Fall 2010

Reverberation reduction in a room for multiple positions

Raghavendra Ravikumar

Department:

REVERBERATION REDUCTION IN A ROOM FOR MULTIPLE POSITIONS

by

RAGHAVENDRA RAVIKUMAR

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

2010

Approved by

Dr. Steven L. Grant, Advisor
Dr. Randy H. Moss
Dr. Sahra Sedighsarvestani

ABSTRACT

Reverberation in a room occurs when the direct path sound from a sound source undergoes multiple reflections from the walls of the room before reaching the listener. An impulse response of the room can be measured called the room impulse response (RIR) which captures the effects of the room. This can be represented digitally on a computer. A filter is designed to cancel the effects of the room using the information in the room impulse response. This filter is called an equalization filter and is usually placed between the source signal and loudspeaker to perform the equalization. The RIR changes for varying source and listener locations, hence an equalization filter designed for one RIR will not perform equalization for multiple positions. This thesis explores methods to perform equalization for multiple positions. One of the simplest methods is spatial averaging equalization, which was used to perform the equalization for multiple positions. Equalizing RIR is only concerned about trying to flatten the frequency spectrum and stabilizing the inverse RIR by looking at its minimum-phase component. Other methods are explored which consider the masking effects of the human auditory system which relates to the perception of sound by the human ear. One such method is impulse response shortening/reshaping which emphasizes the direct path component in the RIR relative to the rest of the components using p-norm and infinity-norm optimization which is an iterative algorithm. This concept is extended for performing reshaping on RIR for multiple positions using the idea in spatial averaging equalization by using RIR's measure for different positions.

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Dr. Steven L. Grant for accepting me as a Research Assistant to work on a very interesting and challenging project. He was very helpful in answering any question, simple or difficult asked to him and shared a great deal of knowledge and experience. Conducting research was always an enjoyable experience under him as his ideas were challenging and stimulated a lot of thought process on some interesting problems in Signal Processing. Research meetings and discussions allowed in identifying areas in which I have faltered and proceed in the right direction. Overall it has been a privilege to work under Dr. Grant who has accomplished and contributed a lot towards research in Signal Processing giving me inspiration to pursue my interest in Signal Processing.

I also thank Dr. Randy H. Moss and Dr. Sehra Sedigh for being part of my thesis committee on a short notice. I thank them for going through my entire thesis and providing helpful comments to improve the entire documentation and presentation of my thesis.

I would also like to thank my research teammate Pratik V. Shah, currently pursuing his PhD in Electrical Engineering under Dr. Grant. He has been helpful in providing the right guidance while understanding any concept and in being systematic. He has also helped me understand my mistakes and supported my views during research meetings.

Research was sponsored by the Leonard Wood Institute in cooperation with the U.S. Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-07-2-0062. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Leonard Wood Institute, the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation heron.

TABLE OF CONTENTS

Page

ABSTRACT iii

ACKNOWLEDGMENTS iv

LIST OF ILLUSTRATIONS vii

LIST OF TABLES ix

SECTION

1. INTRODUCTION 1
    1.1. EQUALIZATION 2
    1.2. SINGLE POINT EQUALIZATION 2
    1.3. MULTIPLE-POINT EQUALIZATION 4
    1.4. REVERBERATION REDUCTION 6
2. EQUALIZATION FOR MULTIPLE POSITIONS 8
    2.1. BACKGROUND 8
    2.2. SPATIAL AVERAGING EQUALIZATION 9
    2.3. MULTIPLE INPUT/OUTPUT INVERSE THEOREM (MINT) 11
        2.3.1. The Principle 12
        2.3.2. Computation of FIR filters for Exact Inversion 13
        2.3.3. Multiple-Input Multiple-Output System 14
        2.3.4. Results 15
3. SHORTENING/RESHAPING OF IMPULSE RESPONSES 19
    3.1. ROOM-REVERBERATION COMPENSATION 19
    3.2. MASKING EFFECTS OF HUMAN AUDITORY SYSTEM 19
    3.3. FORWARD MASKING LEVEL 22
        3.3.1. Masker Level and Signal Delay 23
        3.3.2. Frequency 23
    3.4. FREQUENCY DOMAIN PSYCHOACOUSTICS 23
    3.5. LISTENING ROOM COMPENSATION 24
    3.6. CONCEPT OF IMPULSE RESPONSE RESHAPING/SHORTENING 25
    3.7. INFINITY-NORM OPTIMIZATION 31

