

Blind Adaptive Dereverberation of Speech Signals Using a Microphone Array

A Thesis
Presented to
The Academic Faculty

by

Tariq Saad Bakir

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Electrical and Computer Engineering
Georgia Institute of Technology
April 2004


Copyright © 2004 by Tariq Saad Bakir

Blind Adaptive Dereverberation of Speech Signals Using a Microphone Array

Approved by:



Professor Russell M. Mersereau, Commit-




Professor Biing-Hwang Juang



Professor Thomas Morley



Professor Ye Li



Professor Monson H. Hayes

Date Approved: 3/29/2004

To my parents.

ACKNOWLEDGEMENTS

I would like to begin by thanking the esteemed committee members for willing to take take on the extra burden of reading and discussing my thesis with me.

I am very lucky to have had the opportunity to know Prof. Li. He has contributed significantly to my understanding of the limitations of blind equalization methods through meetings with him. His insight and comments were most helpful in allowing me to explain my results better.

Prof. Juang has been generously providing me with suggestions and feedback since I first met him. His generosity has even extended to lending and purchasing equipment for the audio lab where I performed many of my experiments. His experience with room acoustics, and speech processing helped me tremendously in setting up my experiments. I am very grateful for his invitation to me to join his group meetings where I learned much about some of the cutting edge research going on in speech processing. It is truly an honor for me to have worked with him.

As for Prof. Hayes, I owe much of what I learned about statistical signal processing and adaptive filters to his excellent classes which laid the ground work for my further learning. Prof. Hayes has been part of every major step in my doctoral program; serving as a member in my qualifying exam committee, proposal committee, and thesis reading committee. I am very grateful to all of his suggestions and feedback.

And to Prof. Mersereau, words truly do not convey my appreciation and gratitude for allowing me to work with him. It was his idea to work on the dereverberation problem, an interesting and challenging problem, just like what I requested when I first met during my first week at *Tech*. His guidance and amazing insight helped me overcome many obstacles throughout the period of my research at *Tech*. His meticulous and detailed reading of this thesis, along with all of his corrections and suggestions has allowed me to improve this thesis tremendously. However, I am most grateful for his patience, support and encouragement

of my work over these many years. He has given me freedom to pursue what I felt was the best approach to the problem, even though it seemed like a dead end for a while. Prof. Mersereau's care and concern did not extend to me only, but also to my family during some tumultuous times. To that, my entire family is grateful.

I would also like to thank all the CSIP staff, especially Kay, Charlotte, and Christy for all their help during my years at *Tech*, and to Ms. Marilou Mycko for her help and advise in planning my doctoral program.

I am very thankful to my fellow students at CSIP for the interaction I had with them, and to the many signal processing topics I got acquainted with from their presentations and discussions.

Finally, this thesis is as much of my effort as it is to my family. To my father, I appreciate your advice and support of me, you have kept me on track. To my mother, this thesis is the result of your motivation and support. To my brother Muhannad, I am reminded of the tough times we went through in our graduate work, but I am glad we went through them together. It made it a bit easier. And to my youngest brother Basil, thanks to all your visits to us in Atlanta.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
LIST OF FIGURES	ix
SUMMARY	xiii
I INTRODUCTION	1
1.1 Signal Multipath Propagation	1
1.2 Teleconferencing Requirements	2
1.3 Inversion Based Dereverberation	3
1.4 Applications of Speech Dereverberation	5
1.4.1 Blind Source Separation	5
1.4.2 Speech Recognition	5
1.4.3 Audio Recording	5
1.5 Thesis Scope and Organization	6
II PROBLEM BACKGROUND AND SURVEY	7
2.1 Reverberation In a Closed Room	7
2.2 Related Speech Enhancement Problems	10
2.2.1 Acoustical Echo Cancelation	10
2.2.2 Interference Cancelation	12
2.2.3 Blind Source Separation	13
2.3 Previous Dereverberation Methods	14
2.3.1 Cepstrum-Based Methods	15
2.3.2 Beamforming microphone array methods	17
2.3.3 Model-based methods	18
2.3.4 Channel Inversion Methods	19
III ROOM IMPULSE RESPONSE SIMULATION AND MEASUREMENT	30
3.1 The Physical Modeling of Reverberation	30
3.2 Room Impulse Response Simulation	32
3.2.1 Image Method Simulation Results	36

3.3	Room Impulse Response Measurements	40
3.4	Impulse Response Measurements in Room 352	43
3.4.1	Calibration	44
3.4.2	Measurement	44
3.4.3	Reverberation Time Calculations	45
3.5	Summary	47
IV	SPEECH DEREVERBERATION BASED ON THE RMRE METHOD	48
4.1	Multiple FIR Channel Inversion	48
4.2	The Mutually Referenced Equalizers Method	52
4.3	The Reduced MRE Method	58
4.4	Effect of the Number of Microphones	65
4.5	Experiments and Simulation Results	67
4.5.1	Experiment A: Two Nonminimum Phase Synthetic Channels	67
4.5.2	Experiment B: Two Simulated Channels of Order $M = 199$	72
4.5.3	Experiment C: Two Measured Channels of Order $M = 619$	76
4.5.4	Experiment D: Four Simulated Channels of Order $M = 619$	79
4.5.5	Experiment E: Four Measured Channels of Order $M = 619$	82
4.5.6	Experiment F: Eight Simulated Channels of Order $M = 799$	85
4.6	Further investigation into reducing the number of required equalizers . . .	89
4.7	Summary	93
V	RMRE PERFORMANCE IN THE PRESENCE OF MODELING ER- RORS AND MEASUREMENT NOISE	94
5.1	Model Order Determination	94
5.1.1	Under-modeling of the Room Impulse Response	95
5.1.2	Over-modeling of the Room Impulse Response	102
5.2	Impact of Measurement Noise	104
5.2.1	Constrained Norm RMRE	113
5.3	Summary	116
VI	CONCLUSION	118
	REFERENCES	120

VITA	126
-----------------------	------------

LIST OF FIGURES

Figure 1	Multipath phenomena in a room.	8
Figure 2	Linear system formulation of dereverberation.	8
Figure 3	Echo cancelation system setup.	10
Figure 4	Interference cancelation system setup.	12
Figure 5	Interference cancelation with cross talk system setup.	13
Figure 6	Comparison summary of dereverberation methods.	15
Figure 7	GSC structure.	17
Figure 8	Subchannel approach.	26
Figure 9	Simplified impulse response of a room.	32
Figure 10	The image method of modeling reflection of waves.	34
Figure 11	Multiple images in a 2D plane.	34
Figure 12	(a) Decaying exponential, (b) Decaying exponential with random amplitude. 36	
Figure 13	(a) Decaying exponential, (b) Decaying exponential with Rayleigh distributed random amplitude.	37
Figure 14	(a) Impulse response $r_m = [x, y, z] = [8.25, 8.58, 2.5]ft.$, (b) IEDC.	38
Figure 15	(a) Impulse response for $r_m = [x, y, z] = [8.25, 8.33, 2.5]ft.$, (b) IEDC. . .	38
Figure 16	Difference of two impulse responses.	39
Figure 17	(a) Pole-zero plot of two impulse responses, (b) Magnified portion.	39
Figure 18	(a) Impulse response, (b) IEDC.	40
Figure 19	MLS sequence, N=16.	41
Figure 20	Autocorrelation of MLS sequence.	42
Figure 21	Power spectrum of MLS sequence.	42
Figure 22	(a) Arbitrary shaped impulse response, (b) Resulting estimated impulse response using correlation.	43
Figure 23	Calibration module.	44
Figure 24	Measurement module.	45
Figure 25	Reverberation time calculation module.	46
Figure 26	Comparison of simulated and measured impulse responses, \times is the measured impulse response and o is the simulated impulse response.	46

Figure 27	SIMO block diagram showing unknown channels, the input $s(n)$ and equalizers.	52
Figure 28	Rows of the inverse can be thought of as equalizers.	55
Figure 29	(a) Total number of equalizers taps versus number of channels for $M = 500$, (b) $M = 1000$	66
Figure 30	Impulse responses of two synthetic channels.	68
Figure 31	(a) Frequency Response of channel 1, (b) channel 2.	69
Figure 32	(a) Pole-zero plot of channel 1, (b) channel 2.	69
Figure 33	Learning curve based on RLS adaptive filter.	70
Figure 34	Resulting equalizers at various delays where $v_{1,d}, v_{2,d}$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 99, 198$. The first 99 samples in each sub- plot are $v_{1,d}, d = 0, 99, 198$ and the remaining samples belong to $v_{2,d}, d =$ $0, 99, 198$	70
Figure 35	Result of applying the three different equalizer to the channels.	71
Figure 36	Impulse response of the two channels.	72
Figure 37	(a) Frequency response of channel 1, (b) Frequency response of channel 2.	73
Figure 38	(a) Pole zero plot of channel 1, (b) Pole zero plot of channel 2.	73
Figure 39	Learning curve based on RLS adaptive filter.	74
Figure 40	Resulting equalizers at various delays, where $v_{1,d}, v_{2,d}$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 199, 398$. The first 199 samples in each subplot are $v_{1,d}, d = 0, 199, 398$ and the remaining samples belong to $v_{2,d}, d =$ $0, 199, 398$	75
Figure 41	Result of applying equalizers to the channels.	75
Figure 42	Plot of two reverberant channels.	76
Figure 43	Learning curve based on an RLS adaptive filter implementation.	77
Figure 44	Resulting equalizers at various delays, where $v_{1,d}, v_{2,d}$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 620, 1238$. The first 619 samples in each subplot are $v_{1,d}, d = 0, 620, 1238$ and the remaining samples belong to $v_{2,d}, d = 0, 620, 1238$	77
Figure 45	Result of applying equalizers to the channels.	78
Figure 46	Impulse response of four simulated impulse responses.	79
Figure 47	Learning curve based on an RLS adaptive filter implementation.	80
Figure 48	Resulting equalizers at various delays, where $v_{i,d}, i = 1 : 4$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, \dots, v_{4,d}]$, $d = 0, 414, 826$	80
Figure 49	Result of applying the three equalizers to the channels.	81

Figure 50	Plot of four reverberant channels.	82
Figure 51	Learning curve based on an RLS adaptive filter implementation.	83
Figure 52	Resulting equalizers at various delays, where $v_{i,d}, i = 1 : 4$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, \dots, v_{4,d}]$, $d = 0, 414, 826$	83
Figure 53	Result of applying equalizers to the channels.	84
Figure 54	Plot of eight reverberant channels.	85
Figure 55	Learning curve based on an RLS adaptive filter implementation.	86
Figure 56	Resulting equalizers at various delays, where $v_{i,d}, i = 1 : 8$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, \dots, v_{8,d}]$, $d = 0, 458, 914$	86
Figure 57	Result of applying equalizers to the channels.	87
Figure 58	Original and reconstructed speech signals.	88
Figure 59	Difference between original and reconstructed speech signals.	88
Figure 60	(a) E for the case of R_1 , (b) E for the case of R_2	92
Figure 61	Plot of $J^{11} + (J^{10})^T$	92
Figure 62	Original channels.	96
Figure 63	Original channels, tails portion.	96
Figure 64	RMRE under-modeling RLS learning curve.	97
Figure 65	Resulting equalizers, where $v_{i,d}, i = 1 : 2$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 193, 396$	97
Figure 66	Result of applying equalizers to the channels in the under-modeled case.	98
Figure 67	Original channels.	99
Figure 68	RMRE under-modeling RLS learning curve for $L = 20, \delta = 7$	100
Figure 69	Resulting equalizers for $L = 20, \delta = 7$, where $v_{i,d}, i = 1 : 20$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, \dots, v_{20,d}]$, $d = 0, 106, 210$	100
Figure 70	Result of applying equalizers to the channels for $L = 20, \delta = 7$	101
Figure 71	RMRE under-modeling RLS learning curve in the over-modeling case.	102
Figure 72	Resulting equalizers in the over-modeling case, where $v_{1,d}, v_{2,d}$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 220, 438$	103
Figure 73	Result of applying equalizers to the channels in the over-modeling case.	103
Figure 74	SIMO system with measurement noise.	104
Figure 75	Plot of two measured impulse responses.	106
Figure 76	Learning curve using RLS adaptive filter at SNR=98dB.	107

Figure 77	Resulting equalizers at various delays, where $v_{i,d}, i = 1 : 2$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 199, 398$	107
Figure 78	Result of applying equalizers to the channels at $SNR = 98dB$	108
Figure 79	(a) Original and reconstructed speech signals $SNR = 98dB$ utilizing \mathbf{v}_{199} , (b) difference between original and reconstructed signals, $MSE = 3.24 \times 10^{-4}$	109
Figure 80	(a) Original and reconstructed speech signals $SNR = 98dB$ utilizing \mathbf{v}_0 , (b) difference between original and reconstructed signals, $MSE = 0.00156$	109
Figure 81	Learning curve using RLS adaptive filter at $SNR=65dB$	110
Figure 82	Resulting equalizers at various delays, where $v_{i,d}, i = 1 : 2$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 199, 398$	111
Figure 83	Result of applying equalizers to the channels at $SNR = 65dB$	111
Figure 84	(a) Original and reconstructed speech signals $SNR = 65dB$ utilizing \mathbf{v}_{199} , (b) difference between original and reconstructed signals, $MSE = 0.0378$	112
Figure 85	(a) Original and reconstructed speech signals $SNR = 65dB$ utilizing \mathbf{v}_0 , (b) difference between original and reconstructed signals, $MSE = 0.0475$	112
Figure 86	(a) Middle equalizer at an $SNR=45dB$., (b) Edge equalizer \mathbf{v}_{398} at an $SNR=45dB$	114
Figure 87	(a) Original and reconstructed speech signals, (b) difference between original and reconstructed signals at an $SNR=45dB$, $MSE = 1.69 \times 10^{-5}$	115
Figure 88	(a) Original and reconstructed speech signals $SNR = 24dB$, (b) difference between original and reconstructed signals, $MSE = 0.0011$	115
Figure 89	(a) Middle equalizer with no weighting matrix, (b) with weighting matrix T	116
Figure 90	(a) Original and reconstructed speech signals $SNR = 24dB$ with weighting matrix T , (b) difference between original and reconstructed signals, $MSE = 5.39 \times 10^{-4}$	117

SUMMARY

In this thesis, we present a blind adaptive speech dereverberation method based on the use of a reduced mutually referenced equalizers (RMRE) criterion. The method is based on the idea of the inversion of single-input multiple-output FIR linear systems, and as such requires the use of multiple microphones. However, unlike many traditional microphone array methods, there is no need for a specific array configuration or geometry. The RMRE method finds a subset of equalizers for a given delay in a single step, without the need for the typical channel estimation step. This makes the method practical in terms of implementation and avoids the pitfalls of the more complicated two step dereverberation approach, typical in many inversion methods. Additionally, only the second-order statistics of the signals recorded by the microphones are used, without the need for utilizing higher-order statistics information typically needed when the channels have a nonminimum phase response, as is the case with room impulse responses.

We present simulations and experimental results that demonstrate the applicability of the method when the input is speech, and show that in the noiseless case, perfect dereverberation can be achieved. We also evaluate its performance in the presence of noise, and we present a possible way to modify the proposed RMRE to work for very low SNR values. We also explore the problems when model-order mismatches are present, and demonstrate that the under-modeling of the channel impulse responses order can be combated by increasing the number of microphones. For order over-estimation, we will show that RMRE can handle such errors with no modification.

CHAPTER I

INTRODUCTION

1.1 Signal Multipath Propagation

The problem of acoustical reverberation falls under the more general category of *multipath signal* propagation problems. The multipath phenomenon occurs in many engineering problems and emerges regularly in new applications. Some of the areas where multipath propagation occurs are in seismology [1], wireless communication [2] and sonar processing [3]. For example in some sonar applications, a hydrophone array is used to locate underwater objects by listening to the reflection of a transmitted signal. However, the problem of multipath occurs in shallow water when reflections from the sea bed, in the form of clutter, interfere with the signal reflected from the object of interest. Another example is in wireless communications where the terrain and buildings may create multiple reflections of the radio signals from wireless devices. The common thread in all of all these problems is the reflection (and in some cases diffraction) of an emitted signal by various barriers such as walls, water in the atmosphere or foliage. Thus, the receiver, which may take the form of an antenna or microphone array, will receive the original signal plus echoes of that signal. We use the term echoes here in general sense to mean attenuated and delayed versions of the original signal.

The goal in most multipath problems is to recover, or approximate as close as possible, the original signal by eliminating and reducing the echoes. Some problems are more concerned with obtaining information about the environment that caused the multipath, as is the case in underground exploration. What makes multipath problems more difficult compared to other signal interference problems is that the original signal and its echoes have the same statistical characterization. A common formulation of the multipath problem is to think of it as a signal passing through a system (or a channel) that causes the echoes, and the receiver records the resulting system output. If the output is further assumed to be

the result of a convolutional operation, then the problem of multipath interference becomes more specific and is referred to as a *deconvolution* problem or as an *inverse* problem.

Generally, deconvolution problems can be classified by the amount of *a-priori* knowledge that is available. For example in some problems, the emitted signal is known while other problems may have a complete model of the multipath propagation. Some applications have knowledge of the source location that is emitting the original signal, and other problems have complete specification of the receiver configuration. Problems with no knowledge about the specific emitted input signal nor knowledge about the system are called *blind deconvolution* problems. Those with some partial knowledge are called *semi-blind*. While the deconvolution problem, or more generally the multipath problem, is common in many engineering contexts, the solution to it is very dependent on the problem constraints, assumptions made, and permitted solution complexity.

In this thesis, our focus is on the acoustical multipath phenomena that occur in a closed room. More specifically, the problem is referred to as *reverberation* and the solution is called *dereverberation*. The effect of reverberation on speech, recorded by a microphone at some distance from the speaker, has been experienced by almost everyone at some time or another and is usually described as talking into a barrel. The effect is annoying, reduces intelligibility and causes listening fatigue. Thus, a method to eliminate or reduce reverberation in communication systems is a worthwhile objective with immediate benefits.

1.2 Teleconferencing Requirements

The need for high quality teleconferencing solutions (or more generally videoconferencing) is of major benefit and importance to many businesses and organization. However, the adoption of such systems is still in its infancy due to the quality of the communications. Just as email was the killer application that popularized the internet, the emergence of robust high quality teleconferencing can be the application that drives up the popularity of high speed internet connections. Many large companies such as Lucent, Microsoft and ATT are working actively on such systems, and many experts believe that once teleconferencing solutions reach maturity and improve the audio and image quality, they will reduce the need

for travel and the associated expenses. The teleconferencing problem has many aspects that include the acoustical aspect, video and image processing problems, and the issue of network management in multiparty systems. The interest and demand for teleconferencing is also matched by the parallel increased interest in hands free telephony systems such as the ones used in cars.

The speech reverberation problem is one component of the acoustical problem. While many teleconferencing solutions have been proposed, they all suffer from various acoustical degradation problems such as reverberation. In our view, an effective dereverberation system should meet some basic assumptions. For example, since it is not known beforehand what a speaker will say, the dereverberation algorithm must be blind, meaning it can not require prior knowledge of the input signal. The solution must be adaptive to compensate for any changes that may occur in the room, such as a change in speaker location or orientation or the movement of objects. An additional constraint we imposed on the problem is the need for minimal calibration in terms of knowing where the speaker may be located or the geometric configuration of the microphones. This contrasts, for example, with beamforming dereverberation methods that require the speaker location and microphone array configuration to be known. We do assume the availability of at least two microphones that are separated by some distance. These assumptions and constraints, in addition to more specific ones related to our approach, will be discussed in more detail in the next chapter.

1.3 Inversion Based Dereverberation

The problem of speech dereverberation is a well researched problem going back to late 1960's, as for example described in [4]. Over the decades many methods and algorithms have been proposed and developed. However, as explained earlier, the problem of dereverberation depends on the constraints, and assumptions imposed on the problem. For example, the earliest methods were nonadaptive. The problem remains interesting and challenging to this day and continues to be actively researched. There is no single approach that has proved to be universal in its applicability or superior performance in all situations. As suggested earlier, if the reverberation is considered to be the result of a system (a channel), then it is

natural to aim for the design of an inverse system. However, it is well known that acoustic inversion suffers from many problems. This has been the case for a long time, but some important developments have occurred over the last decade. The first of these was the discovering of the multiple input-output inverse theorem (MINT) developed by [5]. They demonstrated that by using multiple microphones, many of the traditional restrictive and unrealistic assumptions that apply to the single microphone case can be removed. The most important of these is the minimum phase condition on the system. The use of microphone arrays is now a common approach to many acoustical problems such as speaker tracking or dereverberation [6].

While MINT provided a way to find the inverse, it assumed that the degrading system was known or could be estimated. This meant the users had to calculate room impulse responses and design the inverses based on these responses. Unfortunately these impulse responses can vary considerable with slight changes in speaker position or head orientation, and thus any change in the speaker location or room impulse responses required remeasuring and redesigning the inverse. So the question that arises is how can the impulse responses be obtained without requiring the user to do the continuous calibration and measurement procedures ? The answer to this question came from another multipath problem, the problem of channel equalization in wireless communications. The pioneering work of Tong *et. al.* [7] demonstrated that when multiple channels are available, it is possible to estimate the channels by using only the second-order statistics of the outputs, under certain channel assumptions. In other words, blind channel identification was possible without resorting to the higher order statistics methods that had dominated blind identification methods up to that point. Second order identification procedures are more robust, less computationally intensive and lend themselves to adaptive implementations.

This combination of multiple channel inversion and blind channel estimation breakthroughs is what motivated us to choose a microphone array inversion based on second order statistics of the received signal as our speech reverberation solution strategy. While inversion methods are generally more computationally demanding than other methods, we

believe this will be less and less of a problem with continuing increases in the computational power of microprocessors and dedicated signal processing processors. We compare and contrast some of the other approaches to dereverberation in the next chapter and show how they differ from the inversion approach.

1.4 *Applications of Speech Dereverberation*

Speech dereverberation is not limited for use only in teleconferencing. We list some of the other major applications of speech of dereverberation with a brief description of each.

1.4.1 Blind Source Separation

The problem of blind source separation is also commonly referred to as the cocktail party problem. The goal here is to separate (unmix) independent signals that have been mixed by convolutional channels. However, in the presence of reverberation, many blind source separation methods may give less than optimal solutions or even fail as discussed in [8]. By using a dereverberation method as a preprocessor for blind source separation, the performance of unmixing can be improved.

1.4.2 Speech Recognition

In speech recognition, the use of dereverberation can help improve the environmental robustness of a speech recognition system. For example, a user may train the speech recognition software in one room with one specific reverberation level, but later use it in another room. By reducing the reverberation, some measure of robustness in recognition is obtained.

1.4.3 Audio Recording

Professional recording requires expensive custom built rooms that reduce reverberation in addition to high quality audio recording equipment. By eliminating the reverberation, the need for custom built rooms is eliminated. Another related recording application is in surveillance where a microphone must be placed at some distance away from the source.

1.5 Thesis Scope and Organization

In this thesis, our focus is on the adaptive dereverberation of acoustical signals in a closed room using a microphone array and utilizing only second-order statistics of the output. Some of the related problems to investigate are the nature of acoustical reverberation in such rooms and how to model it, the modeling and simulation of room impulse responses, and the effect of noise in the recording equipment on the performance of the proposed method. Our focus is on a single speaker talking at a time located some distance away from the microphones. We do not take into account noise sources due to interference such as fans in the room, since many successful methods have been developed to deal with this problem as described in [9].

- In Chapter 2, we give a survey of the major speech dereverberation methods and contrast their advantages and disadvantages. We also discuss some related speech enhancement problems and clarify how they are different from dereverberation.
- In Chapter 3 we investigate the room impulse response, its statical characterizations and methods for measuring the room impulse response.
- In Chapter 4 we discuss our proposed dereverberation approach based reduced MRE (RMRE) method. Experimental results are provided to demonstrate the applicability of the RMRE approach to the dereverberation problem.
- Chapter 5 discusses the problems of over and under-modeling. Also the impact of measurement noise is evaluated on the proposed method.
- Chapter 6 is a conclusion, with some ideas for further research.

CHAPTER II

PROBLEM BACKGROUND AND SURVEY

In the previous chapter, some of the highlights of the reverberation problem were described. In this chapter, we expand on the problem formulation, problem assumptions and the constraints imposed on our problem solution. Additionally, we discuss the difference between dereverberation and several related speech enhancement problems. Finally, we provide a survey of the major dereverberation approaches found in the literature and compare their advantages and disadvantages.

2.1 Reverberation In a Closed Room

Consider a speaker located some distance away from a microphone in a closed room as shown in Figure 1. Since sound waves travel radially outward from a source, reflections from the walls, floor and furniture will be recorded by the microphone some time after the recording of the direct path signal, a phenomenon called reverberation. This reverberation leads to the acoustical degradation of the transmitted speech signal and causes listening fatigue.

One possible approach to modeling the reverberation phenomenon is by using partial differential equations that describe the propagation and reflection of sound in air. Such a physical and complex approach may be necessary in situations that require the design of unique enclosures and rooms such as concert halls. For most other applications, a much simpler and more effective approach is to treat reverberation as a lumped phenomenon that can be described by a single linear system. With this approach, the reverberation can be completely modeled by the impulse response of the room $h(n)$, also called the acoustical transfer function (ATF). Using this formulation, the reverberation and dereverberation systems can be cast in the standard linear convolutional system formulation shown

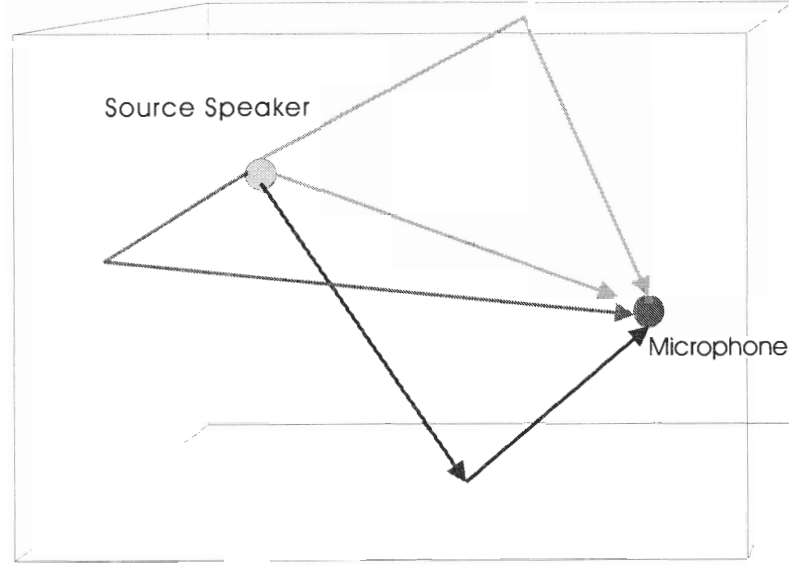


Figure 1: Multipath phenomena in a room.

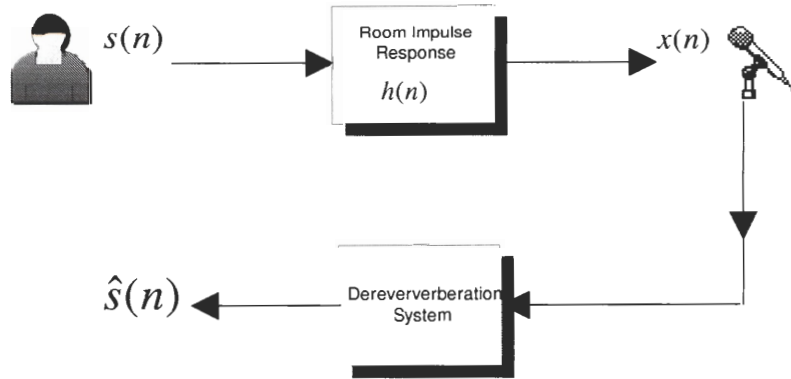


Figure 2: Linear system formulation of dereverberation.

in Figure 2, where $s(n)$ is the clean speech signal (unreverberated) and $x(n)$ is the reverberated speech signal recorded by the microphone and $\hat{s}(n)$ is the dereverberated speech. The primary goal of dereverberation is to undo this convolutional effect, i.e. to perform *deconvolution* on the received reverberated speech signal. The term *deconvolution* has been traditionally used to describe problems that are nonadaptive in nature and is widely used in the signal processing community. The term *equalization* is more frequently used when the deconvolution is preformed in an adaptive setting such as in communication systems and is more popular in the telecommunications community. Since we are considering adaptive deconvolution methods, the two terms mean the same thing for our purposes and we use

them interchangeably. In Chapter 1, we alluded to some key engineering design issues that define this research and differentiate it from other dereverberation methods that we survey in later sections. We list the most important issues next:

1. We assume the availability of multiple microphones, i.e. a microphone array. This is true in many existing teleconferencing systems especially ones that have a camera tracking component as in [6]. By employing multiple microphones, there are several approaches for performing deconvolution, as will be described in the survey of dereverberation methods. The use of multiple microphones adds only a small increase to the overall system cost.
2. The deconvolution must be blind. This is because neither the room's acoustical impulse response nor the input (the clean speech) signals are known *a-priori*. The room's impulse response is assumed to be unknown because the speaker may move from one position to another in the room or the room environment may change (doors opening or closing, chairs moving and so on). By imposing no constraint on knowing the room's impulse response, the proposed system becomes independent of the room geometry, and requires no calibration by the user. The input (speech) signal is assumed to be unknown simply because we cannot predict what users will say. Additionally, speech signals are non-white and non-stationary in nature.
3. In keeping with the minimal calibration theme, no assumptions are made about the microphones locations or the array pattern they form. This means the user can simply place the microphones anywhere in the room at some distance from each other.
4. The proposed method must be adaptive in nature as opposed to a batch processing method. The reason for this is that system must cope with the changes in the room environment such as users leaving or changes in speaker location.
5. Finally, we make the minimal number of assumptions about the nature of speech signals. This means that no need exists to segment speech into segments of various pitch and periodicity. This helps in making the method applicable to other audio

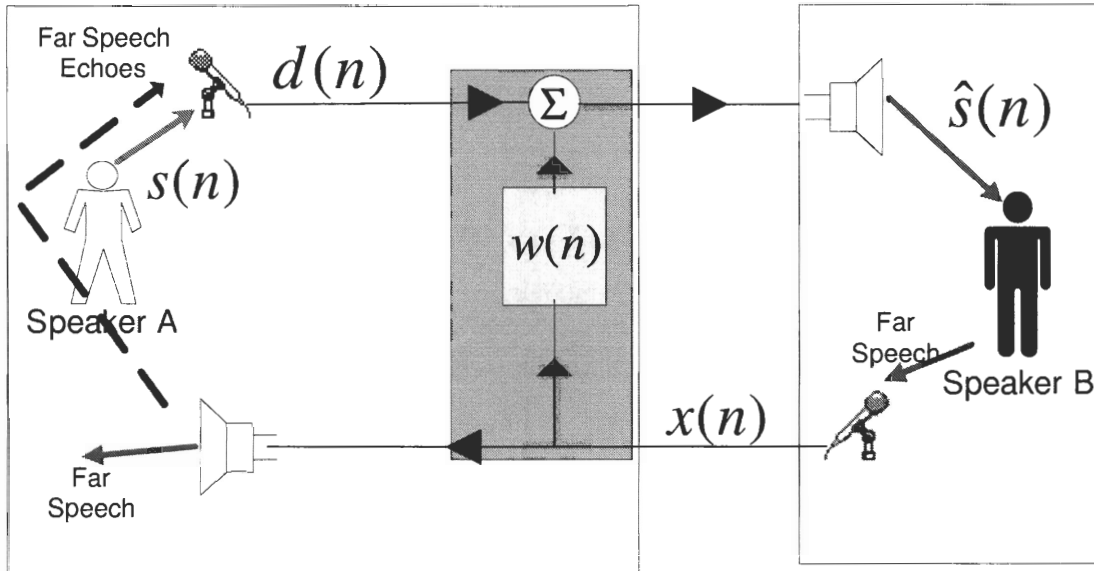


Figure 3: Echo cancellation system setup.

signals such as music and makes it less demanding in terms of human interaction.

2.2 *Related Speech Enhancement Problems*

A number of important and related speech enhancement problems that sometimes get lumped with with the dereverberation problem are noise reduction, echo cancellation and the cocktail party problems. These problems, while related to the general goal of speech enhancement, differ greatly in their basic problem formulation, assumptions and solution approach. In this section we discuss three related problems: the acoustical echo cancellation problem, interference reduction and the cocktail party problem and show why they differ from and must not be confused with the goal of dereverberation.

2.2.1 *Acoustical Echo Cancellation*

The acoustical echo cancellation problem is common in hands free full-duplex telephony systems. The basic cause of the problem is the recording by a microphone the signal emanating from a loudspeaker in a room. To fully understand the echo cause, consider two speakers in two different locations as depicted in Figure 3.

One speaker is located in what is called the near end (speaker A) and the other is in the

far end (speaker B). As speaker A talks into the microphone (the near end speech denoted by $s(n)$), the loudspeaker in room A is broadcasting the speech from the far end. The signal from the loudspeaker gets picked up by the microphone in room A in the form of echoes, and gets transmitted to the user in room B. The purpose of the acoustical canceler function is to reduce this echo [10]. Acoustic cancelers are implemented as adaptive systems and their objective is to approximate the impulse response $h(n)$ between the microphone and loudspeaker in room A. If we denote the far end speech as the signal $x(n)$ and the signal recorded by the microphone as $d(n) = s(n) + x(n) * h(n)$, then the goal of the adaptive filter $w(n)$ would be to minimize the error signal

$$e(n) = d(n) - w(n) * x(n) = s(n) + h(n) * x(n) - w(n) * x(n).$$

The minimum error occurs when the adaptive filter converges to the impulse response $h(n)$. Part of the difficulty of the problem is that the impulse responses can be very long and some nonlinear effects can occur in the loudspeaker system. Much of the research into echo cancellation has focused on ways of modeling the impulse response and developing efficient adaptive structures to speed up the convergence and reduce the computational complexity. Another major problem is the need to detect speech or silence in the communication.

A closely related echo cancellation problem is the one that occurs in telephone networks due to the electric hybrids. The cause of the echoes in this problem is the impedance mismatch in these hybrids which causes the signals to be reflected along the telephone lines. This type of problem is easier than the acoustical echo cancellation problem because the impulse response that models the echo path is much simpler and shorter than the ones associated with room impulse responses.

It is important to highlight the difference between the reverberation and echo cancellation problems since they appear the same at first glance. The main difference lies in the availability of the far speech signal $x(n)$ in acoustical echo cancellation. This means that an error signal can be generated as described above and the problem takes the form of a system identification problem, i.e. that of identifying the impulse response $h(n)$. In the reverberation problem, all that is available is the signal recorded by the microphone. Neither

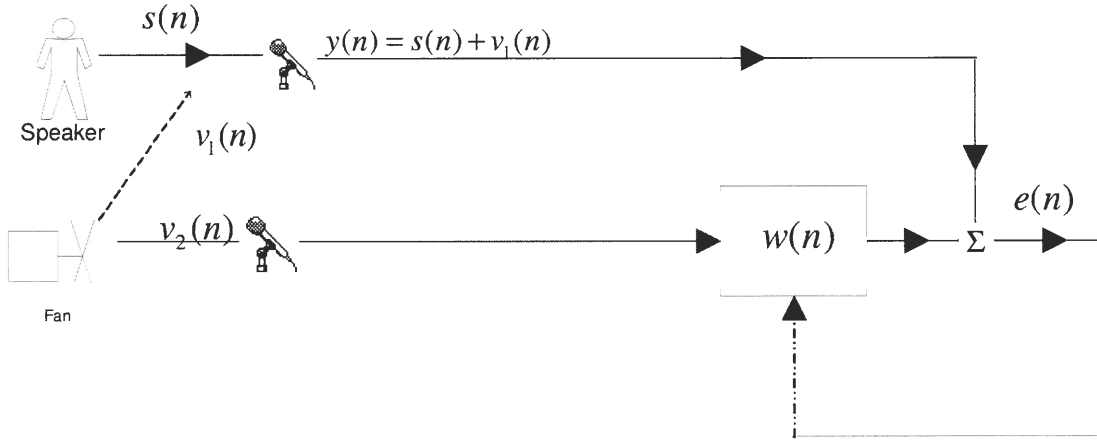


Figure 4: Interference cancellation system setup.

the room impulse response nor the input is known.

2.2.2 Interference Cancellation

The interference cancellation problem is the result of a noise source, such as a fan or ventilation duct, that interferes with the speech recorded by a microphone. The most common way to deal with such a problem is by the use of *reference* microphone that is placed near to the noise source. The *primary* microphone is placed near the talker, as illustrated in Figure 4.

