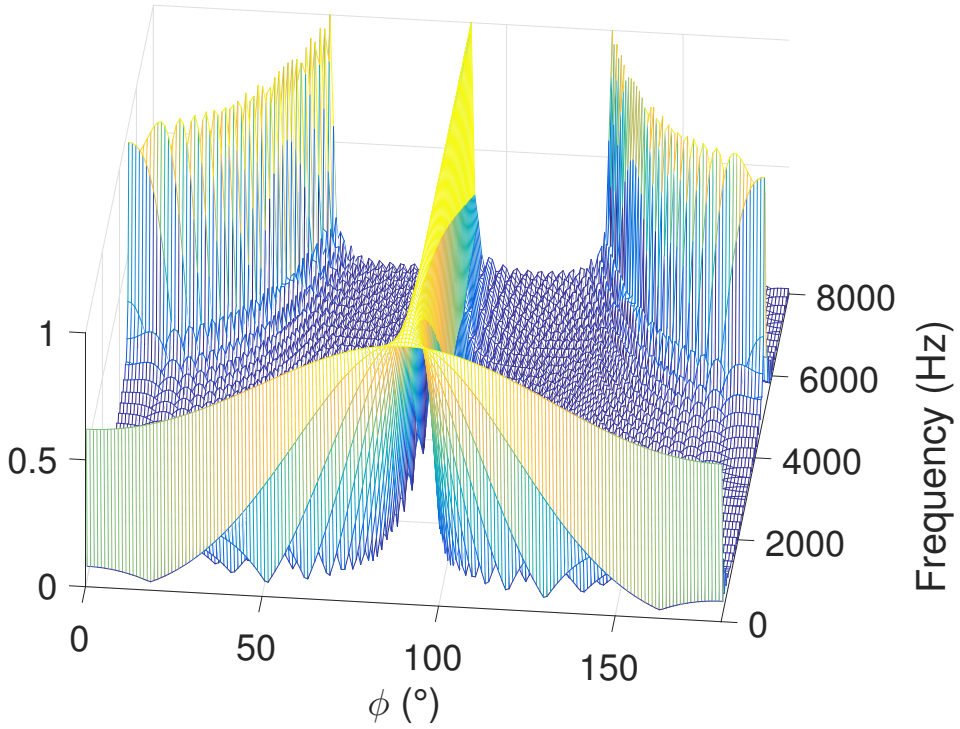


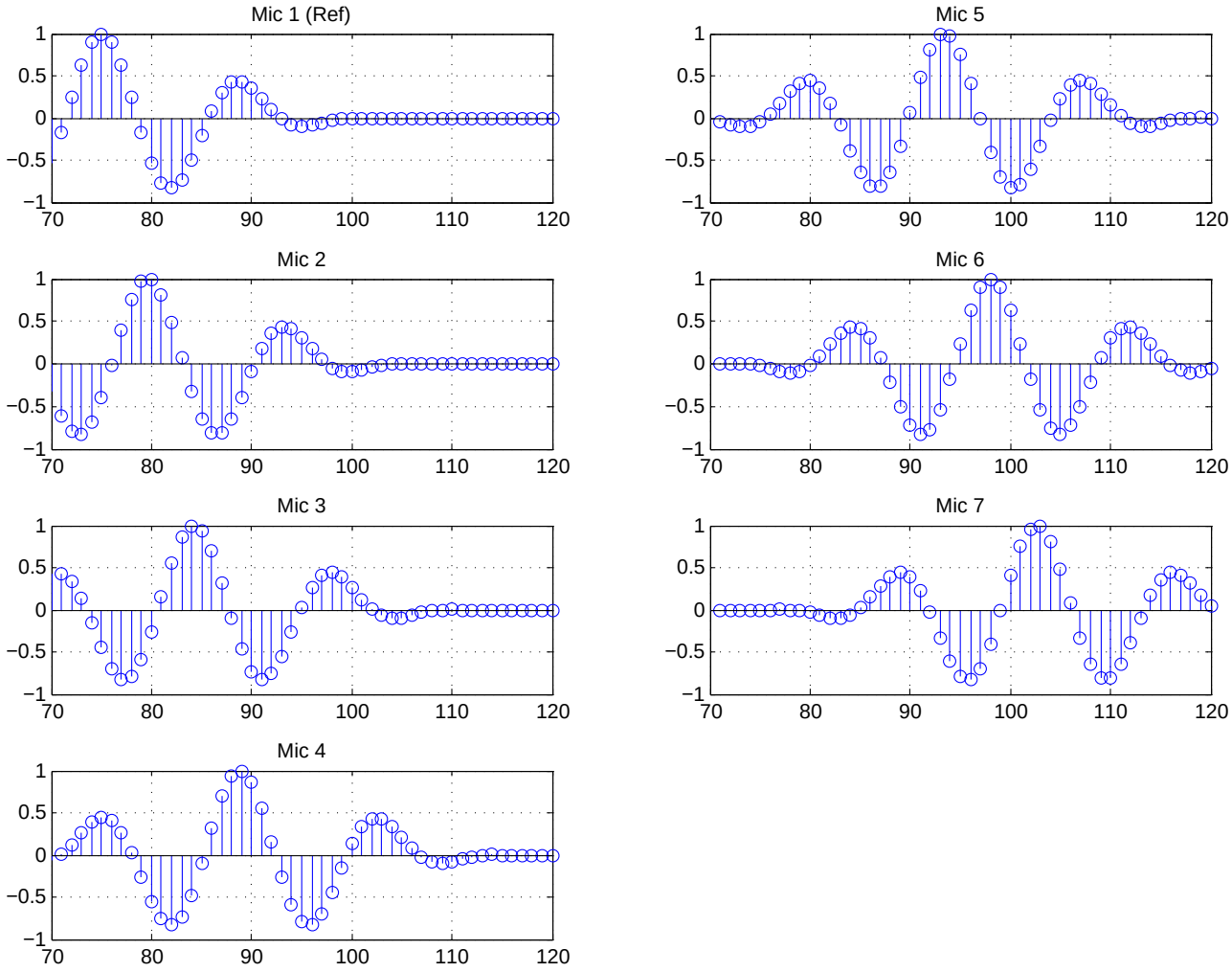
Figure 2.4. Directivity pattern ($M = 27$, $d = 6.67cm$)



2.3 Implementation Challenges

2.3.1 Fractional delays. Equation 2.3 was used to implement Delay-Sum beamforming in the time domain. However, some computational challenges were encountered in the process. While delaying an analytical signal is straightforward, delaying a digital signal is not as simple. This arises from the fact that delaying a sampled signal by a certain amount of time τ does not necessarily correspond to shifting by an integer number of samples. To illustrate this concept, the signals received by a seven-microphone array are simulated and displayed in Figure 2.5.

Figure 2.5. Fractional delay



A closer look at the signals in Figure 2.5 reveals that the distribution of the data points neighboring each peak and trough varies among the seven waveforms. This happens because the samples are not simply shifted by an integer amount. Rather, a *fractional delay filter* was used to delay the reference microphone signal by the time delay which corresponds to the extra distance traveled by the signal of interest to each microphone. According to Bai et al. [9], a fractional delay can be implemented using Lagrange interpolation, through an FIR filter.

Delay-sum beamforming presents an intuitive solution to the source-localization problem. However, though it works in quasi-anechoic laboratory conditions, this algorithm breaks down in reverberant conditions, in the presence of noise and interfering sources, due to the spurious directions introduced by the reverberations. Therefore, the solution for this problem will be addressed using the generalized cross-correlation method. Now, we shift our focus to the time-frequency domain, which is widely used to analyze speech.

2.4 Time-frequency Analysis

Speech usually involves relatively long waveforms, whose frequency content changes fast with time. The obvious way to address such time-varying spectra is to work in the time-frequency domain. A typical technique relies on the Short-Time Fourier Transform (STFT). According to Benesty et al. [5], time-frequency beamforming can be summarized as follows:

1. the signal at each microphone is divided into small time frames,

2. the Fast-Fourier Transform (FFT) is used to transform each frame into the frequency domain,

3. a specific enhancement filter is applied to the complex STFT coefficients, for each frequency,

4. the inverse FFT is applied to bring the enhanced speech signal (beamformer output) back to the time domain.

Many authors describe this frame-based method. In this thesis, I will follow the description presented by Nobutaka Ito [10]. The STFT of a digitized signal, $\alpha[k]$, is implemented through an analysis window $w[k]$ (e.g. Hanning or Hamming) defined over the interval $[-K/2, K/2-1]$, where K is the number of samples per frame:

$$\alpha(\tau,\omega) = \sum_{k=\tau S-K/2}^{\tau S+K/2-1} \alpha[k]w[k-\tau S]e^{-j2\pi k\omega/\omega_s}. \tag{2.7}$$

In Equation 2.7, τ is the frame index, S is known as the frame hop, ω is the angular frequency, and ω_s the sampling frequency. This method of frames will be implied in the following derivations, and it is at the heart of the main algorithm of this research.

2.5 Source Localization with Reverberation

In the literature [7], generalized cross correlation (GCC) is used to estimate the time difference of arrival (TDOA) between two signals. It is simply the cross-correlation of two signals which have been weighted by a function suitable for the application. Suppose we have two microphone signals, $x_1(t)$ and $x_2(t)$. The generalized cross

correlation $R_{12}(\tau)$ between them is

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_{12}(j\omega)X_1(j\omega)X_2^H(j\omega)e^{j\omega\tau}d\omega, \tag{2.8}$$

where $X_1(j\omega)$ and $X_2(j\omega)$ are the discrete Fourier transforms of $x_1(t)$ and $x_2(t)$ respectively, and Ψ_{12} is a weighing function. Since our experiment contains reverberation, we choose the phase transform (PHAT) for the weighing function:

$$\Psi_{12} = \frac{1}{|X_1(j\omega)X_2^H(j\omega)|}. \tag{2.9}$$

The reason for this transform's robustness against reverberation is as follows: since the noise power is proportional to the signal power at a given frequency in the presence of reverberation, we eliminate this dependence by dividing the frequency spectra of the signals by their magnitude, thus only preserving information about the phase [11]. This is exactly what the phase transform does. The algorithmic implementation of this method and the results for localization in reverberation are found in Section 4.2.