3.8. P-NORM OPTIMIZATION 32
3.9. WINDOW FUNCTIONS 35
3.9.1. Reshaping Window. 35
3.9.2. Shortening Window 36
3.10. SIMULATIONS 37
3.10.1. Reshaping 37
3.10.2. Shortening 38
4. RESHAPING IMPULSE RESPONSES FOR MULTIPLE POSITIONS 41
4.1. PURPOSE 41
4.2. METHOD I 44
4.3. METHOD II 49
4.4. COMPARISON OF METHOD I AND METHOD II 54
5. CONCLUSION AND FUTURE WORK 58
APPENDIX 59
BIBLIOGRAPHY 61
VITA 63

LIST OF ILLUSTRATIONS

Figure Page

1.1. Block Diagram of Single Point Pre-Filtering Equalization System 3

1.2. Least Sqaures Equalization Setup 4

1.3. Block Diagram of Multiple-Point Equalization System 5

1.4. Setup for Multi-Channel Least Squares Equalization 6

2.1. Frequency Domain Plots of the Room Impulse Responses 10

2.2. Frequency Domain Plot with only one Room Impulse Response Equalized 10

2.3. Magnitude Responses after Spatial Average Equalization of Responses at the Two Positions 11

2.4. Conventional Inverse Filtering Method 12

2.5. Inverse Filtering Method Based on MINT 12

2.6. De-reverberation using MINT 13

2.7. Inverse Filtering Method for Multiple Input Multiple Output System 14

2.8. Original Two Channel Room Impulse Responses 15

2.9. Equalized Response 16

2.10. Equalized Response for Shorter Filter Lengths 16

2.11. Variation of Error Energies for Different Filter Lengths 17

3.1. Single Channel Setup for Listening Room Compensation 24

3.2. Setup for Listening Room Compensation using Least Squares 25

3.3. Maximization Windows as a Function of Time 27

3.4. Original Response, Shortening Filter and Global Impulse Response (top to bottom) 29

3.5. Decay of Global Impulse Response $g(n)$ 29

3.6. Magnitude Frequency Response of Shortened Global System Response 30

3.7. Logarithm Reciprocal of Window Function w_0(n) (Equation (40)) 36

3.8. Original Filter, Reshaping Filter and Global Impulse Response (top to bottom) 38

3.9. Decay of the Different Responses (reshaping) 39

3.10. Original Filter, Shortening Filter, Global Impulse Response (top-bottom) 39

3.11. Decay of the Different Responses (shortening) 40

4.1. The Original Impulse Response, Reshaping Filter and Global Impulse Response for Location 1 (top to bottom). 42
4.2. Comparison of the Responses in the Logarithmic Scale with the Masking Curve.... 42
4.3. The Test Room Impulse Response, Pre-Filter and their Global Response..... 43
4.4. Illustration of the Global Response $g_{test}(n)$ lying above the Masking Curve..... 43
4.5. Experimental Setup..... 46
4.6. Logarithm Curves for Location 1 46
4.7. Logarithmic Curves for Location 2..... 47
4.8. Logarithmic Curves for Location 3..... 47
4.9. Logarithmic Curves for Location 4..... 48
4.10. Logarithmic Curves for Location 5..... 48
4.11. Logarithmic Curves for Test Location..... 49
4.12. Logarithm Curves for Location 1 51
4.13. Logarithm Curves for Location 2 51
4.14. Logarithm Curves for Location 3 52
4.15. Logarithm Curves for Location 4 52
4.16. Logarithm Curves for Location 5 53
4.17. Logarithm Curves for Reference Location 53
4.18. Comparison for Location 1 54
4.19. Comparison for Test Location 55

LIST OF TABLES

Table Page

4.1. EDM Values of Method I and Method II..... 56

1. INTRODUCTION

An acoustic enclosure can be modeled as a linear system whose characteristics can be described mathematically by a response known as the impulse response, $h(n)$. In case the enclosure happens to be a room, it is called a room impulse response. The room impulse response has a frequency response represented by $H(e^{j\omega})$ which is called the room transfer function. The sound originating from the source reaches the receiver via a direct path and after reflections via a multipath due to the presence of reflecting walls and objects. This phenomenon is described by the room impulse response.

In a reverberant room these reflections cause distortions in the amplitude and phase in the sound received at the microphone when placed at a distance away from the sound source. This causes the human listener to perceive echo and reverberation in the sound and speech signals transmitted from a loudspeaker. This affects the intelligibility of speech and sound at the listener. In other words, the listener is not able to hear the original speech and sound signal. Applications involving multiple loudspeakers require each loudspeaker to be placed at a specific location and sound from each loudspeaker should be distinct. In such applications it is not desirable for the sound to be distorted by the room, since it would render the identification of sound difficult. One such application is Surround Sound which requires each loudspeaker to produce distinct sounds to create a 3-dimensional surround sound experience.