The goal of the adaptive filter $w(n)$ is to get $v_2(n)$ to be as close as possible to $v_1(n)$. The error signal in this case is given by $e(n) = s(n) + v_1(n) - w(n) * v_2(n)$. A more complicated problem is when cross talk occurs, as discussed in [9], between the reference and primary microphones. One way of dealing with this problem is to use a second adaptive filter in the error signal feedback loop as illustrated in Figure 5. The primary microphone in this case records $y_1(n) = s_1(n) + v_1(n)$ and the reference microphone records $y_2(n) = s_2(n) + v_2(n)$. In most cases the amplitude of $s_2(n)$ is much less than that of $s_1(n)$. This means that $s_2(n)$ acts as interference in the reference signal $y_2(n)$. The function of the additional filter $w_2(n)$ is to approximate $s_2(n)$ as accurately as possible by using the signal $e_1(n) = s_1(n)$, thus removing its effect from $y_2(n)$. The error signal $e_2(n)$ will equal $v_2(n)$ if the adaptive filter $w_2(n)$ is effective, and it can be used as the reference signal for $w_1(n)$ as in the simpler no

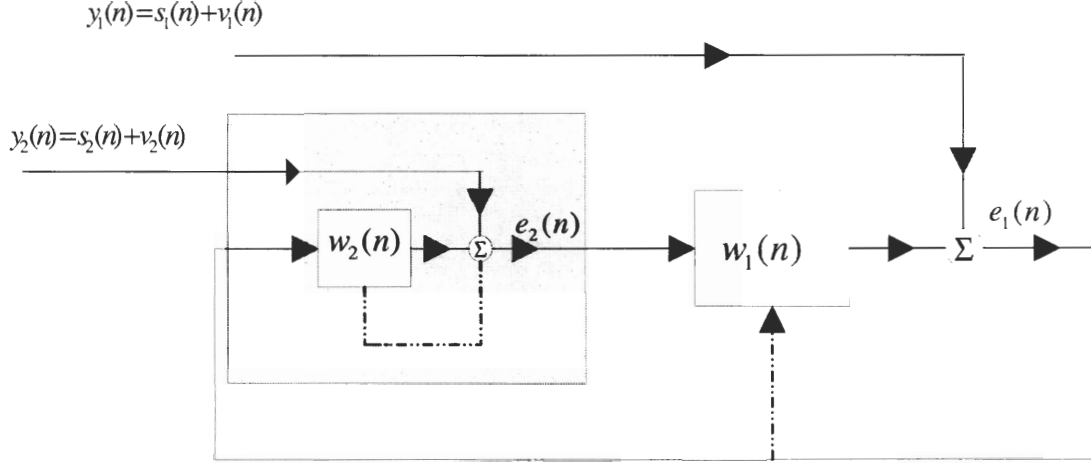


Figure 5: Interference cancellation with cross talk system setup.

cross-talk case discussed earlier.

It is also possible to eliminate the need for a reference channel if the speech periodicity is utilized as detailed in [9], but these methods require pitch extraction and other speech specific processing techniques.

2.2.3 Blind Source Separation

The cocktail party problem, or more formally the blind source separation problem, is one of the most interesting new signal processing problems and has applicability to biomedical, geophysical, image and speech processing. The problem is usually described in terms of a listener in a party trying to concentrate on single distant speaker while hearing the chatter of other speakers. The listener hears a mixture of the speakers in the party and the goal is to undo this mixing and listen to only one speaker. The problem mathematically can be formulated as a multichannel linear convolutional system given by the equation

$$\mathbf{x}(n) = \mathbf{H}\mathbf{s}(n),$$

where $\mathbf{x}(n)$ is the signal recorded by a group of sensors, \mathbf{H} is the unknown channel matrix and $\mathbf{s}(n)$ is a vector of source signals that is also unknown. The problem is a blind since both \mathbf{H} and $\mathbf{s}(n)$ are unknown. Generally, it is required that the number of sensors be equal to or greater than the number of sources, but this is not always necessary. The design

objective is to find a matrix \mathbf{W} that will unmix the effect of \mathbf{H} . To illustrate how this can be done, we consider the simple case of a symmetric and nonsingular \mathbf{H} . We also assume that the sources $\mathbf{s}(n)$ are spatially uncorrelated and that the correlation is given by

$$E[\mathbf{s}(n)\mathbf{s}^T(n)] = \mathbf{I}.$$

The sensors' correlation matrix is then given by

$$E[\mathbf{x}(n)\mathbf{x}^T(n)] = \mathbf{R}_\mathbf{x} = \mathbf{H}\mathbf{R}_\mathbf{s}\mathbf{H}^T$$

which simplifies to $\mathbf{R}_\mathbf{x} = \mathbf{H}^2$ by invoking the symmetry of \mathbf{H} and the unit variance of the sources. Applying an eigenvalue decomposition to the sensor matrix yields

$$\mathbf{R}_\mathbf{x} = \mathbf{V}_\mathbf{x}\mathbf{D}_\mathbf{x}\mathbf{V}_\mathbf{x}^T,$$

which implies that the demixing matrix is given by

$$\mathbf{W} = \mathbf{R}_\mathbf{x}^{-1/2} = \mathbf{V}_\mathbf{x}\mathbf{D}_\mathbf{x}^{-1/2}\mathbf{V}_\mathbf{x}^T.$$

More complicated structures for the mixing matrices and correlated sources naturally requires more sophisticated approaches as described in [8].

It is worthwhile to note that reverberant environments degrade the performance of source separation problems. The reason for this is that blind many source separation methods require that the number of microphones and sensors be known, and the effect of reverberation is to create *spurious* sources. An insightful relation between blind source separation and adaptive beamforming that further elaborates on the problem of reverberation can be found in [11].

2.3 *Previous Dereverberation Methods*

In this section we discuss some of the more popular approaches to dereverberation and discuss the advantages and disadvantages associated with each method. In general there are four broad approaches to accomplish dereverberation, which are

1. Cepstral subtraction based methods.

Dereverberation Method	Advantages	Disadvantages
Cepstral subtraction	Conceptually simple, well researched	Artifacts in enhanced speech, nonadaptive
Beamforming	Robust, adaptive	Requires speaker location, requires array calibration
Model based methods	Good dereverberation results, can be adaptive	Require segmentation, work only for speech
Inversion methods	Most accurate theoretically, multisensor methods make no restrictive assumptions, adaptive in nature	Computationally demanding

Figure 6: Comparison summary of dereverberation methods.

2. Microphone array beamforming methods.
3. Model-based methods.
4. Room impulse response inversion methods. These can be either single microphone or microphone array based.

All of these methods continue to be active areas of research and are used in various applications depending on the exact nature of the dereverberation problem and the system requirements. Figure 6 gives a summary of the pros and cons of each approach.

Our proposed method falls under the fourth category. We discuss some of the basics of each approach, but our focus and most detailed exposition will be limited to the room impulse response inversion methods.

2.3.1 Cepstrum-Based Methods

The cepstrum-based approach is one of the earliest approaches to dereverberation, dating back to the early beginnings of digital signal processing in the late 1960's [4]. Cepstrum-based methods are based on homomorphic filtering theory as explained in [9, 12]. The cepstrum of a signal $s(n)$ is defined as

$$c(n) = 1/2\pi \int_{-\pi}^{\pi} \log |S(e^{jwn})| e^{jwn} dw,$$

where $S(e^{jwn})$ is the Fourier transform of the signal $s(n)$. The nonlinear log function converts the convolutional channel model into an additive model. This can be clearly seen

by noting that

$$y(n) = s(n) * h(n) \xleftrightarrow{FT} Y(w) = S(w)H(w),$$

and by taking the log of the FT domain expression (for simplicity, the phase arguments are ignored) we obtain

$$\log |Y(w)| = \log |S(w)| + \log |H(w)|.$$

Thus, the convolutional nature of the channel is transformed into an additive effect which is more easily compensated for.

Cepstrum-based methods can be either single or multi-microphone in nature. For example in a single microphone approach such as the one described in [13], a procedure was developed to segment the reverberated speech signal. Using these speech segments, an averaging procedure is used to estimate the cepstrum. It turns out that the cepstral components caused by the room impulse response manifest themselves in the form of cepstral peaks. Identifying these peaks and removing them from the cepstrum reduces the reverberation.

Recent cepstrum-based methods use multiple microphones. For example in [14, 15], a two microphone method was used to perform dereverberation of speech signals. This method uses cepstral methods to estimate the phase of the channels which are modeled as an all-pass component and a minimum phase component. Under some channel constraints, such as when assuming FIR channels and no channel zeros on the unit circle, it is possible to find the two components by using phase information only. Cepstral methods have improved in their ability to accomplish dereverberation, but they continue to suffer from major drawbacks. They are rarely able to completely remove room reverberation and they are computationally expensive. This limits their applicability to be mainly used as post processing methods rather than real-time adaptive methods. In addition, audible artifacts can usually be heard in the restored signal due to the differences in the spectral subtraction from one frame of processed speech to another. The assumption of no zeros on the unit circle is also questionable for long acoustical reverberation channels, as we will show in Chapter 3.

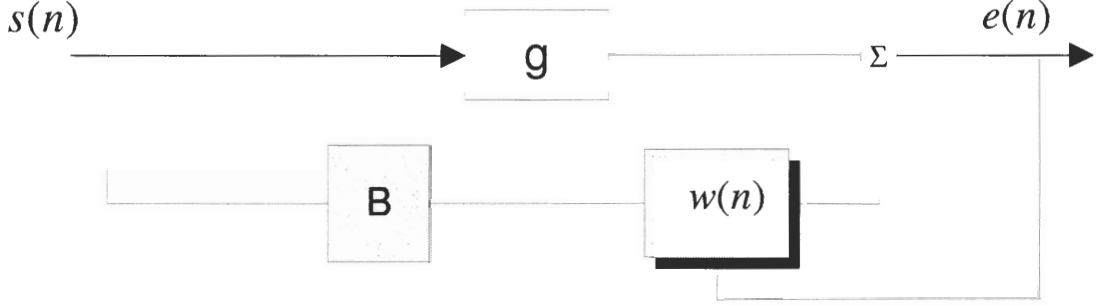


Figure 7: GSC structure.

2.3.2 Beamforming microphone array methods

Beamforming methods are simply spatially selective filters. They can be applied to the speech dereverberation problem by steering a microphone array to a speaker in a known location. A microphone array with proper steering can then attenuate any signal coming from a direction other than the one to which the microphone array is steered. A common type of beamformer that is used to suppress interference is the generalized sidelobe canceler (GSC). The GSC is used when the source location is known. The GSC is composed of a nonadaptive filter portion and an adaptive filter portion as shown in Figure 7. The nonadaptive filter g is steered in the direction of the signal $s(n)$. The adaptive portion is constructed as the cascade of a blocking matrix B and an adaptive filter $w(n)$. The purpose of the blocking matrix B is to stop the desired signal $s(n)$ from feeding into the adaptive portion of the system. Thus, the adaptive filter $w(n)$ will try to match the interference in the adaptive branch to as close as possible to the interference in the nonadaptive branch. This setup is very reminiscent of interference cancellation described earlier. The optimal filter $w(n)$ is found by minimizing the energy of the error signal

$$e(n) = g^T s(n) - Bw^T(n)s(n).$$

The blocking matrix B is usually not unique, and is usually dependent on the problem formulation. Thus, with the GSC-based dereverberation approach, if we consider reverberation to be the interference coming from directions other than the direct path and assume that the speaker's location is known, we can reduce the amount of reverberation.

The GSC and other adaptive beamforming variants are described in [16, 17, 18, 19, 20, 21, 22, 23]. For example, in [23], in addition to using a classic delay-and-sum beamformer, the author adds a post-processing Wiener filter to reduce the effect of noise. In [17] the authors use a subspace approach to minimize the reverberation in a bank teller booth environment by building a microphone array into the frame of the window. In a more recent approach by [24], a frequency-domain approach is used for computational efficiency.

Many of the beamforming subspace methods described have the advantage of being able to handle noise in addition to reducing the reverberation. This makes such methods very applicable for use in noisy environments such the inside of a car as described in [25].

The biggest advantage of the beamforming methods are their simple implementation and robustness with respect to noisy settings. However, as stated earlier, these methods require that the position of the speaker be known in advance and require a pre-designed optimal array geometry.

2.3.3 Model-based methods

One of the more recent approaches to dereverberation can be described as model-based enhancement. These methods utilize the fact that reverberant speech signals possess different statistical properties than clean speech. For example, some methods exploit pitch periodicity as in [26], but these methods are best at handling noisy speech and not reverberation. One of the newest methods is that described in [27] and has sparked the interest of many other researchers. This method is based on the analysis of short reverberant speech segments and identifying regions where the reverberation is interfering with the clean speech. The authors explain that by calculating the linear prediction (LP) residual of the reverberant speech signal, it is possible to identify these regions. The reason for this is attributed to the fact that clean speech has a peaked damped sinusoidal pattern while reverberant speech exhibits a more smeared characteristic. This leads to a reduction in the flatness of the reverberant speech spectrum. The problem is then formulated as a problem of trying to find a way to increase the flatness of the spectrum. The authors suggest using a smoothing procedure based on calculating the entropy in the LP residual. While the authors make it

clear that this method does not remove reverberation completely, it does reduce it more than traditional methods such as cepstral methods. The ideas above have been recently extended by [28] to an adaptive implementation. Instead of relying on entropy calculations, the authors exploit the fact that the LP residual of reverberant speech has a small kurtosis. By maximizing the kurtosis of the LP residual in the reconstructed speech signal, dereverberation is accomplished.

2.3.4 Channel Inversion Methods

2.3.4.1 Multichannel Inversion Methods

Inversion methods eliminate reverberation by undoing, i.e. inverting, the effect of the room impulse response. While these methods are theoretically the most accurate solutions for the deconvolution problem, their usefulness has been limited in the single microphone environment because of the constraints imposed, such as requiring the channel to be minimum phase. In addition, early inversion based methods were not truly blind since they required that the impulse response be known, i.e. they had an impulse response identification step. This required careful measurements and meant the early systems were tied to a given room environment.

The breakthrough in acoustical inversion methods came with the development of the multiple input output theorem (MINT) [5]. The MINT theorem is essentially a restatement of the *Bezout* equation and it shows that by using multiple microphones (channels), it is possible to find an exact inverse for channels modeled as FIR filters based on the *Bezout* equation. One of the most recent methods to follow this approach is discussed in [29]. However, the MINT theorem does not provide a way to accomplish truly blind dereverberation, since the method only provides a way for finding the inverses once the channels have been determined.

2.3.4.2 Blind Channel Identification

The next breakthrough for inversion methods came in the form of new blind channel identification methods for single-input multiple-output (SIMO) FIR channels, using only the second-order statistics (SOS) of the output. These methods were developed in response

to the need to equalize wireless communication channels. Blind equalization has long been a topic of interest in communication theory. The reason is that in wireless environments, the cost of channel training and the rapid changes in the channel make classical channel identification too costly and limited in value. The first generation of blind equalization methods relied on using higher-order statistics (HOS) to estimate the channel. However, these methods suffer from one or more problems of slow convergence, high computational complexity or local minima. It is for this reason that blind system identification and equalization methods using second-order statistics brought on much interest and excitement.

The multichannel single-input multiple-output (SIMO) FIR channel model can arise either in the context of multiple channels, as in our case of multiple microphones, or from the oversampled output of the channels as described, for example, in [30]. The equalization of SIMO FIR channels can fall under one of three broad categories.

1. Blind channel identification followed by inversion.
2. Direct design of equalizers with no channel estimation step.
3. Direct recovery of the original signal.

The methods in the first category are the most common and most studied and understood in terms of performance and limitations. Dozens of such methods exist, such as subspace methods [31, 32, 33], cross-relation methods [34, 35], frequency domain methods [36, 37], outer product method [38] and smoothing methods [39, 40]. The subspace methods decompose the correlation matrix of the output into a signal subspace and a noise subspace. By knowing the eigenvectors associated with the smallest eigenvalues in this matrix and exploiting the fact that the noise subspace is orthogonal to the convolutional channel matrix, the channels can be found. Some methods are a hybrid, such as the linear prediction methods [41, 42, 43] which provides a bridge between the first and second categories.

The third approach while is the most direct, usually requires that the input signal come from a finite alphabet set as described for example in [44] or be an independent, identically distributed (i.i.d.) input. For signal processing applications, such as speech signals, no such constraints can be imposed and so we will not discuss these methods further.

In the next section, we discuss the first SOS blind identification algorithm and explain its relation to the well known ESPRIT spectrum estimation method.

2.3.4.3 The TXK Algorithm

As noted earlier, the key step to most channel inversion methods is how to identify the channel, i.e. the room impulse response in our case. An added complication is that room impulse responses are generally nonminimum phase. To see why this is a complication, consider the output $y(n)$ of a linear system $h(n)$ excited by a stationary signal $s(n)$. The output spectrum contains no phase information about the system since the spectrum of the output is given by

$$Y(e^{j\omega}) = |H(e^{j\omega})|^2 S(e^{j\omega}).$$

Thus, the only way to obtain a channel estimate $\hat{H}(e^{j\omega})$ is to make the minimum phase assumption. If the use of higher-order statistics (HOS) is possible, then several methods can be used to obtain $\hat{H}(e^{j\omega})$ without the minimum phase assumption as detailed in [45, 46, 47]. However, it is well known that obtaining HOS information from short data segments is not robust and tends to suffer severe performance limitations in an adaptive implementation. This means that HOS methods are of limited use in practical blind equalization and channel identification problems.

An important first step in the quest for SOS blind identification was proposed by Gardner who noted that if the output $y(n)$ was *cyclostationary*, then identifying the channel reduces to finding the GCD of the various cyclic spectra as described in [48]. However, Gardner's proposed method still requires a training sequence (a predetermined input signal $s(n)$), meaning that it is not blind.

The first published result on the feasibility of blind channel identification, based solely on the availability of SOS output (receiver) information, was performed by Tong, Xu, Kailath [7] and their method has become known as the TXK algorithm. The authors showed that by having a single-input multiple-output FIR channel (SIMO) model, and assuming the input $s(n)$ was independent and identically distributed (i.i.d.), it is possible to identify all the FIR channels in the SIMO system. A SIMO system is easily obtained from SISO system either

by having temporal or spatial oversampling. For example, having multiple microphones or antennas instead of one creates a SIMO model. The TXK algorithm starts by considering the output of an M sensor SIMO system as given by:

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ \vdots \\ x_M(n) \end{bmatrix} = \sum_{i=0}^L \begin{bmatrix} h_1(i) \\ h_2(i) \\ \vdots \\ h_M(i) \end{bmatrix} s(n-i) \quad (1)$$

or more compactly in matrix form,

$$\mathbf{x}(n) = H\mathbf{s}(n),$$

where the channel matrix H is Toeplitz. It is important to note that in addition to having M sensors, the impulse responses are modeled as FIR filters of order L , or equivalently as FIR filters with $L + 1$ taps. A critical design parameter that is not explicitly shown in the above equation is the output data window size K , also called the smoothing factor [49]. The parameter K must be chosen so that H is full rank, i.e. tall and skinny. The value of K is chosen based on prior knowledge about the estimated channel order. This concept of knowing how long the channels are is important to all SOS blind identification methods. We elaborate on this point in subsequent chapters. The TXK method notes that the correlation matrix $R_{\mathbf{x}}$ of the output data at lag zero is given by

$$R_{\mathbf{x}}(0) = HE[\mathbf{s}(n)\mathbf{s}^T(n)]H^T = H I H^T,$$

where I is the identity matrix. Note that since H is full column rank, the correlation matrix $R_{\mathbf{x}}(0)$ will be also full rank. Assuming that H is real, it can be calculated up to an orthogonal transformation Q such that the estimated channel is given by $F = HQ$ since $FF^T = HQQ^TH^T = HH^T$. The problem is then how to determine Q to resolve the uniqueness problem. The key to the solution is to realize that the correlation matrix at lag one can be easily found and is given by

$$R_{\mathbf{x}}(1) = HE[\mathbf{s}(n+1)\mathbf{s}^T(n)]H^T = H J H^T,$$

where J is the backward shift matrix with ones along the subdiagonal. The main difference between the two lag matrices is that $R_{\mathbf{x}}(1)$ has a rank that is one less than $R_{\mathbf{x}}(0)$, i.e. $R_{\mathbf{x}}(1)$ has a nullspace of size one. This nullspace plays a crucial role in identifying the orthogonal matrix Q . Since F can be calculated, its pseudoinverse can be also calculated and is given by $F^\dagger = Q^T H^\dagger$. Pre-and post-multiplying $R_{\mathbf{x}}(1)$ by F^\dagger yields

$$C = F^\dagger R_{\mathbf{x}}(1) (F^\dagger)^T = Q^T J Q,$$

or equivalently as $CQ^T = Q^T J$. Denoting the last column of Q^T by q_m , the following relation holds

$$C \begin{bmatrix} q_1 & q_2 & \dots & q_m \end{bmatrix} = \begin{bmatrix} q_2 & q_2 & \dots & \mathbf{0} \end{bmatrix}$$

$$Cq_m = \mathbf{0},$$

and hence the nullspace of C gives the last row of Q up to a scale factor. The remaining columns of Q^T can be found recursively using the relationship

$$Cq_{k-1} = q_k$$

where $k = 0 : m - 1$.

The TXK algorithm uses the same approach of considering different snapshots as the ESPRIT algorithm developed by [50]. The ESPRIT algorithm was proposed for the spectral estimation of the sum of sinusoidal signals in white noise. To see the similarity between the two approaches, consider a signal $\mathbf{x}(n)$ composed of the sum of p sinusoids in white noise $w(n)$ as given by

$$\mathbf{x}(n) = S\mathbf{z}(n),$$

where S is a Vandermonde matrix and \mathbf{z} is a vector of complex coefficients. This means the correlation of $\mathbf{x}(n)$ is given by

$$R_x(0) = SPS^H,$$

where P is a diagonal matrix. In additive noise $\mathbf{w}(n)$ of unit variance, the signal model is

$$\mathbf{y}(n) = \mathbf{x}(n) + \mathbf{w}(n),$$

and the correlation is given by

$$R_y(0) = SPS^H + I.$$

If we consider the next frame of the signal $\mathbf{y}(n+1)$, the correlation with the previous frame is then given by

$$R_y(1) = SP\Theta^H S^H + J,$$

where Θ is a diagonal matrix that contains the phase shift due to the one sample displacement. Thus, the signal correlations $\mathbf{x}(n)$ at lag zero and one are given by

$$R_x(0) = R_y(0) - I = SPS^H$$

$$R_x(1) = R_y(1) - J = SP\Theta^H S.$$

Subtracting these two correlations from each other yields

$$SP(I - \Theta^H I)S^H = \mathbf{0},$$

which is the equivalent of a generalized eigenvalue problem in which the diagonal of Θ plays the role of generalized eigenvalues λ_i . Finding these eigenvalues yields the frequencies that compose the Vandermonde matrix $S(n)$. In the TXK algorithm, the channel matrix H plays a role similar to S and the idea of using the correlation matrices at two different lags for the identification of the null space and decomposing the signal space into two orthogonal components.

As important as the TXK algorithm was and its results were, it raised some critical questions. For example what conditions must the channels meet to be identifiable, or what conditions must be imposed on the input? Is there an explicit SIMO system structural property that allows identifiability? The answer is yes to both of these questions as we explain in the next section.

2.3.4.4 Channel Identifiability Conditions

Blind SOS channel identification is only possible under some conditions. The first requirement is that the channels have no common zeros, which is the same as requiring that the channels be coprime. This condition is equivalent to requiring that the composite channel matrix be of full column rank as explained in [51, 52]. The second condition for identifiability is that the input signal be persistently exciting, a condition required by

many classical system identification methods as described in [53]. Thus, the identifiability conditions for a SIMO system can be summarized as:

1. The channel matrix H must be of full column rank.
2. The channels $h_1(n), h_2(n), \dots, h_M(n)$ must be coprime, meaning that the channels must not share any common zeros.
3. The input signal $s(n)$ must be persistently exciting.

The persistent excitation condition is required for a unique solution in the finite output data case and when the input data is considered to be an unknown deterministic signal as detailed in [49]. The coprime channels requirement raises an interesting question in the case of acoustical channels. If two microphones are placed in close proximity together, would we expect their impulse responses to be very similar and violate the coprimeness condition? A preliminary answer to this question is no. This is because while the impulse responses may be very similar to each other, i.e. the coefficients of the FIR filters may be very nearly equal, a polynomial with slightly perturbed coefficients compared to another polynomial will not have near equal roots. We elaborate more on this issue in the next chapter when we discuss acoustical impulse responses.

The next class of methods we discuss are the subspace methods. What these methods reveal is that it is the structure of the SIMO system output that allows for identifiability.

2.3.4.5 Subspace methods

As explained in the previous section, the TXK method assumes that the input signal is i.i.d. and makes no explicit use of the data output structure. The subspace approach is exemplified by the subchannel matching approach described in [40] and [54]. In the simplest case of two channels as shown in Figure 8, two unknown channels $h_1(n), h_2(n)$ with no common zeros are to be identified. It is clear that in this case the error $e(n)$ is given by

$$e(n) = x_1(n) * v_1(n) - x_2(n) * v_2(n) = s(n) * (h_1(n) * v_1(n) - h_2(n) * v_2(n)).$$

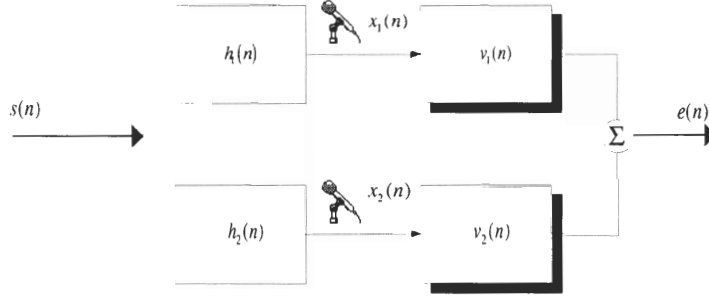


Figure 8: Subchannel approach.

Thus, for the error to be zero, the filter $v_1(n)$ must be proportional to $h_2(n)$ and $v_2(n)$ be proportional to $h_1(n)$. A more compact way to write the above relation is

$$e(n) = \begin{bmatrix} \mathbf{x}_1^T & -\mathbf{x}_2^T \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = 0.$$

The subchannel approach can be easily extended for multiple channels. For example in the three channel case, the pairwise relationships might be given by

$$\begin{bmatrix} \mathbf{x}_1^T & -\mathbf{x}_2^T & 0 \\ \mathbf{x}_3^T & 0 & -\mathbf{x}_1^T \\ \mathbf{x}_3^T & 0 & -\mathbf{x}_2^T \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_2 \end{bmatrix} = \mathbf{0}.$$

2.3.4.6 Linear Prediction approach

The linear prediction approach to the problem of SOS blind deconvolution developed by [43] is one of the more interesting and versatile approaches to the problem of blind SOS identification. Linear prediction is a well established method for deconvolution problems, especially in seismology where the inverse filters are also referred to as spiking filters.

The linear prediction formulation once again begins by considering a SIMO FIR channel system output as given in Equation (1). The z -transform of the SIMO impulse response has the form of a vector given by

$$\mathbf{h}(z) = \begin{bmatrix} h_1(z) \\ \vdots \\ h_M(z) \end{bmatrix},$$

where the individual entries $h_1(z), \dots, h_M(z)$ are the z -transforms of the individual impulse responses $h_1(n), \dots, h_M(n)$. The generalized Bezout equation states that it is possible to find another vector of polynomials $\mathbf{g}(\mathbf{z})$ such that the equation

$$\mathbf{g}^T(z)\mathbf{h}(z) = 1 \quad (2)$$

holds under the assumption that $\mathbf{h}(\mathbf{z})$ is irreducible, i.e. no common zeros between the various impulse responses. Thus, it is possible to consider the vector polynomial $\mathbf{g}(\mathbf{z})$ as acting as an inverse filter for $\mathbf{h}(\mathbf{z})$. This is a sharp contrast to the single channel case where an FIR filter requires an IIR filter for exact inversion. Also, note that the output of a SIMO system can be expanded into the sum of individual vectors as given by:

$$\mathbf{x}(n) = \begin{bmatrix} x_1(n) \\ \vdots \\ x_M(n) \end{bmatrix} = \sum_{i=0}^L \begin{bmatrix} h_1(i) \\ \vdots \\ h_M(i) \end{bmatrix} s(n-i) = \begin{bmatrix} h_1(0) \\ \vdots \\ h_M(0) \end{bmatrix} s(n) + \dots + \begin{bmatrix} h_1(L) \\ \vdots \\ h_M(L) \end{bmatrix} s(n-L), \quad (3)$$

where M, L are the number of channels and channel order respectively. If the inverse filter $\mathbf{g}(\mathbf{z})$ is applied to $\mathbf{x}(n-k)$ then the result is $s(n-k)$. Using this result, we can rewrite Equation (3) in term of the previous output samples as given by:

$$\mathbf{x}(n) = \begin{bmatrix} h_1(0) \\ \vdots \\ h_M(0) \end{bmatrix} s(n) + \underbrace{\begin{bmatrix} h_1(1) \\ \vdots \\ h_M(1) \end{bmatrix} \underbrace{\mathbf{g}^T(z)\mathbf{x}(n-1)}_{s(n-1)} + \dots + \begin{bmatrix} h_1(L) \\ \vdots \\ h_M(L) \end{bmatrix} \underbrace{\mathbf{g}^T(z)\mathbf{x}(n-L)}_{s(n-L)}}_{\hat{\mathbf{x}}(n)}. \quad (4)$$

Equation (4) has the form of a linear prediction term $\hat{\mathbf{x}}(n)$ plus an innovation term (the prediction error term). The blind linear prediction identification algorithm utilizes this fact and under the assumption that input $s(n)$ is i.i.d., the vector $\mathbf{h}_0 = [h_1(0), \dots, h_M(0)]^T$ can be identified using the correlation:

$$E \left[(\mathbf{x}(n) - \hat{\mathbf{x}}(n)) (\mathbf{x}(n) - \hat{\mathbf{x}}(n))^T \right] = \mathbf{h}_0 \mathbf{h}_0^T.$$

Once \mathbf{h}_0 is found, it is possible to obtain the input sequence directly using the relation

$$\hat{s}(n) = \frac{1}{\|\mathbf{h}_0\|} (\mathbf{x}(n) - \hat{\mathbf{x}}(n)) = \frac{1}{\|\mathbf{h}_0\|} \mathbf{h}_0 s(n),$$

where $\hat{s}(n)$ is the estimated input signal. Once $s(n)$ is estimated, a nonblind system identification formulation can be used to find the channels if required.

2.3.4.7 *Direct Inversion*

The combination of the MINT theorem and the blind SIMO FIR channel identification methods has been applied to the dereverberation problem. For example in the method described in [54], the authors used a blind procedure to identify two acoustical channels. Another method based on a cross relation method is described in [55] and the method was extended recently in [56]. While all these methods can estimate the acoustical channels, they all require a two-stage process for dereverberation; a channel identification stage and then an inversion stage based on the MINT theorem or some variant of it. This approach is not optimal since it requires a two-stage process. An error in the estimated channels leads to the design of inverse filters that do not truly invert the system. Also, the two-stage process means that tracking and convergence issues in an adaptive system become critical. It is for this reason we have avoided a two-stage procedure and prefer to have a single stage method for dereverberation.

As suggested by the linear prediction approach, it is possible to obtain $s(n)$ directly once an inverse has been found. This led to the development of new methods that attempt to bypass the channel identification stage and attempt to find the inverse FIR filters directly as described in [57, 58, 59, 60, 61, 62, 63, 64, 65]. For example in [63], the cross correlation function needed to find the MSE equalizer is found directly (with some assumptions on the nature of the input). In [58], the authors formulate blind equalization as a generalized eigenvalue problem and develop a criterion for finding the optimal delay equalizer under the assumption that the input signal is constant modulus. In [60], the authors adapt some of the blind array processing methods based on Capon's method to find the equalizers.

A very interesting method, with links to linear prediction, that finds the equalizers directly is the mutually referenced equalizers (MRE) method described in [59, 62, 64]. This method finds all the equalizers at all possible delays. The method has a simple adaptive formulation and makes only very mild assumptions on the nature of the input. In fact, the

MRE method has proven to be very versatile and has been applied to other applications such as multichannel image deblurring as described in [66]. It is exactly because of these advantages of the MRE method that we propose a method based on a modified MRE method for the speech dereverberation problem. However, before discussing the MRE and the required modifications, we discuss the issue of room impulse response modeling, measurement and related acoustical issues in the next chapter.

CHAPTER III

ROOM IMPULSE RESPONSE SIMULATION AND MEASUREMENT

In this chapter, we continue our investigation of the room impulse response. While the modeling and simulation of acoustical reverberation is a complicated and difficult problem, the focus here will be only on those aspects that are of most relevance to this research; namely the simulation and measurement of room impulse responses. It is important to note that not all reverberation is undesirable. In concert halls, a certain amount of reverberation enhances the sound quality and contributes to the richness of the sound. However, when speech is involved, the reverberation *blurs* and *smears* the speech signal.

3.1 *The Physical Modeling of Reverberation*

The simplest physical explanation for reverberation is based on the reflection and absorption of acoustical waves by various surfaces in a closed room. Because of the nonlinear absorption and reflection from nonuniform surfaces, the acoustic waves are attenuated and diffused. As explained in Chapter 1, the phenomenon of reverberation can be viewed either as a lumped process modeled by a linear system or as a physical model governed by differential equations that describe the propagation of the sound waves. The simplest physical model is the propagation of a sinusoidal wave of frequency w in a rectangular room (with rigid walls) of dimension L_x, L_y, L_z . The wave equation in this case simplifies to the well known Helmholtz equation given by:

$$\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2} + k^2 p = 0, \quad (5)$$

where p is the sound pressure and k is the wavenumber. The above equation is separable, and utilizing the boundary conditions that must be satisfied, the general solution at point

(x, y, z) at time t is given by:

$$p(x, y, z, t) = A \cos \left[\frac{n_x \pi x}{L_x} \right] \cos \left[\frac{n_y \pi y}{L_y} \right] \cos \left[\frac{n_z \pi z}{L_z} \right] \exp(j\omega t), \quad (6)$$

where A is a constant dependent on the boundary conditions [67]. The wave number k parameter is related to the wavenumber in each variable by

$$k = \pi \left[\left(\frac{n_x}{L_x} \right)^2 + \left(\frac{n_y}{L_y} \right)^2 + \left(\frac{n_z}{L_z} \right)^2 \right]^{\frac{1}{2}},$$

and each wavenumber k defines an oscillation *mode*. When a sound source excites the room and the associated modes, the modes resonate at their corresponding frequency. Once the excitation terminates, these modes decay at different rates depending on location, time, boundary conditions, and the associated wavenumber. The superposition of all these decaying modes is what creates the reverberation. However, this complex phenomena of reverberation is usually quantified with a scalar quantity called the reverberation time, denoted by T_{60} . Sabine's pioneering work on room acoustics and the quantification of reverberation led him to define the reverberation time T_{60} as the time required for the initial sound pressure p_0 to decay by $60dB$. Obtaining such a quantity requires the measurement of the reverberation time, but a simple approximation exists for rectangular rooms of moderate size with rigid walls, and is given by:

$$T_{60} = K \frac{V}{\Sigma A_i}, \quad (7)$$

where K is a proportionality constant that depends on the reflectivity of the room surfaces, V is the room volume, and A_i is the effective surface area of the i -th absorbing surface in the room. The parameter A_i is the product of the absorbing surface size S_i and the absorption coefficient α_i . In the idealized case of hard walls, T_{60} is approximated by:

$$T_{60} = 0.161 \frac{V}{\Sigma A_i}. \quad (8)$$

Equation (7) can be extended to larger rooms if the absorption of the air in the room is taken into account. In this case, the reverberation time is given by:

$$T_{60} = 0.161 \frac{V}{\Sigma A_i + mV},$$

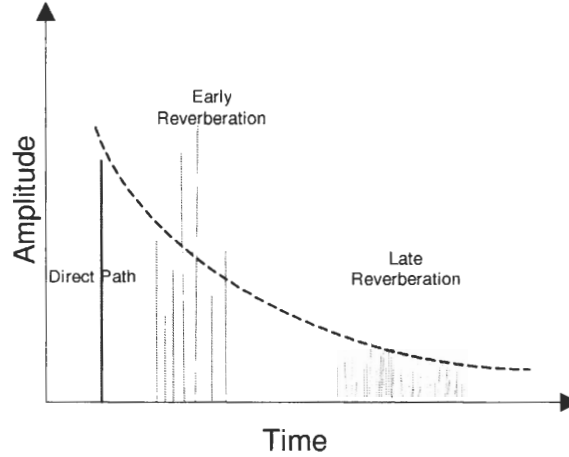


Figure 9: Simplified impulse response of a room.

where m is the absorption coefficient of air at a given temperature. One interesting point to note is that none of these reverberation time equations depend on the locations of the source or microphone inside the room. This is a consequence of the fact that the reverberation time T_{60} is an average measure of the decay time of all the various decaying modes that result from the solution of the Helmholtz equation. In addition, the fact that reverberation at a given location in a room is the superposition of all the decaying modes, it is reasonable to model the phenomenon as a linear system defined by the convolution of the impulse response between a source and receiver.

In the next section, we discuss the simulation of the room impulse responses. It turns out that the most effective room impulse response simulation methods rely on the physical modeling of the reverberation described above.

3.2 Room Impulse Response Simulation

The typical shape (of a simplified) room impulse response shown in Figure 9. As illustrated, the typical impulse response can be considered to be made up of three segments described as the direct path, the early reverberation, and the late reverberation. The approximate shape of the impulse response follows that of a decaying exponential. The direct path signal is not always (but frequently is) the maximum of the impulse response since the early reverberation can add up constructively and create a larger amplitude. The early

reverberation is associated with those surfaces closest to the microphone, such as tables. The late reverberation forms the tail of the impulse response, and involves a dense collection of smaller valued echoes that are the result of the diffusivity of the reflected echoes.