2.6 Minimum Variance Distortionless Response Beamformer

Now, we address the issue of speech enhancement in the time-frequency domain using the Minimum Variance Distortionless Response (MVDR) beamformer, as presented by Habets et al. [12]. First, we consider an array of M microphones. From the basic theory of linear systems, we can say that the signal, $y_n(t)$, recorded at the n^{th} microphone is simply the signal of interest (SOI), $s_0(t)$, convolved with the impulse response between the SOI and the n^{th} microphone, with the addition of some incoherent noise and coherent noise (interference) contained in the term $v_n(t)$. This is

expressed concisely in Equation 2.10

$$y_n(t) = g_n * s_0(t) + v_n(t), \quad n = 1, 2, ..., M. \tag{2.10}$$

By taking the FFT of equation 2.10, we acquire the frequency-domain version of the array model, and also the inputs for the MVDR beamformer:

$$\mathbf{Y}(j\omega) = \mathbf{G}(j\omega)S_0(j\omega) + \mathbf{V}(j\omega), \tag{2.11}$$

where the boldface notation indicates compacting the spectra for all the microphones into one vector. Therefore the beamformer output, $Z(j\omega)$, is the sum of all the inputs which have been weighted by the beamformer weights contained in the vector $\mathbf{W}$:

$$Z(j\omega) = \mathbf{W}^H(j\omega)\mathbf{Y}(j\omega) \tag{2.12}$$

$$= \mathbf{W}^H[\mathbf{G}(j\omega)S_0(j\omega) + \mathbf{V}(j\omega)]. \tag{2.13}$$

The MVDR algorithm requires the following stipulations on the output: minimal noise and interference with a fixed gain, Q, to the SOI. That is, we have the following constrained optimization problem:

$$\min \mathbf{W}^H\mathbf{\Phi}_v\mathbf{W} \quad \text{s.t. } \mathbf{W}^H\mathbf{G} = Q, \tag{2.14}$$

where $\mathbf{\Phi}_v$ is the power spectral density (PSD) matrix of the noise. It can be shown that the solution to this problem is the modified famous Capon beamformer [12]:

$$\mathbf{W}_{\text{MVDR}}(j\omega) = Q^*(j\omega) \cdot \frac{\mathbf{\Phi}_v^{-1}(j\omega)\mathbf{G}(j\omega)}{\mathbf{G}^H(j\omega)\mathbf{\Phi}_v^{-1}(j\omega)\mathbf{G}(j\omega)}. \tag{2.15}$$

The novelty in this form presented by Habets et al. [12] is the dual usage of the beamformer. In the case where Q is set to equal the complete frequency response

between the SOI and reference microphone of the array, the beamformer behaves as a noise reduction filter. However, when Q is set to equal the direct-path portion of the same response, the beamformer then behaves as a combined noise reduction and dereverberation filter. These two modes will be elucidated and illustrated in the experimental portion of this thesis.

The MVDR beamformer is usually derived (see, for example, [13]) using the manifold vector, $\mathbf{d}$. The latter contains information about the geometry of the array using the time delays for each microphone. In the frequency domain, the manifold vector is:

$$\mathbf{d}(j\omega) = [e^{-j\omega\tau_1} ... \, e^{-j\omega\tau_M}]^T. \tag{2.16}$$

Such derivations present the solution to the MVDR beamformer in the following form:

$$\mathbf{W}_{MVDR}(j\omega) = \frac{\mathbf{\Phi}_v^{-1}(j\omega)\mathbf{d}(j\omega)}{\mathbf{d}^H(j\omega)\mathbf{\Phi}_v^{-1}(j\omega)\mathbf{d}(j\omega)}, \tag{2.17}$$

In contrast, our approach relies on the experimentally measured impulse responses between the SOI and each microphone ($\mathbf{G}(j\omega)$), which requires no a priori knowledge of the geometry of the array.

2.7 Performance Metrics

In order to gauge the performance of the MVDR beamformer, we use three metrics proposed by Habets et al. [12]: the input signal-to-noise ratio, iSNR, the output SNR, oSNR, and the array gain, $\mathscr{A}_g$. Each of these metrics may be computed "locally"— as a function of frequency—or "globally"—by integrating over all frequencies.

The local iSNR is

$$\text{iSNR}(j\omega) = \frac{|Q(j\omega)|^2 \phi_s(j\omega)}{\phi_{vref}(j\omega)}, \tag{2.18}$$

whereas the global iSNR is

$$\text{iSNR} = \frac{\int_{-\pi}^{\pi} |Q(j\omega)|^2 \phi_s(j\omega) d\omega}{\int_{-\pi}^{\pi} \phi_{vref}(j\omega) d\omega}, \tag{2.19}$$

where $\phi_s(j\omega)$ is the PSD of the SOI and ϕ_{vref} is the PSD of the reference microphone noise signal. In other words, the iSNR is the ratio of the power of the beamformer input SOI and the power of the input noise.

Similarly, the second metric is calculated locally as follows:

$$\text{oSNR}(j\omega) = \frac{|\mathbf{W}_{\text{MVDR}}^H(j\omega)\mathbf{G}(j\omega)|^2 \phi_s(j\omega)}{\mathbf{W}_{\text{MVDR}}^H(j\omega)\mathbf{\Phi}_v(j\omega)\mathbf{W}_{\text{MVDR}}(j\omega)}. \tag{2.20}$$

The global version of the oSNR becomes:

$$\text{oSNR} = \frac{\int_{-\pi}^{\pi} |\mathbf{W}_{\text{MVDR}}^H(j\omega)\mathbf{G}(j\omega)|^2 \phi_s(j\omega) d\omega}{\int_{-\pi}^{\pi} \mathbf{W}_{\text{MVDR}}^H(j\omega)\mathbf{\Phi}_v(j\omega)\mathbf{W}_{\text{MVDR}}(j\omega) d\omega}. \tag{2.21}$$

As its name implies, the oSNR is simply the ratio of the power of the beamformer output SOI and the power of the output noise.

Finally, the array gain is the ratio of the global output SNR and the global input SNR:

$$\mathscr{A}_g = \frac{\text{oSNR}}{\text{iSNR}}. \tag{2.22}$$

It is a measure of how effectively the array enhances the signal-to-noise ratio by reducing the noise.

Note that in a real-world situation, we generally do not have access to the original SOI. Therefore, we cannot compute the SNRs directly. Instead, we may

approximate the SOI with the data actually recorded by the microphone array. The metrics that use this approximation are the input data-to-noise ratio (iDNR) and the output data-to-noise ratio (oDNR).

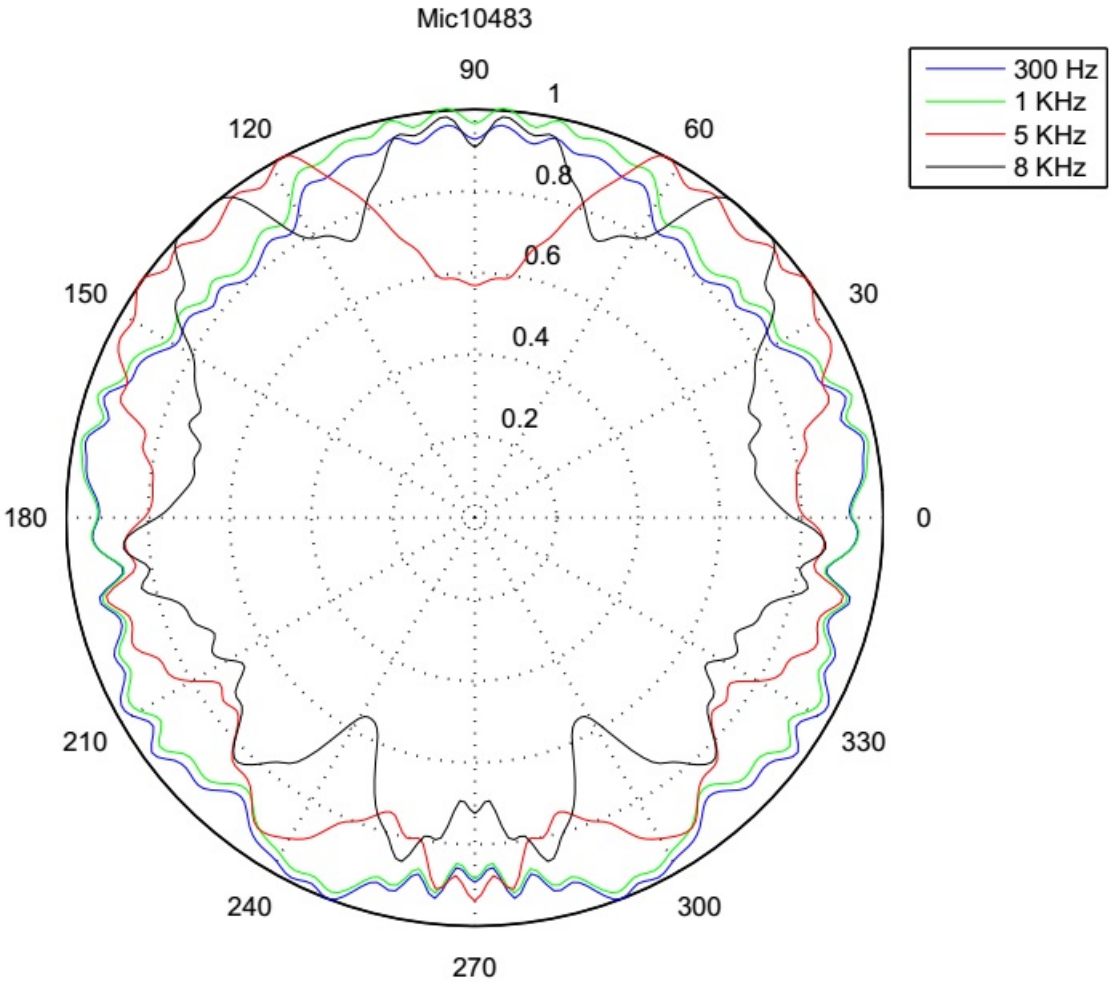
Chapter 3: EXPERIMENT

3.1 Microphone Calibration

A proper implementation of the delay-sum beamformer in the time domain (Section 2.1) requires one to account for the angular sensitivity of the array microphones. In order to determine the directivity response of each of our thirty-two microphones, a visit was arranged to the Johnston Space Center Audio Development Laboratory (ADL). In their 15'x15'x15' anechoic chamber, we positioned each microphone on a remotely controllable turntable. A signal containing four sine waves (with frequencies of 300 Hz, 1 kHz, 5 kHz, and 8 kHz) was synthesized using the Audacity software, and played through a speaker facing the microphone. The signal was recorded through the microphone using the Orion DAC, Ardour recording software, and an Asus laptop, then the microphone was rotated through an angle of 3 degrees and the process was repeated.