There are methods of acoustically reducing the reverberation and echo by using sound absorbing foams on the walls, curtains or panels in the room. They help in absorbing the reflections thereby removing the distortions in the speech signal due to the room. However, these foams are highly expensive and to install them in room of large sizes would add to the setup cost of the audio system. A cost effective option is to analyze the room impulse response and the speech signal and remove the reverberation and echo electronically. One of way of achieve this is by cancelling the effects of the room in the sound and speech signal. This method is called room equalization. In another method the aspects of the human auditory system are used such that the direct path component of the room impulse response is made to sound louder than the other components.

1.1. EQUALIZATION

The sound recorded at the microphone can be considered to be a combination of sound originating directly from the source and many planes waves due to multiple reflections of the original sound wave from the walls. These travel in different directions encountering the walls at different angles of incidence. In the time domain these reflections are perceived as echoes and reverberation which are delayed attenuated versions of the original source signal. The process of equalization involves reducing the effects of reflection from the wall surface by using an inverse filter designed to compensate for the unevenness in the room transfer function at the microphone position. This equalization filter is applied to the source signal before it is transmitted into the room. If $h_{eq}(n)$ is the equalization filter for the room impulse response $h(n)$, then for perfect equalization $h_{eq}(n) \otimes h(n) = \delta(n)$ where $\otimes$ is the convolution operator and $\delta(n) = 1,\ n = 0\,;\ 0,\ n \neq 0$ is the Kronecker delta function. The problems associated with this are however, (i) the room response is usually not invertible (not minimum phase), (ii) designing an equalization filter for a specific position will introduce poor equalization performance at other positions in the room. This means the equalization filter that is designed to equalize the response for one position will not work for responses recorded at other positions. This is because the sound pressure is different at different points in the room. Thus based on the requirements, equalization is done for single point and multiple points.

1.2. SINGLE POINT EQUALIZATION

In a single point equalization system, equalization is between a single source and single receiver and is usually done by pre-filtering as shown in Figure 1.1. The function $H(z)$ is the room transfer function between the source and receiver, $F(z)$ is the equalization filter and $X(z)$ and $Y(z)$ are the input and output signals, respectively expressed in the z-domain. Output signal is expressed as $Y(z)=H(z)F(z)X(z)$. The perfect equalization filter is the inverse filter $F(z)=H^{-1}(z)$. This inverts both the magnitude and phase of the frequency response. However, if $H(z)$ is non-minimum phase the filter becomes unstable.

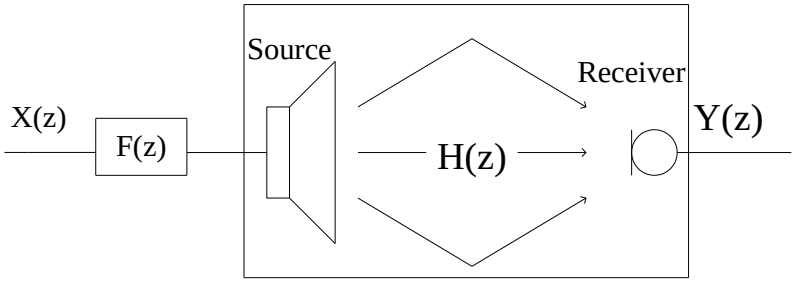

Figure 1.1. Block Diagram of Single Point Pre-Filtering Equalization System

Inverse Filter Based on Least-Square Error. To solve the problem of unstable inverses, inverse filtering is done by using the least-squares method. If $x(k)$ is the input signal, the output signal can be written as $y(k) = \sum_{n=0}^{N-1} h(n) x_f(k-n)$ where,

$$x_f(k) = \sum_{n=0}^{L-1} f(n) x(k-n) \tag{1}$$

which is the pre-filtered input signal. The squared error between the delayed original signal $x(k\text{-}d)$ and equalized signal $y(k)$ is given by,

$$\varepsilon = \sum_{k=0}^{\infty} e^2(k) = \sum_{k=0}^{\infty} \left[x(k-d) - y(k) \right]^2 \tag{2}$$

d is used to model the delay in the input signal. After solving for the minimization, the equalization filter can be calculated using the matrix equation,

$$\mathbf{f} = (\mathbf{Y}^{\mathbf{T}}\mathbf{Y})^{-1}\mathbf{Y}\mathbf{x} \tag{3}$$

Where,

$$\mathbf{Y} = \begin{bmatrix} y(0) & & \\ y(1) & y(0) & \\ \vdots & y(1) & \ddots \\ y(m) & \vdots & y(0) \\ 0 & y(m) & y(1) \\ \vdots & 0 & \vdots \\ \vdots & \vdots & y(m) \end{bmatrix} \tag{4}$$

The above matrix has L rows and m+1 columns and m+1 is the length of y. The vector **x** contains the input signal and is given by,

$$\mathbf{x} = \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(m) \end{bmatrix}$$

The setup for this technique is shown in Figure 1.2.