The modeling of realistic and accurate sounding impulse responses is a non-trivial mathematical endeavor, but some well established modeling methods have been developed. The earliest reverberation model was proposed by Schroeder, and consisted of a bank of comb filters connected in parallel, which in turn was connected in series to a group of allpass filters connected in series. The purpose of the comb-filter was to create the decaying exponential shape associated with a typical impulse response, while the allpass filter bank purpose was to create the smear and blur effect associated with a typical impulse response. While the model captures the overall features of reverberation, it usually lacks sufficient echo density in the tail. Other more natural sounding filtering approaches for creating reverberation impulse responses, such as the Jot reverberator, are described in [68]. However, in most of these models, it is difficult to relate the filter's parameters to the physical room that is being simulated. A more successful and realistic approach to modeling the room impulse response is *auralization*. Auralization is the use of the physical model of sound propagation and reflection, along with room size, and absorption parameters to simulate the room impulse response. The image method, developed by Allen and Berkley [69], is one of the most frequently used methods for generating the impulse response in a room between an arbitrarily positioned source and microphone. The method calculates the impulse response by using the wave propagation equation, in conjunction with the image method for the modeling of the sound reflections. The image method, popular in optics, is illustrated in Figure 10. A source wave is reflected by a wall at an angle θ and is recorded by the microphone. To the microphone, the acoustical reflection is equivalent to receiving a direct path signal from a virtual source, i.e. an image of the original source. The same principle can be extended to multiple images as illustrated in Figure 11. A FORTRAN77 implementation of the image method is given in [69], and Matlab implementations of the image method can be found in [6, 70]. The input parameters to the image method are the:

- Room dimension

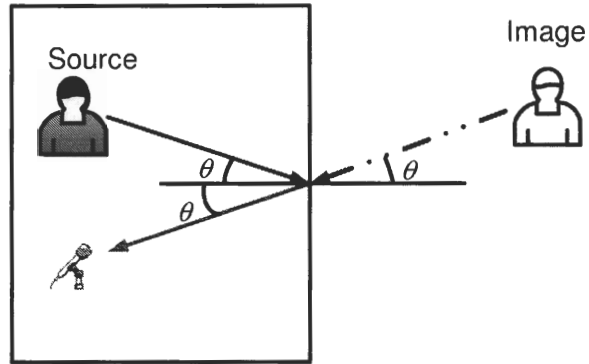


Figure 10: The image method of modeling reflection of waves.

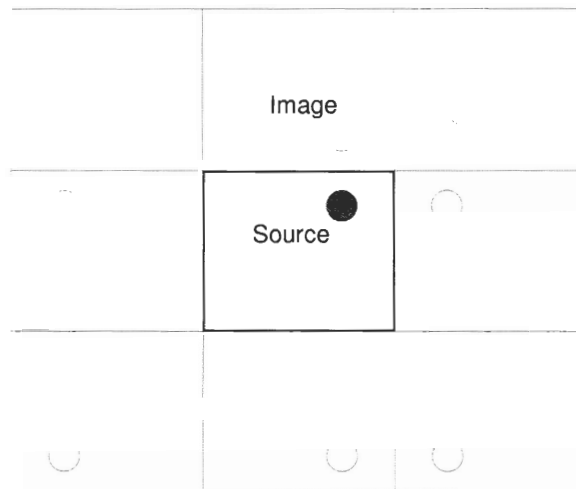


Figure 11: Multiple images in a 2D plane.

- Source location
- Microphone location
- Six reflection coefficients of the various walls and surfaces
- Sampling rate

The reverberation time T_{60} is not a parameter that is defined in the image model, but rather it is a quantity that must be calculated from the room impulse response. The backward integration method proposed by Schroeder [71] is the most frequently used method because of its robustness. Statistically, the room impulse response can be modeled as:

$$h(n) = e^{-n/\tau} r(n) + \sigma_w w(n), \quad (9)$$

where τ is a decay time constant and $\sigma_w w(n)$ is white gaussian noise with standard deviation σ_w to model the background noise in a room. The variable $r(n)$ is a random variable that models the amplitude of the recorded echoes. A good choice for $r(n)$ is the Rayleigh distribution defined by:

$$r(n) = \frac{n}{\sigma^2} e^{-n^2/2\sigma^2},$$

where σ^2 is the average power of the received signals. The Rayleigh distribution is frequently used to model the amplitude of received signals in a multipath environment. Since the T_{60} time is a measure of the time needed for the initial energy to decrease by 60dB, the squared impulse response given by:

$$f(n) = (h(n))^2 = \left(e^{-n/\tau} r(n) + \sigma_w w(n) \right)^2,$$

and $f(n)$ is utilized to create a new curve called the *integrated energy decay curve* (IEDC). Noting that $f(n)$ is in reality a continuous signal $f(t)$, the IEDC is defined as a tail integration (or more accurately as a summation in the discrete case) of $f(t)$,

$$IEDC(t) = \int_t^{t_\infty} f(u) dt.$$

While the echoes theoretically never decay to zero, there comes a time when the ambient noise becomes larger than the echoes, and the value of t_∞ is chosen to be that time. The

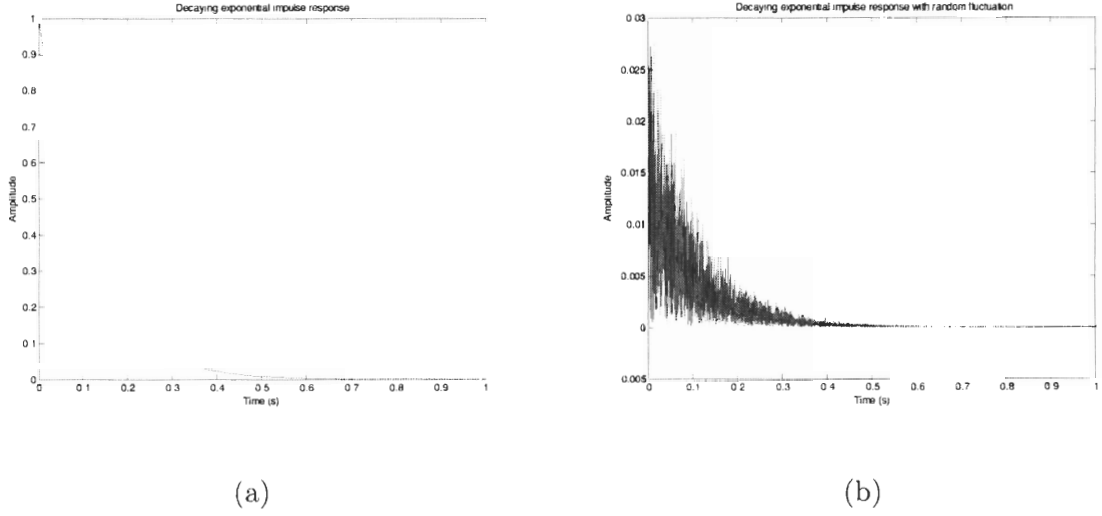


Figure 12: (a) Decaying exponential, (b) Decaying exponential with random amplitude.

IEDC, under the assumptions above, resembles a linear curve when plotted as a logarithmic quantity. In the noiseless and pure decaying exponential case, the curve will be linear, but under more realistic conditions, it exhibits a fluctuation about the ideal case. In such a case, a linear interpolation is performed on the IEDC curve to obtain T_{60} . Figure 12 shows a pure decaying exponential impulse response and decaying exponential with a Rayleigh distribution random amplitude model as given in Equation (9). The IEDC curves shown in Figure 13 were obtained using a modified version of the original Schroeder implementation described in [72].

The case of random amplitude shows the IEDC curve deviating towards the end. A more realistic impulse response will exhibit more deviations throughout as will be shown later.

3.2.1 Image Method Simulation Results

We consider the simulation of a few room impulse responses using the image method for later comparison with the measured impulse responses of room 352 in the GCATT building. The room size is $[x, y, z] = [18, 21, 9]ft = [5.49, 6.40, 2.74]m$.

For the first simulation, the sampling rate is chosen to be 8kHz; the reflection coefficients are chosen to be 0.5 for the walls and 0.4 for both the floor and ceiling. The source signal

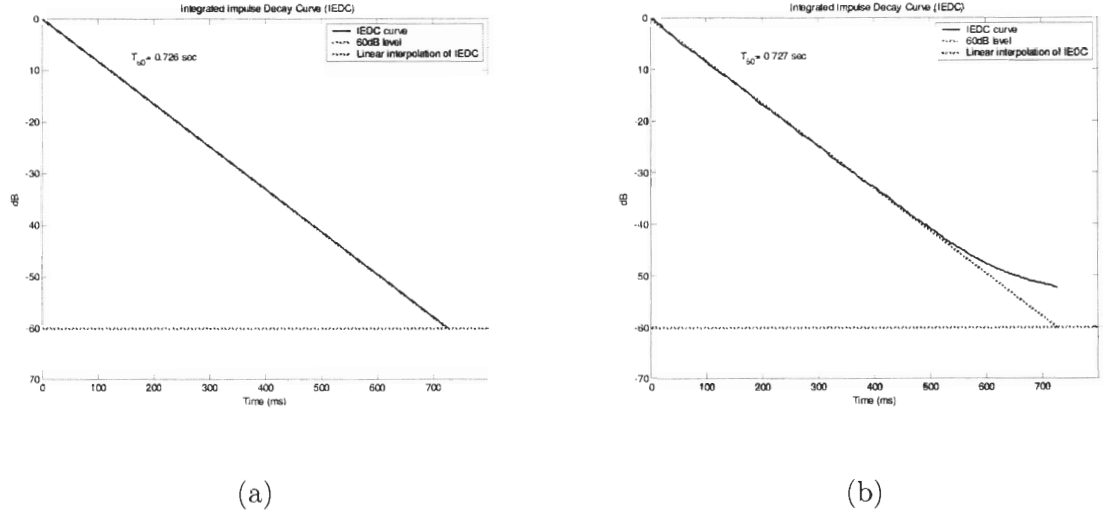
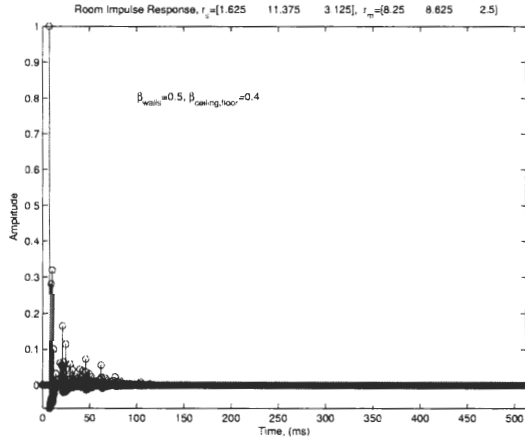


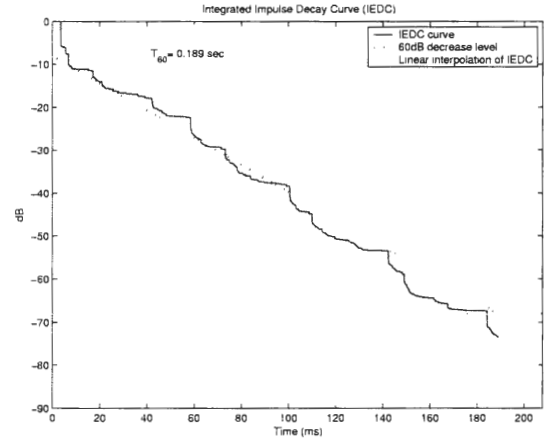
Figure 13: (a) Decaying exponential, (b) Decaying exponential with Rayleigh distributed random amplitude.

is defined to be at $r_s = [x, y, z] = [1.67, 11.17, 3.17]ft$ and the microphone is chosen to be at $r_m = [x, y, z] = [8.25, 8.58, 2.5]ft$. The resulting impulse response, along with the corresponding IEDC curve are shown in Figure 14. Next, we consider a small change in the position of the microphone to $r_m = [x, y, z] = [8.25, 8.33, 2.5]ft$, and once again the resulting impulse response, along with the corresponding IEDC curve are shown in Figure 15. Note that the reverberation time changed only slightly, and the impulse response has a few main differences in the early reverberations. Figure 16 shows the difference between the above two impulse responses. The pole-zero plot of the impulse responses in Figures 14 and 15 is shown in Figure 17 (the impulse responses were truncated to 1000 samples). While the pole-zero plots seem very similar, zooming in as in Figure 17b shows that it is unlikely the zeros of the two impulse responses will coincide and violate the common zeros condition discussed in Chapter 2 for the SOS blind equalization and identification of SIMO channels. Also, separating the microphones by a larger distance creates a greater difference in the impulse responses and reduces the possibility of common zeros even further.

The reverberation time is more dependent on the reflection coefficients than on the distance between the speaker and microphone in a room. To illustrate this, we consider the simulation of a room with the reflection coefficients chosen to be 0.7 for the walls and

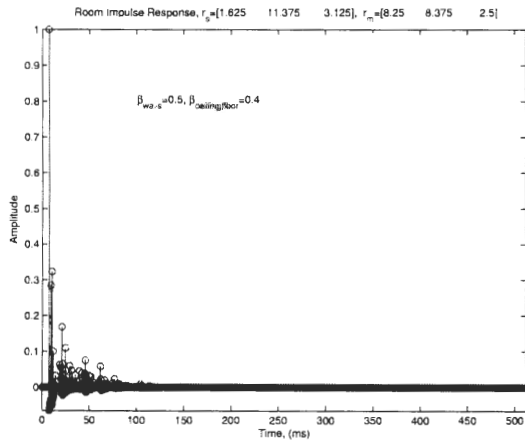


(a)

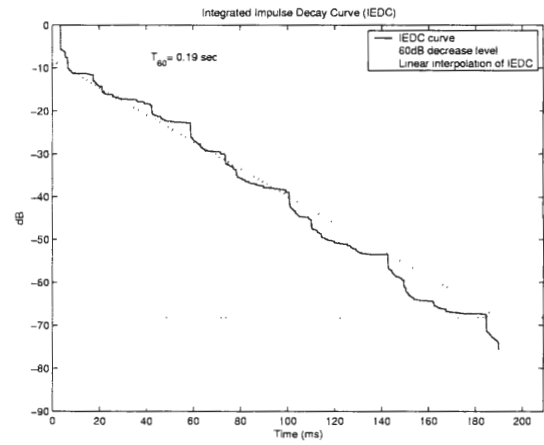


(b)

Figure 14: (a) Impulse response $r_m = [x, y, z] = [8.25, 8.58, 2.5]ft.$, (b) IEDC.



(a)



(b)

Figure 15: (a) Impulse response for $r_m = [x, y, z] = [8.25, 8.33, 2.5]ft.$, (b) IEDC.

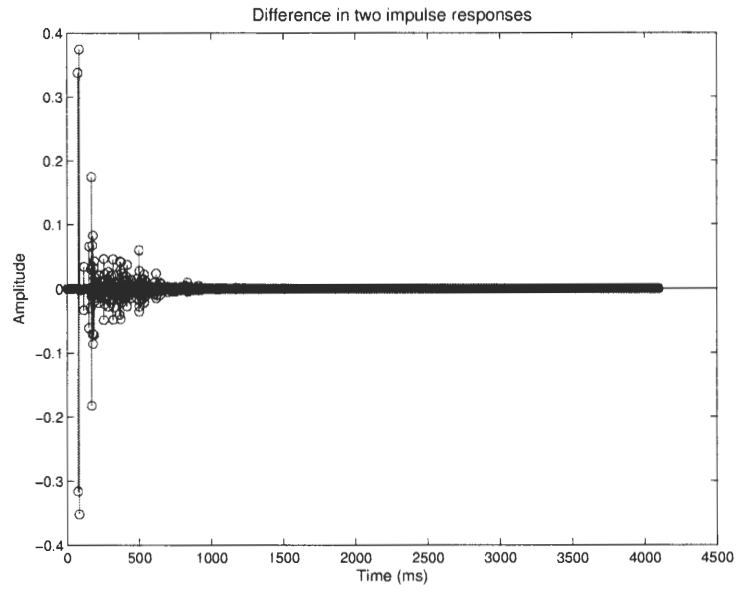


Figure 16: Difference of two impulse responses.

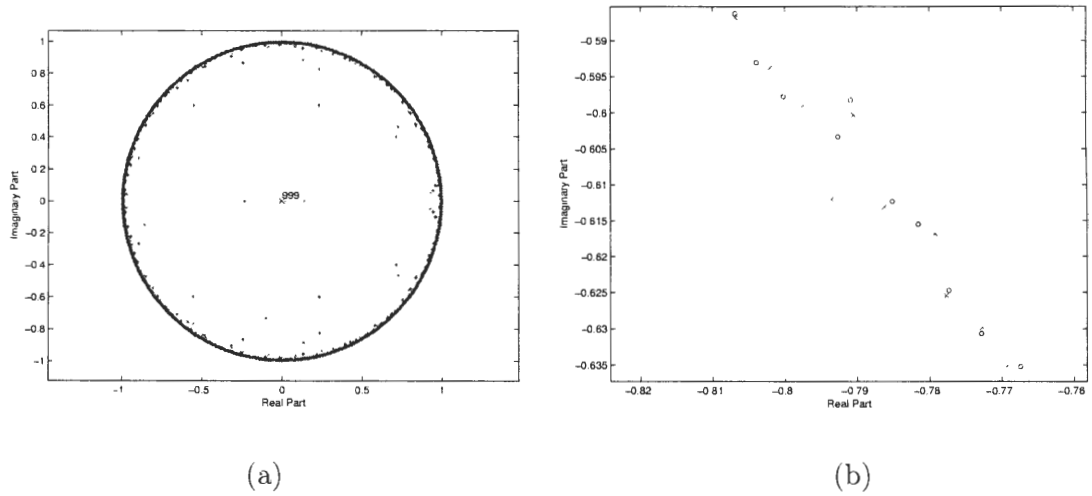


Figure 17: (a) Pole-zero plot of two impulse responses, (b) Magnified portion.

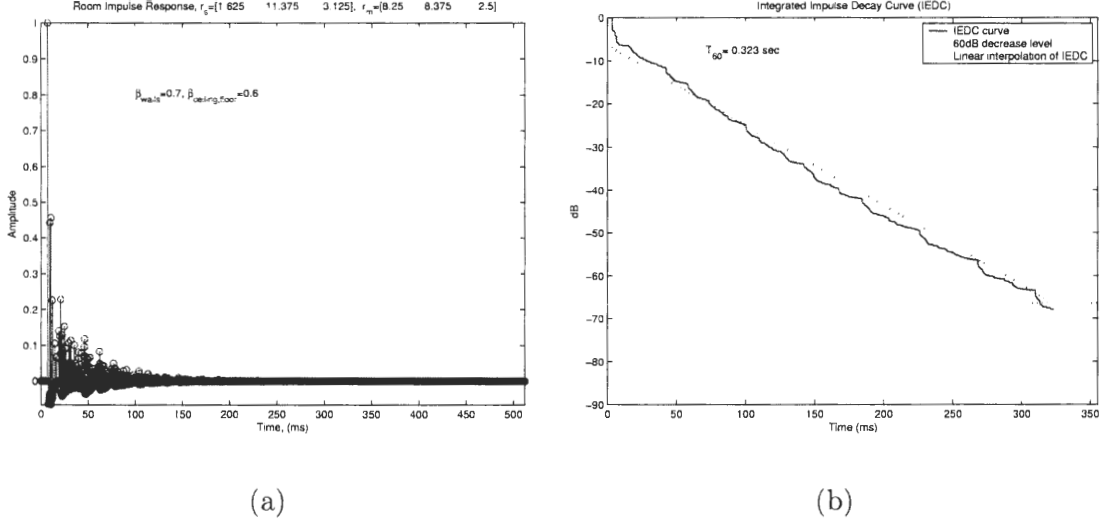


Figure 18: (a) Impulse response, (b) IEDC.

0.6 for both the floor and ceiling. The microphone is placed again at $r_m = [8.25, 8.33, 2.5]$. The new reverberation time is calculated to be $T_{60} = 0.323\text{sec}$, and the resulting impulse response and IEDC curves are shown in Figure 18.

3.3 Room Impulse Response Measurements

The simplest, and most obvious, way to obtain the room impulse response is by creating an impulsive signal at some distance away from a microphone, and then recording the resulting impulse response. Three problems arise with this type of approach. The first is the difficulty in creating an exact delta function even with the use of an exploding balloon or gunshot. The second problem is the poor SNR such a method would yield due to the small amount of energy in the delta function. The third drawback is the difficulty in repeating the experiment with an identical input. For example, no two balloons would pop in the same way.

Another method of obtaining an impulse response, albeit indirectly, is by using a sinusoidal sweep signal. This approach gives the Fourier transform of the impulse response over a certain frequency range. While this method is widely used, it suffers from nonlinearity problem inherent in the speakers and some of the recording equipment [67].

The third approach involves using a maximum length sequence (MLS) as the excitation

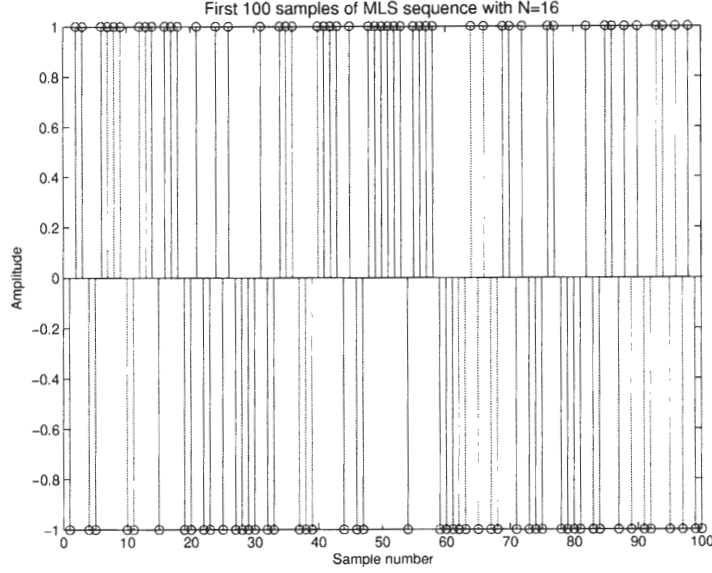


Figure 19: MLS sequence, $N=16$.

signal. MLS signals are deterministic in nature but appear as a random binary sequence. The generation of an MLS sequence depends on defining a primitive polynomial $m(x)$ (in $GF(2)$) of order N . This is equivalent to having a sequence of shift registers in cascade that implements a difference equation with the output taken *mod*2. Depending on the value of N chosen, an MLS sequence with a period of $2^N - 1$ samples is generated. For example with $N = 16$, a sequence of 65535 samples is generated. When played back at a sampling rate of $8kHz$, this corresponds to an audio signal that is approximately 8.1s long. What makes the MLS sequence $s(n)$ very useful in measuring the room impulse response is that it has correlation properties similar to white noise. For example, the auto-correlation of $s(n)$ is:

$$r(n) = \frac{1}{N} \sum_{i=0}^N s(i)s(i+n) \approx \delta(n).$$

Thus, the sequence $s(n)$ has a flat power spectrum. For example, for $N = 16$, Figure 19 shows the first 100 samples of $s(n)$. Figure 20 shows the autocorrelation of the sequence $s(n)$ and how it approximates an impulse, and Figure 21 shows the power spectrum of the sequence $s(n)$. If we denote the impulse response by $h(n)$, the recorded signal $x(n)$ is given by $x(n) = s(n) * h(n)$. Cross-correlation of $x(n)$ with the input $s(n)$ gives the room impulse

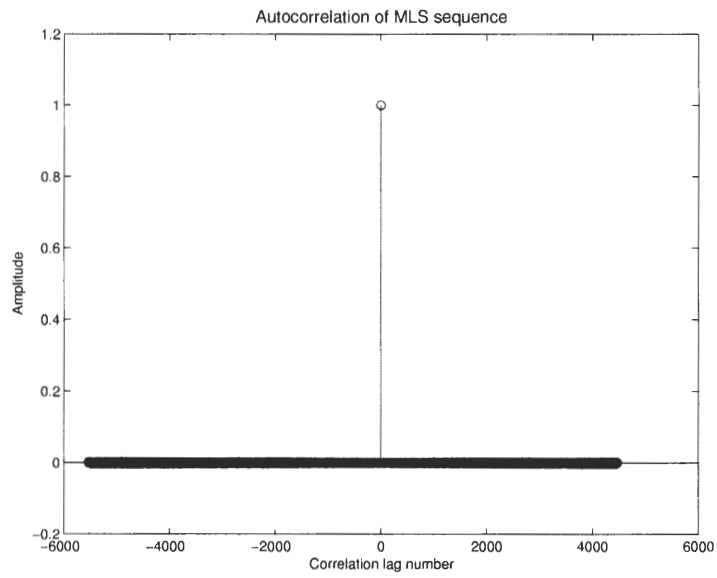


Figure 20: Autocorrelation of MLS sequence.

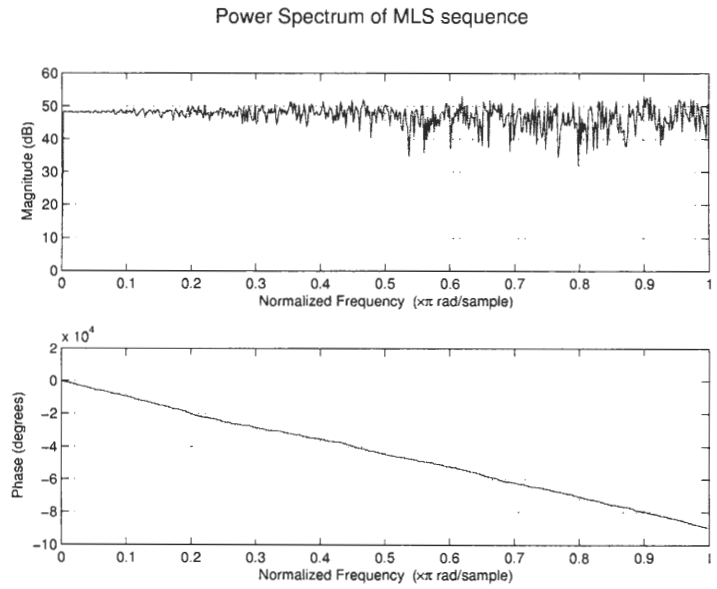


Figure 21: Power spectrum of MLS sequence.

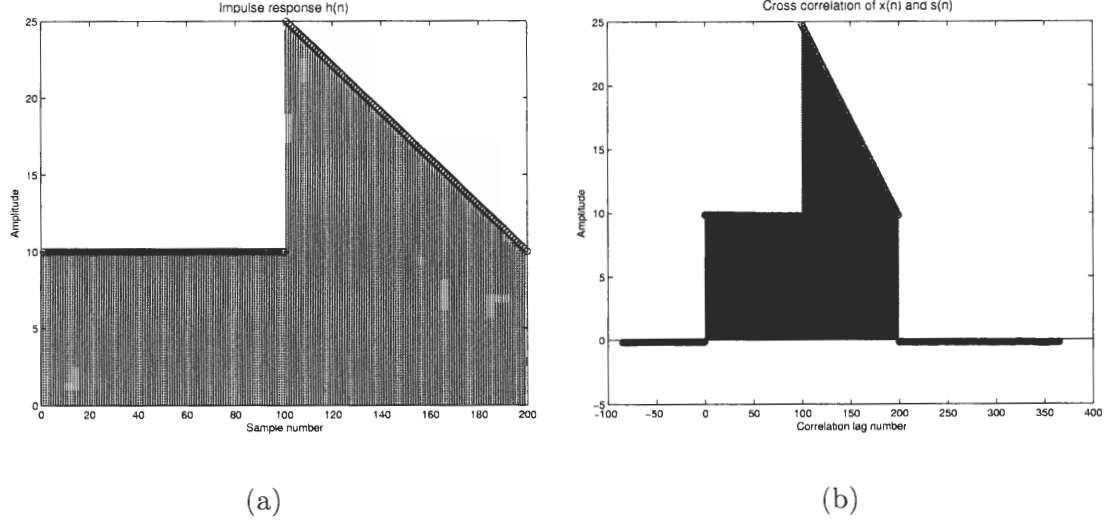


Figure 22: (a) Arbitrary shaped impulse response, (b) Resulting estimated impulse response using correlation.

response $h(n)$,

$$\frac{1}{N} \sum_{i=0}^{N-1} s(n+i)x(i) = \frac{1}{N} \sum_{i=0}^{N-1} s(n+i) (h(n) * s(n)) = h(n).$$

As a simple example, consider the arbitrary impulse response $h(n)$ shown in Figure 22a.

An MLS signal with $N = 16$ was filtered with $h(n)$ and the result $x(n)$ was cross correlated with $s(n)$. The resulting cross correlation is shown in Figure 22b.

Note that in practice, the procedure of exciting the room with $s(n)$ may be repeated several times to improve the SNR. While the procedure described above is simple from a theoretical standpoint, its implementation is nontrivial in a real environment because of the nonlinearities in speakers and the ambient noise we had in our audio lab.

3.4 Impulse Response Measurements in Room 352

While the theory and mathematics behind room impulse response measurement may seem simple and easy to implement, the actual realtime implementation is very difficult and involved. We opted for the use of a commercial package called RAE [73] to do our room impulse response measurements. While many packages exist for this task, we have found the low price, flexible hardware requirements and good results obtained with RAE to be unmatched. The closest room to an anechoic chamber available to us was Room 352 in

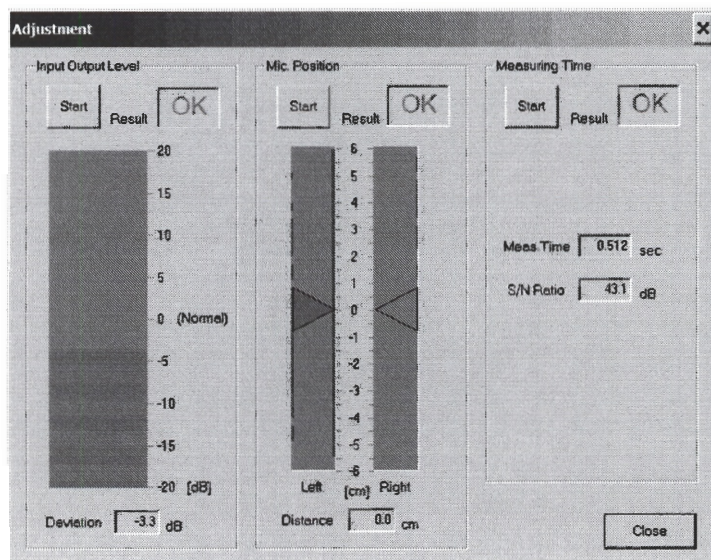


Figure 23: Calibration module.

the GCATT building. However, in the course of our experiments we have discovered some flaws in the room; specifically we have found that the ventilation system was louder than normal, and even when we turned it off, there was a humming sound emanating from the ceiling. However, we still were able to manage to do some experiments on measuring the impulse response.

3.4.1 Calibration

The first step in doing the measurement is the calibration of the required speaker volume and the length of the MLS excitation signal. This process is automated as shown in Figure 23. As can be seen in Figure 23, we consistently got an SNR of approximately 40dB. Anechoic chambers have an SNR range between 60dB and 80dB.

3.4.2 Measurement

Once the calibration has been performed, the impulse response is measured using the module shown in Figure 24.

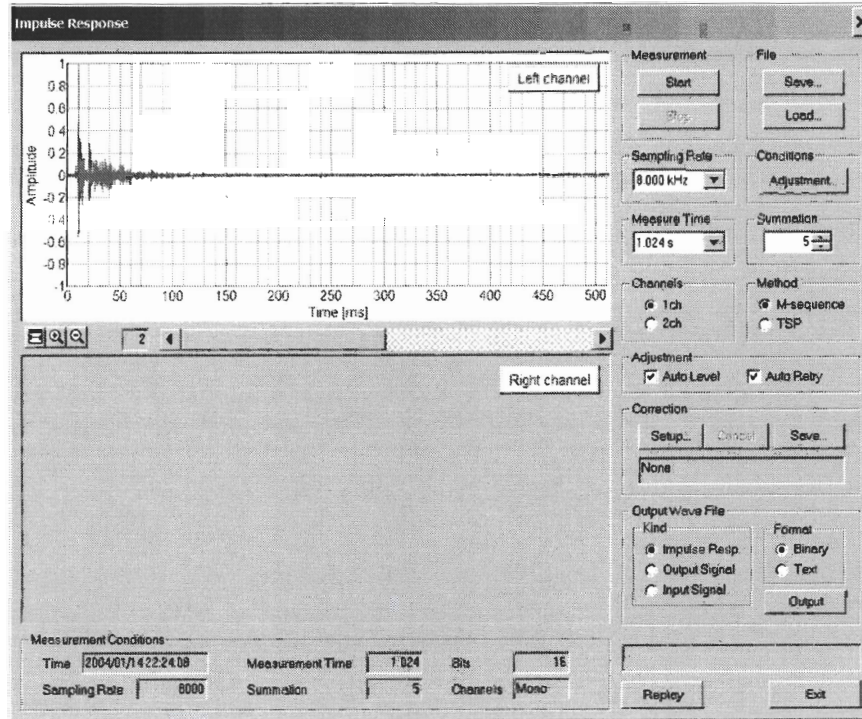


Figure 24: Measurement module.

3.4.3 Reverberation Time Calculations

After the measurement of the impulse response, the reverberation time calculation module can be used to obtain the reverberation time of the measured impulse response as shown in Figure 25. The measured T_{60} time calculated by RAE was found to be 0.16s, which matches well with the estimated T_{60} time of approximately 0.19s found from the simulations. However, we want to emphasize that while the reverberation times are close, the impulse responses will not be equal on a sample by sample basis. Figure 26 shows the simulated and measured impulse responses (adjusted for peak amplitudes to match for easier comparison). The measured impulse response exhibits more peaks due to the many reflecting surfaces found in Room 352, especially the large wooden table located near the center of the room.

These comparisons allow us to use the simulated impulse response in the next chapter when it is convenient, especially when considering many microphones.

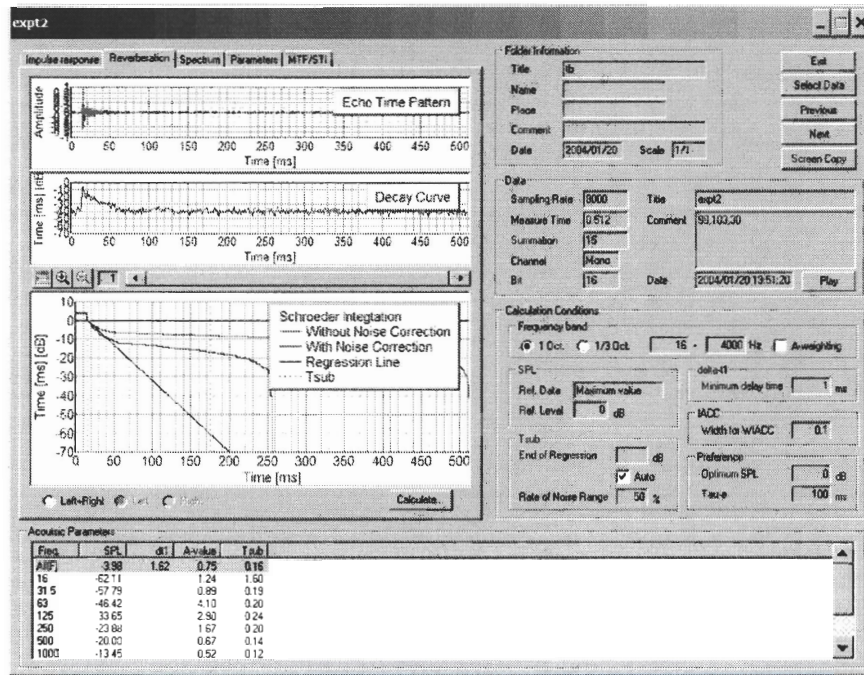


Figure 25: Reverberation time calculation module.

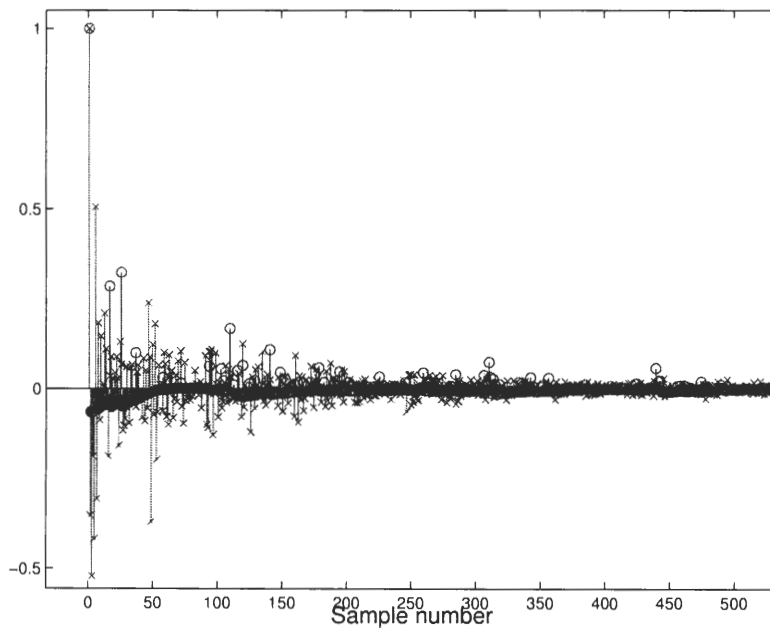


Figure 26: Comparison of simulated and measured impulse responses, \times is the measured impulse response and o is the simulated impulse response.

3.5 Summary

In this chapter, we have discussed the image method and compared it to actual measured impulse responses, showing that the image method gives a good approximation to the measured impulse response in terms of reverberation time. We also described how the reverberation time is measured and why it is an important quantity for room acoustics.

CHAPTER IV

SPEECH DEREVERBERATION BASED ON THE RMRE METHOD

In this chapter, we discuss the dereverberation approach based on the mutually referenced equalizers (MRE) method and the reduced MRE (RMRE) method. We point out the limitations of the MRE method and why modifications are necessary. Experimental results are provided to demonstrate the applicability of the RMRE approach to the dereverberation problem. Finally, we provide some theoretical results that shed light on the RMRE method and possible ways of further reducing computational cost.