After exporting the .wav files, the directivity patterns of each microphone were extracted at the four frequencies and normalized to the highest amplitude. The patterns were then interpolated for all angles between 0 and 360 degrees with a 1° resolution. A sample directivity pattern is shown in Figure 3.1

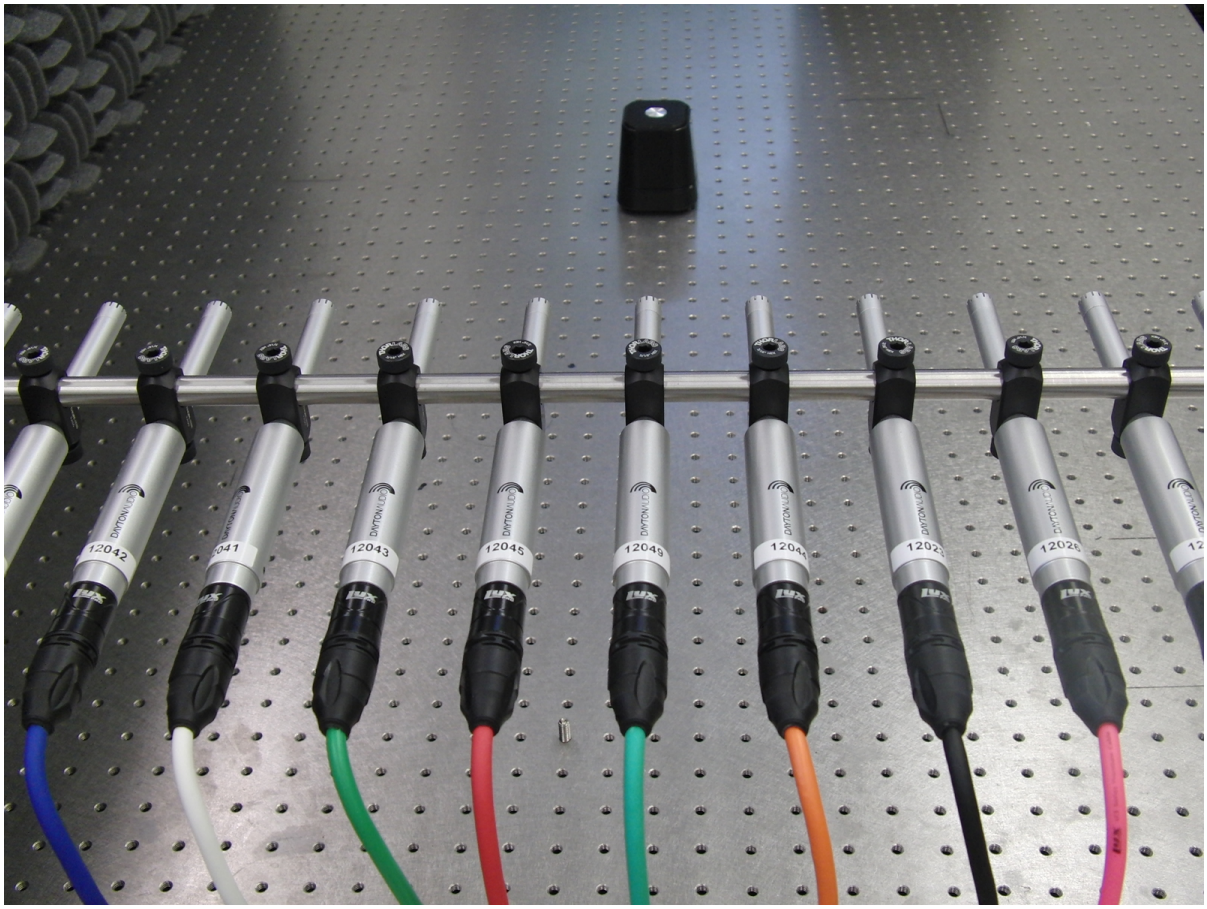
Figure 3.1. Directivity pattern for Microphone 10483



3.2 Preliminary Measurements

In order to implement a general framework for beamforming, we performed preliminary measurements in a quasi-anechoic enclosure. The latter was set up on top of an optical table using foam panels to minimize reverberation. An image of the setup is displayed in Figure 3.2 below.

Figure 3.2. Preliminary experiment setup



In this setup, a microphone array was constructed on an acoustically treated optical table, and a signal of interest (SOI) was placed at a given azimuthal angle, ϕ. The experiment was recorded using the Orion 32-channel analog-to-digital converter and a laptop. Sum-delay beamforming (Section 2.1) was implemented in Matlab in order to recover the direction of arrival (DOA) of the SOI. A sample analysis of this preliminary experiment is shown in Section 4.1.

3.3 Experimental Setup

We simulated a spaceship deck using a vacant hallway in the Physics Department (Broussard Hall) on the ULL campus. The dimensions of the hallway are, approximately, 20.28 m (L) x 2.98 m (W) x 3.71 m (H). For the array, we used Dayton EMM-6 electret microphones (frequency response 18 Hz – 20 kHz, sensitivity 10 mV/Pa), which were previously calibrated (see Section 3.1). A uniform linear array (ULA) of 30 microphones was set up along a longitudinal wall of the hallway; the spacing between adjacent sensors was 6.67 cm. The ULA microphones were connected to a 32-channel microphone preamplifier and analog-to-digital converter interface (the *Antelope MP 32/Orion 32* system). We sampled the experiment at 44,100 Hz. The signal of interest (SOI)—recorded separately—was played through an Avantone MixCube passive reference monitor. This particular loudspeaker was chosen because it does not have crossover electronics that may introduce distortions in the SOI. The SOI, "Orion, turn on the heater!" represents something an astronaut would utter, containing a keyword—"Orion"—that could be used to trigger the "Smart Spaceship" system.

In order to acquire the impulse response between the SOI and each microphone, an exponentially swept sine wave whose frequency ranged from 100 Hz to 8 kHz was played through the same loudspeaker, and the microphone inputs recorded. Additionally, a sound sample of actual ambient noise from the International Space Station [4] was played through a three-way large bookshelf loudspeaker (Sony SS-B3000). Two separate mono amplifiers (AudioSource 5.2) were used to provide consistent gain levels to the exponential sine sweeps and ISS noise. To simulate

interfering sources (e.g. other astronauts, onboard systems, etc.), two interfering signals were played from different directions simultaneously (a radio station and a French poem being read by one of the authors) using independent audio players. A schematic diagram of the experiment is shown in Figure 3.3, and a real picture of the experiment is shown in Figure 3.4.

Figure 3.3. Experimental setup — schematic

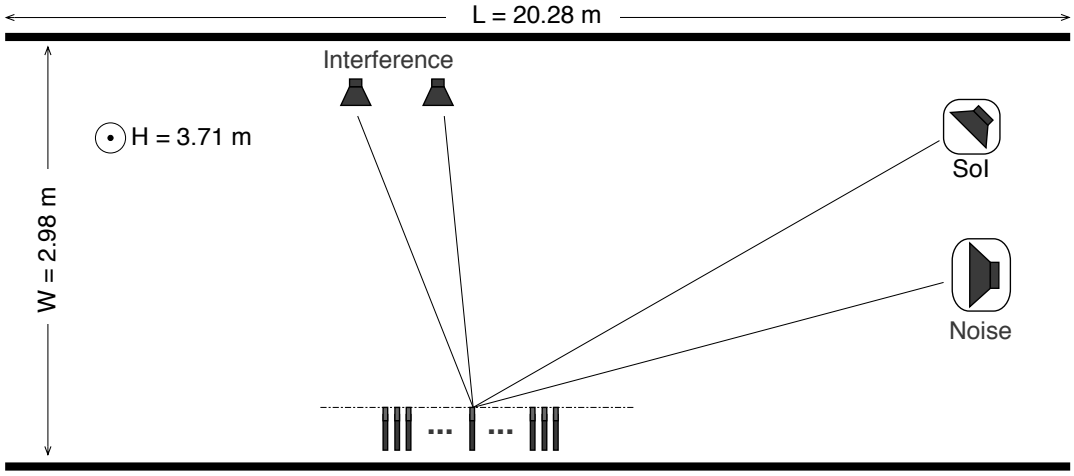


Figure 3.4. Experimental setup

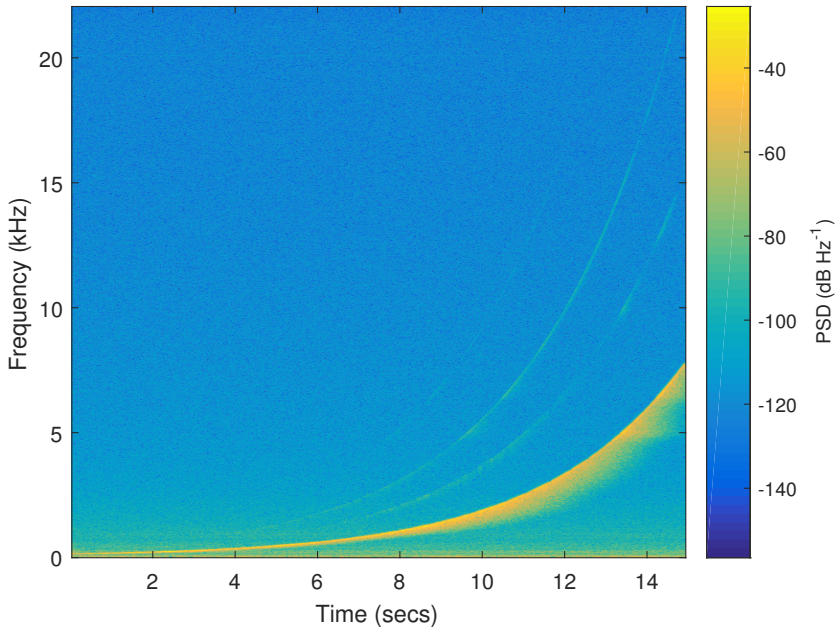
3.4 Impulse Response of the Hall

The impulse response of the hallway at the location of each microphone was calculated using an exponential swept-sine technique presented by Farina [14]. Initially, a 15-second test signal, $x(t)$, that consists of a sinusoid with exponentially increasing frequency, ranging from 100 Hz (f_1) to 8 kHz (f_2), was synthesized in Matlab, using the following equation:

$$x(t) = \sin\left[\omega_1 T \, \frac{e^{\frac{t}{T}\ln\left(\frac{\omega_2}{\omega_1}\right)} - 1}{\ln\left(\omega_2/\omega_1\right)}\right] \tag{3.1}$$

where $\omega_1 = 2\pi f_1$, $\omega_2 = 2\pi f_2$, and T is the sweep duration in seconds. A sequence of the test signal followed by silence was constructed. The sequence was played through the speaker placed at the same location of the SOI and recorded by each microphone of the ULA. A spectrogram of the input at one of the microphones is displayed in Figure 3.5 below.

Figure 3.5. Exponential sine sweep



This spectrogram illustrates the advantage of Farina's method over other methods for finding the impulse responses. The sine sweep excitation can be seen in yellow in the spectrogram. However, we can also observe fainter shadows of this main sweep. These shadows are a result of the nonlinearities present in the system, most likely due to the speakers used. Therefore, this method allows us to extract the linear part of a system's impulse response from the nonlinear components.