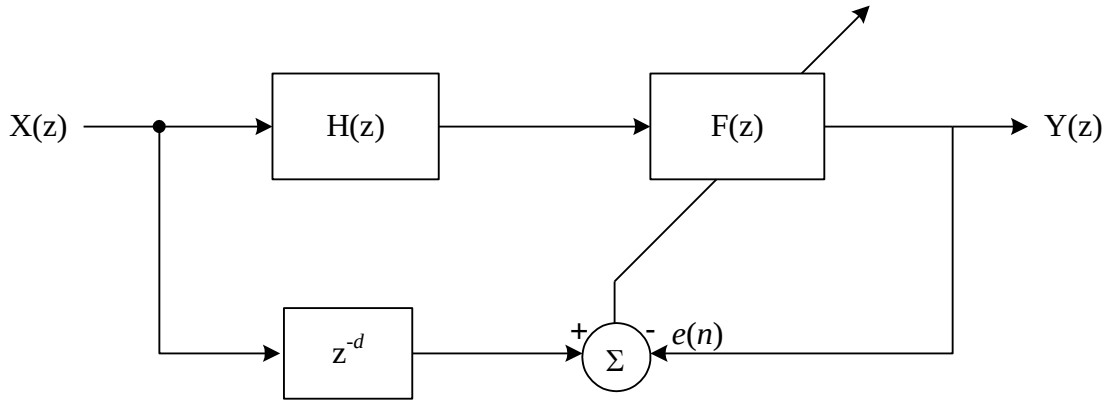

Figure 1.2. Least Sqaures Equalization Setup

However, due to changing impulse responses at different positions with respect to a fixed sound source it desired to perform equalization at multiple positions. Section 2 discusses two standards methods used practically to perform multiple point equalization.

1.3. MULTIPLE-POINT EQUALIZATION

The Figure 1.3 shows a multiple-point equalization system where there is a single source and microphones at multiple positions. It uses a single inverse filter, $F(z)$. The functions, $H_i(z)$ and $Y_i(z)$ are the room transfer functions and output at each microphones, respectively. The number of microphones is M. A perfect equalization filter cannot be achieved for all microphone positions because the room transfer functions have different phase responses. However, the sections that follow discuss methods that achieve reasonable equalization.

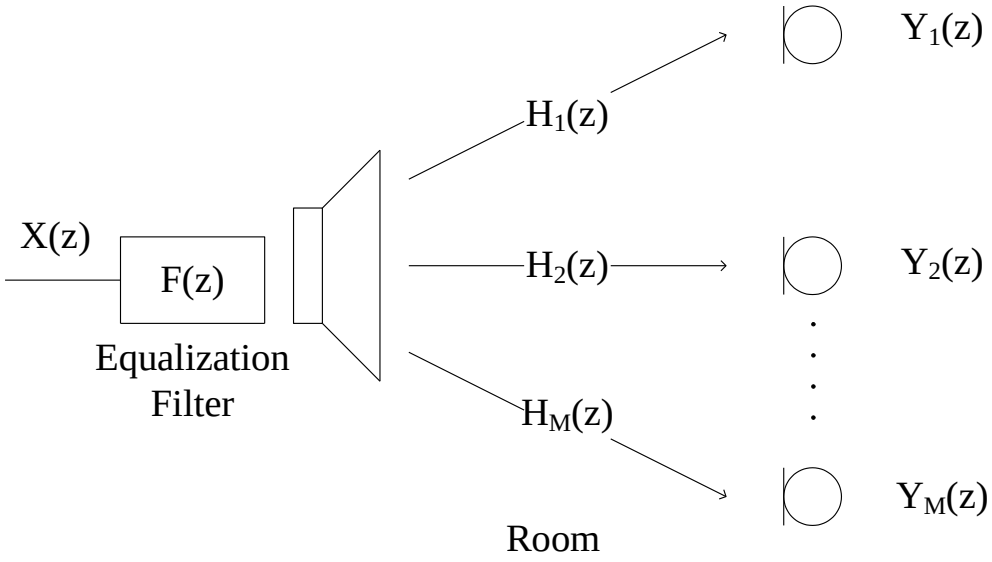

Figure 1.3. Block Diagram of Multiple-Point Equalization System

Least-Squares Method. In the time domain, the relationship between the input and output signal can be written as,