4.1 *Multiple FIR Channel Inversion*

Before discussing the MRE method, we give a brief exposition of the underlying principle of single-input multiple-output (SIMO) FIR channel equalization (inversion). The solution to the problem depends on finding a solution to the Bezout equation, which states that given two polynomials $H_1(z)$ and $H_2(z)$, there exists a pair of polynomials $G_1(z)$ and $G_2(z)$ such that $H_1(z)G_1(z) + H_2(z)G_2(z) = 1$, under the assumption that no common zeros exist between $H_1(z)$ and $H_2(z)$, i.e. that they are coprime. The relationship is extendable to L polynomials, and the generalized Bezout equation in this case is given by

$$H_1(z)G_1(z) + \dots + H_L(z)G_L(z) = z^{-d}, \quad (10)$$

where d is a positive integer. A proof of Equation (10) is given in [74]. One remarkable feature of the Bezout equation is that it does not require that the roots of the polynomials $H_i(z)$, $i = 1, 2$ lie inside the unit circle (no minimum-phase requirement), as is required in the single channel (SISO) case.

In the case of two microphones, let the (FIR) impulse response between the speaker and the microphones be denoted by the polynomials $H_1(z)$ and $H_2(z)$, and the order of

both polynomials be M , i.e. each channel is modeled as an FIR filter with $M + 1$ taps. The Bezout equation can be written in this case as a dot product of two vector polynomials given by:

$$\begin{bmatrix} G_1(z) & G_2(z) \end{bmatrix} \begin{bmatrix} H_1(z) \\ H_2(z) \end{bmatrix} = z^{-d}. \quad (11)$$

In the time domain, the relation in Equation (11) can be written as a linear system of equations given by:

$$[\mathbf{H}_1, \mathbf{H}_2] \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \mathbf{e}_d,$$

where the matrices $\mathbf{H}_i, i = 1, 2$ are convolution matrices

$$\mathbf{H}_i = \begin{bmatrix} h_i(0) & & & & \\ \vdots & h_i(0) & & & \\ h_i(M) & \vdots & \ddots & & \\ & h_i(M) & & h_i(0) & \\ & & & \vdots & \\ & & & & h_i(M) \end{bmatrix}_{(M+N) \times N},$$

and the vector \mathbf{e}_d is a unit vector with a 1 in the d -th position. One point that we have not yet specified is: what is the minimum required order for the polynomials $G_i(z), i = 1, 2$? It can be shown that the minimum required order is $M - 1$, i.e. the FIR filters $g_1(n), g_2(n)$ have to be one point shorter than the impulse responses (channels). A formal proof of this fact and the more general L -channel case requires some knowledge of matrix polynomial properties, but we provide here our alternative (and simpler) plausibility demonstration based solely on the rank properties of matrices.

Lemma 1: Given L channels $H_i(z)$ each of order M , the minimum equalizer $G_i(z)$ order is:

$$\left\lceil \frac{M-1}{L-1} \right\rceil.$$

Proof: The Bezout equation in the time domain has the form

$$[\mathbf{H}_1, \dots, \mathbf{H}_L] \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_L \end{bmatrix} = \mathbf{e}_d \quad (12)$$

$$\mathbf{H}_{(M+N) \times LN} \mathbf{g} = \mathbf{e}_d,$$

where the rank of \mathbf{H} is $M + N$. Multiplying by \mathbf{H}^T , we obtain

$$\mathbf{H}^T \mathbf{H}_{(M+N) \times LN} \mathbf{g} = \mathbf{H}^T \mathbf{e}_d, \text{ or equivalently}$$

$$R_{(LN \times LN)} \mathbf{g} = \mathbf{h}_d \quad .$$

Since the rank of the deterministic correlation matrix R is $LN - (M + N)$, this implies the solution \mathbf{g} will have $LN - (M + N) + 1$ zero elements. Note that N (a user defined parameter) must be chosen to satisfy:

$$N \geq \left\lceil \frac{M-1}{L-1} \right\rceil ,$$

so that equation (12) will have a solution. ■

It is possible to find equalizers of higher order that are also solutions, but in this case, the solution is no longer unique. As a simple example, consider $H_1(z) = 1 + 8z^{-1} + 4z^{-2} - 2z^{-3}$, $H_2(z) = -3 + 3z^{-1} - 7z^{-2} + z^{-3}$. The roots of $H_1(z), H_2(z)$ are $[-7.43, -0.88, 0.31]$ and $[0.42 + 1.42i, 0.42 - 1.42i, 0.15]$ respectively. Then the solution to Equation (11) for $d = 0$ is:

$$[G_1(z), G_2(z)] = \begin{bmatrix} 0.092 - 0.24z^{-1} + 0.034z^{-2}, & -0.3 - 0.14z^{-1} + 0.068z^{-2} \end{bmatrix} .$$

As noted, higher order solutions exist, one example of which is given by:

$$\begin{bmatrix} G_1(z) \\ G_2(z) \end{bmatrix}^T = \begin{bmatrix} 0.09 - 0.23z^{-1} + 0.022z^{-2} + 0.019z^{-3} - 0.0027z^{-4} \\ -0.3 - 0.14z^{-1} + 0.087z^{-2} + 0.012z^{-3} - 0.0054z^{-4} \end{bmatrix}^T .$$

It is generally desirable to have the lowest order solution, but in some circumstances higher order solutions provide robustness against model order mismatches. We discuss this issue in the next chapter.

It is important to emphasize that in the two channel example, the polynomials $G_1(z), G_2(z)$ must work as a pair to produce the desired impulse at a given delay d . Applying $G_1(z)$ to $H_1(z)$ alone does not yield a useful result. For equalization purposes, a result of the form $\alpha\delta(n - n_0)$, for a nonzero scale factor α , would be still considered as an exact equalizer since the result is only scaled and delayed. In the next sections we explain the basics of the MRE method and point out to its advantages and limitations for the dereverberation problems.

4.2 The Mutually Referenced Equalizers Method

The mutually referenced equalizers (MRE) method [59] is a method developed for the direct blind equalization of wireless communication channels modeled as FIR filters. As with many of the SIMO deconvolution methods, it only uses second-order statistics of the channel outputs. The method is based on the formulation of a constrained multidimensional MSE criterion. Minimizing this criterion gives the equalizers for all possible delays.

What makes the method applicable to the dereverberation problem is the fact that it makes very mild assumptions about the nature of the input signal and channels. Although the problem is formulated by assuming a PAM/QAM channel, the method does not assume that the input signal is white. This feature makes the algorithm applicable to speech signals. Additionally, the ability to implement the MRE method adaptively makes it ideal for situations where the channels or input are time varying.

If we assume the existence of L channels, and that each channel is modeled as an FIR filter of order M , as depicted in Figure 27, then the output of the channels at time n is given by the linear convolutional model in Equation (13)

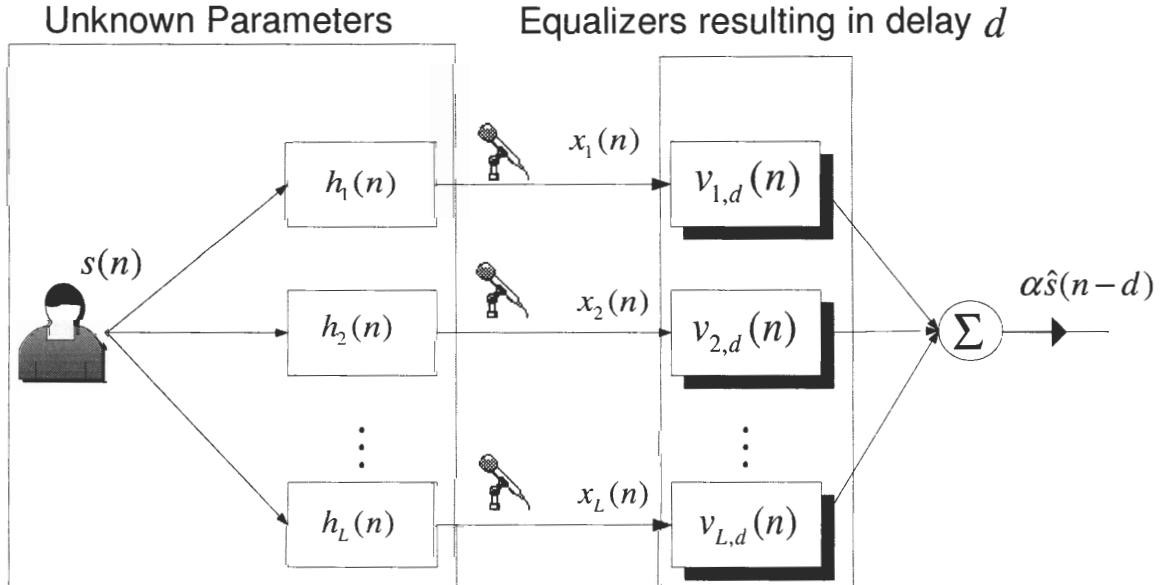


Figure 27: SIMO block diagram showing unknown channels, the input $s(n)$ and equalizers.

$$\begin{bmatrix} \mathbf{x}_1(n) \\ \vdots \\ \mathbf{x}_L(n) \end{bmatrix} = \begin{bmatrix} H_1 \\ \vdots \\ H_L \end{bmatrix} \mathbf{s}(n) \equiv \mathbf{x}_n = H\mathbf{s}(n) \quad (13)$$

where

$$\mathbf{x}_i(n) = \begin{bmatrix} x_i(n) \\ \vdots \\ x_i(n - N + 1) \end{bmatrix}, \mathbf{s}(n) = \begin{bmatrix} s(n) \\ \vdots \\ x_i(n - K + 1) \end{bmatrix},$$

$$H_i = \begin{bmatrix} h_i(0) & \cdots & h_i(M) & & \\ & \ddots & \ddots & \ddots & \\ & & h_i(0) & \cdots & h_i(M) \end{bmatrix}.$$

The matrix H is a composite channel matrix (usually referred to as a Sylvester matrix) of dimension $LN \times K$ where $K = M + N$ and N is a user defined parameter that defines the output's data window size. Notice how the channel order M is implicitly defined by K . As discussed in the previous section, it is possible to find a set of equalizers, i.e. inverses, denoted by $v_{i,d}(n)$ in Figure 27. If these inverses are found, the resulting output will be a delayed and attenuated version of $s(n)$ denoted by $\alpha \hat{s}(n - d)$.

To make the above definitions more concrete and help us explain the development of the MRE method, we consider the simple example of a two microphone system ($L = 2$) with each channel being three taps long ($M = 2$). We define the output over a window of $N = 3$ samples which implies that the number of columns is $K = M + N = 5$. The system

output in this case is given by

$$\begin{bmatrix} \underbrace{\begin{bmatrix} x_1(n) \\ x_1(n-1) \\ x_1(n-2) \end{bmatrix}}_{\text{Channel 1 output}} \\ \underbrace{\begin{bmatrix} x_2(n) \\ x_2(n-1) \\ x_2(n-2) \end{bmatrix}}_{\text{Channel 2 output}} \end{bmatrix}_{[6 \times 1]} = \begin{bmatrix} \underbrace{\begin{bmatrix} h_1(0) & h_1(1) & h_1(2) & 0 & 0 \\ 0 & h_1(0) & h_1(1) & h_1(2) & 0 \\ 0 & 0 & h_1(0) & h_1(1) & h_1(2) \end{bmatrix}}_{\text{Channel 1}} \\ \underbrace{\begin{bmatrix} h_2(0) & h_2(1) & h_2(2) & 0 & 0 \\ 0 & h_2(0) & h_2(1) & h_2(2) & 0 \\ 0 & 0 & h_2(0) & h_2(1) & h_2(2) \end{bmatrix}}_{\text{Channel 2}} \end{bmatrix}_{[6 \times 5]} \underbrace{\begin{bmatrix} s(n) \\ s(n-1) \\ s(n-2) \\ s(n-3) \\ s(n-4) \end{bmatrix}}_{\text{Input signal}} \quad (14)$$

To be able to find a left inverse for H , it is required that H be of full column rank, i.e. that the number of rows be greater than or equal the number of columns. The full column rank condition is another way of stating the coprimeness condition since a Sylvester matrix is of full column rank if the polynomials generating it are coprime. This makes picking the data window size N critical and closely tied to knowing the order of the channels. If N is chosen correctly as prescribed in the proof of Lemma 1, then the channel matrix is overdetermined and it is possible to find an inverse (a pseudoinverse). Notice that in a single channel deconvolution problem, the channel matrix would be underdetermined (corresponding to a short and wide matrix). We next describe how the MRE method can obtain the equalizers (inverse) using only knowledge of \mathbf{x}_n .

Denote the (pseudo)inverse of H by the matrix $V^T = H^\dagger$. If H is $m \times n$, then V^T will be $n \times m$ and the product of $V^T H$ will be equal to $I_{n \times n}$, where I is the identity matrix. Thus, V^T acts as what is commonly referred to in communication theory as a zero-forcing equalizer [75]. If we denote each row of V^T by $\mathbf{v}_i^T, i = 0 : m - 1$ then the product can be written as in Equation (15):

$$\mathbf{v}_i^T H = \begin{bmatrix} \mathbf{v}_{1,i}^T & \dots & \mathbf{v}_{L,i}^T \end{bmatrix} \begin{bmatrix} H_1 \\ \vdots \\ H_L \end{bmatrix} = \mathbf{e}_i^T, \quad (15)$$

where \mathbf{e}_i is an impulse vector with a 1 at the i -th position, and $\mathbf{v}_{j,i}^T, j = 1 : L$ are the

individual components of the equalizers as illustrated in Figure 27. This important idea of considering the various rows of V^T to be equalizers with different delays is illustrated in Figure 28.

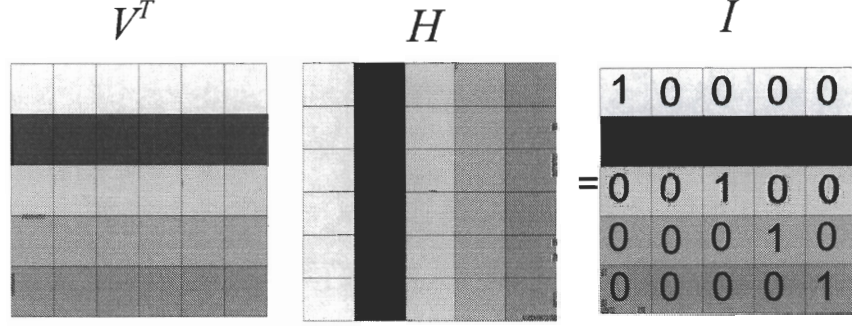


Figure 28: Rows of the inverse can be thought of as equalizers.

The set of $K - 1$ equalizers can be *referenced* with respect to each other to form a set error relations of the form:

$$J_i = \left(\underbrace{\mathbf{v}_i^T H \mathbf{s}(n)}_{\mathbf{e}_i^T} - \underbrace{\mathbf{v}_{i+1}^T H \mathbf{s}(n+1)}_{\mathbf{e}_{i+1}^T} \right)^2, \quad (16)$$

each of which will be zero if the equalizers are exact. If the various error functions for the equalizers are incorporated into a sum of squares of the form

$$J_{MRE} = J_1 + \dots + J_{K-1}, \quad (17)$$

then minimizing J_{MRE} will yield a set of optimal equalizers. This simple, yet powerful, point of view of the equalizers, and the fact that the input is a time shifted version of a common signal from one sample to the next lies at the heart of the MRE method.

Building on our previous example, V^T is given by

$$V^T = H_{[5 \times 6]}^\dagger = \begin{bmatrix} [\mathbf{v}_0^T]_{[1 \times 6]} \\ \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \mathbf{v}_3^T \\ \mathbf{v}_4^T \end{bmatrix}. \quad (18)$$

Since the matrix V^T is an inverse for H , the product $V^T H = I$. Thus, $[\mathbf{v}_0^T]_{[1 \times 6]} H = [1, 0, 0, 0, 0]$ which implies that \mathbf{v}_0 is the zero delay equalizer. This extends to $[\mathbf{v}_1^T]_{[1 \times 6]} H = [0, 1, 0, 0, 0]$ and so on. In general \mathbf{v}_j^T is the j^{th} delay equalizer. The goal of the MRE method to find these equalizers in H^\dagger from knowledge only of the outputs \mathbf{x}_n . The key to the solution is to realize that $\mathbf{v}_0^T \mathbf{x}_n = \mathbf{v}_0^T H \mathbf{s}_n = [1, 0, 0, 0, 0] \mathbf{s}_n = s(n)$ and $\mathbf{v}_1^T \mathbf{x}_{n+1} = \mathbf{v}_1^T H \mathbf{s}_{n+1} = [0, 1, 0, 0, 0] \mathbf{s}_{n+1} = s(n)$. Subtracting these two relations yields $\mathbf{v}_0^T \mathbf{x}_n - \mathbf{v}_1^T \mathbf{x}_{n+1} = 0$. Additional relationships can found by noting that $\mathbf{v}_1^T \mathbf{x}_n - \mathbf{v}_2^T \mathbf{x}_{n+1} = 0$, $\mathbf{v}_2^T \mathbf{x}_n - \mathbf{v}_3^T \mathbf{x}_{n+1} = 0$, $\mathbf{v}_3^T \mathbf{x}_n - \mathbf{v}_4^T \mathbf{x}_{n+1} = 0$. The result is a total of 4 equations. Thus, the MRE criterion (which is the sum of the MSE of the above difference relations) is given by:

$$\begin{aligned}
J_{MRE}(\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4) &= E |\mathbf{v}_0^T \mathbf{x}[n] - \mathbf{v}_1^T \mathbf{x}[n+1]|^2 + E |\mathbf{v}_1^T \mathbf{x}[n] - \mathbf{v}_2^T \mathbf{x}[n+1]|^2 \\
&+ E |\mathbf{v}_2^T \mathbf{x}[n] - \mathbf{v}_3^T \mathbf{x}[n+1]|^2 + E |\mathbf{v}_3^T \mathbf{x}[n] - \mathbf{v}_4^T \mathbf{x}[n+1]|^2 \\
&= [\mathbf{v}_0^T R_x[0] \mathbf{v}_0 + \mathbf{v}_1^T R_x[0] \mathbf{v}_1 - 2\mathbf{v}_0^T R_x[1] \mathbf{v}_1] \\
&+ [\mathbf{v}_1^T R_x[0] \mathbf{v}_1 + \mathbf{v}_2^T R_x[0] \mathbf{v}_2 - 2\mathbf{v}_1^T R_x[1] \mathbf{v}_2] \\
&+ [\mathbf{v}_2^T R_x[0] \mathbf{v}_2 + \mathbf{v}_3^T R_x[0] \mathbf{v}_3 - 2\mathbf{v}_2^T R_x[1] \mathbf{v}_3] \\
&+ [\mathbf{v}_3^T R_x[0] \mathbf{v}_3 + \mathbf{v}_4^T R_x[0] \mathbf{v}_4 - 2\mathbf{v}_3^T R_x[1] \mathbf{v}_4].
\end{aligned} \tag{19}$$

If the quadratic cost function above is differentiated w.r.t. each equalizer \mathbf{v}_j and each derivative is set to zero, the resulting linear system of equations take the form

$$\underbrace{\begin{bmatrix} R_x[0] & -R_x^T[1] & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -R_x[1] & 2R_x[0] & -R_x^T[1] & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -R_x[1] & 2R_x[0] & -R_x^T[1] & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -R_x[1] & 2R_x[0] & -R_x^T[1] \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -R_x[1] & R_x[0] \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} \mathbf{v}_0 \\ \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \mathbf{v}_4 \end{bmatrix}}_{\mathbf{v}} = \mathbf{0}. \tag{20}$$

This linear system is a block tridiagonal system that only depends on the lag zero and lag one correlation matrices. If we assume that \mathbf{s}_n is white noise for the time being, then we

can derive a simple expressions for $R_x[0], R_x[1]$;

$$\begin{aligned}
R_x[0]_{[6 \times 6]} &= E[\mathbf{x}_n \mathbf{x}_n^T] = E[H \mathbf{s}_n \mathbf{s}_n^T H^T] = H H^T \\
R_x[1]_{[6 \times 6]} &= E[\mathbf{x}_{n+1} \mathbf{x}_n^T] = E[H \mathbf{s}_{n+1} \mathbf{s}_n^T H^T] = H \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} H^T.
\end{aligned} \tag{21}$$

In general the k^{th} lag correlation matrix can be written as $R[k] = H J^k H^T$ and $R[-k] = R^T[k]$. Notice that the backward shift matrix J of size $p \times p$ is zero for $k > p - 1$ (nilpotent matrix). Thus, the equalizers can be found by utilizing only outputs of the channels to form the block correlation matrix.

4.3 The Reduced MRE Method

While the MRE method is well suited for wireless channels that are often modeled as FIR filters with a small number of taps, the MRE method is not very well suited for longer channels. The reason for this is that the MRE method finds the equalizers for all possible delays. While this feature makes the solution more robust against noise as explained in [59, 64], it makes the solution too computationally expensive. For example, two channels each with 100 taps would mean that the minimum size for the channel matrix H is 198×198 . This means the total number of taps required is $198^2 = 39204$. Clearly, this makes the MRE method as originally proposed impractical for longer channels.

For the acoustical dereverberation problem, it would be sufficient to have only a subset of all possible equalizers. The simplest approach would be to reformulate the MRE criterion with a reduced number of equalizers. In our illustrative example, let us assume we are only interested in obtaining the zero delay equalizer and the lag one equalizer. This modifies the MSE criterion to have a single difference relation

$$J_{MRE}(\mathbf{v}_0, \mathbf{v}_1) = E |\mathbf{v}_0^T \mathbf{x}[n] - \mathbf{v}_1^T \mathbf{x}[n+1]|^2. \quad (22)$$

Expanding this equation and differentiating w.r.t to each equalizer yields:

$$\begin{bmatrix} R_x[0] & -R_x^T[1] \\ -R_x[1] & R_x[0] \end{bmatrix} \begin{bmatrix} \mathbf{v}_0 \\ \mathbf{v}_1 \end{bmatrix} = \mathbf{0}. \quad (23)$$

While a solution to the above system exists, it does not correspond to an equalizer for the complete channel matrix H . The interesting thing about the solution that comes from this reduced MRE criterion is the structure of the result:

$$\begin{bmatrix} \mathbf{v}_0^T \\ \mathbf{v}_1^T \end{bmatrix} H.$$

In the ideal case, the result would be two scaled impulses $\delta(n)$ and $\delta(n-1)$. For the reduced MRE however, the solution is more similar to:

$$\begin{bmatrix} \mathbf{v}_0^T \\ \mathbf{v}_1^T \end{bmatrix} H = \begin{bmatrix} a & b & c & d & e & f & \cdots & g & 0 \\ 0 & a & b & c & d & e & f & \cdots & g \end{bmatrix},$$

where the elements on each diagonal have the same value. As a simple example consider a two channel system defined by

$$H = \begin{bmatrix} 1 & -1 & 5 & 0 & 0 \\ 0 & 1 & -1 & 5 & 0 \\ 0 & 0 & 1 & -1 & 5 \\ 3 & -2 & -2 & 0 & 0 \\ 0 & 3 & -2 & -2 & 0 \\ 0 & 0 & 3 & -2 & -2 \end{bmatrix}.$$

Then a solution to (23) yields

$$H_{[2 \times 6]}^\dagger H_{[6 \times 5]} = \begin{bmatrix} 0.1071 & -0.4200 & 0.6922 & -0.5247 & -0.0000 \\ -0.0000 & 0.1071 & -0.4200 & 0.6922 & -0.5247 \end{bmatrix}.$$

The above solution is unacceptable as an equalizer because it would mean each recovered sample $\hat{s}(n)$ would be the result of FIR filtering the original speech signal with an FIR filter with coefficients $[a, b, \dots, g]$, and note that these coefficients are unknown when the equalizers are calculated. Hence, the problem here is that we have multiple diagonals when we desire only two impulses as defined by

$$H_{[2 \times 6]}^\dagger H_{[6 \times 5]} = \alpha \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

The problem can be traced to the lack of constraints in the reduced criterion. To understand this better consider

$$\mathbf{s}_n = \begin{bmatrix} \boxed{\begin{matrix} s_n \\ s_{n-1} \\ s_{n-2} \\ s_{n-3} \end{matrix}} \\ s_{n-4} \end{bmatrix} \quad \mathbf{s}_{n+1} = \begin{bmatrix} s_{n+1} \\ \boxed{\begin{matrix} s_n \\ s_{n-1} \\ s_{n-2} \\ s_{n-3} \end{matrix}} \end{bmatrix},$$

where the common samples are boxed. The goal is to find a \mathbf{v}_0 and \mathbf{v}_1 such that

$$\mathbf{v}_0^T H \begin{bmatrix} \boxed{s_n} \\ \boxed{s_{n-1}} \\ \boxed{s_{n-2}} \\ \boxed{s_{n-3}} \\ s_{n-4} \end{bmatrix} = \mathbf{v}_1^T H \begin{bmatrix} s_{n+1} \\ \boxed{s_n} \\ \boxed{s_{n-1}} \\ \boxed{s_{n-2}} \\ \boxed{s_{n-3}} \end{bmatrix}.$$

Clearly one solution would be for $\mathbf{v}_0^T H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$ and $\mathbf{v}_1^T H = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix}$. However, another possibility is for $\mathbf{v}_0^T H = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \end{bmatrix}$ and $\mathbf{v}_1^T H = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \end{bmatrix}$. In other words, because we have a set of shared samples, the solution is not unique.

Our initial approach to deal with this problem was to reduce the number of common samples. The simplest way to do this would be to consider

$$\mathbf{s}_n = \begin{bmatrix} s_n \\ s_{n-1} \\ s_{n-2} \\ s_{n-3} \\ s_{n-4} \end{bmatrix}, \mathbf{s}_{n+4} = \begin{bmatrix} s_{n+4} \\ s_{n+3} \\ s_{n+2} \\ s_{n+1} \\ s_n \end{bmatrix}.$$

The only possible solution now is for $\mathbf{v}_0^T H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$ and $\mathbf{v}_4^T H = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix}$. Now we it is possible to obtain the edge equalizers without the problem of multiple diagonals, an idea hinted to by [59]. The modified error criterion for the previous example problem is $J_{MRE}(\mathbf{v}_0, \mathbf{v}_4) = E |\mathbf{v}_0^T \mathbf{x}[n] - \mathbf{v}_4^T \mathbf{x}[n+4]|^2$. Once again by expanding and differentiating we get a system of linear equations given by

$$\begin{bmatrix} R_x[0] & -R_x^T[4] \\ -R_x[4] & R_x[0] \end{bmatrix} \begin{bmatrix} \mathbf{v}_0 \\ \mathbf{v}_4 \end{bmatrix} = \mathbf{0}.$$

In the general problem, the problem of multiple diagonals is eliminated if we consider the input samples at $\mathbf{s}(n)$ and $\mathbf{s}(n+K-1)$, since these share only the sample at time n . In such a case, the edge equalizers (delay zero and delay $K-1$) can be found and these form a reduced set of equalizers as stated in Lemma 2.

Lemma 2: Assume that K is known (or estimated) and assume that $\mathbf{v}_0^T \mathbf{x}_n = \mathbf{v}_{K-1}^T \mathbf{x}_{n+(K-1)}$ holds for all n . Defining the reduced error criterion

$$J_{RMRE}(\mathbf{v}_0, \mathbf{v}_{K-1}) = E \left| \mathbf{v}_0^T \mathbf{x}[n] - \mathbf{v}_{K-1}^T \mathbf{x}[n + K - 1] \right|^2,$$

and minimizing this criterion yields the linear system

$$\begin{bmatrix} R_x[0] & -R_x^T[K-1] \\ -R_x[K-1] & R_x[0] \end{bmatrix} \begin{bmatrix} \mathbf{v}_0 \\ \mathbf{v}_{K-1} \end{bmatrix} = \mathbf{0}.$$

The equalizers $\mathbf{v}_0^T, \mathbf{v}_{K-1}^T$ will be the zero delay and delay $K-1$ equalizers such that,

$$\underbrace{\begin{bmatrix} \mathbf{v}_0^T \\ \mathbf{v}_{K-1}^T \end{bmatrix}}_{V^T}_{[2 \times LN]} [H]_{[LN \times K]} = \alpha \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}_{[2 \times K]}, \quad (24)$$

where α is nonzero a scale factor.

Proof:

Let

$$\tilde{S}_n = [s_{n+(K-1)}, s_{n+(K-2)}, \dots, s_n, s_{n-1}, \dots, s_{n-(K-1)}]^T,$$

and note that the difference relationship

$$\mathbf{v}_0^T \mathbf{x}[n] - \mathbf{v}_{K-1}^T \mathbf{x}[n + K - 1]$$

may be written as

$$\begin{bmatrix} 1 & 0 \end{bmatrix} V^T \mathbf{x}_n = \begin{bmatrix} 0 & 1 \end{bmatrix} V^T \mathbf{x}_{n+(K-1)}.$$

Expanding this expression yields

$$\begin{aligned} \begin{bmatrix} 1 & 0 \end{bmatrix} V^T H \mathbf{s}_n &= \begin{bmatrix} 0 & 1 \end{bmatrix} V^T H \mathbf{s}_{n+(K-1)} \\ \begin{bmatrix} 1 & 0 \end{bmatrix} V^T H \begin{bmatrix} \mathbf{0}_{K-1} & I_K \end{bmatrix} \tilde{S}_n &= \begin{bmatrix} 0 & 1 \end{bmatrix} V^T H \begin{bmatrix} I_K & \mathbf{0}_{K-1} \end{bmatrix} \tilde{S}_n. \end{aligned}$$

We omit \tilde{S}_n since it appears on both sides of the above equation, and we can consider only

$$\begin{aligned} \begin{bmatrix} 1 & 0 \end{bmatrix} V^T H \begin{bmatrix} \mathbf{0}_{K-1} & I_K \end{bmatrix} &= \begin{bmatrix} 0 & 1 \end{bmatrix} V^T H \begin{bmatrix} I_K & \mathbf{0}_{K-1} \end{bmatrix} \\ \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1K} \end{bmatrix} \begin{bmatrix} \mathbf{0}_{K-1} & I_K \end{bmatrix} &= \begin{bmatrix} W_{21} & W_{22} & \cdots & W_{2K} \end{bmatrix} \begin{bmatrix} I_K & \mathbf{0}_{K-1} \end{bmatrix} \end{aligned}$$

which is satisfied if

$$\begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1K} \\ W_{21} & W_{22} & \cdots & W_{2K} \end{bmatrix}$$

has the form

$$\alpha \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix}. \blacksquare$$

The proof of Lemma 2 is similar in construction to the general case given in [59], and Lemma 2 shows that instead of needing to obtain all the equalizers, we can obtain a partial set consisting of the two edge equalizers. While this approach seems promising in terms of reducing the overall number of required equalizers, it suffers from a lack of robustness. The main problem is that the edge equalizers are very susceptible to noise and the nonwhite nature of the input signal. The reason for this is that they rely on the lag zero correlation matrix and lag $K - 1$ correlation matrix and there are fewer samples available to estimate $R[K - 1]$. To overcome this problem, we propose that for the speech dereverberation problem, to use an MRE criterion with three equalizers corresponding to two edge equalizers and a “middle” (central) equalizer $\mathbf{v}_{K/2}^T$ as proposed by us in [76]. Additionally, in Chapter 5, we will see that the middle equalizer tend to cause less noise amplification resulting in reconstructed speech signals with less hissing. Thus, by using three equalizers instead of two, it is possible to obtain the “middle” equalizer with additional robustness added to the problem. Of course the disadvantage is that we have increased the number of required equalizers from two to three.

From our previous example, if we only seek the equalizers $\mathbf{v}_0^T, \mathbf{v}_{K/2}^T, \mathbf{v}_{K-1}^T$, then the reduced MRE (RMRE) criterion can be formulated as:

$$J_{RMRE}(\mathbf{v}_0, \mathbf{v}_{K/2}, \mathbf{v}_K) = E(|\mathbf{v}_0^T \mathbf{x}[n] - \mathbf{v}_{K/2}^T \mathbf{x}[K/2]|^2 + |\mathbf{v}_{K/2}^T \mathbf{x}[K/2] - \mathbf{v}_{K-1}^T \mathbf{x}[K-1]|^2). \quad (25)$$

The resulting block correlation matrix R in this case is given by:

$$\begin{bmatrix} R_x[0] & -R_x^T[\frac{K}{2}] & \mathbf{0} \\ -R_x[\frac{K}{2}] & 2R_x[0] & -R_x^T[\frac{K}{2}] \\ \mathbf{0} & -R_x[\frac{K}{2}] & R_x[0] \end{bmatrix} \quad (26)$$

Similar to Lemma 2, Lemma 3 demonstrates minimizing Equation (25) yields the two edge equalizers and a central equalizer. The proof is similar in construction to that of Lemma 2.

Lemma 3. Consider the reduced MRE (RMRE) criterion defined in (25), and the resulting block correlation matrix R in this case is given by Equation (26). Minimizing this criterion will result in equalizers such that:

$$\begin{bmatrix} \mathbf{v}_0^T \\ \mathbf{v}_{K/2}^T \\ \mathbf{v}_{K-1}^T \end{bmatrix} H = \alpha \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \cdots & 0 & 1 & 0 & \cdots \\ 0 & \cdots & 0 & 1 \end{bmatrix},$$

where α nonzero scale factor.

Proof.

Let $\tilde{S}(n)$ be defined as in Lemma 2, and define $\tilde{S}_2(n) = [s_{n+(K-2)}, \dots, s_{n+K}, \tilde{S}(n)]$.

The second difference relation can be written as:

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} V^T H \mathbf{s}(n) = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} V^T H \mathbf{s}(n + (K)/2)$$

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} V^T H \begin{bmatrix} \mathbf{0}_{K-2} & \mathbf{0}_{K-1} & I_K \end{bmatrix} \tilde{S}_2(n) = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} V^T H \begin{bmatrix} \mathbf{0}_{K-2} & I_K & \mathbf{0}_{K-1} \end{bmatrix} \tilde{S}_2(n).$$

The same formulation can be applied to the first difference relation to obtain:

$$\begin{bmatrix} 0 & 1 & 0 \end{bmatrix} V^T H \begin{bmatrix} \mathbf{0}_{K-2} & I_K & \mathbf{0}_{K-1} \end{bmatrix} \tilde{S}_2(n) = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} V^T H \begin{bmatrix} I_K & \mathbf{0}_{K-1} & \mathbf{0}_{K-2} \end{bmatrix} \tilde{S}_2(n).$$

Omitting the dependence on $\tilde{S}(n)$ we obtain the following pair of relations:

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} V^T H \begin{bmatrix} \mathbf{0}_{K-2} & \mathbf{0}_{K-1} & I_K \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} V^T H \begin{bmatrix} \mathbf{0}_{K-2} & I_K & \mathbf{0}_{K-1} \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 & 0 \end{bmatrix} V^T H \begin{bmatrix} \mathbf{0}_{K-2} & I_K & \mathbf{0}_{K-1} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} V^T H \begin{bmatrix} I_K & \mathbf{0}_{K-1} & \mathbf{0}_{K-2} \end{bmatrix},$$

which are satisfied only if

$$\begin{bmatrix} \mathbf{v}_0^T \\ \mathbf{v}_{K/2}^T \\ \mathbf{v}_{K-1}^T \end{bmatrix} H$$

has the form

$$\alpha \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \cdots & 0 & 1 & 0 & \cdots \\ 0 & \cdots & 0 & 1 \end{bmatrix} \cdot \blacksquare$$

As in the case of the full adaptive MRE formulation, it is possible implement the RMRE method adaptively. We first assume that the input signal $s(n)$ satisfies the property that $E[s(n)s(n+m)] = 0$ for some integer m . For the case of white noise as input, it clearly satisfied for $m = 1$. For a nonwhite persistently exciting input signal, such as speech, the relation $E[s(n)s(n+m)] = 0$ will be true for some “relatively” large value of m . Also, we remind the reader that typical room impulse responses are at least hundreds of taps long, which implies that the value of $K = M + N$ will be larger than the channel order. The first step towards an adaptive formulation of the RMRE is to find and define a new vector \mathbf{z}_n such that

$$E[\mathbf{z}_n \mathbf{z}_n^T] = \begin{bmatrix} R_x[0] & -R^T[\frac{K}{2}] & \mathbf{0} \\ -R[\frac{K}{2}] & 2R_x[0] & -R^T[\frac{K}{2}] \\ \mathbf{0} & -R[\frac{K}{2}] & R_x[0] \end{bmatrix},$$

and this will be the case if \mathbf{z}_n is defined as

$$\mathbf{z}_n = \begin{bmatrix} \mathbf{x}_n \\ -\mathbf{x}_{n+\frac{K}{2}} + \mathbf{x}_{n+\frac{K}{2}+K} \\ \mathbf{x}_{n+2K} \end{bmatrix}.$$

While it is possible to implement the MRE and RMRE using the LMS algorithm, the performance and convergence rate will be very slow for long channels and nonwhite input. The optimal convergence rate is achieved with the RLS algorithm. As described in [59], an efficient way to implement is by using a linear prediction framework. This essentially is equivalent to constraining the first coefficient of the \mathbf{v}_0 equalizer to be one. To see this consider \mathbf{z}_n again, but this time \mathbf{x}_n has been partitioned into the first sample $x_1(n)$ and the remaining samples as $\tilde{\mathbf{x}}_n$

$$\mathbf{z}_n = \begin{bmatrix} x_1(n) \\ \tilde{\mathbf{x}}_n \\ -\mathbf{x}_{n+\frac{K}{2}} + \mathbf{x}_{n+\frac{K}{2}+K} \\ \mathbf{x}_{n+2K} \end{bmatrix} = \begin{bmatrix} x_1(n) \\ \tilde{\mathbf{z}}_n \end{bmatrix}$$

Thus if the error signal is $e(n)$ is defined to be $x_1(n) - \tilde{\mathbf{v}}^T \tilde{\mathbf{z}}_n$ in the RLS algorithm, and we see that it the same approach linear prediction follows. Another possibility would be the

use of a combination of the backward and forward prediction error terms like in the Burg method.