The impulse response was extracted as follows: we assume that for microphone i, the signal recorded, $y_i(t)$, is a convolution of the input signal with the impulse response of the hallway-microphone system, $g_i(t)$, that is:

$$y_i(t) = x(t) * g_i(t), \tag{3.2}$$

where $*$ is the convolution operation. To extract the impulse response for each microphone, we suppose that an inverse filter $f(t)$ exists such that

$$x(t) * f(t) = \delta(t), \tag{3.3}$$

where $\delta(t)$ is the Dirac delta function, i.e. the unit impulse. If we convolve each side of Equation 3.2 with this inverse filter and exploit the commutativity of the convolution operation, we get:

$$\begin{aligned} y_i(t) * f(t) &= x(t) * g_i(t) * f(t) && (3.4)\\ &= x(t) * f(t) * g_i(t) && (3.5)\\ &= \delta(t) * g_i(t) && (3.6)\\ &= g_i(t). && (3.7) \end{aligned}$$

The fact that $\delta(t)$ is the identity function for the convolution is used in the last step. In short, to acquire the impulse response for each microphone, we convolve the signal recorded by that microphone with our inverse filter, $f(t)$. $f(t)$ is derived by Novak et al. [15] to be

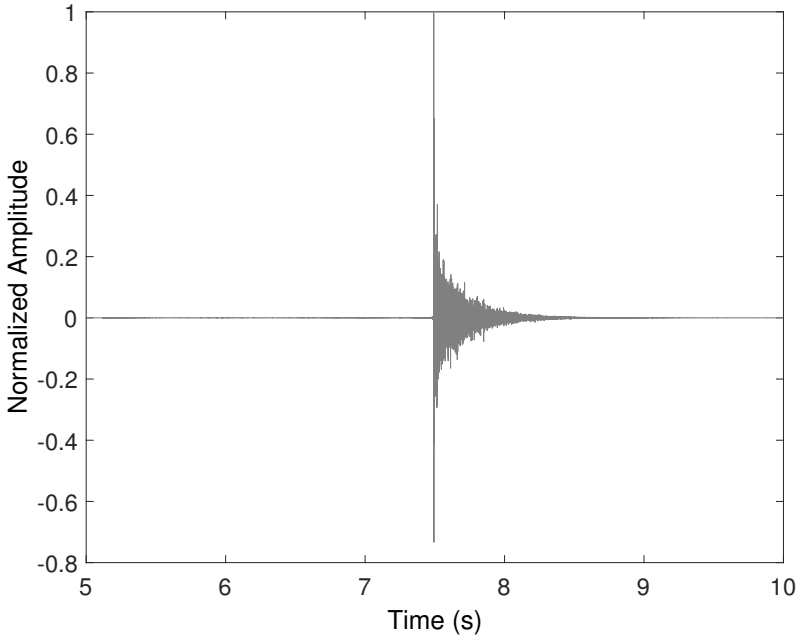
$$f(t) = \frac{f_1}{L} e^{-\frac{t}{L}} x(-t), \tag{3.8}$$

where $x(-t)$ is the time-reversed input signal, $x(t)$, and the parameter L is:

$$L = \frac{T}{\ln\left(\frac{\omega_2}{\omega_1}\right)}. \tag{3.9}$$

A sample impulse response acquired using this method is shown in Figure 3.6.

Figure 3.6. Sample impulse response

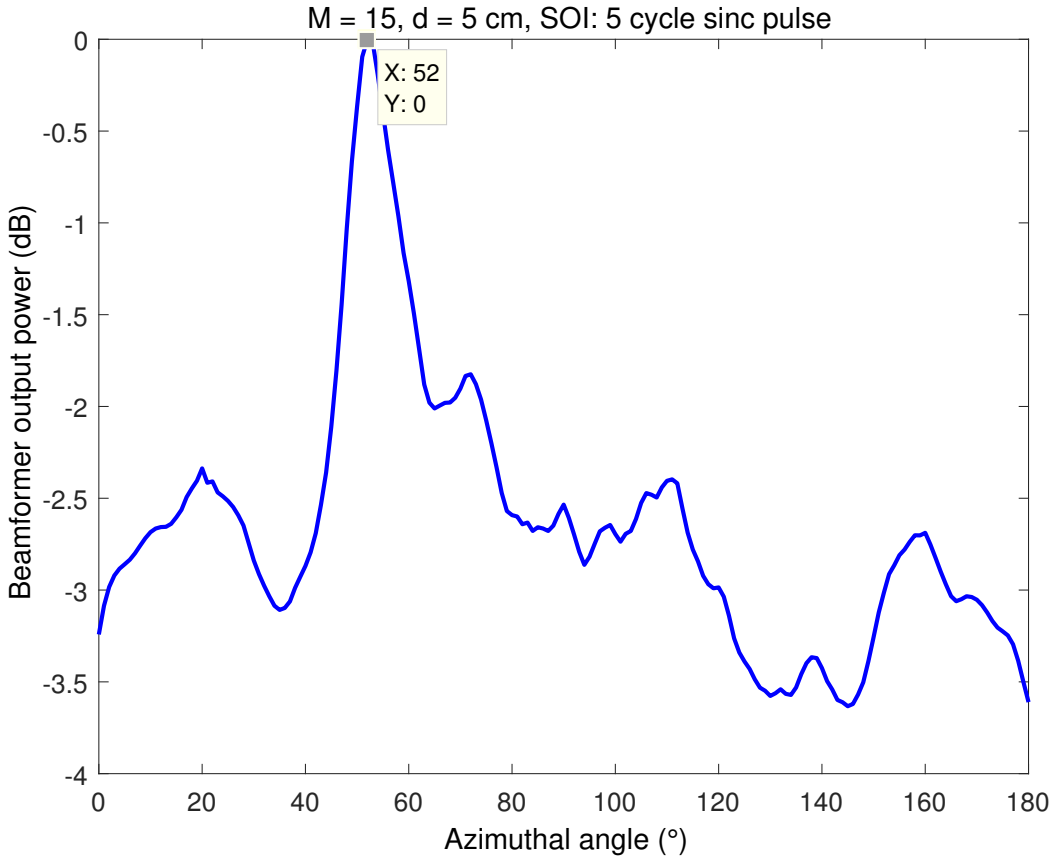


Chapter 4: RESULTS

4.1 Source Localization with Reverberation Suppression

A sample analysis of the preliminary experiment described in Section 3.2 is addressed here. In the case where the array consisted of fifteen microphones equally spaced by 5 cm and the SOI consisted of a 5-cycle sinc pulse arriving from a direction of $53^\circ \pm 2^\circ$, we recovered the beamformer output power as a function of azimuthal angle shown in Figure 4.1.

Figure 4.1. Delay-sum sample result



This beamformer output shows that the individual signals of the microphone array add up most coherently when the array is electronically steered to an azimuthal

angle of 52°. We may deduce that the DOA of the SOI is 52°.

4.2 Source Localization in the Presence of Reverberation

Since the delay-sum algorithm in the time domain fails in the presence of reverberation, we revert to the GCC-PHAT algorithm (see Section 2.5). The method of frames outlined in Section 2.4 is used to implement this algorithm, as suggested by [7]. For a given short-time frame, r, and an a pair of microphone signals, we compute the cross-spectrum, and average over the total number of frames, N:

$$C_{12}[k] = \frac{1}{N}\sum_{r=1}^{N} X_{1,r}[k]X_{2,r}^{H}[k]. \tag{4.1}$$

Then, the phase weighting function becomes:

$$\Psi_{12}[k] = \frac{1}{|C_{12}[k]|}. \tag{4.2}$$

Finally, the generalized cross-correlation between the two microphones, $R_{12}[\tau]$, is found by taking the inverse Fourier Transform of the product between C and Ψ, namely:

$$R_{12}[\tau] = \text{IFFT}\left(\frac{C_{12}[k]}{|C_{12}[k]|}\right) \tag{4.3}$$

$$= \frac{1}{K}\sum_{k=0}^{K-1} \frac{C_{12}[k]}{|C_{12}[k]|} e^{jk\frac{2\pi}{K}\tau}. \tag{4.4}$$

By implementing the GCC-PHAT algorithm in Matlab, we were able to use it to estimate the DOA of the SOI as well as the two interfering sources, henceforth called *Radio* and *Hugo* (refer to Section 3.3 for details of the experimental setup). Three pairs of microphones were used to test the algorithm for each source; the estimated DOA and its difference from the experimentally measured DOA are tabulated below. For the first run of the experiment, where the walls of the hallway were not treated with any foam

panels, such that the T_{60} reverberation time was approximately 1.4 seconds, the results are as follows:

Table 4.1. GCC-PHAT results — $T_{60} = 1.4$ s

Signal	Mic 1	Mic 2	DOA	Angle Diff.	Percent Error
SOI	1	14	51.53	7.06	15.89
SOI	6	14	48.37	3.90	8.78
SOI	11	14	46.11	1.65	3.71
Noise	1	14	35.15	12.48	55.03
Noise	6	14	33.10	10.43	46.02
Noise	11	14	27.65	4.98	21.95
Radio	1	14	108.12	12.12	12.62
Radio	6	14	105.07	9.06	9.44
Radio	11	14	101.10	5.09	5.31
Hugo	1	14	NAN	NAN	NAN
Hugo	6	14	119.41	-4.46	-3.60
Hugo	11	14	103.36	-20.51	-16.56

It can be seen from Table 4.1 that, using the pair of microphones (11, 14), the algorithm estimates the true location of the source with a reasonable accuracy. However, it fails to converge for the *Hugo* interference. For the second run of the experiment, where sections of the hallway walls were lined with absorbing foam panels to attenuate the reverberation (T_{60} of 1.1 seconds), the results are:

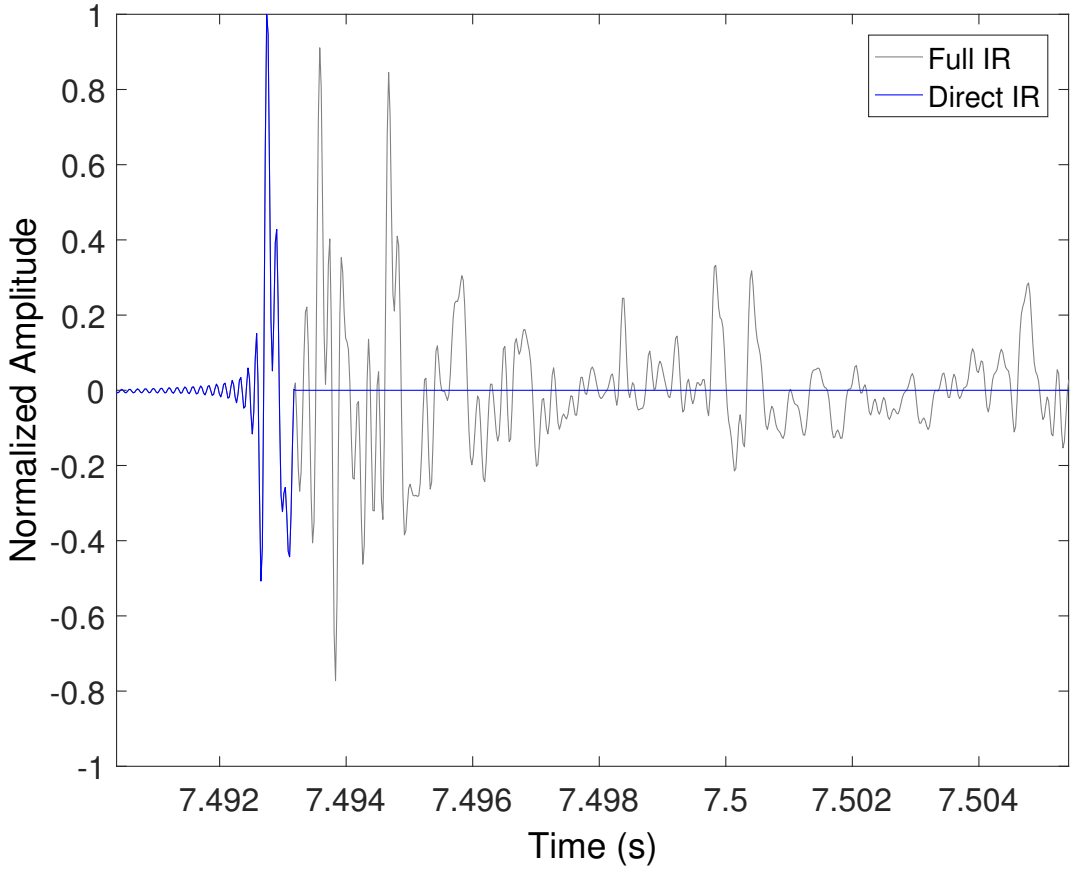
Table 4.2. GCC-PHAT results — $T_{60} = 1.1$ s

Signal	Mic 1	Mic 2	DOA	Angle Diff.	Percent Error
Radio	1	14	103.36	12.12	12.62
Radio	6	14	105.07	9.06	9.44
Radio	11	14	101.10	5.09	5.31
Hugo	1	14	127.18	3.31	2.67
Hugo	6	14	125.29	1.42	1.14
Hugo	11	14	122.63	-1.24	-1.00

4.3 Impulse Response Implementation in MVDR

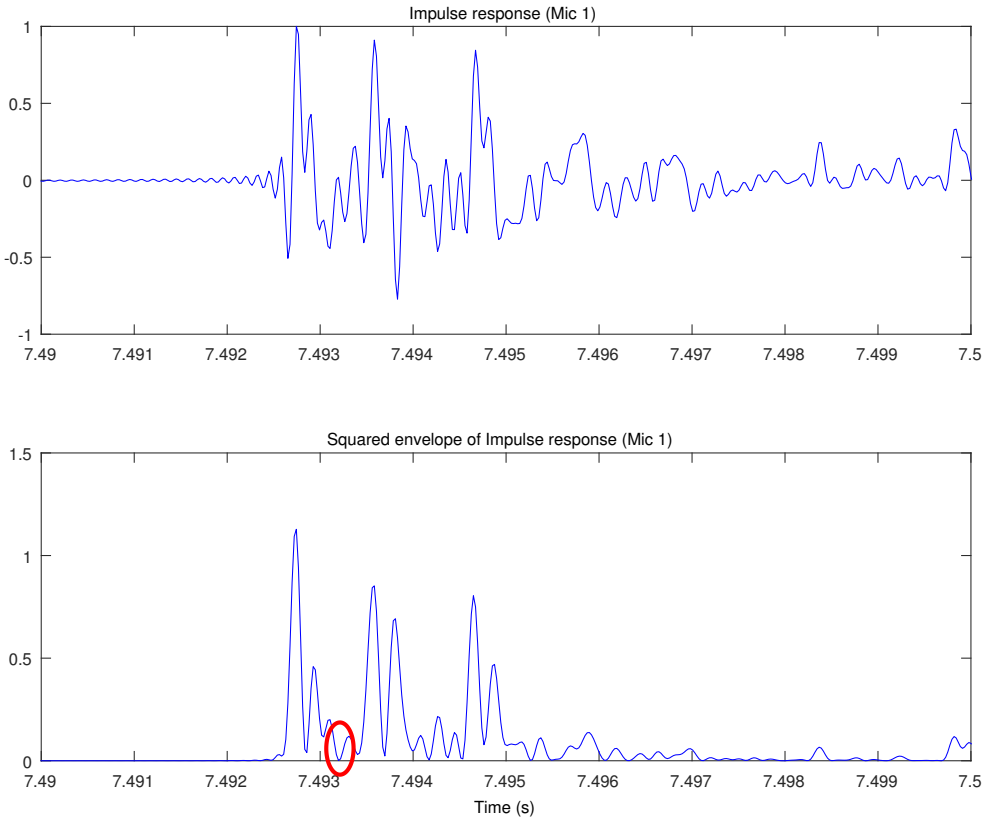
By considering Equation 2.15, we find that the solution to the MVDR constrained optimization problem calls for two usages of the gain factor, Q. The first usage — denoising — requires the full frequency response, G, between the SOI and reference microphone. G is found by taking the FFT of the complete impulse response, seen in grey color in Figure 4.2. The second usage — denoising and dereverberation — requires the direct-path frequency response, G_d, between the SOI and reference microphone. G_d is found by taking the FFT of the direct-path impulse response, shown in blue in Figure 4.2.

Figure 4.2. Full and direct-path impulse response



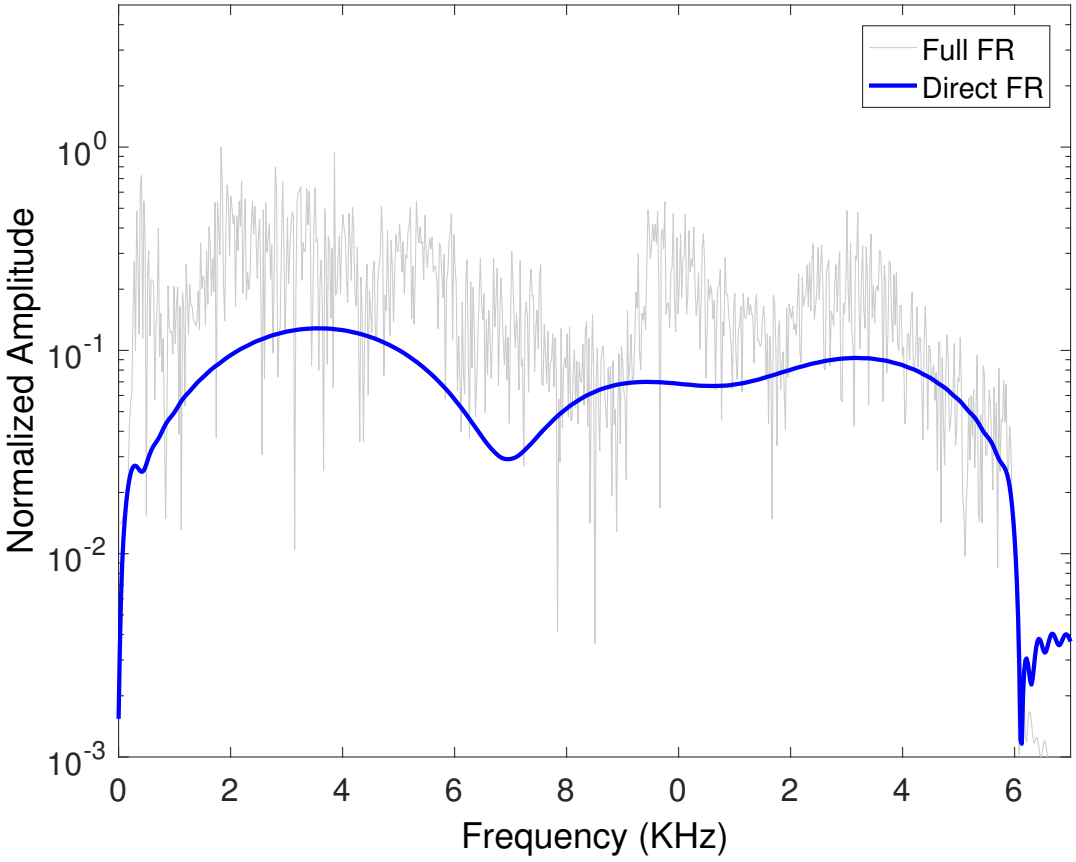
In order to determine the direct-path version of each impulse response, the envelope of the analytic signal was plotted and the cutoff was chosen as the function's first zero. An example analysis is shown in Figure 4.3.

Figure 4.3. Direct-path analysis



This subjective criterion matched the expected direct-path time of flight when ray tracing was applied to the experimental setup. Finally, both the full and direct-path frequency responses of the reference microphone are shown in Figure 4.4.

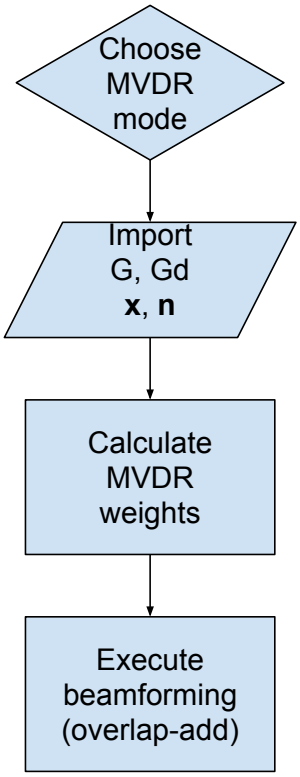
Figure 4.4. Full and direct-path frequency response



4.4 MVDR Algorithm

The modified MVDR solution presented by Habets et al. [12] was implemented in Matlab according to the process shown in Figure 4.5

Figure 4.5. MVDR algorithm process



Initially, the user is asked to choose whether denoising or a combination of denoising and derverberation is desired. Then, the full and direct-path frequency responses from the SOI to each microphone is loaded, and the experimental data are imported. Next, the program computes the MVDR weights using Equation 2.15. Finally, the beamforming is executed using overlapping frames in the STFT domain.

4.4.1 Computation of the PSD matrices. In order to compute the noise PSD matrix in Equation 2.15, two methods were used and compared. First, the Spectrum Element Method was used: the array noise spectra are estimated using Welch's power spectral density estimate, then the noise PSD matrix at each frequency

is simply the outer product of the array spectra at that frequency. That is,

$$\mathbf{\Phi}(j\omega) = V(j\omega)V^H(j\omega). \tag{4.5}$$

Note that $\mathbf{\Phi}$ is of size M x M, where M is the number of microphones in the array. The second method of computing this matrix is the expectation value method. This method consists of segmenting the data matrix into short-time frames (see Section 2.4 for details), computing the PSD matrix at each frame, and finally averaging all the frames. This can be expressed mathematically as follows:

$$\mathbf{\Phi}(j\omega) = E[V(j\omega)V^H(j\omega)] \tag{4.6}$$

$$= \frac{1}{K}\sum_{i=1}^{K} V_i(j\omega)V_i^H(j\omega), \tag{4.7}$$

where K is the number of frames. Results of both methods are shown in Section 4.5.

4.4.2 MVDR algorithm testing via synthetic inputs. In order to gain a better grasp on the MVDR beamformer's effect on the input signals, we resort to "synthetic" inputs obtained by convolving the pure SOI with the individual impulse responses and adding the individual noise signals, measured by each microphone. This intermediary step allows us to study the effects of varying the reverberation and noise levels on the efficiency of the MVDR algorithm. Additionally, this serves as a validation for our algorithm; since we are using the analytical solution to the constrained optimization problem posed in Equation 2.14, we expect to recover the exact, uncorrupted SOI as the output of our synthetic beamformer. This is indeed what we observe in the results presented below.