4.4 *Effect of the Number of Microphones*

One design advantage in the acoustical dereverberation problem is the ability to increase the number of microphones available. This adds minimal cost to the overall system. The advantage of doing so lies in the fact that as the number of channels L increases, the total number of taps required is reduced for certain values of L . In our previous two channel examples, given $H_1(z), H_2(z)$ each of order M , then each equalizer $V_{1,d}(z), V_{2,d}(z)$ for a given delay d had to be at least of order $M - 1$. Recall that it is the pair of equalizers working together according to the Bezout equation that permits $V_{1,d}(z)H_1(z) + V_{2,d}(z)H_2(z) = z^{-d}$ or in vector notation as $\begin{bmatrix} V_{1,d}(z) & V_{2,d}(z) \end{bmatrix} \begin{bmatrix} H_1(z) \\ H_2(z) \end{bmatrix} = z^{-d}$. As the number of channels increases, the number of taps in the equalizers is reduced for certain values of L . The reason for this is that the number of rows in the H matrix grows much faster (number rows $=LN$) than the number of columns (number of columns $K = M + N$) as new channels are added. Since L is increasing, the window length N can be made smaller. For our reduced RMRE criterion, the total number of taps in all three equalizers is given by

$$T = 3L \left\lceil \frac{M-1}{L-1} \right\rceil$$

where T is the total number of taps in all three equalizers, M is the channel order, and L is number of channels. In the two channel case, where each channel is of order M , equalizers $V_{1,d}(z), V_{2,d}(z)$ had to be at least of order $M - 1$ to satisfy

$$V_{1,d}(z)H_1(z) + V_{2,d}(z)H_2(z) = z^{-d}.$$

For a three channel setup

$$V_{1,d}(z)H_1(z) + V_{2,d}(z)H_2(z) + V_{3,d}(z)H_3(z) = z^{-d},$$

each filter $v_{i,d}$ is required to be $\left\lceil \frac{M-1}{2} \right\rceil$ taps long.

The limiting case is when the number of microphones equals the channel order, and in this case a single tap is allocated to each equalizer. The total number of equalizers versus

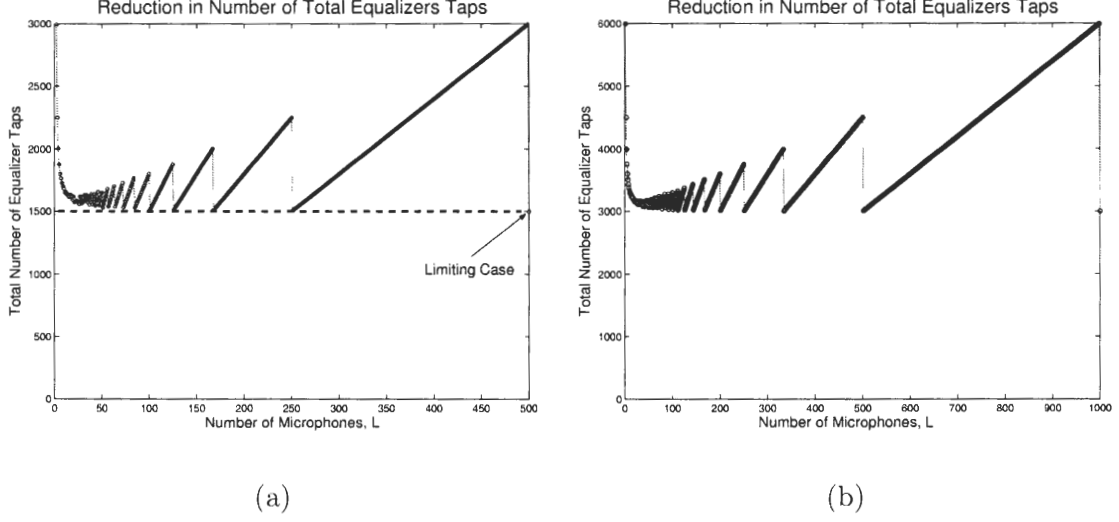


Figure 29: (a) Total number of equalizers taps versus number of channels for $M = 500$, (b) $M = 1000$.

the number of microphones is shown in Figure 29 for $M = 500$ and $M = 1000$. It can be seen that increasing the number of microphones to twenty, the value of T is close to the limiting case. Notice that certain values of L are no better than using two microphones in terms of reducing the total number of taps. For the $M = 500$ channels case, even in the case of four microphones, there is a one-third reduction in the total number of taps when compared to the two-microphone case. A similar conclusion can be drawn to the $M = 1000$ case, which leads us to suggest using no more than a dozen microphones in order to achieve a significant reduction in the total number of taps for typical room impulse responses.

4.5 Experiments and Simulation Results

In this section, we present a number of experimental results. The purpose of each experiment is to demonstrate some interesting point about the RMRE method or provide a demonstration of how the various parameters and conditions influence the algorithm. In all of these experiments, the sampling rate used is 8000Hz, and the input signal $s(n)$ is a clean speech signal recorded by a speaker that is very close to the microphone. The T_{60} time for the simulated channels, i.e. ones generated using the image method, is approximately 190ms, in keeping with the T_{60} time for Room 352.

4.5.1 Experiment A: Two Nonminimum Phase Synthetic Channels

The goal of the first experiment goal is to demonstrate the nonminimum phase response of the channels does not limit our ability to find a solution. In order to *exaggerate* the zeros' locations in the channels, two synthetic channels of order $M = 99$ are shown in Figure 30, along with their frequency response and pole-zero plots shown in Figures 31 and Figure 32, respectively. We use the term synthetic here to emphasize that the impulse responses in this experiment were *not* generated using the image method. The approach we used to generate the impulse responses in this experiment was by multiplying a decaying exponential signal $e^{-\alpha n}$ with a normally distributed random sequence with zero mean and unit variance. Since the channel order is $M = 99$, the equalizers will be specified to be the minimum order required, i.e equalizers have an order of 98. Since we are finding a total of three equalizer pairs $v_{1,d}, v_{2,d}$, the total number of unknown taps is $(3)(99 \times 2) = 594$.

Applying the RMRE adaptive RLS implementation, we obtain the learning curve shown in Figure 33. One notable feature of the RMRE approach (in the absence of noise and model order mismatches) is the bend, or knee, in the learning curve. We have found this behavior to be characteristic, and the bend occurs roughly after m iterations where m is the number of taps in all three equalizers. For example, in Figure 33, the curve starts at around sample iteration number 250 and the bend occurs at around sample 900, which implies the dip occurs after approximately 650 samples. The total number of all taps (in all three equalizers) is 594. The resulting equalizers are shown in Figure 34. In order to

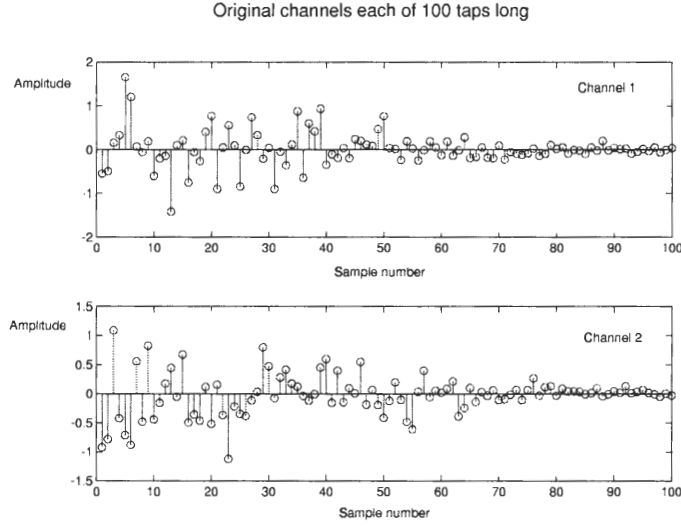


Figure 30: Impulse responses of two synthetic channels.

reduce the number of (sub)plots and emphasize that it is the equalizer pair that results in equalization, we denote \mathbf{v}_d as the composite delay d equalizer formed from the concatenation of $v_{1,d}$ and $v_{2,d}$. Thus, $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$ and in Figure 34, the first 99 samples in each subplot are $v_{1,d}, d = 0, 99, 198$ and the remaining samples belong to $v_{2,d}, d = 0, 99, 198$. The most interesting feature about the equalizers is that they have the same sort of decreasing amplitude shape as the channels. The result of applying the equalizers to the two channels is shown in Figure 35, and as expected, three scaled impulses at various delays are obtained.

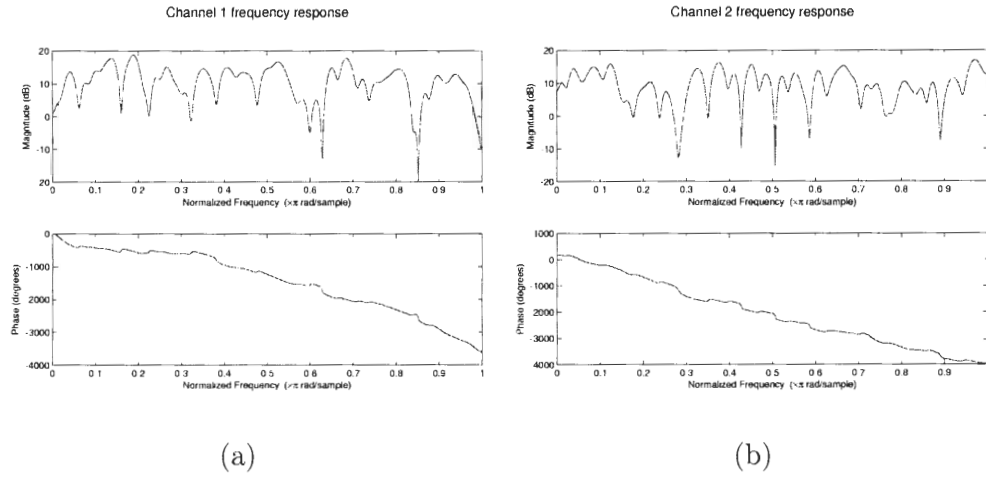


Figure 31: (a) Frequency Response of channel 1, (b) channel 2.

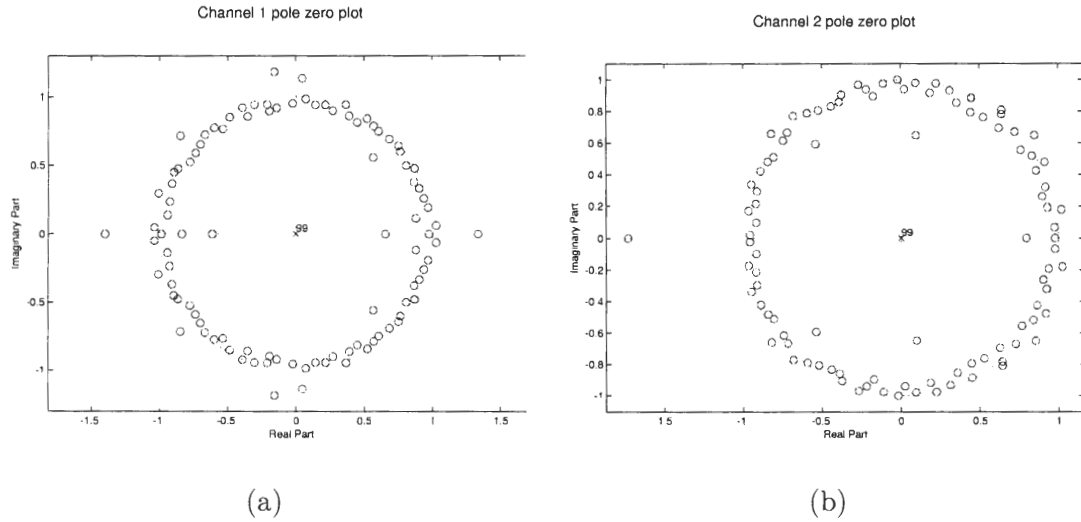


Figure 32: (a) Pole-zero plot of channel 1, (b) channel 2.

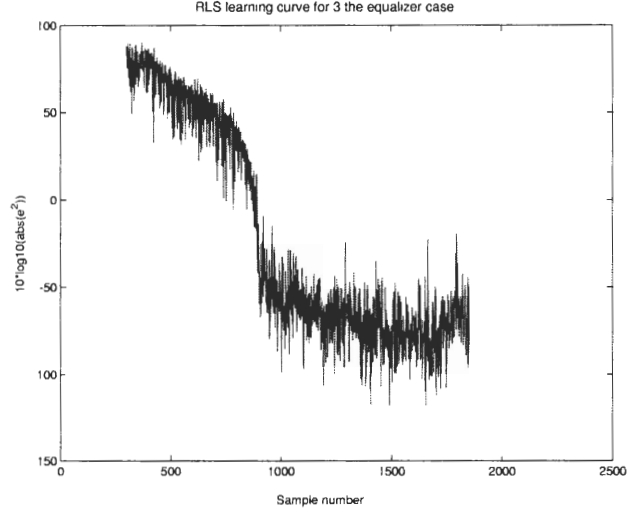


Figure 33: Learning curve based on RLS adaptive filter.

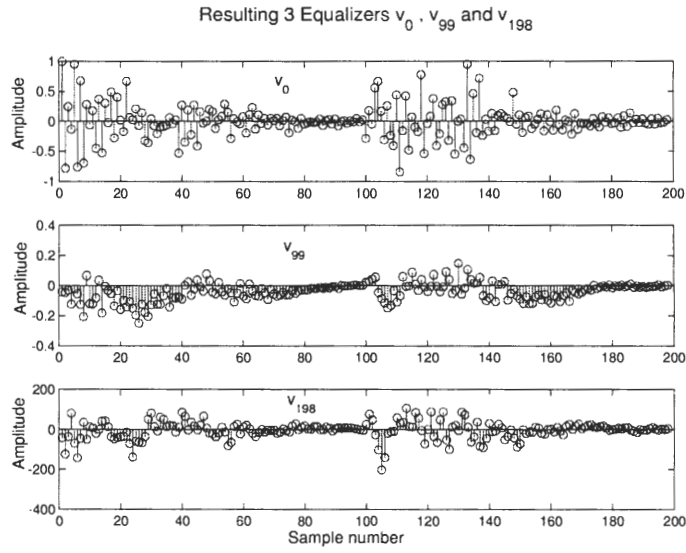


Figure 34: Resulting equalizers at various delays where $v_{1,d}, v_{2,d}$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 99, 198$. The first 99 samples in each subplot are $v_{1,d}, d = 0, 99, 198$ and the remaining samples belong to $v_{2,d}, d = 0, 99, 198$.

Result of applying 3 equalizers to the channel matrix H

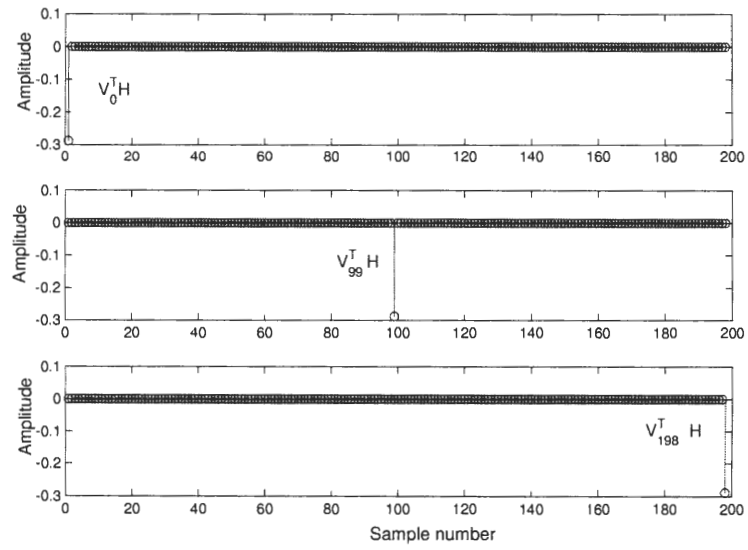


Figure 35: Result of applying the three different equalizer to the channels.

4.5.2 Experiment B: Two Simulated Channels of Order $M = 199$.

In this experiment, we utilize the image method to generate our impulse responses and consider a slightly longer channel. Figure 36 shows the two simulated channels using the image method. The original impulse responses were longer, but in this example we truncated the simulated impulse response after 200 taps. The room size in this experiment is $[10, 15, 12.5]ft$, the source is located at $r_s = [2.5, 8.3, 3.3]ft$, and the microphones were chosen to be at $[4.2, 0.8, 5]ft$ and $[5.8, 0.8, 5]ft$. The effective reverberation time T_{60} in this example is 135ms. The reverberant speech signals were generated by filtering the clean speech signal with each impulse response (creating two reverberant speech signals). Figure

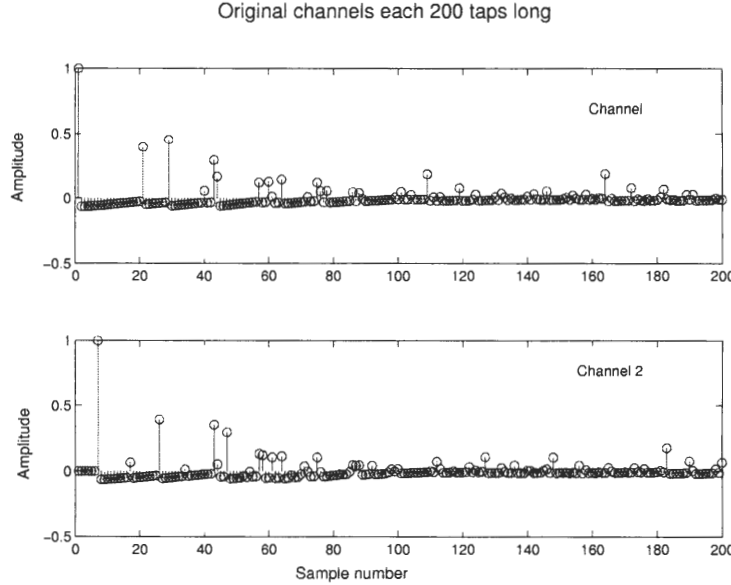


Figure 36: Impulse response of the two channels.

37 shows the frequency response of the two channels and Figure 38 shows their pole-zero plot. Note these channels are again nonminimum phase, and that most of the zeros lay near the unit circle. Using the adaptive formulation of the reduced MRE, we obtain the learning curve shown in Figure 39. Notice once again that the algorithm converges approximately after 1200 iterations, which is close to the number of total equalizer taps of 1194. The resulting equalizers at various delays are shown in Figure 40 and the result of applying the equalizers to the channel matrix H is shown in Figure 41. We see from

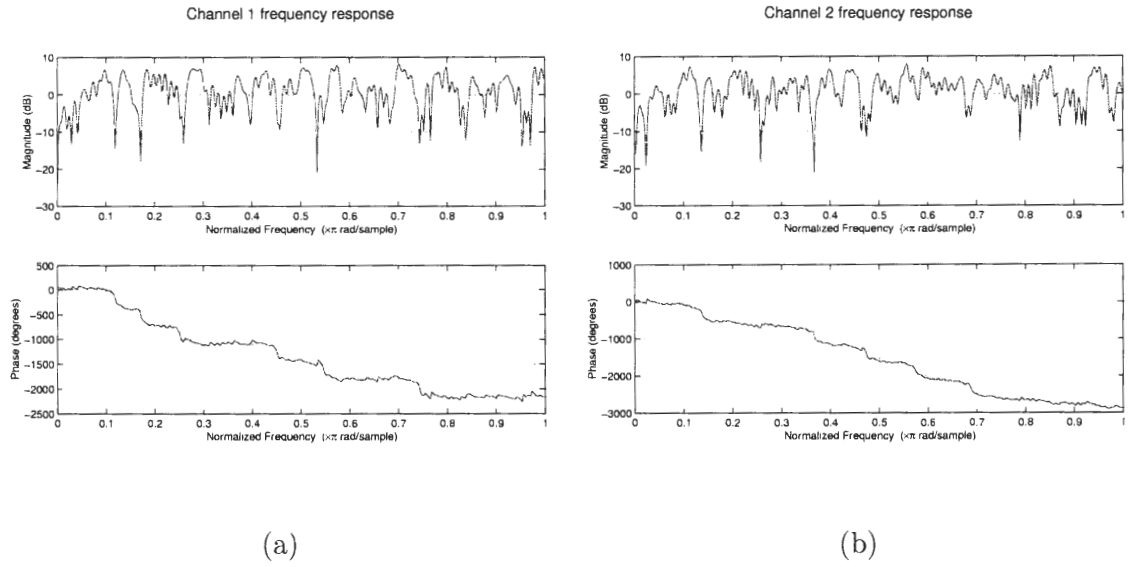


Figure 37: (a) Frequency response of channel 1, (b) Frequency response of channel 2.

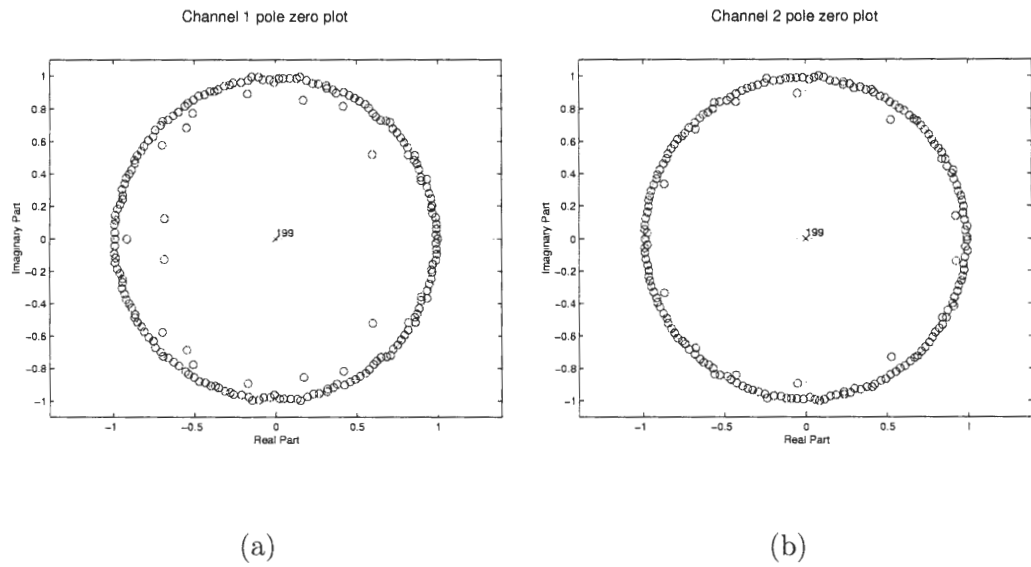


Figure 38: (a) Pole zero plot of channel 1, (b) Pole zero plot of channel 2.

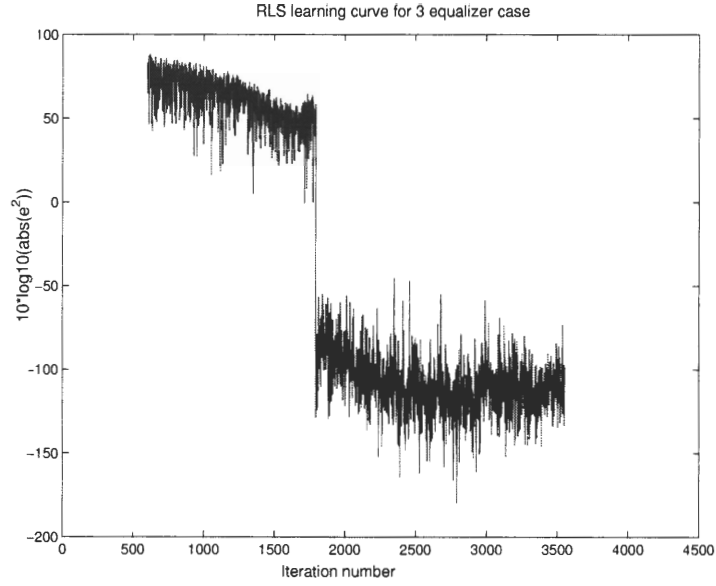


Figure 39: Learning curve based on RLS adaptive filter.

Figure 41 that the equalizers do invert the channel yielding the impulses at two edges and a central impulse. By applying any one of these three equalizers to the reverberant speech signal, the reverberation is removed. However, we will see in chapter 5 that in the presence of noise, the middle equalizer is the best one for use in dereverberation.

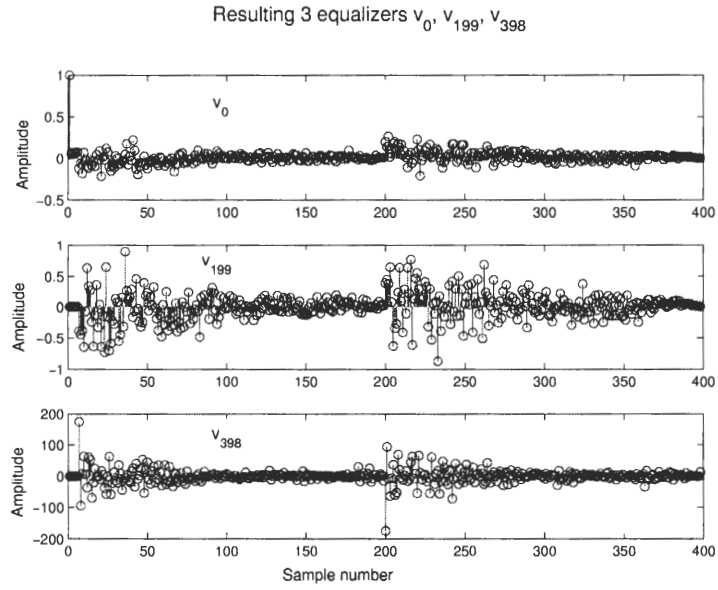


Figure 40: Resulting equalizers at various delays, where $v_{1,d}, v_{2,d}$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 199, 398$. The first 199 samples in each subplot are $v_{1,d}$, $d = 0, 199, 398$ and the remaining samples belong to $v_{2,d}$, $d = 0, 199, 398$.

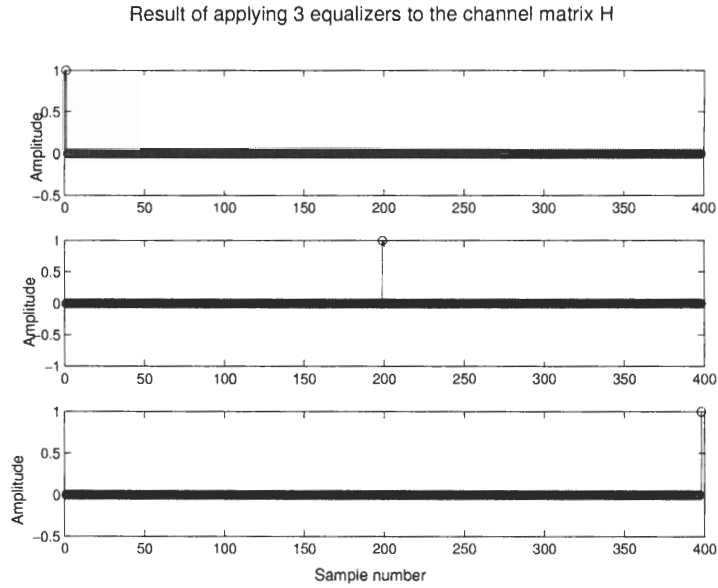


Figure 41: Result of applying equalizers to the channels.

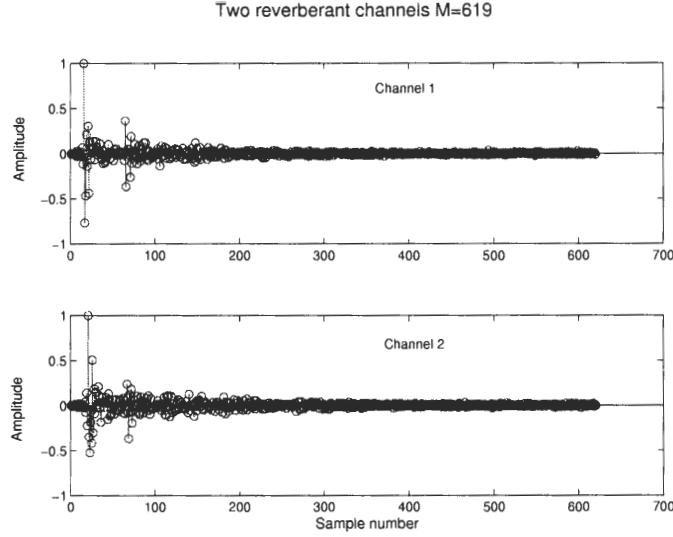


Figure 42: Plot of two reverberant channels.

4.5.3 Experiment C: Two Measured Channels of Order $M = 619$

In this experiment instead of using the simulated impulse responses, we use some of our measured impulse responses (as described in Chapter 3) and shown in Figure 42, and the impulse responses are much longer than the previous experiment. The reverberation time for these impulse responses is approximately 160ms. Using the adaptive RLS implementation of the RMRE method we obtain the learning curve shown in Figure 43. The resulting equalizers are shown in Figure 44, and the resulting delta functions in Figure 45. It can be seen that the equalizers do indeed result in three impulses at predefined delays, and we see that the RMRE finds a solution in this case, just like in the simulated impulse responses example.

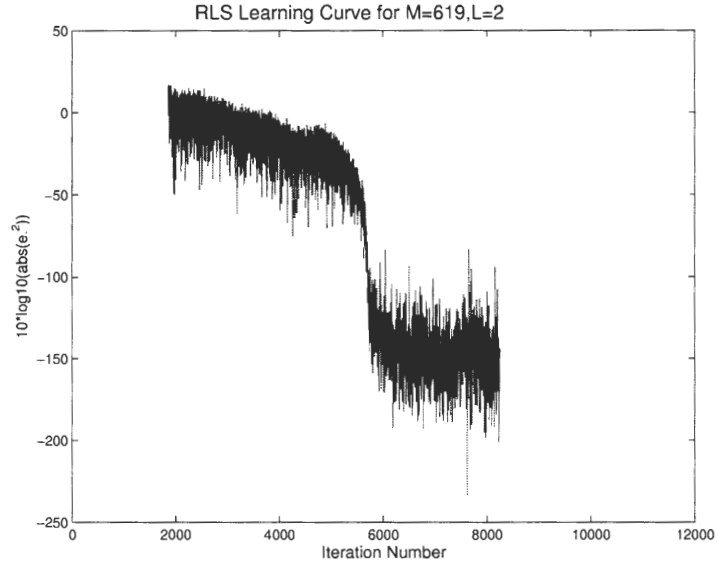


Figure 43: Learning curve based on an RLS adaptive filter implementation.

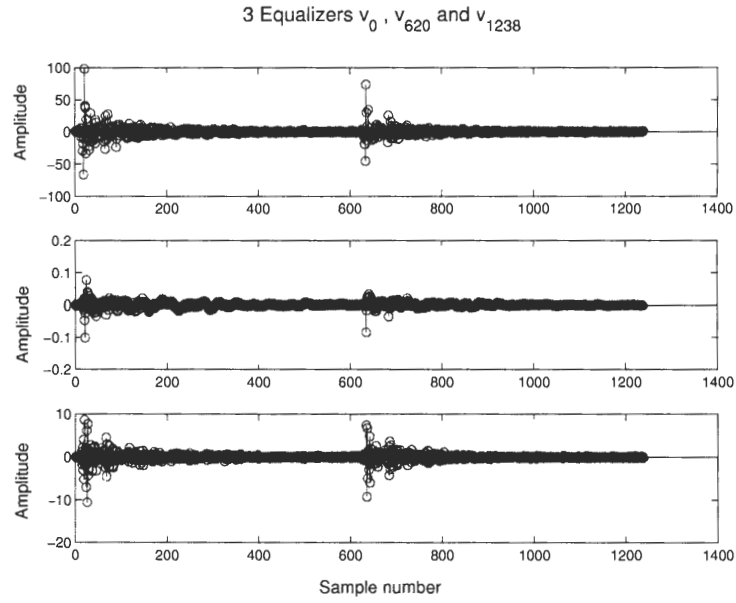


Figure 44: Resulting equalizers at various delays, where $v_{1,d}, v_{2,d}$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 620, 1238$. The first 619 samples in each subplot are $v_{1,d}, d = 0, 620, 1238$ and the remaining samples belong to $v_{2,d}, d = 0, 620, 1238$.

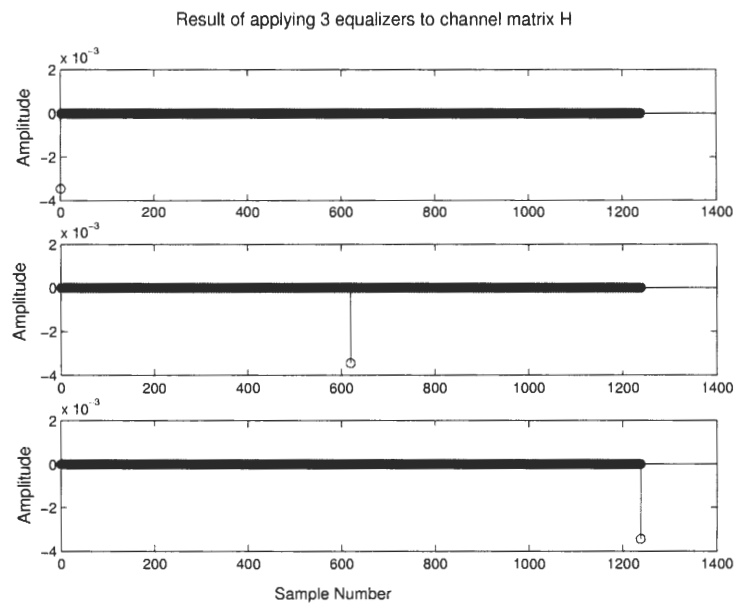


Figure 45: Result of applying equalizers to the channels.

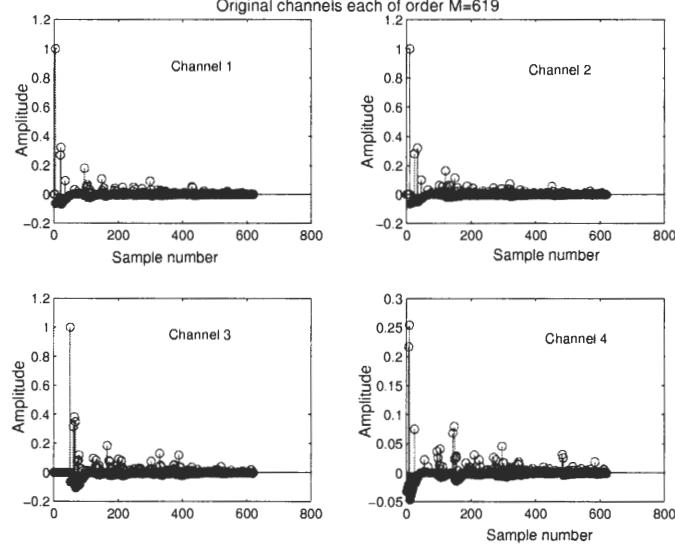


Figure 46: Impulse response of four simulated impulse responses.

4.5.4 Experiment D: Four Simulated Channels of Order $M = 619$

In this experiment, we increase the number of microphones to four and use the image method to simulate four impulse responses shown in Figure 46. The room size in this case is the same as that of Room 352 ([18, 21, 9]ft.) The source was located at [1.67, 11.17, 3.17]ft, and the microphones were chosen to be located at

$$\mathbf{r}_m = \begin{bmatrix} 5.33 & 6.1 & 2.5 \\ 8.25 & 8.5 & 2.5 \\ 8.25 & 1.5 & 2.5 \\ 5.33 & 12 & 2.5 \end{bmatrix} \text{ ft.}$$

The reverberation time T_{60} was approximately 190ms for these four channels. Using the adaptive RLS implementation of the RMRE method we obtain the learning curve shown in Figure 47. The resulting equalizers are shown in Figure 48. Notice that in this case we are concatenating four filter $v_{i,d}, i = 1 : 4$ but they are of shorter length individually compared to the two-channel case. The resulting Delta functions are shown in Figure 49.

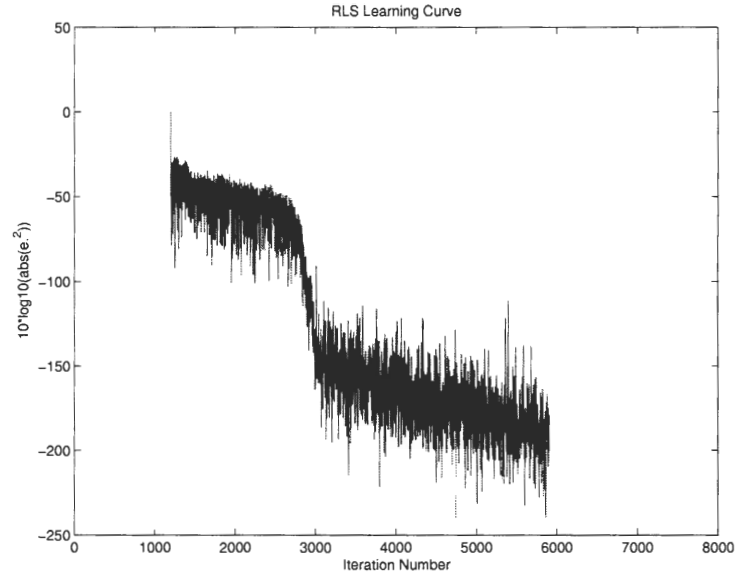


Figure 47: Learning curve based on an RLS adaptive filter implementation.

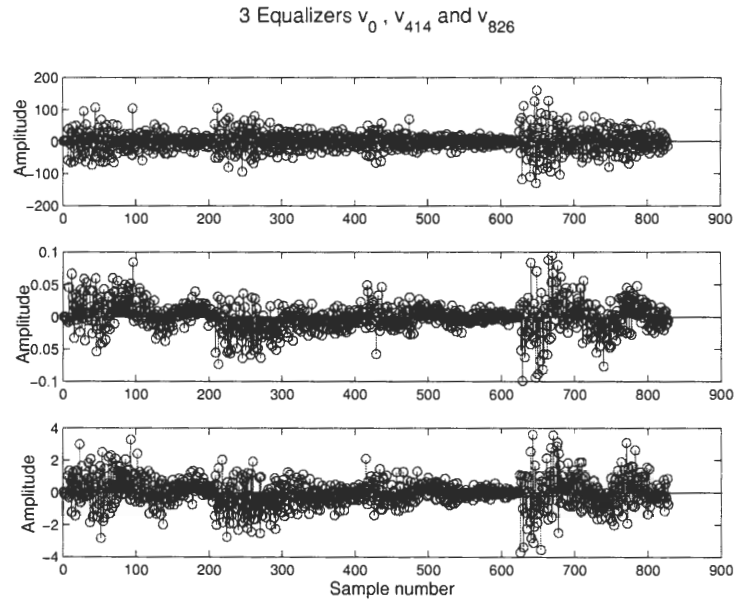


Figure 48: Resulting equalizers at various delays, where $v_{i,d}, i = 1 : 4$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, \dots, v_{4,d}]$, $d = 0, 414, 826$.

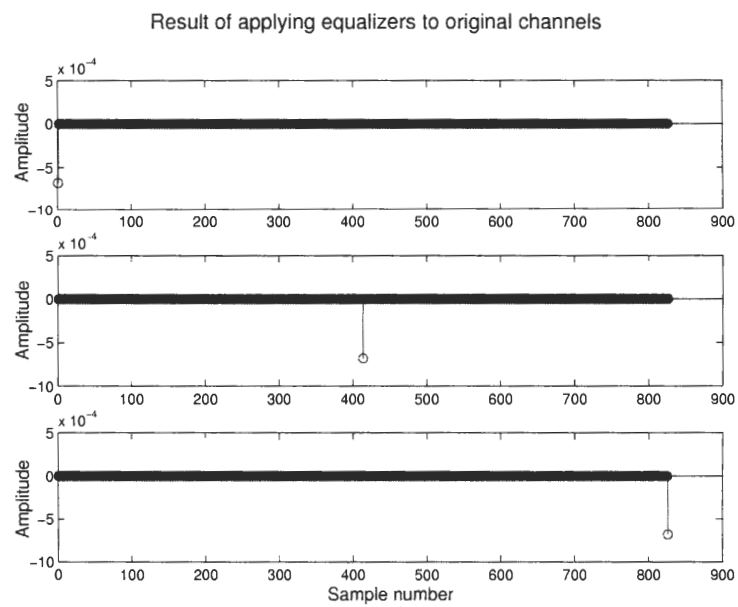


Figure 49: Result of applying the three equalizers to the channels.

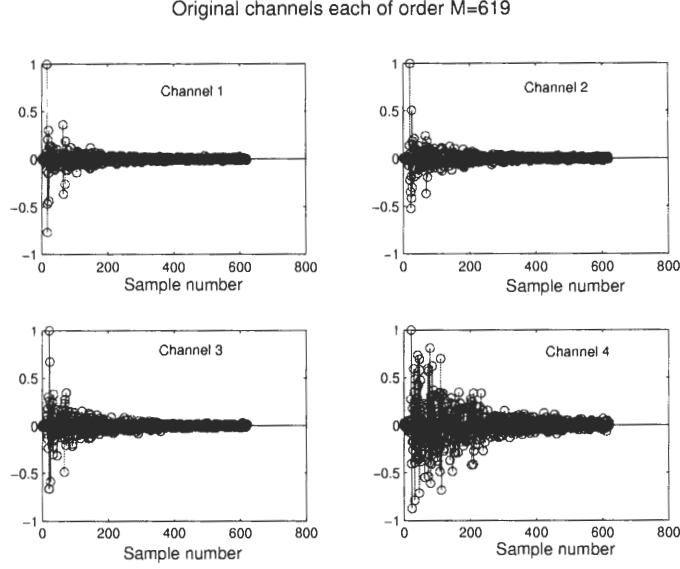


Figure 50: Plot of four reverberant channels.

4.5.5 Experiment E: Four Measured Channels of Order $M = 619$

In this experiment, we use four measured channels as described in Chapter 3, truncated to order $M = 619$, as shown in Figure 50. The microphones were placed at

$$\mathbf{r}_m = \begin{bmatrix} 5.33 & 6.1 & 2.5 \\ 8.25 & 8.5 & 2.5 \\ 8.25 & 1.5 & 2.5 \\ 1.67 & 14.16 & 0.83 \end{bmatrix} \text{ ft.}$$

and the source was at $[1.67, 11.17, 3.17]$. The reverberation time for these channels was approximately 160ms. Using the adaptive RLS implementation of the RMRE method we obtain the learning curve shown in Figure 51. The resulting equalizers are shown in Figure 52, and the resulting Delta functions in Figure 53. Once again, it can be seen that the equalizers do indeed result in impulses at predefined delays.

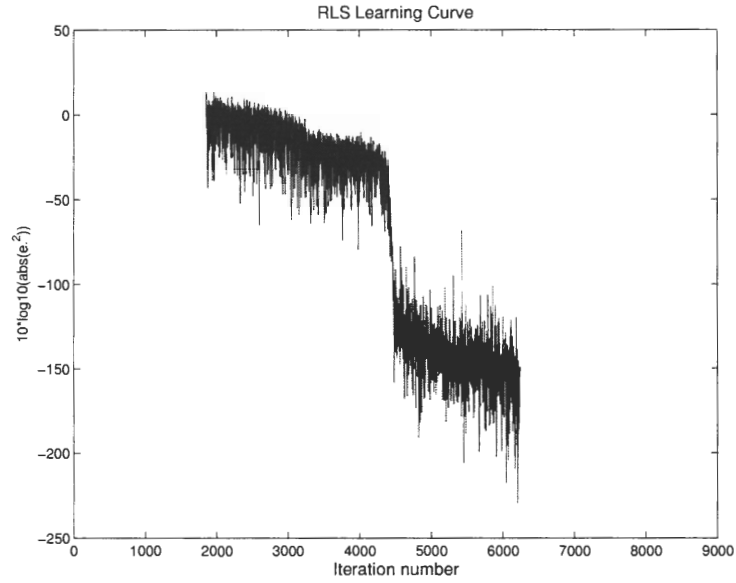


Figure 51: Learning curve based on an RLS adaptive filter implementation.

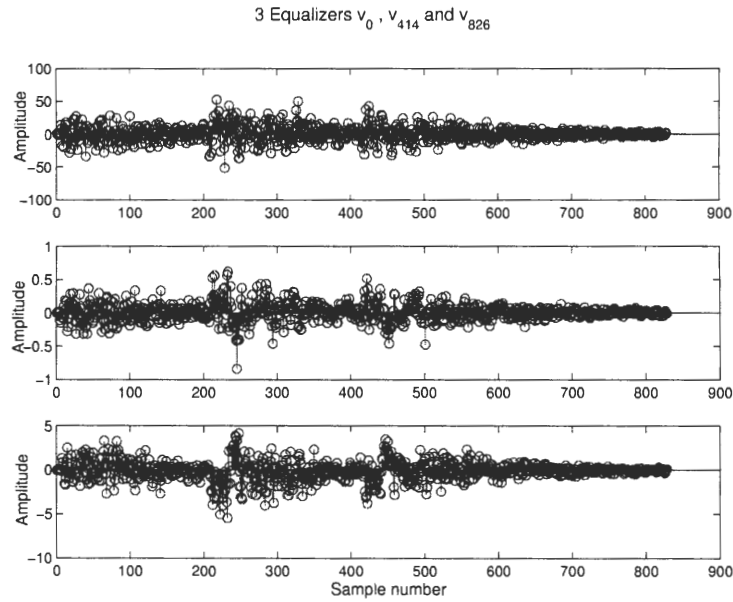


Figure 52: Resulting equalizers at various delays, where $v_{i,d}, i = 1 : 4$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, \dots, v_{4,d}]$, $d = 0, 414, 826$.

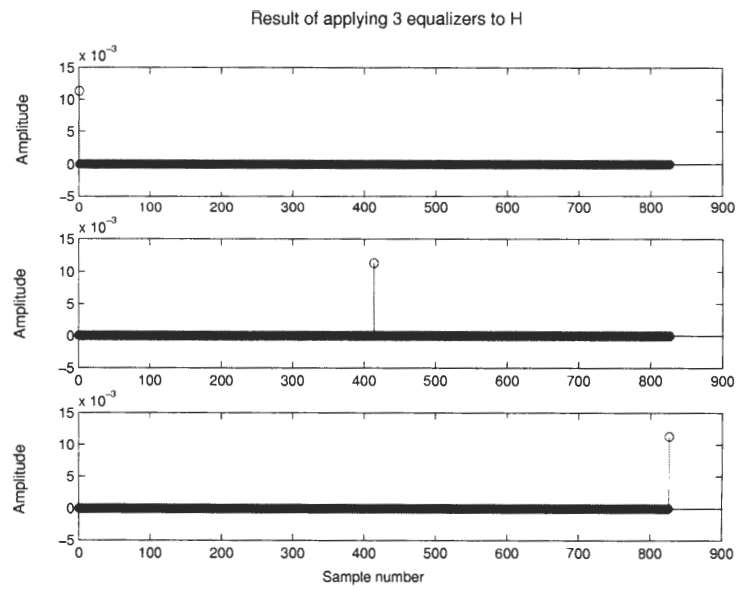


Figure 53: Result of applying equalizers to the channels.

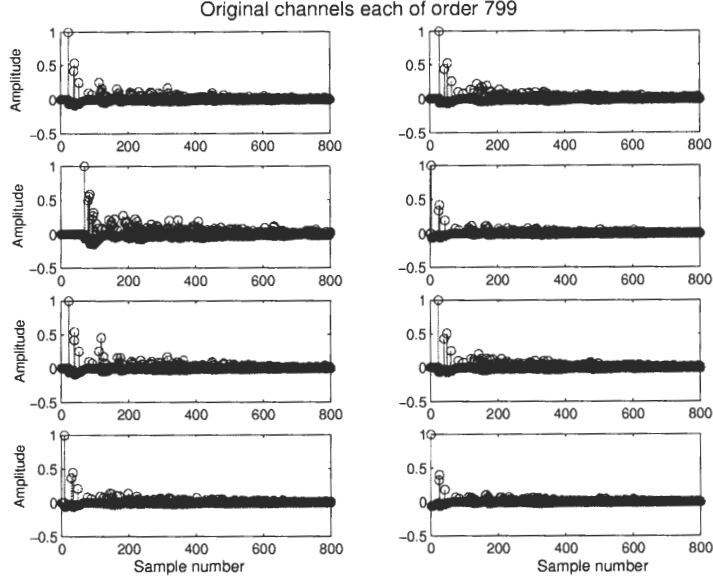


Figure 54: Plot of eight reverberant channels.

4.5.6 Experiment F: Eight Simulated Channels of Order $M = 799$

In this final experiment, we simulate channels that are in a more reverberant environment by changing the reflection coefficients of the image method from the previous examples. The resulting impulse response are shown in Figure 54 and have an approximate reverberation time of approximately 280ms. The microphones were located at

$$\mathbf{r}_m = \begin{bmatrix} 5.33 & 6.1 & 2.5 \\ 8.25 & 8.5 & 2.5 \\ 8.25 & 1.5 & 2.5 \\ 5.33 & 12 & 2.5 \\ 5 & 5.8 & 2.5 \\ 7.9 & 9.1 & 2.5 \\ 6 & 9.7 & 2.5 \\ 5.1 & 12 & 2.5 \end{bmatrix} \text{ ft.}$$

Using the adaptive RLS implementation of the RMRE method we obtain the learning curve shown in Figure 55. The resulting equalizers are shown in Figure 56, and the resulting Delta functions in Figure 57. A plot of the original (clean) speech signal and the

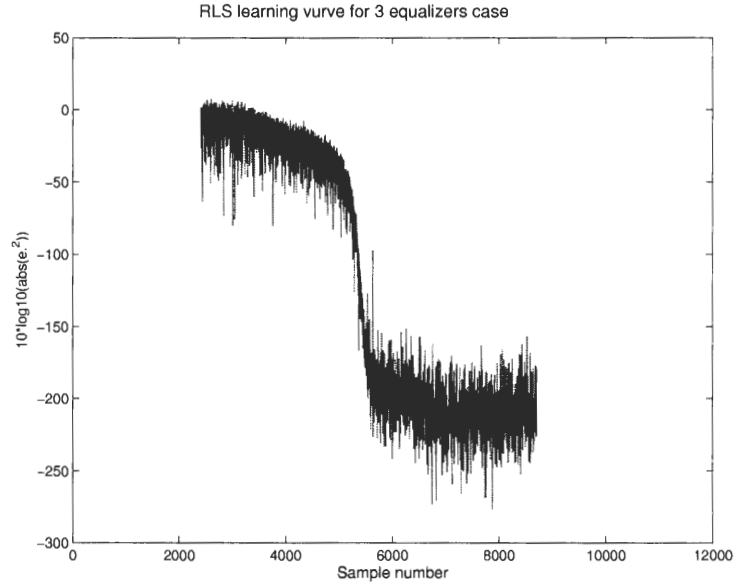


Figure 55: Learning curve based on an RLS adaptive filter implementation.

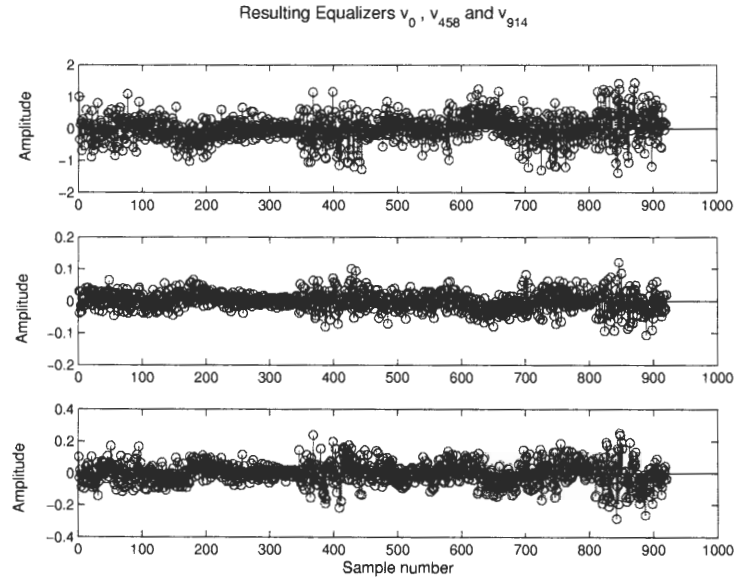


Figure 56: Resulting equalizers at various delays, where $v_{i,d}, i = 1 : 8$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, \dots, v_{8,d}]$, $d = 0, 458, 914$.

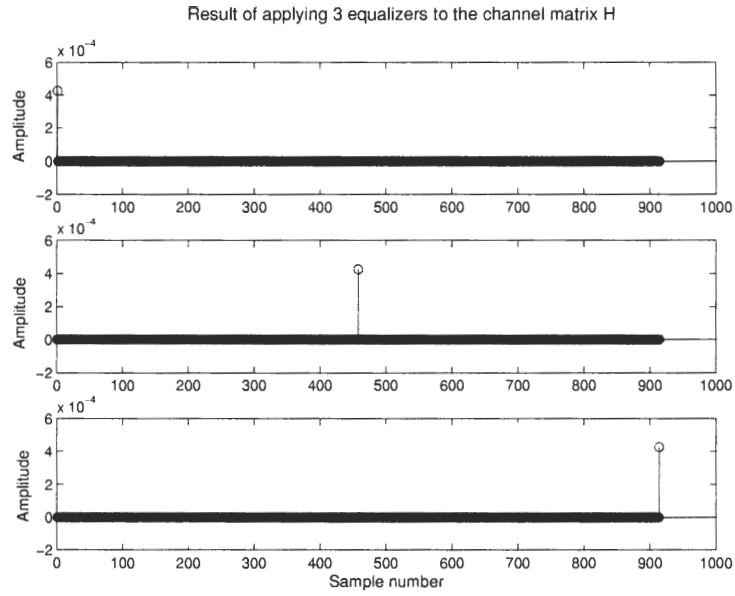


Figure 57: Result of applying equalizers to the channels.

dereverberated speech (after proper scaling) signal are shown in Figure 58, and Figure 59 shows the difference between the two signals, which is very small and primarily due to rounding errors.

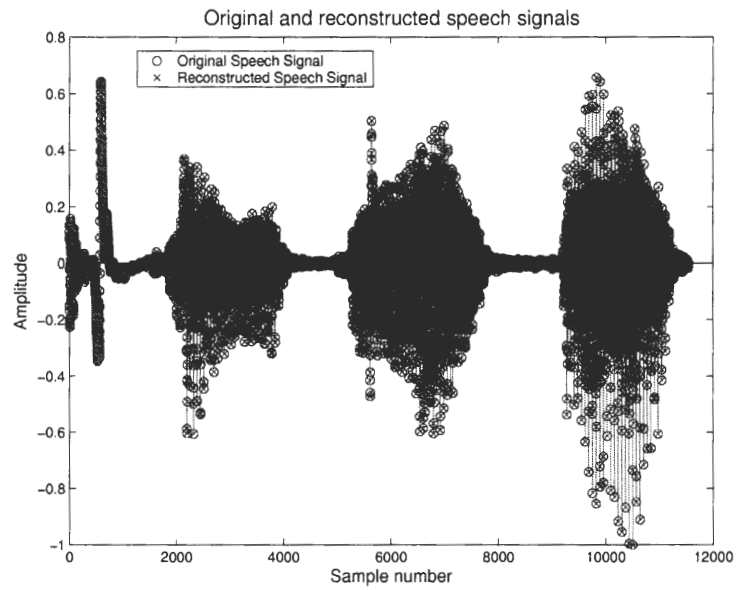


Figure 58: Original and reconstructed speech signals.

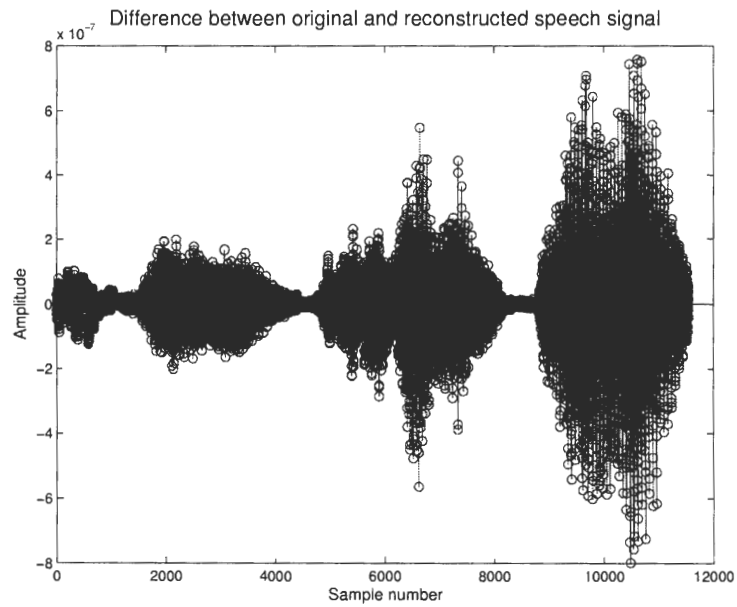


Figure 59: Difference between original and reconstructed speech signals.

4.6 Further investigation into reducing the number of required equalizers

We demonstrated in the previous sections how it is possible to obtain a partial set of equalizers depending on how the MRE criterion is formulated. The case of three equalizers seems to provide a good compromise between robustness and computational complexity associated with the long channels associated with the acoustical dereverberation problem. The question that we continue to investigate in this section is how to modify and reformulate the MRE method with all its advantages to obtain an adaptive implementation using only two equalizers other than the edge equalizers. While we have not been able to obtain a robust adaptive implementation of the ideas presented here, we feel that they do provide some theoretical insight into the problem and further the understanding of the MRE and RMRE approach. We assume that the input is white for the sake of simplicity.

Recall in the case of $J_{MRE}(\mathbf{v}_0, \mathbf{v}_1) = E |\mathbf{v}_0^T \mathbf{x}[n] - \mathbf{v}_1^T \mathbf{x}[n+1]|^2$, differentiating this criterion and setting the result equal to zero yields the system of equations given by

$$\begin{bmatrix} R_x[0] & -R_x^T[1] \\ -R_x[1] & R_x[0] \end{bmatrix} \begin{bmatrix} \mathbf{v}_0 \\ \mathbf{v}_1 \end{bmatrix} = \mathbf{0}.$$

As explained before, the solution yielded equalizers that produced multiple diagonals instead of impulses. Our aim now is to show how it is possible to obtain the middle/central equalizers instead of the edge equalizers. Let $\mathbf{v}_a, \mathbf{v}_b$ denote the two central equalizers and let us define a new criterion of the form

$$J(\mathbf{v}_a, \mathbf{v}_b) = E |\mathbf{v}_a^T \mathbf{x}[n] - \mathbf{v}_b^T \mathbf{x}[n+1]|^2 \quad s.t. \quad \mathbf{v}_a^T R_x[k] \mathbf{v}_b = 0, \quad (27)$$

where $R_x[k]$ is some correlation matrix for some lag k to be determined. In the case of white noise as input, $R_x[k]$ has an especially simple formula given by $R_x[k] = HJ^k H^T$ where J^k as defined before is the backward shift matrix raised to the k -th power. The idea behind the constraint can be seen by considering the expression $\mathbf{v}_a^T R_x[k] \mathbf{v}_b = \mathbf{v}_a^T H J^k H^T \mathbf{v}_b = \mathbf{v}_b^T H (J^k)^T H^T \mathbf{v}_a = 0$. This constraint is essentially a statement of Q orthogonality that is commonly used to describe conjugate directions. By imposing this constraint on the equalizers, we force the result of $\mathbf{e}_a^T = \mathbf{v}_a^T H$ to be orthogonal to J^k and $\mathbf{e}_b^T = \mathbf{v}_b^T H$ to

be orthogonal to $(J^k)^T$. The intersection of these two conditions results in equalizers that make \mathbf{e}_a and \mathbf{e}_b the remaining basis for $(J^k)^T + J^k$.

To better explain the idea behind the constraint, consider $L = 2$ channels, each with 10 ($M = 9$) taps and a data window of 11 samples for each channel. This implies that our channel correlation matrix will have the form $H = \begin{bmatrix} [H_1] \\ [H_2] \end{bmatrix}_{[22 \times 20]}$. Thus, the criterion we seek to minimize is

$$\min e^2 = (\mathbf{v}_a^T \mathbf{x}[n] - \mathbf{v}_b^T \mathbf{x}[n+1])^2 \text{ s.t. } \mathbf{v}_a^T R_x[-11] \mathbf{v}_b = 0.$$

The value of $k = -11$ was picked based on the fact that

$$\mathbf{v}_a^T R_x[-11] \mathbf{v}_b = \underbrace{\mathbf{v}_a^T H}_{e_a^T} J^{11} \underbrace{H^T \mathbf{v}_b}_{e_b} = \underbrace{\mathbf{v}_b^T H}_{e_b^T} (J^{11})^T \underbrace{H^T \mathbf{v}_a}_{e_a}.$$

This means the only possible solution would require \mathbf{e}_a and \mathbf{e}_b be the remaining basis for $(J^{11})^T + J^{11}$.

This constraint can be easily incorporated by using the Lagrange multipliers method. The resulting cost function is now

$$\zeta^2 = e^2 + \lambda \mathbf{v}_a^T R_x[-11] \mathbf{v}_b = \mathbf{v}_a^T R[0] \mathbf{v}_a + \mathbf{v}_b^T R[0] \mathbf{v}_b - 2 \mathbf{v}_a^T R[-1] \mathbf{v}_b + \lambda \mathbf{v}_a^T R_x[-11] \mathbf{v}_b$$

where λ is any nonzero scale factor. Differentiating the result w.r.t. to each equalizer and setting the result to zero yields the following linear system:

$$\begin{bmatrix} 2R[0] & -2R[-1] + \lambda R_x[-11] \\ -2R[1] + \lambda R_x[11] & 2R[0] \end{bmatrix}_{[44 \times 44]} \begin{bmatrix} \mathbf{v}_a \\ \mathbf{v}_b \end{bmatrix} = \mathbf{0}.$$

Notice how the R matrix in this case relies on correlation matrices at lag zero, one and eleven and how the off-diagonal terms contain the sum of two correlation matrices. The problem with the above correlation matrix is that it is difficult to find a rank one update for an adaptive formulation. That is to say, there exists no $\mathbf{u}(n)$ such that $R(n) = R(n-1) + \mathbf{u}(n) \mathbf{u}^T(n)$. It is possible to find a rank two update, and it is also possible to find a rank one update formula of the form $R(n) = R(n-1) + \mathbf{u}(n) \mathbf{v}^T(n)$. As an example of a rank

one update, to obtain $R = \begin{bmatrix} R(0) & -R(-1) + R(-11) \\ -R(1) + R(11) & R(0) \end{bmatrix}$, requires writing R as the product of

$$\begin{aligned} \begin{bmatrix} R(0) & -R(-1) + R(-11) \\ -R(1) + R(11) & R(0) \end{bmatrix} &= E \left(\begin{bmatrix} X_n + X_{n+25} \\ -X_{n+1} + X_{n+61} \end{bmatrix} \begin{bmatrix} X_n + X_{n+50} & -X_{n+1} + X_{n+36} \end{bmatrix} \right) \\ &= \begin{bmatrix} R(0) + R(25) + R(-25) + R(50) & -R(-1) + R(-11) - R(-24) + R(36) \\ -R(1) + R(11) - R(-49) + R(61) & R(0) - R(-35) - R(60) + R(25) \end{bmatrix} \\ &= \begin{bmatrix} R(0) & -R(-1) + R(-11) \\ -R(1) + R(11) & R(0) \end{bmatrix}. \end{aligned}$$

Modifying the classic RLS adaptive filter requires rewriting the updated inverse using the matrix inversion lemma. The main problem with the block correlation matrix above is that it is very sensitive with respect to its structural properties. For example, consider two correlation matrices defined by

$$\begin{aligned} R_1 &= \begin{bmatrix} 2R[0] & -2R[-1] + R[-11] \\ -2R[1] + R[11] & 2R[0] \end{bmatrix}_{[44 \times 44]} \\ R_2 &= \begin{bmatrix} 1.0001 \times 2R[0] & -2R[-1] + R[-11] \\ -2R[1] + R[11] & 2R[0] \end{bmatrix}_{[44 \times 44]}, \end{aligned}$$

where R_2 is the likely result of a rank one nonsymmetric update. To obtain the central equalizers for each case, we solve the system of equations defined by $R_1 \mathbf{v} = R_1 \begin{bmatrix} \mathbf{v}_a \\ \mathbf{v}_b \end{bmatrix} = 0$

and $R_2 \mathbf{v} = R_2 \begin{bmatrix} \mathbf{v}_a \\ \mathbf{v}_b \end{bmatrix} = 0$. The result $E = \begin{bmatrix} \mathbf{v}_a^T \\ \mathbf{v}_b^T \end{bmatrix} H$ is shown in Figure 60. Notice how the second set of impulses associated with R_2 is badly scaled with scale factor of 1×10^{-12} .

Understanding how the constraint allows us to obtain the central equalizers sheds light on the possibility of finding only one central/middle equalizer. The simplest approach would be to find the null space of $(R_x[11] + R_x^T[10]) = \text{null} \left(H \left(J^{11} + (J^{10})^T \right) H^T \right)$. Looking at an image of $J^{11} + (J^{10})^T$ (dark cells indicate ones), we can see that the only component orthogonal to this matrix is an impulse at delay 10. In the case of constant modulus input

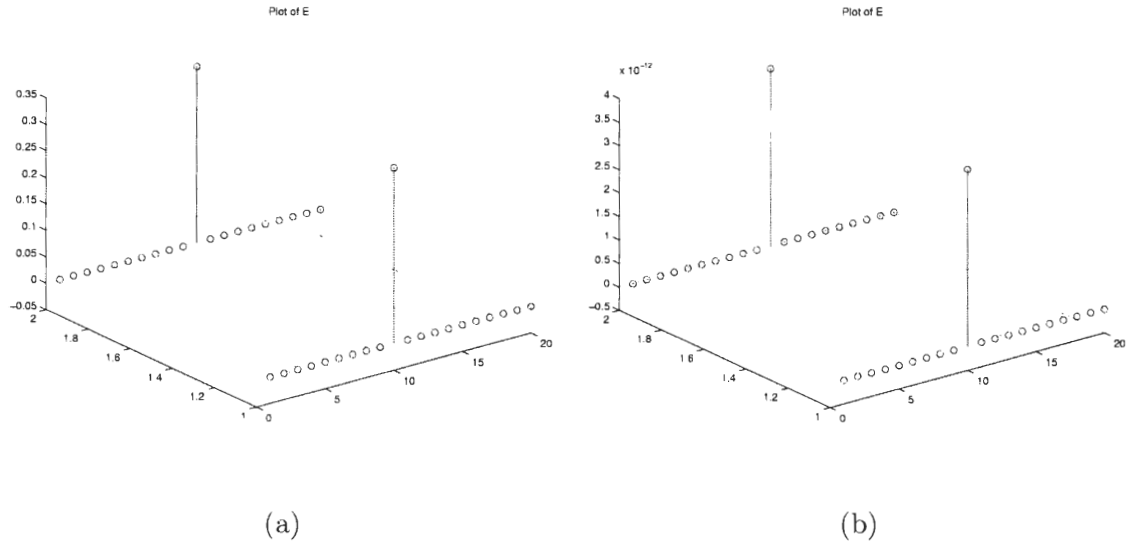


Figure 60: (a) E for the case of R_1 , (b) E for the case of R_2 .

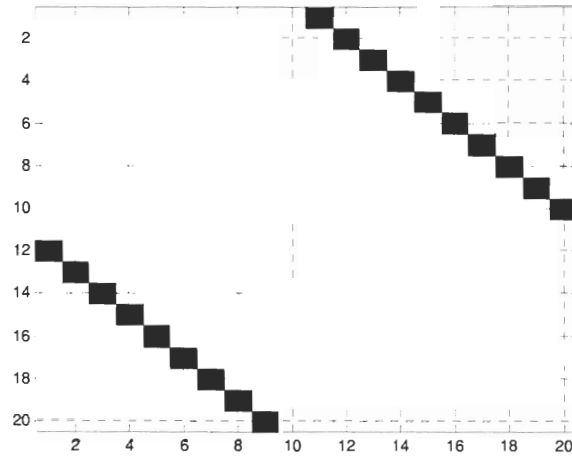


Figure 61: Plot of $J^{11} + (J^{10})^T$

signals, there exists a criterion described in [58] which finds the optimal delay value. Some of these ideas and issues discussed here may provide the basis for future research.

4.7 *Summary*

In this chapter, we explained the basics of the MRE method, and we demonstrated theoretically and experimentally the applicability of the RMRE method for accomplishing blind dereverberation of speech signals. Our experimental results were performed on both synthetic and measured room impulse responses using speech as an input signal. Finally, we discussed some theoretical results and associated problems that can be used to further reduce the computational cost of the RMRE.

CHAPTER V

RMRE PERFORMANCE IN THE PRESENCE OF MODELING ERRORS AND MEASUREMENT NOISE

In this chapter, we discuss some of the remaining issues that are required for a real-world implementation of our proposed method described in Chapter 4. Some of the issues, such as the need for model order determination have been mentioned in the previous chapters. In particular, we discuss the issues of under and over-modeling of the room impulse responses, the impact of measurement noise, and possible methods to improve robustness in the presence of noise. Experimental results and simulations are included where appropriate to help us clarify and explain some of our observations and results.

5.1 Model Order Determination

So far, we have assumed that the channel, i.e. impulse response, order is known or can be accurately estimated by some pre-processing step. However, in any practical implementation, the channel orders must be estimated. While there are many classical methods to accomplish this, such as AIC and MDL described in [53], there is some evidence, for example in [77], that these methods are not accurate in the context of blind identification and equalization. Additionally, not all blind identification and equalization methods using second-order statistics are robust with respect to errors in the channel order estimates as explained in [78, 79]. For example, subspace and related methods such as the TXK algorithm are not forgiving with respect to model order mismatches. One advantage of the MRE method, and by extension the RMRE method, is that it has demonstrated some robustness with respect to certain types of channel order estimation errors [80]. However, we must point out that the model order determination of FIR filters in a SIMO system remains an open problem that requires further study. The problem is even more complicated when the input is non-stationary and non-white, as in the case of speech, and when the channels

exhibit the long tails typical of acoustical channels.

5.1.1 Under-modeling of the Room Impulse Response

The problem of under-modeling arises when the true order of the FIR channels in a SIMO system is underestimated. For example, let us consider the case of a two-channel system, with FIR channel impulse responses $h_1(n), h_2(n)$ of order $M = 199$, i.e. 200 taps long. From the Bezout equation and inversion principles discussed in Chapter 4, any equalizer pair $v_{1,d}, v_{2,d}$ for a specified delay d must have an order of at least $M = 198$, i.e. $v_{1,d}(n), v_{2,d}(n)$ must be each at least 199 taps long. If the estimated order is $\hat{M} = 189$, then the original impulse responses can be thought of as being composed of two parts; the primary (or main) $\hat{M} = 189$ segment, denoted by $h_{i,p}(n), i = 1, 2$, and a tail segment denoted by $h_{i,t}(n), i = 1, 2$ compromising the truncated end of the impulse responses. The output of the SIMO system is then given by:

$$x_i(n) = (h_{i,p}(n) + h_{i,t}(n)) * s(n) = x_{i,p}(n) + x_{i,t}(n),$$

for $i = 1, 2$. Thus, it can be seen that $x_{i,t}(n)$ behaves as colored noise when under-modeling occurs. However, more ominously, the Bezout equation does not state that an inverse of order less than $m = 198$ can be found to invert the pair of channels $h_1(n)$ and $h_2(n)$. Thus, if the estimated order is $\hat{M} = 189$, and we attempt to find equalizers of order 188, the resulting equalizers will not be inverses. To further illustrate these concepts, consider the two impulse responses given in Figure 62 and a magnified portion of the tail shown in Figure 63. Instead of estimating the true order $m = 199$, we underestimate the order to $\hat{M} = 198$, and attempt to find the equalizers. The resulting learning curve is shown in Figure 64, and clearly since no equalizer exists for this case, the algorithm does not converge. For the sake of completeness, Figure 65 shows the resulting equalizers and Figure 66 shows the result of applying the equalizers to the original channels, which is supposed to at least *resemble* impulses at various delays. These results indicate that the RMRE method fails in the case of under-modeling, a not-so-surprising result considering the lack of an inverse in such a case.

One possible way to combat the under-modeling breakdown of the RMRE method is by

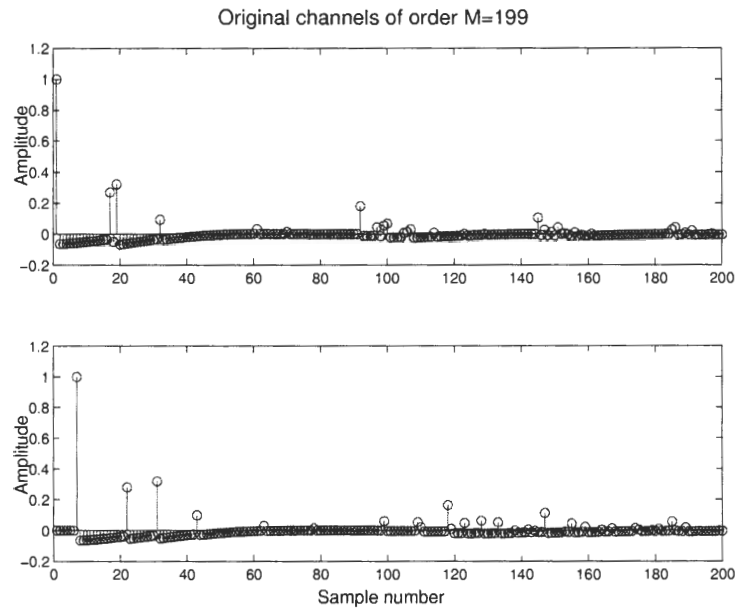


Figure 62: Original channels.

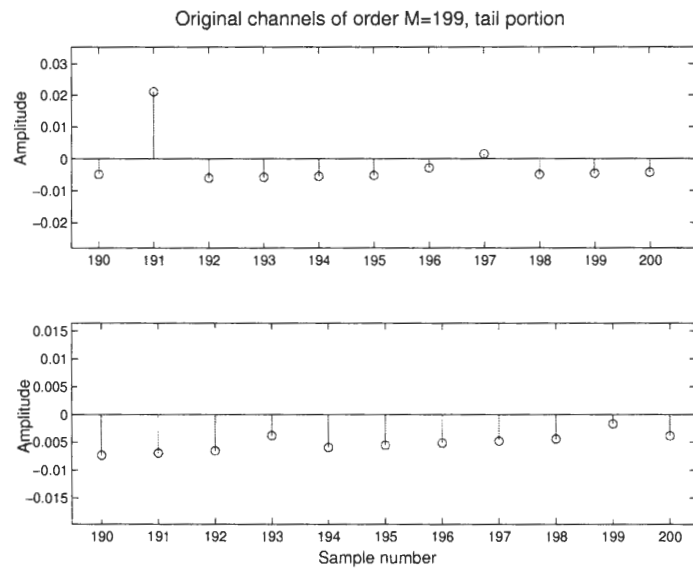


Figure 63: Original channels, tails portion.