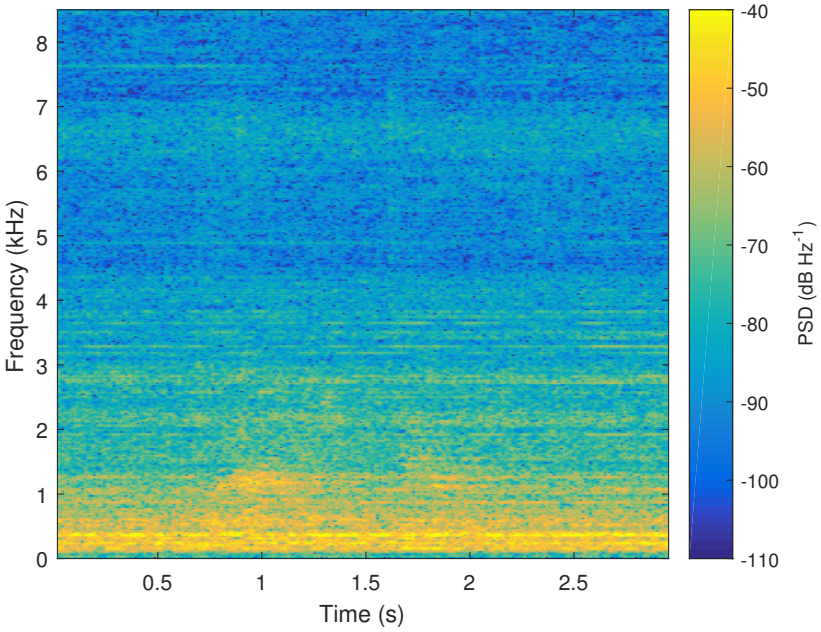
4.4.3 Variations on a synthetic theme. In a real-life situation, since the astronaut will be moving about the spaceship, we may not know the exact impulse response from the speaker to each microphone. Rather, we might average the impulse responses which are pre-measured from various positions. To simulate this situation, we tested the MVDR algorithm with the following variation: instead of propagating the SOI using the impulse responses measured at the location of the SOI, we do it using those measured at either Radio or Hugo. To compute the MVDR weights, we also vary the impulse responses used.

4.5 MVDR Results

Note that in the following sections, the spectrograms are produced using 4096-point, Hanning-windowed frames with an overlap of 75%.

4.5.1 Experimental results — expectation value method. By employing the expectation value method to compute the noise covariance matrices in the MVDR algorithm (see Section 4.4.1), we are able to extract the SOI from the surrounding noise, in reverberant conditions. First, we display the spectrogram of the signal recorded at the reference microphone of the array, which we choose to be the central microphone.

Figure 4.6. Signal spectrogram at reference microphone



Next, we show the spectrogram of the output of the MVDR beamformer for each of its modes: denoising only (Figure 4.7) and denoising and dereverberation (Figure 4.8).

Figure 4.7. MVDR output ($Q = G^{ref}$)

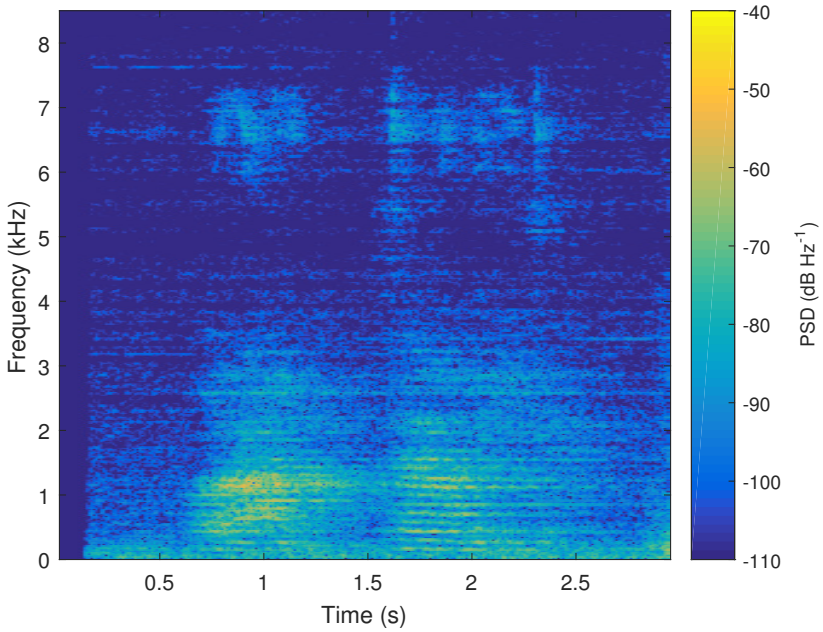
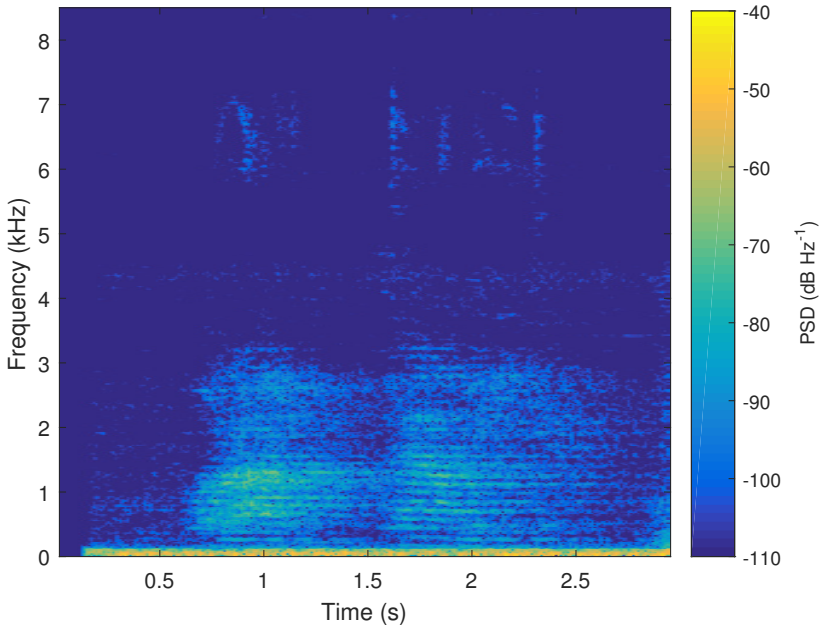


Figure 4.8. MVDR output ($Q = G_d^{ref}$)



Whereas the SOI is buried in noise in Figure 4.6, clear characteristics of speech,

namely vocal formants, can be observed in both spectrograms of the output. This visual representation of the output, along with the auditory inspection of the output corroborate the success of the MVDR algorithm. Furthermore, we quantify this performance with the metrics presented by Habets et al. [12] in Table 4.4. These

Table 4.3. Global performance metrics — expectation value method

	iSNR	oSNR	$\mathscr{A}_g$
$Q = G^{ref}$	0.111	0.235	2.13
$Q = G_d^{ref}$	0.111	2.14	19.4

metrics show that the overall gain provided by the array is 2.13 when the MVDR beamformer operates in the noise reduction mode, and 19.4 in the noise reduction and dereverberation mode.

We also examine the local performance metrics for denoising at 1 kHz and 6.7 kHz.

Table 4.4. Local performance metrics — denoising

f(Hz)	1074	6726
iSNR	0.0012	0.0003
iDNR	5.9468	0.0381
oSNR	0.0191	0.0145
oDNR	0.7136	0.3419

4.5.2 Experimental results — spectral element method. Similarly, using the spectral element method described in Section 4.4.1, we show the results for denoising (Figure 4.9) and denoising and dereverberation (Figure 4.10).

Figure 4.9. MVDR output ($Q = G^{ref}$)

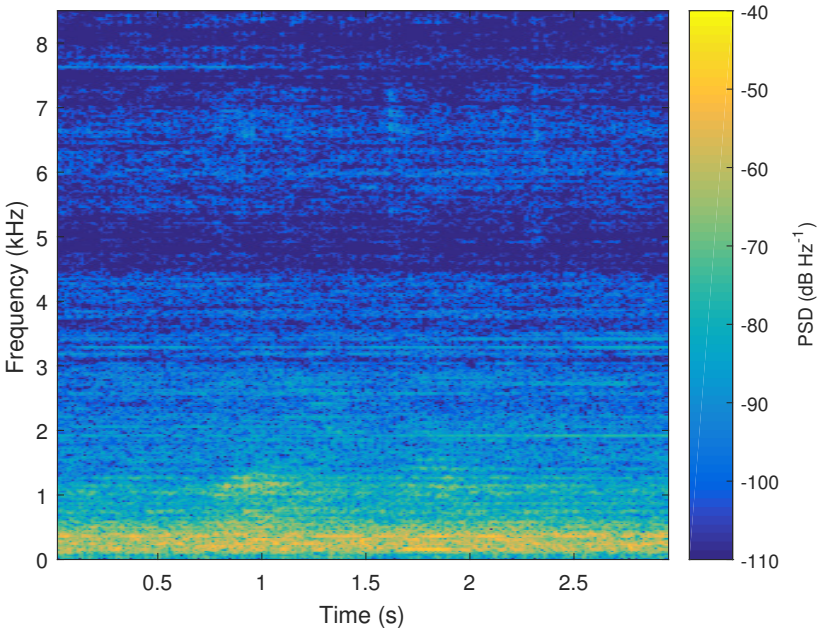
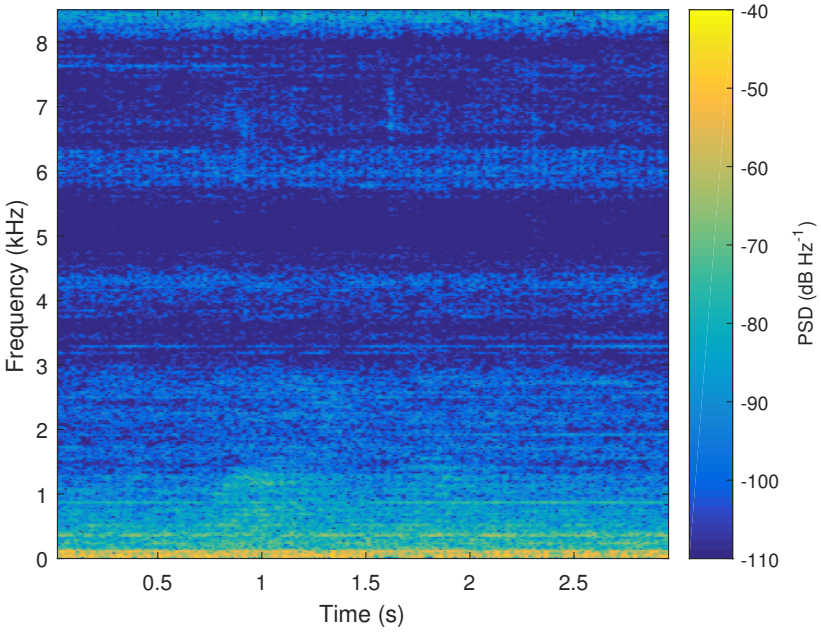


Figure 4.10. MVDR output ($Q = G_d^{ref}$)



This method is not as effective as the expectation value method at extracting

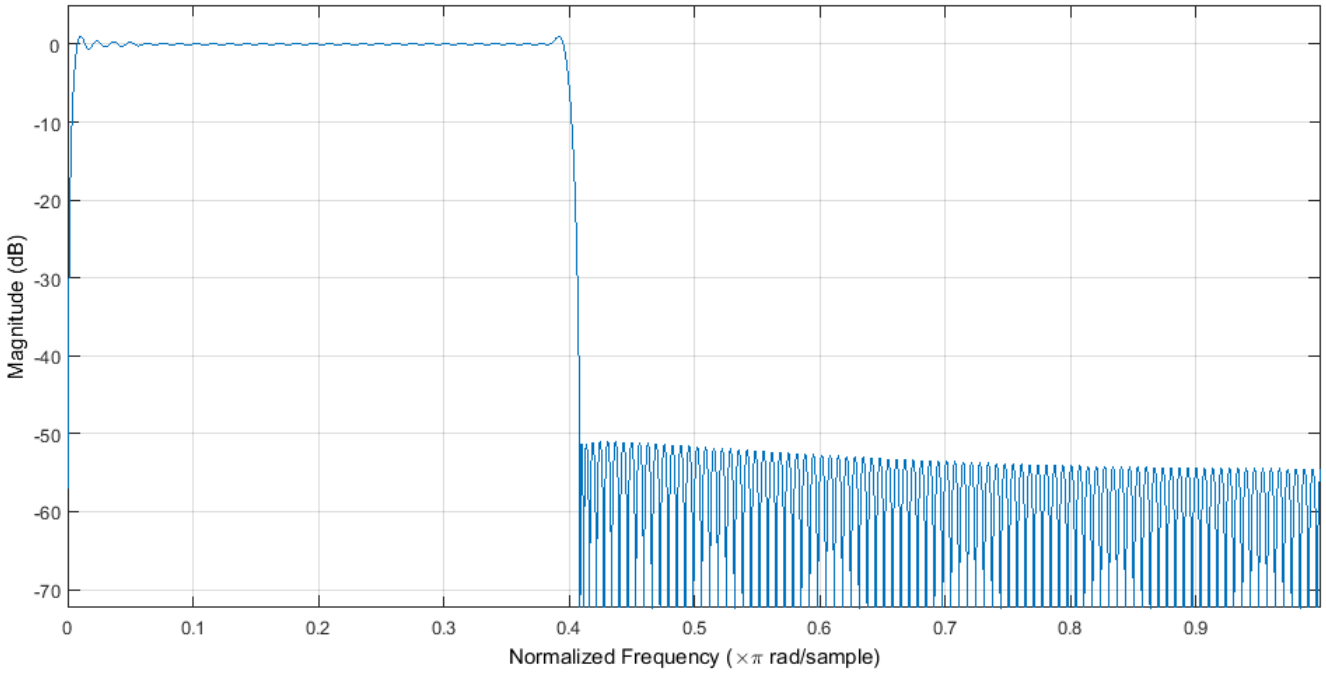
the SOI. One may observe in Figure 4.10 that most of the spectral content of the SOI remains obscured by noise at the output. Metrics for the performance of this method are presented in Table 4.5.

Table 4.5. Performance metrics — spectral element method

	iSNR	oSNR	$\mathscr{A}_g$
$Q = G^{ref}$	0.111	0.0035	0.0319
$Q = G_d^{ref}$	0.111	0.0042	0.0376

4.5.3 Bandpass filtering results. Since our calculated impulse responses are defined over the 100 Hz to 8 kHz range, they are not fully immersed in the experimental data (sampled at 44,100 Hz). Therefore, in order to further optimize the MVDR beamformer, various bandpass filters were designed and applied to the signals collected by the array. Here, we display one such filter: a 100-order least-squares linear-phase FIR bandpass filter with cutoff frequencies of 100 Hz and 8 kHz.

Figure 4.11. Least-squares bandpass FIR filter



The results for incorporating this filter in the beamformer are displayed in the next two figures.

Figure 4.12. MVDR with bandpass filter ($Q = G^{ref}$)

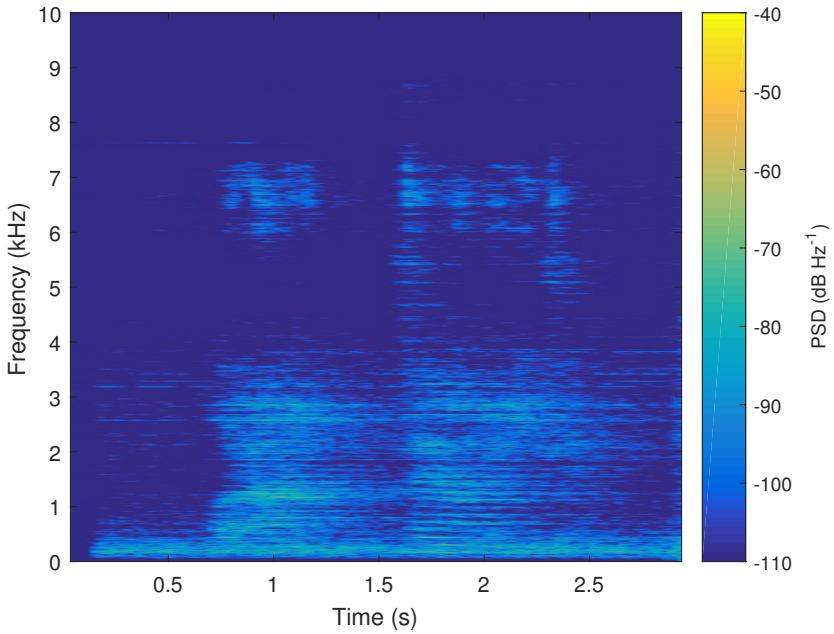
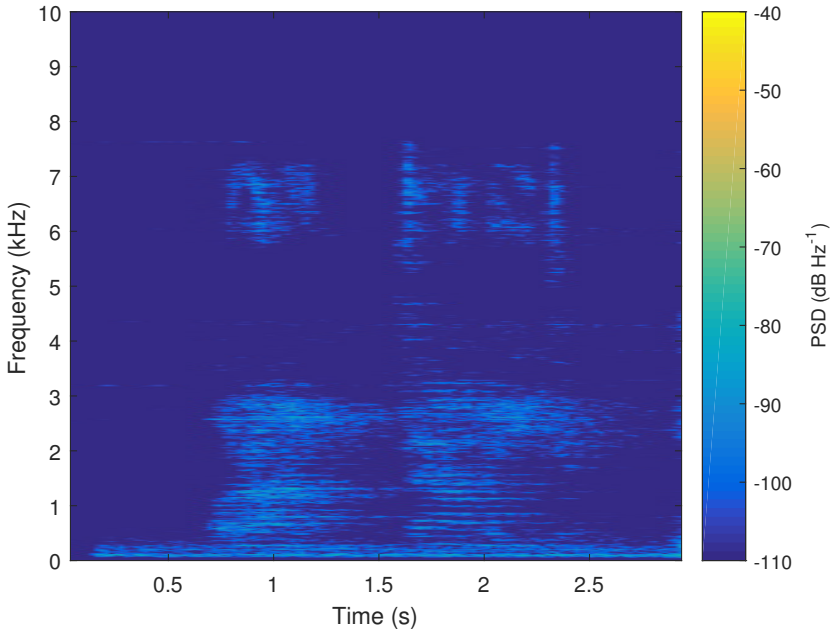


Figure 4.13. MVDR with bandpass filter ($Q = G_d^{ref}$)



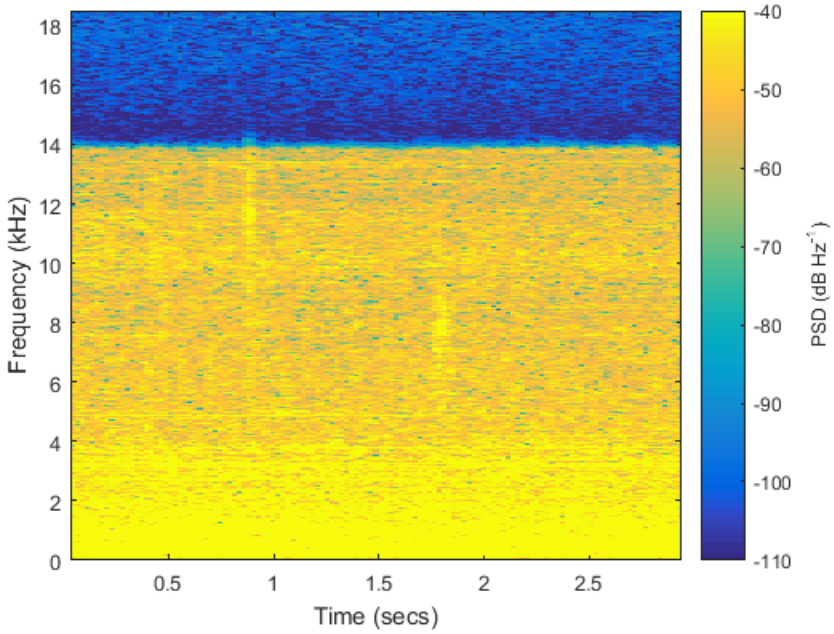
Metrics for these results are shown in Table 4.6

Table 4.6. Performance metrics — MVDR with bandpass filter

	iSNR	oSNR	$\mathscr{A}_g$
$Q = G^{ref}$	0.110	0.240	2.203
$Q = G_d^{ref}$	0.110	0.0451	0.414

4.5.4 Synthetic results. As discussed in Section 4.4.2, the synthetic MVDR program can be used as a validation of the algorithm. Here, we show a sample run of the program with the following recipe as the input to the beamfomer: the SOI and interfering sources convolved with the impulse responses between the radio source and each microphone, combined with the ISS ambient noise. To illustrate the efficacy of the algorithm, the input noise is amplified by a factor of 20. Accordingly, here is the spectrogram of the synthetic input to the MVDR beamformer:

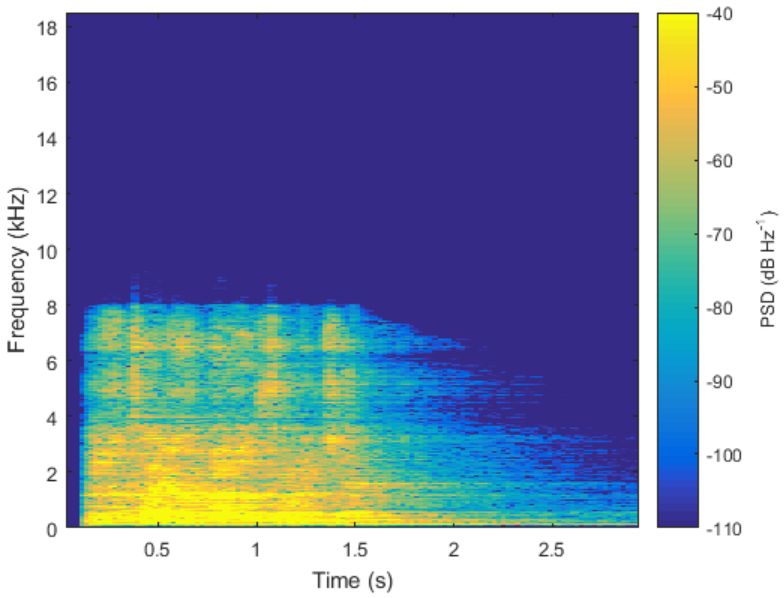
Figure 4.14. MVDR synthetic input



One may note that the SOI is indistinguishable from the noise in Figure 4.14.