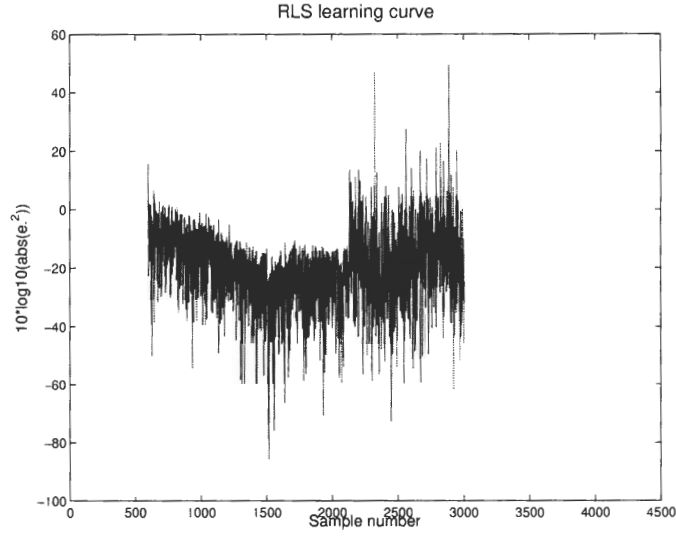


Figure 64: RMRE under-modeling RLS learning curve.

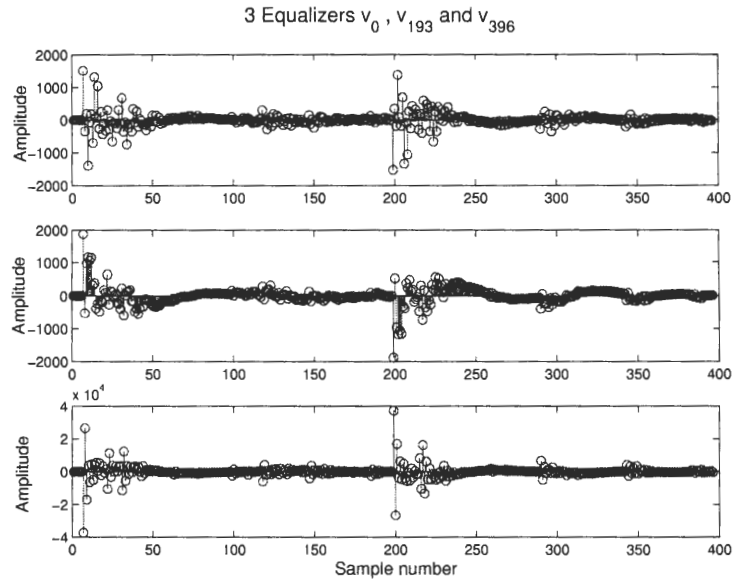


Figure 65: Resulting equalizers, where $v_{i,d}, i = 1 : 2$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 193, 396$.

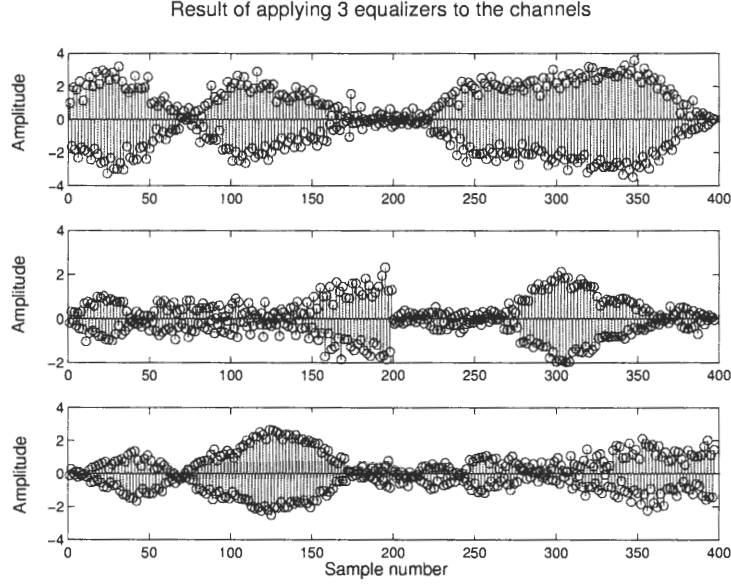


Figure 66: Result of applying equalizers to the channels in the under-modeled case.

increasing the number of microphones. We have already demonstrated in Chapter 4 that increasing the number of microphones helps us achieve some reduction in the total number of required equalizers taps. Recall that the order of the L filters $v_{i,d}(n), i = 1 : L$ (for a delay of d samples) is given by

$$\left\lceil \frac{M-1}{L-1} \right\rceil,$$

where M is the order of the room impulse responses. If we underestimate the order M by δ , then the number of taps is given by

$$\left\lceil \frac{M-1-\delta}{L-1} \right\rceil = \left\lceil \frac{M-1}{L-1} - \frac{\delta}{L-1} \right\rceil = \left\lceil C \frac{r}{L-1} - \frac{\delta}{L-1} \right\rceil,$$

where C and r are the integer and remainder of the parts of the division of $(M-1)/(L-1)$. As long as $\frac{r}{L-1} > \frac{\delta}{L-1}$, then the operation of rounding up to the nearest integer will not be affected by the order under-modeling, and the equalizers can be found. For example, consider the example of $L = 20$ channels, each with $M = 199$ and (synthetic) impulse responses shown in Figure 67. Assume that we underestimate the true number of taps by 8, i.e. $\delta = 7$, then

$$\left\lceil \frac{199}{19} - \frac{7}{19} \right\rceil = \left\lceil 10 \frac{9}{19} - \frac{7}{19} \right\rceil = 11,$$

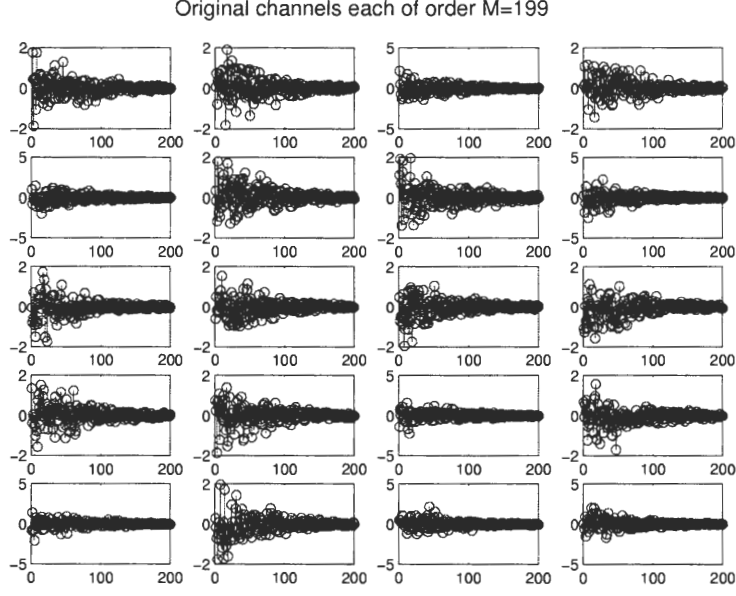


Figure 67: Original channels.

which is the same the number of taps for the equalizers as if we had not under-estimated the true channel order. Figure 68 shows the learning curve for the case when under-modeling by $\delta = 7$ occurs, and the resulting equalizers along and the result of applying the equalizers to the channels are shown in Figure 69 and Figure 70 respectively. Thus, we see that under-modeling can be partially alleviated by the use of more microphones.

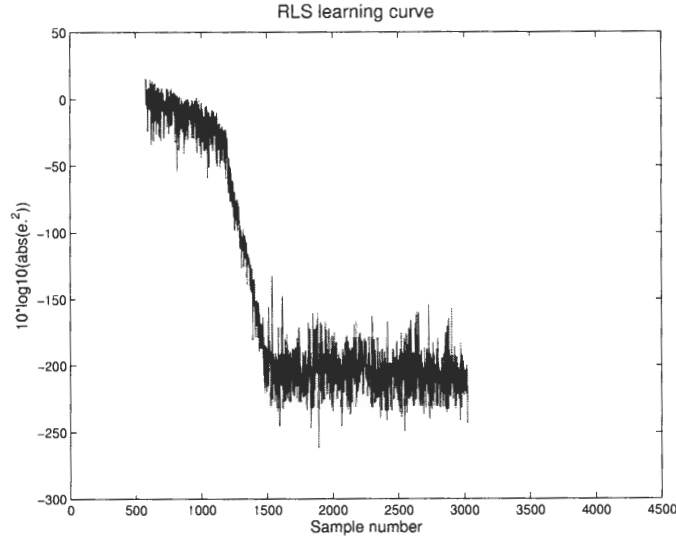


Figure 68: RMRE under-modeling RLS learning curve for $L = 20, \delta = 7$.

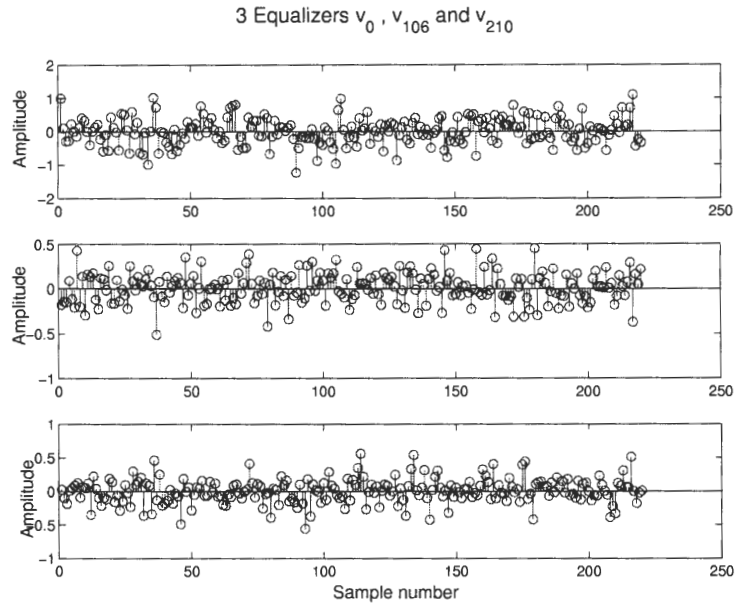


Figure 69: Resulting equalizers for $L = 20, \delta = 7$, where $v_{i,d}, i = 1 : 20$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, \dots, v_{20,d}]$, $d = 0, 106, 210$.

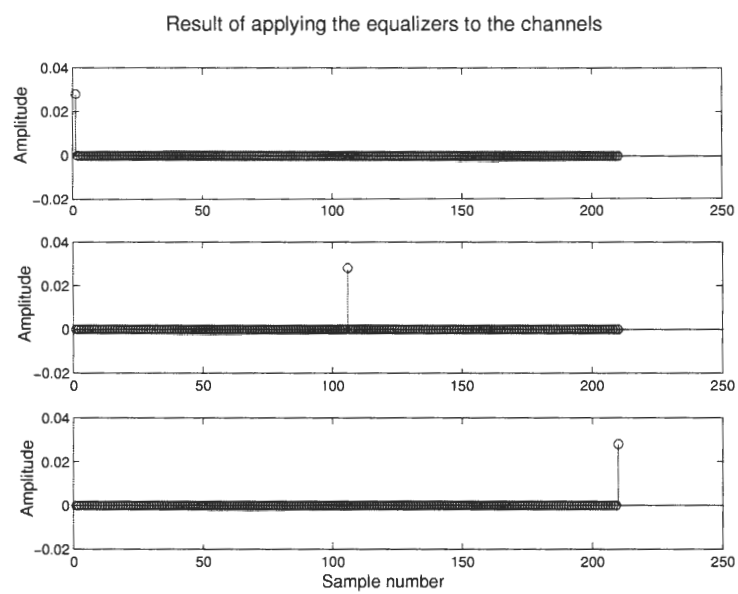


Figure 70: Result of applying equalizers to the channels for $L = 20, \delta = 7$.

5.1.2 Over-modeling of the Room Impulse Response

The over-modeling situation occurs when the order of the channels is over-estimated. The disadvantage of over-modeling is the increase in the number of parameters, and hence the computational cost. However, the over-modeling situation is handled easily by the RMRE, since the Bezout equation permits the existence of higher order equalizers, albeit ones that are non-unique. However, we do not wish to imply that the gross over-modeling can be allowed. For example, using the channels from the example in the previous section, we over estimate the channels by forty taps, and the resulting learning curve is shown in Figure 71. The resulting equalizers along with the result of applying the equalizers

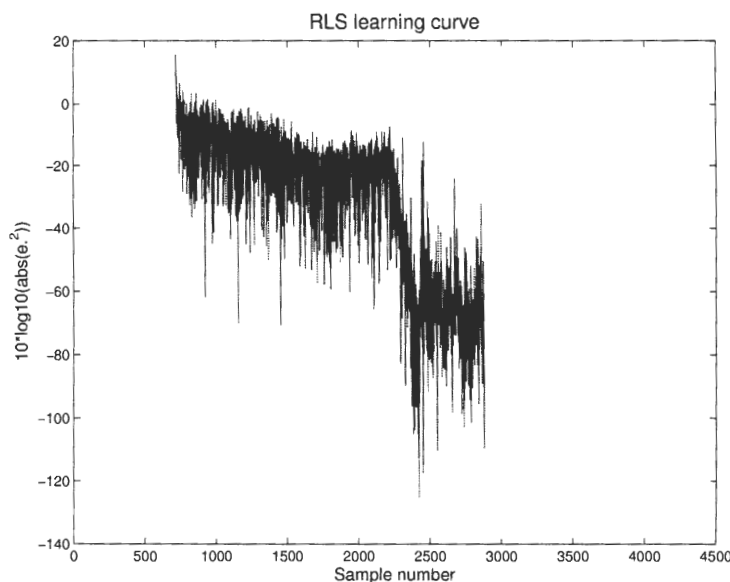


Figure 71: RMRE under-modeling RLS learning curve in the over-modeling case.

to the channels are shown in Figure 72 and Figure 73. As can be seen in Figure 71, a longer number of iteration is required for convergence, and since more parameters exist, computational cost also has increased. In addition, it means the equalizers calculated will cause a larger lag and delay than is necessary in the reconstructed (dereverberated) speech signal.

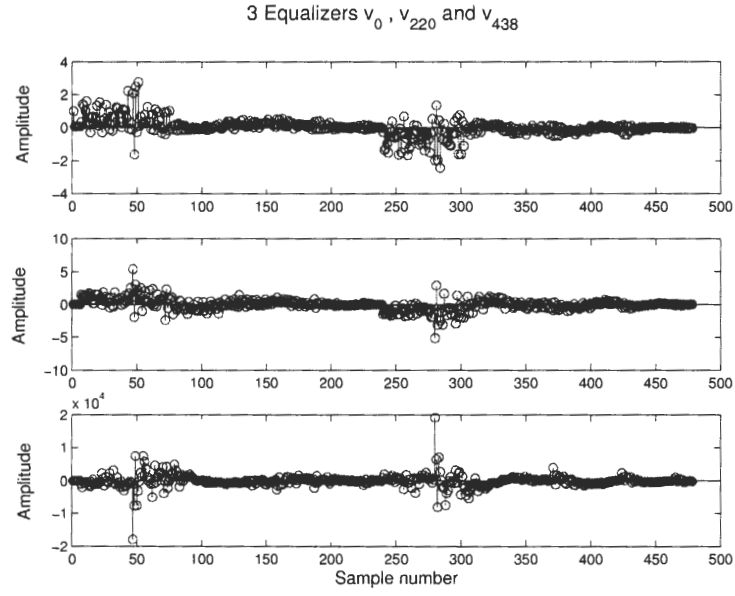


Figure 72: Resulting equalizers in the over-modeling case, where $v_{1,d}$, $v_{2,d}$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 220, 438$.

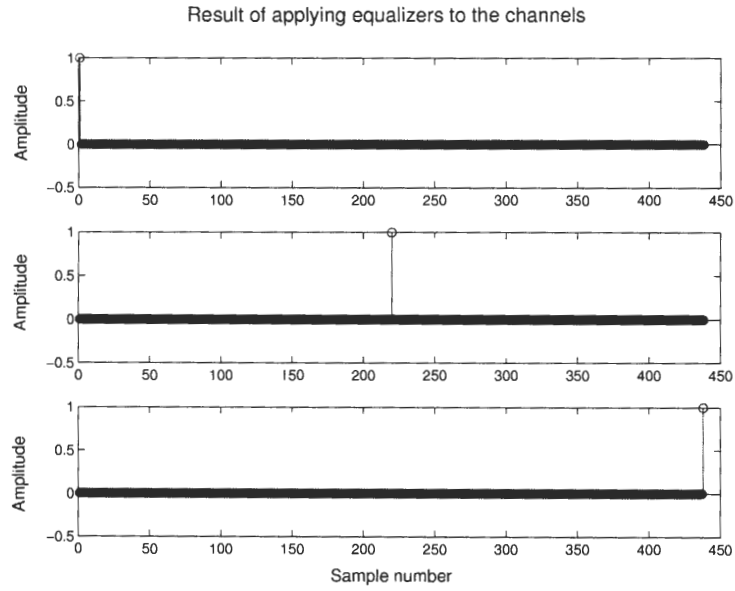


Figure 73: Result of applying equalizers to the channels in the over-modeling case.

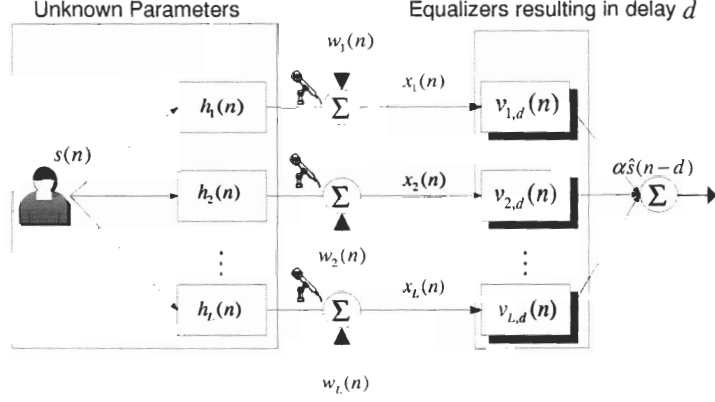


Figure 74: SIMO system with measurement noise.

5.2 Impact of Measurement Noise

So far, our focus has been solely on the problem of dereverberation, and we have not discussed the issue of measurement noise. Generally, inversion methods are notorious for their sensitivity to noise and usually require regularization. The MRE criterion combated the noise problem by considering the entire set of all possible equalizers, which added a certain measure of robustness to the method, as detailed in [59]. This result is interesting, in part, because the MRE method is based on a zero-forcing criterion, which is known to be less than optimal in presence of noise. However, for the RMRE criterion, the robustness inherent from the computation of all the equalizers is not an option, and in addition the long room impulse responses complicate the problem further. In this section, we explain the problem of noise in a SIMO FIR system, and then discuss the performance of the RMRE method. Finally, we explain one possible way to enhance the robustness of the RMRE method in the presence of noise.

Consider the SIMO system with measurement noise $w_i(n)$, $i = 1 : L$, as shown in Figure 74. We assume that the measurement noise is white Gaussian noise with zero mean and a known variance σ_w^2 . If the RMRE approach is applied directly with no modifications to account for the noise, the required equalizers will be found but noise amplification will occur. The cause of the noise amplification is not because of division by small values of the frequency response, as is the case in single channel inversion, but instead it is due to

the *coloration* and increase in the variance of the noise as the result of applying the filters $v_{i,d}(n), i = 1 : L$. To illustrate this, assume a two-channel system for which the equalizer pair $v_{1,d}(n), v_{2,d}(n)$ at delay d of minimal order $M - 1$ have been calculated. In this case, the estimated input signal $\hat{s}(n)$ is given by

$$\begin{aligned}\hat{s}(n) &= v_{1,d}(n) * ((s(n) * h_1(n)) + w_1(n)) + v_{2,d}(n) * ((s(n) * h_2(n)) + w_2(n)) \\ &= \alpha \left[s(n-d) + \underbrace{v_{1,d}(n) * w_1(n)}_{i_1(n)} + \underbrace{v_{2,d}(n) * w_2(n)}_{i_2(n)} \right],\end{aligned}\quad (28)$$

where α is non-zero scale factor. The quantities $i_1(n), i_2(n)$ are the sources of the noise amplification. To see why, let us assume that $w_1(n)$ is white gaussian noise with zero mean and variance σ_w^2 . Then the variance of the signal $i_1(n)$ is given by

$$\sigma_{i_1}^2 = \sigma_w^2 \sum_{k=0}^{M-1} (v_{1,d}(k))^2,$$

and a similar expression applies to the variance of $i_2(n)$. If we define $i(n)$ to be $i(n) = i_1(n) + i_2(n)$, then the variance of $i(n)$ will be given by

$$\sigma_i^2 = \sigma_w^2 \left[\sum_{k=0}^{M-1} (v_{1,d}(k))^2 + \sum_{k=0}^{M-1} (v_{2,d}(k))^2 + \sum_{k=0}^{M-1} (v_{1,d}(k)v_{2,d}(k))^2 \right].$$

An interesting question is: does increasing the number of microphones, and consequently decreasing the length of the equalizers help in reducing the magnitude of noise amplification? The answer is that for the room impulse response case, using fewer microphones is better. Since the room impulse responses have shapes similar to decaying exponentials (as explained in Chapter 4), the inverses also tend to have a decaying exponential form, for example as was shown in experiment A in Chapter 4. Increasing the number of microphones makes the equalizers exhibit a more uniform variation and resemble less the decaying exponential shape as, for example, in experiment F in Chapter 4. This in turn makes the norms of the filters $v_{i,d}, i = 1 : L$ larger.

The RMRE method, as proposed, gives good dereverberation results at an SNR of approximately 98dB. The SNR of high quality recording equipment and sound cards, such as the Soundblaster Audigy 2 Platinum, is around 106dB. To demonstrate the performance of the RMRE at $SNR = 98dB$, we use the two measured channel impulse responses described

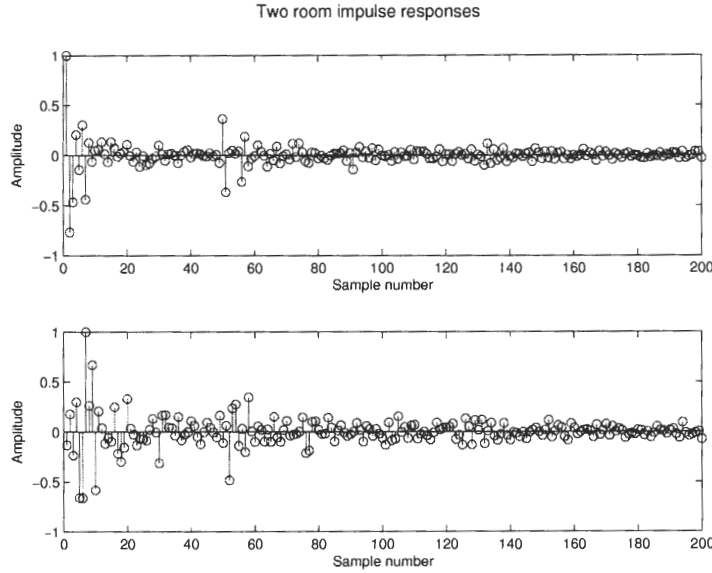


Figure 75: Plot of two measured impulse responses.

in Experiment E of Chapter 4, and specify $M = 199$ as shown in Figure 75. The direct application of the adaptive RMRE method described in the previous chapter yields the learning curve shown in Figure 76. Notice how the convergence is slower and the sharp decrease in the curve is now replaced by a more gradual transition. The resulting equalizers are shown in Figure 77. The norms of the equalizers $\mathbf{v}_0, \mathbf{v}_{199}, \mathbf{v}_{398}$ are $[40.26, 2.973, 6.135]$. In Figure 78 we apply the three equalizers to the impulse responses, and we can see that that the results resemble impulses, especially for the middle equalizer \mathbf{v}_{199} . Using the middle equalizer \mathbf{v}_{199} for dereverberation, we plot the original and restored speech signals along with the difference between the two in Figure 79. The MSE of the difference is 3.24×10^{-4} .

In the previous chapter, we claimed that middle equalizer performs better than the edge equalizers in the presence of noise. This is illustrated in Figure 80 where the MSE of the zero delay equalizer is larger and is equal to $MSE = 0.00156$. Listening to the reconstructed speech signal from the \mathbf{v}_0 equalizer reveals a significantly more noticeable hissing sound associated with the amplification of the noise caused by the larger norm of \mathbf{v}_0 .

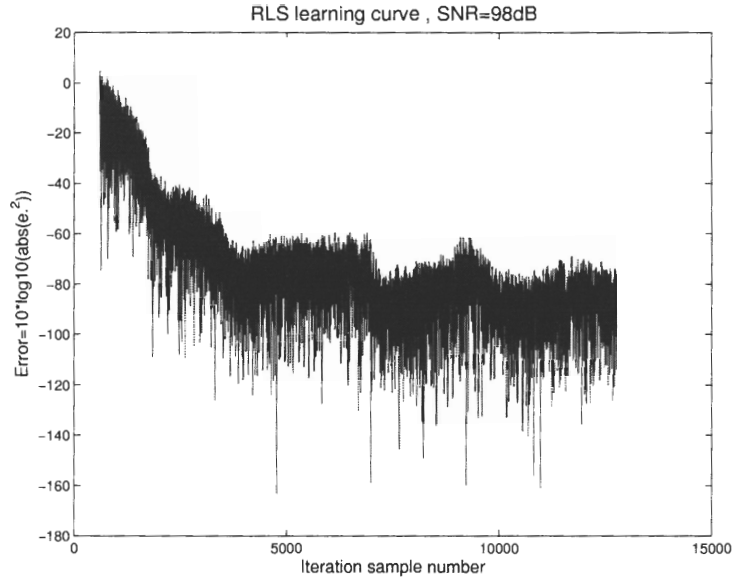


Figure 76: Learning curve using RLS adaptive filter at SNR=98dB.

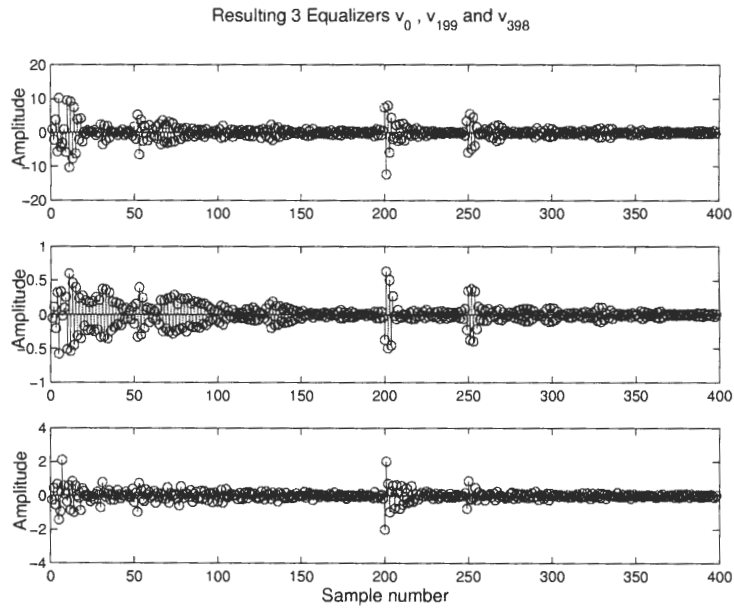


Figure 77: Resulting equalizers at various delays, where $v_{i,d}, i = 1 : 2$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 199, 398$.

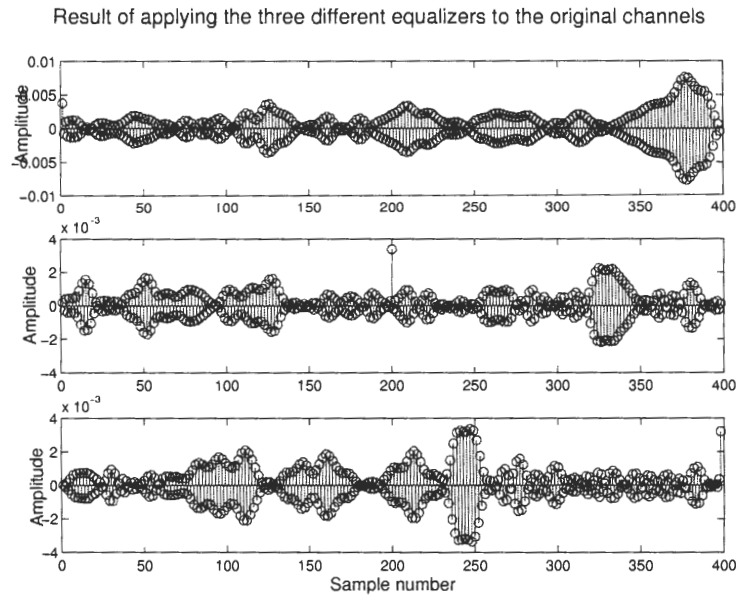


Figure 78: Result of applying equalizers to the channels at $SNR = 98dB$.

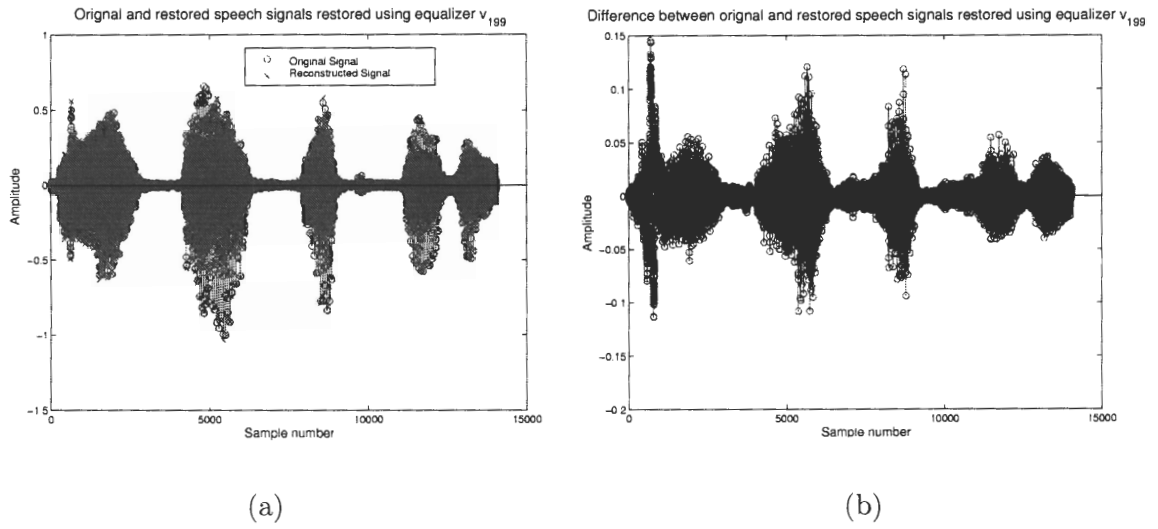


Figure 79: (a) Original and reconstructed speech signals $SNR = 98dB$ utilizing v_{199} , (b) difference between original and reconstructed signals, $MSE = 3.24 \times 10^{-4}$.

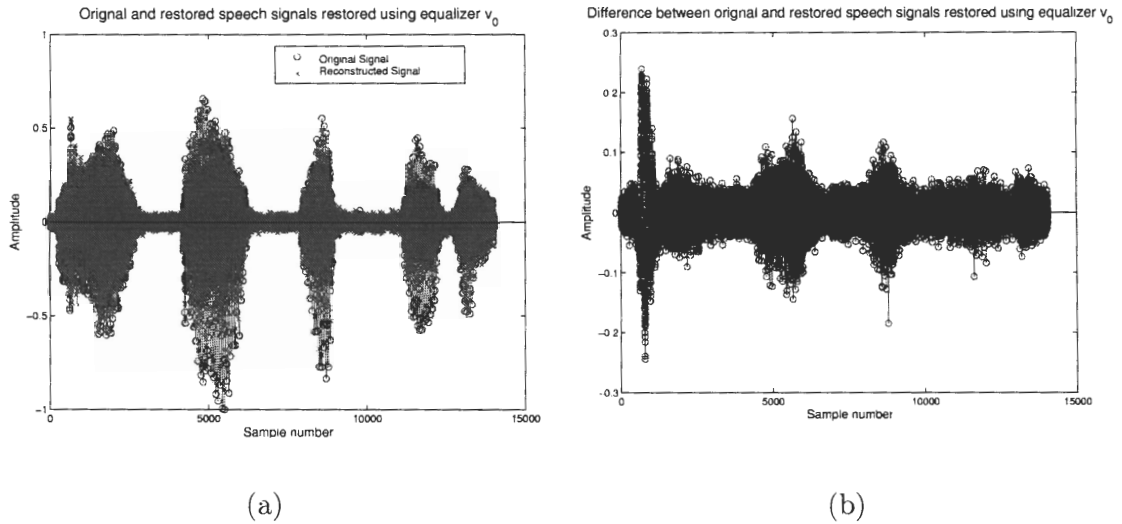


Figure 80: (a) Original and reconstructed speech signals $SNR = 98dB$ utilizing v_0 , (b) difference between original and reconstructed signals, $MSE = 0.00156$.

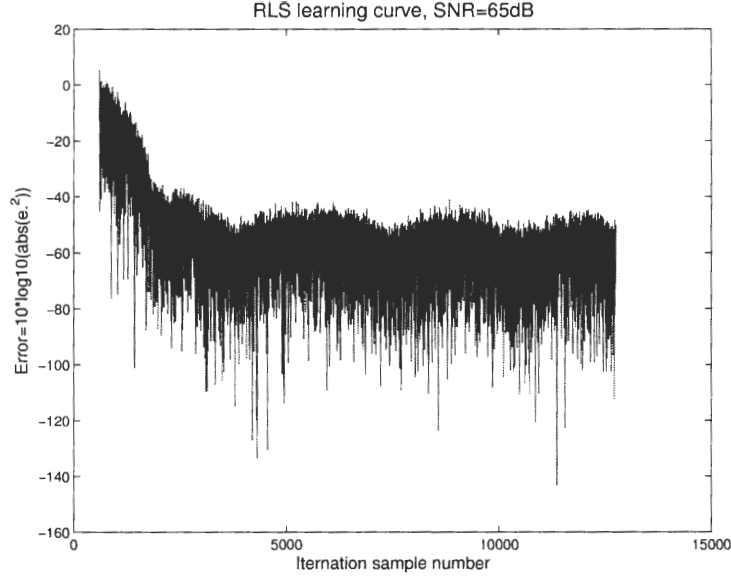


Figure 81: Learning curve using RLS adaptive filter at SNR=65dB.

The RMRE method also gives acceptable dereverberation results at an SNR of approximately 65dB. To demonstrate this, we use the same channels as the $SNR = 95dB$ experiment shown in Figure 75. The direct application of the adaptive RMRE method described in the previous chapter yields the learning curve shown in Figure 81. The resulting equalizers are shown in Figure 82, and the norms of the equalizers $\mathbf{v}_0, \mathbf{v}_{199}, \mathbf{v}_{398}$ are [16.17, 1.211, 2.832]. In Figure 83 we apply the three equalizers to the impulse responses, and in this case, we see less of a resemblance to the desired impulses. Using the middle equalizer \mathbf{v}_{199} for dereverberation, we plot the original and restored speech signals along with the difference between the two in Figure 84, and the MSE of the difference is 0.0378. Once again, the delay zero equalizer reconstruction is shown in Figure 85 where the MSE of the zero delay equalizer is larger and is given by $MSE = 0.0475$. Listening to this reconstructed speech signal from reconstruction using \mathbf{v}_0 reveals a significantly more noticeable hissing sound associated with the amplification of the noise caused by the larger norm of \mathbf{v}_0 .

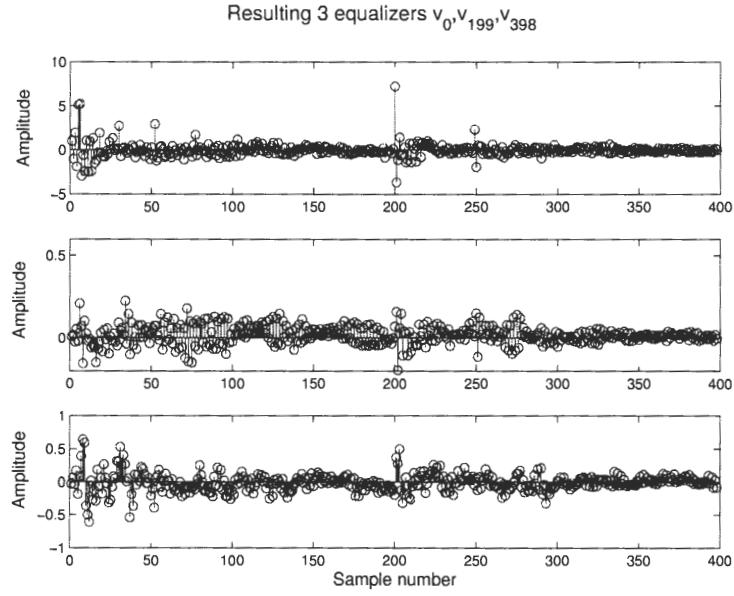


Figure 82: Resulting equalizers at various delays, where $v_{i,d}, i = 1 : 2$ are concatenated to form $\mathbf{v}_d = [v_{1,d}, v_{2,d}]$, $d = 0, 199, 398$.