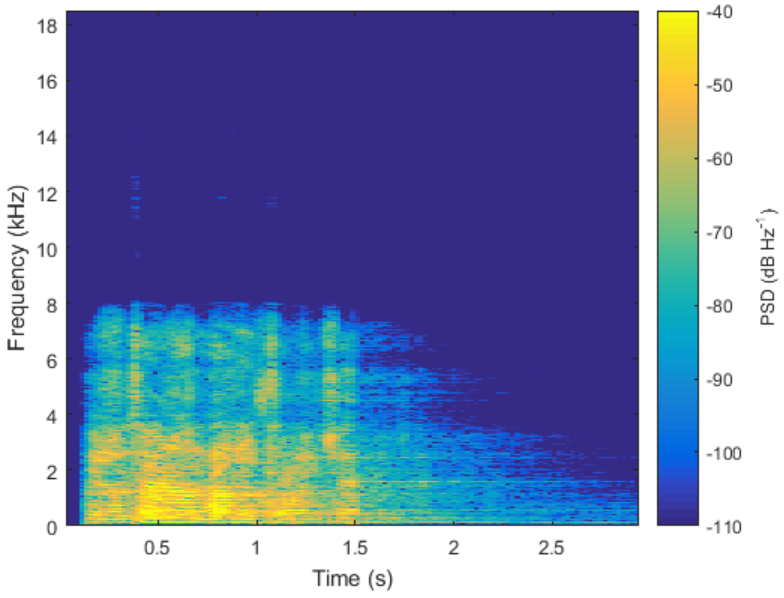
The result of denoising the input is shown next. (Figure 4.15)

Figure 4.15. MVDR synthetic output ($Q = G^{ref}$)



As for the outcome of denoising as well as dereverberating the input, it is shown in Figure 4.16.

Figure 4.16. MVDR synthetic output ($Q = G_d^{ref}$)



The performance metrics for these results are displayed in Table 4.7. Note that

Table 4.7. Performance metrics — Synthetic MVDR

	iSNR	oSNR	$\mathscr{A}_g$
$Q = G^{ref}$	6.529	1.38e11	2.11e10
$Q = G_d^{ref}$	6.529	1.54e8	2.36e7

the exceedingly elevated array gain factors in these results are due to the fact that the noise is almost completely eliminated at the output of the MVDR beamformer.

4.5.5 Variations on a synthetic theme — results. In the case where the SOI is propagated using the impulse responses which were measured at the location of *Radio*, and deconvolved using the impulse responses measured at the location of the SOI, we get the following results:

Figure 4.17. MVDR synthetic variation (Radio/SOI) ($Q = G^{ref}$)

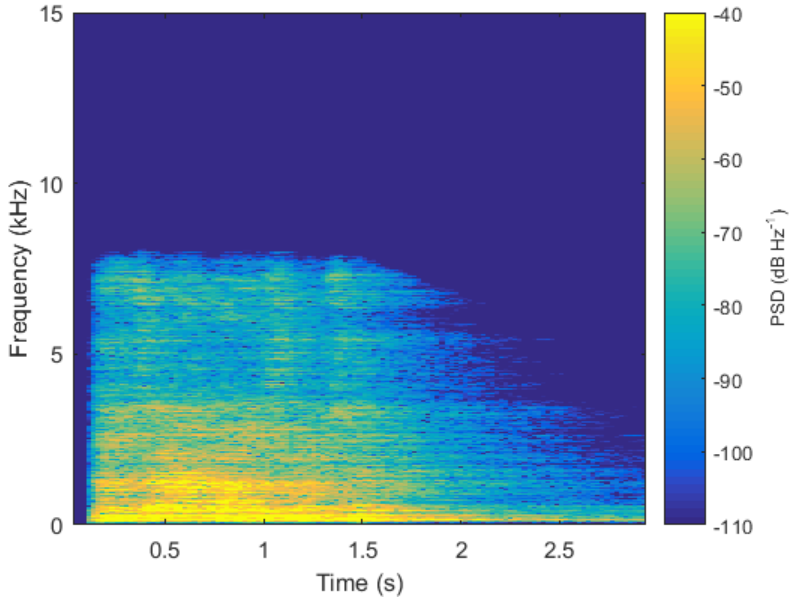
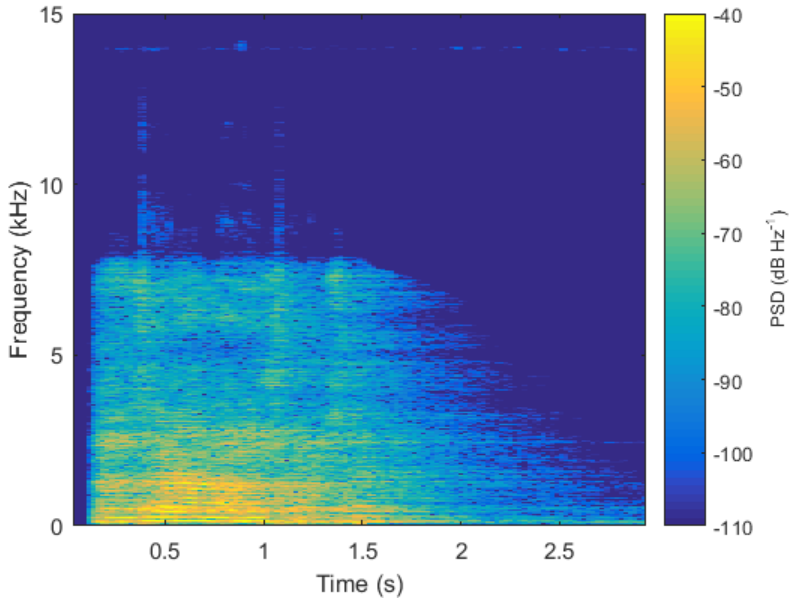


Figure 4.18. MVDR synthetic variation (Radio/SOI) ($Q = G_d^{ref}$)



Finally, the performance metrics for this variation are shown in Table 4.8. Based

Table 4.8. Performance metrics — Synthetic Variation MVDR

	iSNR	oSNR	$\mathscr{A}_g$
$Q = G^{ref}$	5.906	2.26e10	3.82e9
$Q = G_d^{ref}$	5.906	2.19e7	3.70e6

on these results, we find that the MVDR beamformer is an efficient processor for speech enhancement. Its efficiency is quantified by the array gain factor and validated by resorting to synthetic signals.

4.6 Conclusion

In this thesis, we have applied microphone array signal processing to data from an experiment which simulates an astronaut speaking in a spacecraft, in the presence of noise, reverberation and interference. Using MVDR beamforming, we were able to

enhance the astronaut's speech. Furthermore, using the GCC-PHAT algorithm, we localized various sound sources in the experiment.

4.7 Future Work

In the future, we would like to implement a post-filter to the MVDR beamformer in order to further enhance the SNR of the microphone array. In parallel, we would like to solve the constrained optimization imposed by the MVDR beamformer (Equation 2.14) using a machine learning algorithm called gradient descent. Instead of computing the closed-form analytical solution to the problem (Equation 2.15), gradient descent iteratively minimizes the cost function, that is the mean squared error between the beamformer output and the ideal output. A comparison between the two methods may provide a computational advantage.

Bibliography

[1] J. Goodman and F. Grosveld, "Acoustics and noise control in space crew compartments," 2015.

[2] J. de Bedout, M. Franchek, R. Bernhard, and L. Mongeau, "Adaptive-passive noise control with self-tuning helmholtz resonators," *Journal of Sound and Vibration*, vol. 202, no. 1, pp. 109 – 123, 1997.

[3] C. Liu, J. Luo, and Y. Lai, "Acoustic metamaterials with broadband and wide-angle impedance matching," *Phys. Rev. Materials*, vol. 2, p. 045201, Apr 2018.

[4] "Ambient noise of the international space station." `https://soundcloud.com/colchrishadfield/space-station-noise`. Accessed: 2017-10-24.

[5] J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*. 2011.

[6] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[7] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University Providence, 2000.

[8] I. McCowan, "Microphone arrays: A tutorial," *Queensland University, Australia*, pp. 1–38, 2001.

[9] M. R. Bai, J.-G. Ih, and J. Benesty, *Acoustic Array Systems: Theory, Implementation, and Application*. John Wiley & Sons, 2013.

[10] N. Ito, *Robust microphone array signal processing against diffuse noise*. PhD thesis, University of Tokyo, 2012.

[11] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 3, pp. 1526–1540, 2004.

[12] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the mvdr beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, 2010.

[13] R. M. Toroghi, Y. Oualil, and D. Klakow, "A multi-stage, multi-channel processing system for overlapping speech separation in a real scenario," in *Speech Communication; 11. ITG Symposium; Proceedings of*, pp. 1–4, VDE, 2014.

[14] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio Engineering Society Convention 122*, Audio Engineering Society, 2007.

[15] A. Novak, L. Simon, F. Kadlec, and P. Lotton, "Nonlinear system identification using exponential swept-sine signal," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 8, pp. 2220–2229, 2010.

Achi, Peter Y. Bachelor of Science, University of Louisiana at Lafayette, Spring 2011; Master of Music, University of Louisiana at Lafayette, Spring 2013; Master of Science, University of Louisiana at Lafayette, Spring 2018.
Major: Physics
Title of Thesis: Speech Enhancement Techniques for Large Space Habitats Using Microphone Arrays
Thesis Director: Andi G. Petculescu
Pages in Thesis: 61; Words in Abstract: 174

Abstract

The astronauts' ability to communicate easily among themselves or with the ship's computer should be a high priority for the success of missions. Long-duration space habitats — whether spaceships or surface bases — will likely be larger than present-day Earth-to-orbit/Moon transfer ships. Hence an efficient approach would be to free the crew members from the relative burden of having to wear headsets throughout the spacecraft. This can be achieved by placing microphone arrays in all crew-accessible parts of the habitat. Processing algorithms would first localize the speaker and then perform speech enhancement. The background "noise" in a spacecraft is typically fan and duct noise (hum, drone), valve opening/closing (click, hiss), pumps, etc. We simulate such interfering sources by a number of loudspeakers broadcasting various sounds: real ISS sounds, a continuous radio stream, and a poem read by one author. To test the concept, we use a linear 30-microphone array driven by a zero-latency professional audio interface. Speaker localization is obtained by time-domain processing. To enhance the speech-to-noise ratio, a frequency-domain minimum-variance approach is used.

Biographical Sketch

Peter Y. Achi grew up in Lebanon, and moved to the United States to continue his education. He earned a B.S. degree in physics in 2011 and a M.M. degree in piano performance in 2013 from the University of Louisiana at Lafayette. He completed his M.S. degree in physics in 2018 at the same university, under the mentorship of Dr. Andi Petculescu.