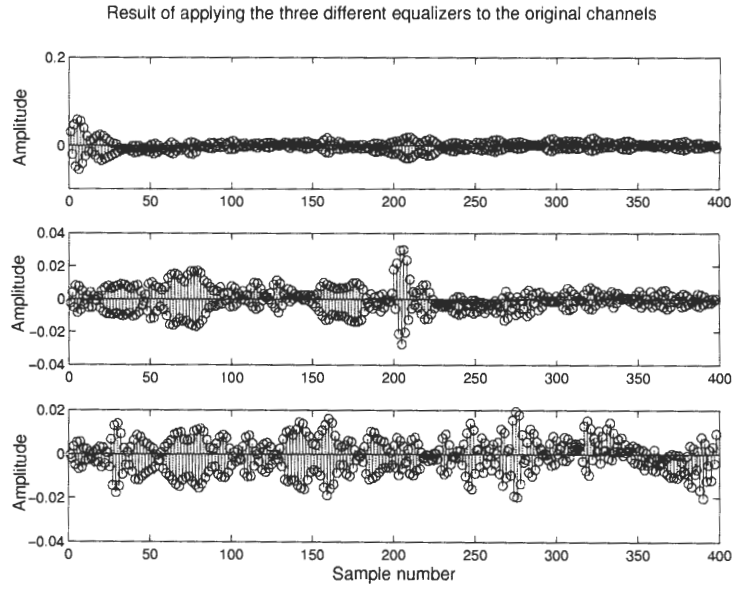
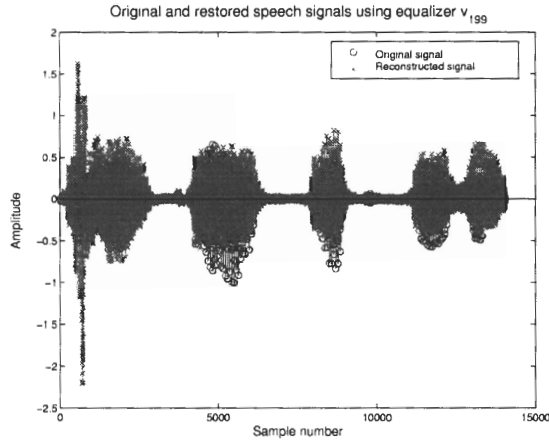
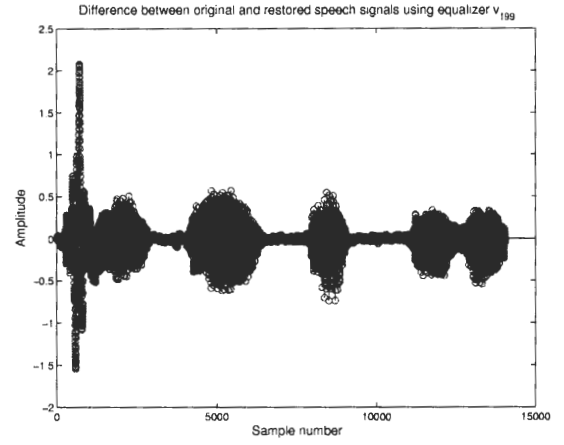


Figure 83: Result of applying equalizers to the channels at $SNR = 65dB$.

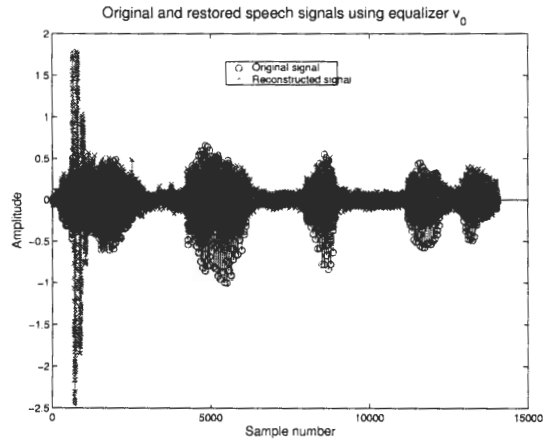


(a)

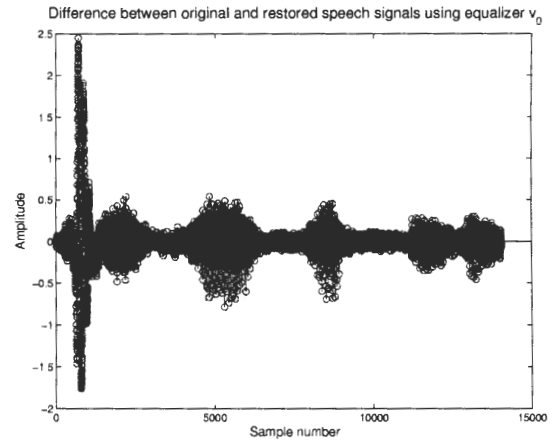


(b)

Figure 84: (a) Original and reconstructed speech signals $SNR = 65dB$ utilizing \mathbf{v}_{199} , (b) difference between original and reconstructed signals, $MSE = 0.0378$.



(a)



(b)

Figure 85: (a) Original and reconstructed speech signals $SNR = 65dB$ utilizing \mathbf{v}_0 , (b) difference between original and reconstructed signals, $MSE = 0.0475$.

5.2.1 Constrained Norm RMRE

While some other dereverberation methods deal with robustness by implementing a noise reduction step and then a dereverberation step, or use a Wiener filter as some of the beamforming methods described in Chapter 2 do, we have opted to seek a way to reformulate the RMRE criterion so that noise immunity is part of the method, and maintain a single stage dereverberation approach. Before discussing our specific modification, we want to emphasize that *our goal is not the removal of measurement noise*, but rather is a dereverberation approach that will work in the presence of noise.

From the above discussion, one remedy to the problem of noise amplification is to constrain the energy of the equalizers, possibly by imposing a unit norm constraint. The RMRE criterion defined in Equation (25) can be reformulated in terms of a constrained (quadratic) minimization of the form

$$\min \mathbf{v}^T R \mathbf{v} = \mathbf{v}^T (R_{\mathbf{x}} + R_{\mathbf{w}}) \mathbf{v} \quad \text{s.t. } \mathbf{v}^T \mathbf{v} = 1, \quad (29)$$

where $R_{\mathbf{x}}$ is given by

$$\begin{bmatrix} R_x[0] & -R_x^T\left[\frac{K}{2}\right] & \mathbf{0} \\ -R_x\left[\frac{K}{2}\right] & 2R_x[0] & -R_x^T\left[\frac{K}{2}\right] \\ \mathbf{0} & -R_x\left[\frac{K}{2}\right] & R_x[0] \end{bmatrix},$$

and $R_{\mathbf{w}}$, the noise correlation, is given by

$$R_{\mathbf{w}} = \sigma_w^2 \begin{bmatrix} I & & \\ & 2I & \\ & & I \end{bmatrix}.$$

Minimizing the criterion in Equation (29) is equivalent to finding the minimum eigenvalue associated with the Rayleigh quotient $\frac{\mathbf{v}^T R \mathbf{v}}{\mathbf{v}^T \mathbf{v}}$. To demonstrate the performance of this approach, we consider the previous system with the two channels shown in Figure 62. Finding the minimum eigenvalue and associated eigenvector gives us all three equalizers $\mathbf{v}_0, \mathbf{v}_{200}, \mathbf{v}_{398}$, and Figure 86 shows the middle equalizer, along with one of the edge equalizers for comparison at an SNR=45dB. Notice how the edge equalizer has a larger amplitude (and hence norm). The result of applying the middle equalizer to the noisy reverberated

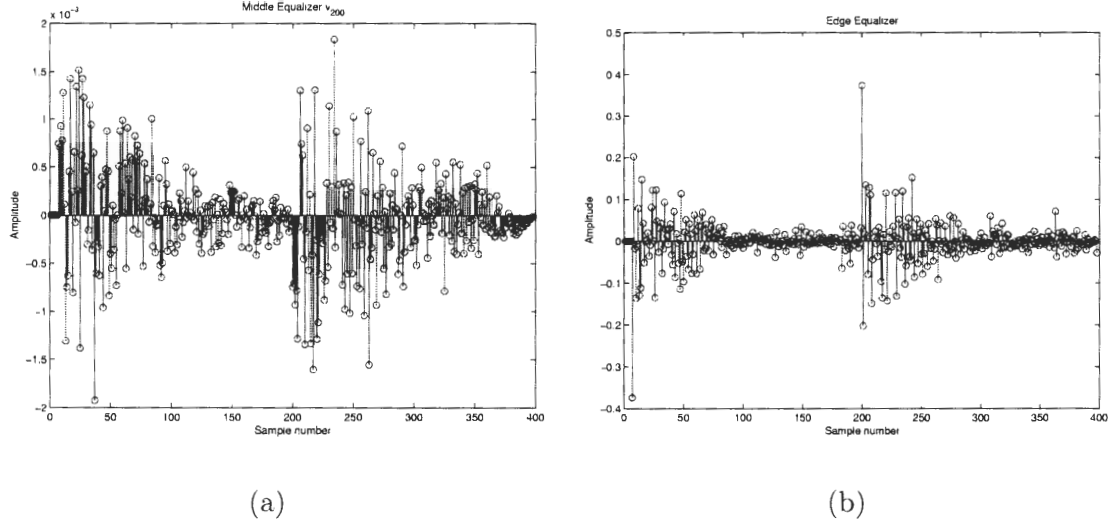


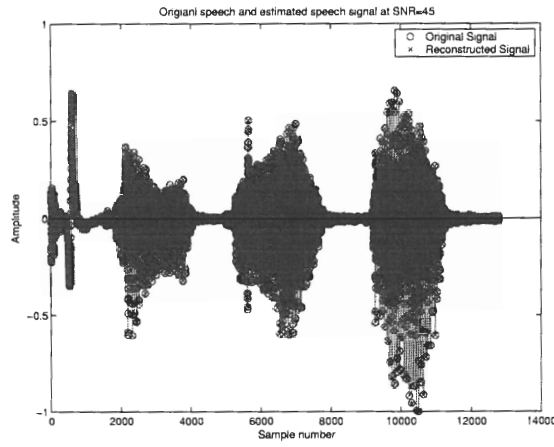
Figure 86: (a) Middle equalizer at an $SNR=45dB$., (b) Edge equalizer \mathbf{v}_{398} at an $SNR=45dB$.

speech recording $x_1(n), x_2(n)$ is shown in Figure 87, and the mean square error in this case is $MSE = 1.69 \times 10^{-5}$. However, the approach described above breaks down at lower SNR , for example at an $SNR = 24dB$, the noise amplification becomes unacceptable and the MSE increases dramatically, as illustrated in Figure 88 where the $MSE = 0.0011$.

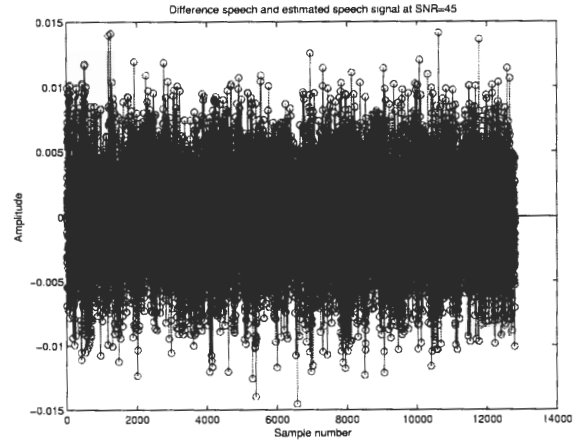
For the low SNR case, we have found that applying a weighting matrix T to the noise correlation matrix helps in reducing the MSE . To explain this idea, we consider the modified criterion

$$\min \mathbf{v}^T R \mathbf{v} = \mathbf{v}^T (R_{\mathbf{x}} + T R_{\mathbf{w}}) \mathbf{v} \quad \text{s.t. } \mathbf{v}^T \mathbf{v} = 1, \text{ where } T = \begin{bmatrix} I & & \\ & (\frac{M+1}{2}) I & \\ & & I \end{bmatrix}.$$

The role of the block diagonal matrix T is to further increase the penalty on the energy in the middle equalizer and dampen its oscillatory behavior. We arrived at the scale factor $(M + 1)/2$ through trial and error, and found that it yields the MSE for a broad range of SNR values. Additionally, our experiments with various permutations of the weighting blocks in T showed that other structures increase the MSE . Figure 89 shows a comparison between the middle equalizer magnitude in the un-weighted and weighted criterion case. Using the middle equalizer with a weighting matrix T applied to the previous case of an

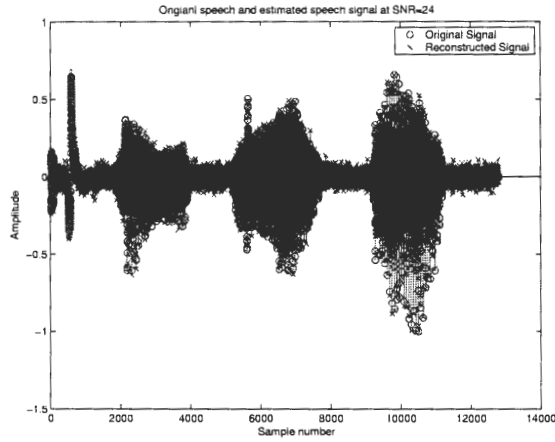


(a)

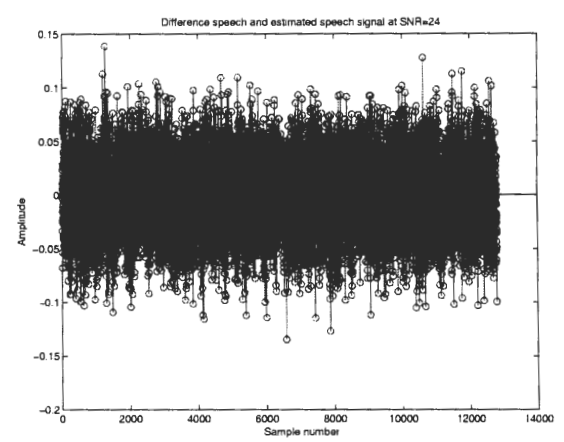


(b)

Figure 87: (a) Original and reconstructed speech signals, (b) difference between original and reconstructed signals at an $SNR=45dB$, $MSE = 1.69 \times 10^{-5}$.



(a)



(b)

Figure 88: (a) Original and reconstructed speech signals $SNR = 24dB$, (b) difference between original and reconstructed signals, $MSE = 0.0011$.

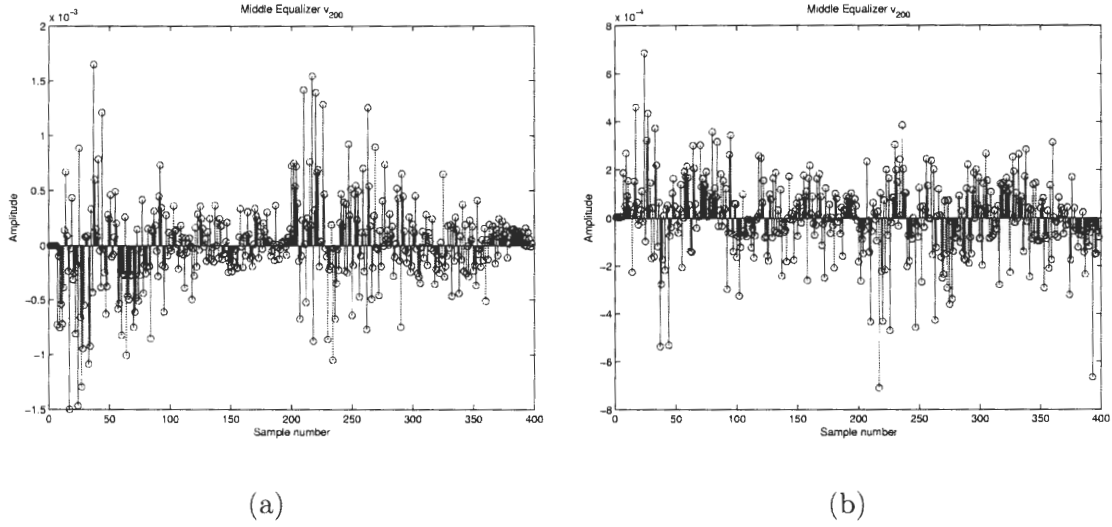
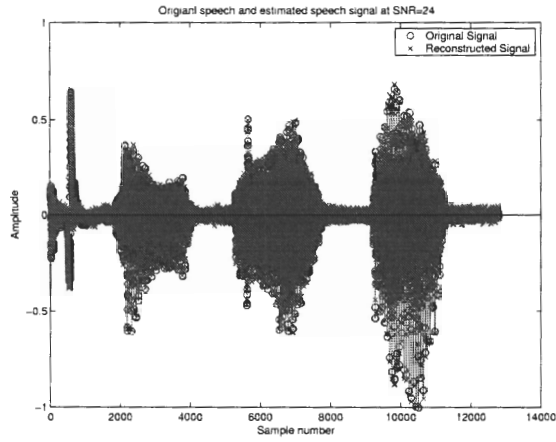


Figure 89: (a) Middle equalizer with no weighting matrix, (b) with weighting matrix T .

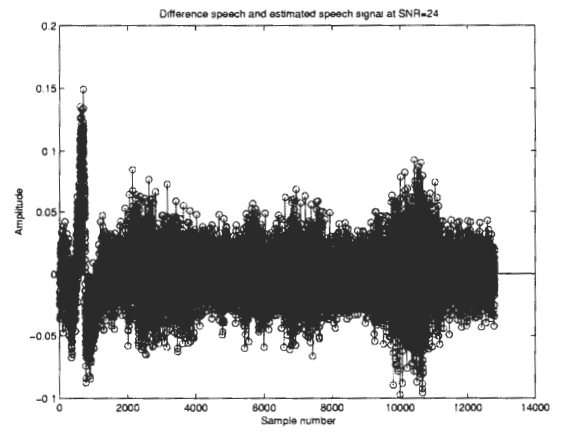
$SNR = 24dB$, the $MSE = 5.39 \times 10^{-4}$ in this case, and Figure 90 shows the resulting reconstructed speech signal, along with the difference between the original and reconstructed speech signals. Notice how there is less noise amplification especially in the silent regions of the speech signal. Listening to the reconstructed signal from the approach involving the matrix T leads to less noise amplification and little residual reverberation in it.

5.3 Summary

In this chapter, we discussed and investigated some of the issues that are needed for any real world implementation of the RMRE method. We demonstrated that under-modeling problems can be combated by increasing the number of microphones, and that over-modeling poses to no problem to the RMRE approach. Additionally, we investigated the effect of measurement noise on the RMRE, and proposed an approach to reformulate the problem in the presence of noise.



(a)



(b)

Figure 90: (a) Original and reconstructed speech signals $SNR = 24dB$ with weighting matrix T , (b) difference between original and reconstructed signals, $MSE = 5.39 \times 10^{-4}$.

CHAPTER VI

CONCLUSION

In this thesis, we have demonstrated the benefit of using the RMRE method for the dereverberation of speech signals in a closed room environment. The RMRE is a modified MRE criterion, adapted for long acoustical impulse responses. The RMRE method finds a subset of three equalizers instead of all possible delay equalizers. The middle equalizer tends to be the most robust one in noisy environments. The ability of the proposed method to adaptively find the equalizers directly without the requirement for a channel impulse response estimation stage, and having no requirement on microphone locations to be in any predefined array geometry makes the method appealing when minimal calibration is required. The RMRE method is able to remove reverberation completely in the noiseless case, and a significant amount of reverberation can be removed at high SNR values. We also proposed a constrained norm RMRE approach that is able to handle lower SNR levels without having the noise amplification problem associated with the standard RMRE method.

Some further issues are worth investigating for both theoretical and practical reasons. The problem of under-modeling may benefit from further investigation to see if the RMRE may be modified to explicitly handle under-modeling errors. In our opinion, this problem is intimately related to robustness of the RMRE in the presence of colored noise.

For any practical real world implementation, where hardware costs must be minimized, it is desirable to find less computationally demanding adaptive algorithms. Advances in the acoustical echo cancelation problem provide many approaches and methods that are applicable to the dereverberation problem. For example, many fast RLS implementations have been developed over years with varying degrees of stability and computational complexity [81]. Affine projection adaptive (APA) filters may also be of value. APA is usually described as having a computational complexity and convergence rate that lies between the

LMS and RLS adaptive methods [10]. Subband adaptive filters are also appealing because of their inherent parallelism. A major problem with critically sampled subband adaptive methods is the spectral leakage that occurs due to the analysis filter bank. One of the earliest studies and remedies of this problem was done in [82] and suggests the use of “cross term” adaptive filters. However this only adds more adaptive filters that need to be adapted and complicates implementation. Some recent research on the use of subband filters for blind channel identification has been proposed by [56]. However, this method requires the use of single sided subbands analysis filter bank leading to complex adaptive filters. A more promising approach in our view is that based on using oversampled filterbanks as described in [83]. While this approach may not be as optimal as critically sampled subband methods, it may add only a small amount of extra processing.

The implementation of a separate noise reduction stage, as a preprocessor for the RMRE method may be worth a second look if it is desired to implement dereverberation in a very noisy environment. Also, the use of the RMRE method as a preprocessor for blind source separation methods, and the performance of the RMRE in a multiple speaker environment are problems that may provide further research topics.

REFERENCES

- [1] E. A. Robinson, Time Series Analysis and Applications, Boston, IRHD, 1981.
- [2] J. G. Proakis, Digital Communications, Boston, McGraw-Hill, 2001.
- [3] R. O. Nielsen, Sonar Signal Processing, Boston, Artech House 1991.
- [4] A. Oppenheim, R. Schafer, and T. Stockham, "Nonlinear filtering of multiplied and convolved signals", Proceeding of the IEEE, Vol. 56, pp. 1264-1291, Aug. 1968.
- [5] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol 36, No. 2, pp. 145-152, Feb. 1988.
- [6] Y. Arden Huang, Real-Time Source Localization with Passive Microphone Arrays, School of Electrical and Computer Engineering, Georgia Institute of Technology 2001.
- [7] L. Tong, G. Xu, and T. Kailath, "A new approach to blind identification and equalization based on multipath channel", In Proc. of the 25th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 1991.
- [8] M. Z. Ikram, Multichannel Blind Separation of Speech Signals in a Reverberant Environment, School of Electrical and Computer Engineering, Georgia Institute of Technology 2001.
- [9] J. R. Deller, J. L. Hansen, and J. G. Proakis, Discrete-Time processing of Speech Signals, IEEE Press, Piscataway NJ, 2000.
- [10] S. L. Gay and J. Benesty, Acoustic Signal Processing For Telecommunication, Kluwer Academic Publishers, 2000.
- [11] Shoko Araki, Shoji Makino, Tsuyoki Nishikawa, Hiroshi Saruwatari; Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech, Proc. ICASSP'01, pp. 2737 - 2740, May 2001.
- [12] T. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, Prentice Hall, Upper Saddle River NJ, 2002.
- [13] B. D. Bees, M. Blostein and O. Kabal, "Reverberant Speech Enhancement Using Cepstral Processing", IEEE International Conference on Acoustics, Speech, and Signal Processing 1991, pp. 977-980.
- [14] S. Subramaniam, A. P. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation", IEEE Trans. Speech and Audio Processing, Vol. 4, No. 5, Sep. 1996.
- [15] A. P. Petropulu and S. Subramaniam, "Cepstrum based deconvolution for speech dereverberation", IEEE International Conference on Acoustics, Speech, and Signal Processing 1994, Vol I, pp. 9-12.

- [16] J.L. Flanagan, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, 78, pp. 1508-1518, Nov. 1985.
- [17] S. Affes and Y. Grenier, "A Signal Subspace Tracking Algorithm for Microphone Array Processing of Speech," *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 5, pp. 425-437, Sept. 1997.
- [18] J. L. Flanagan, D. Berkley, G. Elko, J. West, and M. Sondhi, "Autodirective microphone systems", *Acoustica*, Vol. 73, pp 58-71, 1991.
- [19] Y. Mahieux, G Le Tourneau, and A. Saliou, "A microphone array for multimedia workstations," *J. Audio Eng Soc.*, Vol. 44, pp. 365-372, May 1996.
- [20] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with post filtering," *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 3 pp. 240-259, May 1998.
- [21] W. Kellerman, "A self steering digital microphone array," *Proceedings of the Int. Conf. Acoust., Speech and Audio Processing 1991*, pp. 3581-3584, May 1991.
- [22] J. Gonzalez-Rodrigues, J. L. Bote, and J O. Garcia, "Speech dereverberation and noise reduction with a combined microphone array applications," *Proceedings of the Int. Conf. Acoust., Speech and Audio Processing 1998*, pp. 3613-3616.
- [23] R. Zelinski, "A microphone array with adaptive post filtering for noise reduction in reverberant rooms," *Proceedings of the Int. Conf. Acoust., Speech and Audio Processing 1998*, pp. 2578-2581.
- [24] S. Doclo and, M. Moonen, "Combined Frequency-Domain dereverberation and noise reduction technique for multi-microphone speech and enhancement," *Proc. of the International Workshop on Acoustic Echo and Noise Control 2001*, pp. 31-34.
- [25] S. Affes and Y. Grenier, "Test of adaptive beamformers for speech acquisition in cars," in *Proc. 5th Int. Conf. Signal Processing Applications and Technology 1994*, Vol. I, pp. 154-15.
- [26] J. H. Chen and A. Gesbo, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech Audio Processing*, Vol 3, pp. 59-71, Jan. 1995.
- [27] B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Processing*, Vol 8, No. 3, pp. 267-281, May 2000.
- [28] B. W. Gillespie, H. S. Malvar, D. F. Florencio, "Speech dereverberation via maximum kurtosis subband adaptive filtering," *IEEE International Conference on Signal Processing 2001*, Vol. 6, pp. 3701-3704.
- [29] A. Gonzalez and J. Lopez, "Fast Transversal Filters for Deconvolution in Sound Reproduction," *IEEE Trans. on Speech and Audio Processing*, Vol 9, No. 4, pp. 429-440, May 2001.
- [30] W. Gardner, *Cyclostationarity in Communications and Signal Processing*, New York, IEEE Press, 1994.

- [31] G. Xu, H. Liu, L. Tong and T. Kailath, "A Least-squares Approach to Blind Channel identification," IEEE Trans. Signal Processing, Vol 43, No. 12, pp. 2982-2993, Dec. 1995.
- [32] M. Frikel, V. Barroso and J. Xavier, "Blind Recursive Estimation of SIMO Channels," 2nd IEEE Workshop on Signal Processing Advances in Wireless Communications 1999, pp. 235-238.
- [33] Moulines, P. Duhamel, J. Cardoso, and S. Mayrargue, "Sub-space methods for the blind identification of multichannel FIR filters," IEEE Trans. Signal Processing, Vol. 43, pp. 516-525, Feb. 1995.
- [34] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," IEEE Trans. Signal Processing, Vol. 43, pp. 2982-2993, Dec. 1995.
- [35] H. Liu, G. Xu, and L. Tong, "A deterministic approach to blind identification of multichannel FIR systems," Proceedings of the Int. Conf. Acoust., Speech and Audio Processing 1994, Vol. 4, pp. 581-584.
- [36] Y. Li and Z. Ding, "Blind channel identification based on second-order cyclostationary statistics," Proceedings of the Int. Conf. Acoust., Speech and Audio Processing 1993, Vol. 4, pp. 81-84.
- [37] C. Papadias and T. M. Slock, "Fractionally Spaced Equalization of Linear Polyphase Channels and Related Blind techniques Based on Multichannel Linear Prediction," IEEE Trans. Signal Processing, Vol 47, No. 3, pp. 641-654, March 1999.
- [38] Z. Ding, "An outer product decomposition algorithm for multichannel blind identification," in Proc. 8th IEEE Workshop on Signal Processing Applications (ISSAP) 1996, pp. 132-135.
- [39] Q. Zhao and L. Tong, "Adaptive blind Channel Estimation by Least Squares Smoothing," IEEE Trans. Signal Processing, Vol 47, No. 11, pp. 3000-3012, Nov. 1999.
- [40] Z. Ding and G. Y. Li, Blind Equalization and Identification, New York, Marcel Dekker, 2001.
- [41] K. Abed-Meraim, E. Moulines and P. Loubaton, "Prediction error method for Second-Order blind Identification," IEEE Trans. Signal Processing, Vol 45, No. 3, pp. 694-705, March 1997.
- [42] D. Slock, "Blind fractionally-spaced equalization, perfect-reconstruction filter-banks and multichannel linear prediction," Proceedings of the Int. Conf. Acoust., Speech and Audio Processing 1994 , Vol. 4, pp. 585-588.
- [43] D. Slock, "Blind Fractionally-Spaced Equalization, Perfect Reconstruction Filter Banks and Multichannel Linear Prediction," Proceedings of the Int. Conf. Acoust., Speech and Audio Processing 1994, pp. 585-588.
- [44] A. J. Van der Veen, S. Talwar, and A. Paulraj, "Blind identification of FIR channels carrying multiple finite alphabet signals," Proceedings of the Int. Conf. Acoust., Speech and Audio Processing 1995, pp. 1213-1216.

- [45] G. Giannakis and J. Mendel, "Identification of nonminimum phase systems using higher order statistics," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. 37, pp. 7360377, Mar. 1989. Apr. 1992.
- [46] J. K. Tugnait, "Identification of nonminimum phase linear stochastic systems," *Automatica*, Vol. 22, pp. 454-464, 1986.
- [47] J. K. Tugnait, "Blind equalization and estimation of digital communication FIR channels using cumulant matching," *IEEE Trans. Comm.*, Vol. 43, pp. 1240-1245, Feb 1995.
- [48] W. A. Gardner, "A new method of channel identification," *IEEE Trans. on Communications*, 39(6), June 1991.
- [49] H. Liu, G. Xu, L. Tong, T. Kailath "Recent Developments in Blind Channel Equalization: From Cyclostationarity to Subspaces," *Signal Processing*, 50(1996) 83-99.
- [50] Paulraj, A., R. Roy, T. Kailath, "Estimation of Signal Parameters via Rotational Invariance Techniques-ESPRIT," *Proc. XiXth Asilomar Conf. on Circuits, Systems, and Computers*, Pacific Grove, CA, November 1985, pp. 83-89.
- [51] E. Serpendin and G. Giannakis, "A Simple Proof of a Known Blind Identifiability Result," *IEEE Trans. Signal Processing*, Vol 47, No. 2, pp. 591-593, Feb. 1999.
- [52] A. Scaglione, S. Barbarossa and G. Giannakis, "Inverting Overdetermined Toeplitz System With Application to Blind Block-Adaptive Equalization," *Record of the Thirty-Second Asilomar Conference on Signals, Systems and Computers 1998*, Vol. 2, pp. 1134-1137.
- [53] L. Ljung, *System Identification: Theory for the User*, Prentice Hall, 1999.
- [54] M. Gurelli and C. L. Nikias, "EVAM: An Eigenvector-Based Algorithm for Multichannel Blind Deconvolution of Input Colored Signals," *IEEE Trans. Signal Processing*, Vol 43, No. 1, pp. 134-149, Jan. 1995.
- [55] P. Zhao and J. P. Reilly, "Exponentially Decaying Time-Recursive blind Deconvolution Algorithm for Speech Dereverberation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (ASSP) 1995*, pp. 127-130.
- [56] J. P. Reilly, M. Seibert, M. Wilbur and N. Ahmadvand, "The single sided Subband Decomposition: Application to the Decimation of Large Problems," To be submitted.
- [57] G. Giannakis and C. Tepedelenlioglu, "Direct Blind Equalizers of Multiple FIR Channels: A Deterministic Approach," *IEEE Trans. Signal Processing*, Vol 47, No. 1, pp. 62-74, Jan. 1999.
- [58] H. Luo and R. Liu, "Blind equalizers for Multipath Channels With Best Equalization Delay," *IEEE International Conference on Acoustics, Speech, and Signal Processing 1999*, Vol. 5, pp. 2511-2514.
- [59] D. Gesbert, P. Duhamel and S. Mayrargue, "On-Line Multichannel Equalization Based on Mutually Referenced Filters," *IEEE Trans. Signal Processing*, Vol 45, No. 9, pp. 2307-2317, Sept. 1997.

- [60] M. Tsatsanis and Z. Xu, "Constrained Optimization Method for Direct Blind Equalization," *IEEE Trans. Signal Processing*, Vol 17, No. 3, pp. 424-433, Sept. 1999.
- [61] J. Ayadi and D. Slock, "Multichannel estimation by blind MMSE ZF equalization," *SPAWC 1999*, pp. 251-254, 1999.
- [62] D. Gesbert, A. van der Veen and A. Paulraj, "On the equivalence of Blind Equalizers Based on MRE and Subspace intersections," *IEEE Trans. Signal Processing*, Vol 47, No. 3, pp. 856-859, March 1999.
- [63] J. Shen and Z. Ding, "Direct Blind MMSE Channel Equalization based on Second-Order Statistics," *IEEE Trans. Signal Processing*, Vol 48, No. 4, pp. 1015-1022, April 2000.
- [64] K. Abed-Merain and Y. Hua, "Blind Equalization Methods In Colored Noise Field," *Information, Decision and Control 1999*, pp. 477-481.
- [65] G. Giannakis, S. Halford, "Blind Fractionally Spaced Equalization of Noisy FIR Channels: Direct and Adaptive Solutions," *IEEE Trans. Signal Processing*, Vol 45, No. 9, pp. 2277-2292, Sept. 1997.
- [66] Wirawan, P. Duhamel and H. Maitre, "Multi-Channel High Resolution Blind Image Restoration," *IEEE International Conference on Acoustics, Speech, and Signal Processing 1999*, Vol. 6, pp. 3229-3232.
- [67] H. Kuttruff, *Room Acoustics Fourth Edition*, London, UK, Spon Press, 2000.
- [68] M. kahrs and K. Brandenburg, *Applications of Digital Signal Processing to Audio and Acoustics*, Kluwer Academic Publishers, 1998.
- [69] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, Vol. 65, No. 4, pp. 943-950, April 1979.
- [70] Tsan-Ming Wu, *Statistical Impulse Response Modeling and Dereverberation for Room Acoustics*, School of Electrical and Computer Engineering, Georgia Institute of Technology 2000.
- [71] M. R., "New method of measuring the reverberation time," *J. Acoust. Soc. Am.* 37, 1965, pp. 409-412.
- [72] Christopher Brown, <http://www.mathworks.com/matlabcentral>, File name t60.m, April 2001.
- [73] Real Time Analyzer, <http://www.ymec.com>, Yoshimasa Electronics Inc, 2003.
- [74] T. Kailath, *Linear systems*, Englewood Cliffs, N.J., Prentice-Hall, 1980.
- [75] E. A. Lee, D. G. Messerschmitt, *Digital Communication*, Kluwer Academic Publishers, 1993.
- [76] Tariq Bakir and Russell M. Mersereau, "Blind Adaptive Dereverberation of Speech Signals Using Microphone Arrays," *Eleventh Annual Adaptive Sensor Array Processing Workshop (ASAP)*, March 2003.

- [77] A. Liavas and P. Regalia, "On the behavior of Information Theoretic Criteria for Model Order Selection," *IEEE Trans. Signal Processing*, Vol. 49, No. 8, pp. 1689-1695, Aug. 2001.
- [78] J. Dalmas, H. Gazzah, A. Liavas and P. Regalia, "Statistical Analysis of Some Second Order methods for Blind Channel Identification/Equalization with Respect to channel Undermodelling," *IEEE Trans. Signal Processing*, Vol 48, No. 7, pp. 1984-1998, July 2000.
- [79] A. Liavas, P. Regalia and J. Delmas, "Blind Channel Approximation: Effective Channel Order Determination," *IEEE Trans. Signal Processing*, Vol. 47, No. 12, pp. 3336-3344, Aug. 2001.
- [80] A. Liavas, P. Regalia and J. Delmas, "On the Robustness of Linear Prediction Method for Blind Channel Identification with Respect to Effective Channel Undermodelling/Overmodelling," *IEEE Trans. Signal Processing*, Vol. 48, No. 15, pp. 1477-1481, May 2000.
- [81] Z. Liu, "QR Methods of $O(N)$ Complexity in Adaptive Parameter Estimation," *IEEE Trans. Signal Processing*, Vol. 43, No. 3, pp. 720-729, March 1995.
- [82] A. Gilloire and M. Vetterli, "Adaptive Filtering in Subbands: Analysis, Experiments, and Application to Acoustic Echo Cancellation," *IEEE Trans. Signal Processing*, Vol. 40, No. 8, pp. 1862-1875, Aug. 1992.
- [83] Z. Cvetkovic and M. Vetterli, "Oversampled Filter Banks," *IEEE Trans. Signal Processing*, Vol. 46, No. 5, pp. 1245-1255, May 1998.

VITA

Tariq Saad Bakir, born May 1975 Blacksburg VA, graduated from Blacksburg High School in 1993. In Fall 1993, he attended Auburn University and graduated with a B.S. in Electrical Engineering (Summa Cum Laude) in 1997. He obtained an M.S. in Electrical Engineering in Fall 1998 from Auburn University, and the thesis title was "A filter design method for minimizing blurring in a region of interest in low resolution MRE images" supervised by Prof. Stanley Reeves. He joined the Georgia Institute of Technology Ph.D. program in Fall 1998 where worked under the supervision of Prof. Russell Mersereau. His research interests are in image and signal enhancement and reconstruction, array processing, statistical and adaptive signal processing, smart antennas and wireless communication systems.