$$y_i(k) = \sum_{n=0}^{N-1} h_i(n) x_f(k-n) \tag{5}$$

Where,
$$x_f(k) = \sum_{n=0}^{L-1} f(n) x(k-n) \tag{6}$$

$h_i(n)$ is the i[th] room impulse response. The filter coefficients represented in f(n), $n=0,1,..L$-1 are used to minimize the cost function, ε. This cost function is the sum of squares of the error between the delayed input signals $x(k-d_i)$ and output signals $y_i(k)$.

$$\varepsilon = \sum_{i=1}^{M} \sum_{k=0}^{\infty} e_i(k) = \sum_{i=1}^{M} \sum_{k=0}^{\infty} \left[x(k-d_i) - y_i(k) \right]^2 \tag{7}$$

The modeling delays d_i (i=1,…M) are set differently reflecting the difference in the propagation times of the direct sound in each of the room impulse responses in the system. The equalization filter tries to recover the waveforms of the original source signals. The setup is shown below in Figure 1.4.

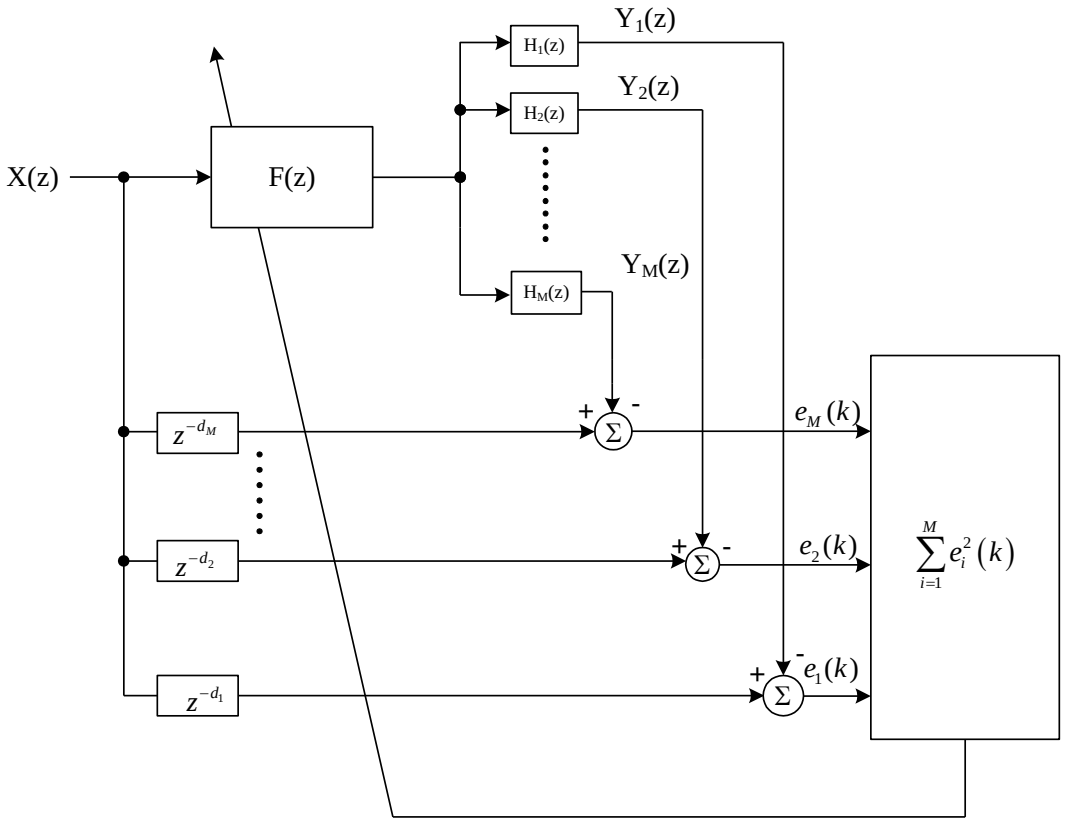

Figure 1.4. Setup for Multi-Channel Least Squares Equalization

The least squares method appears to be a reasonable approach mathematically but, it does not reflect the physical characteristics of the room impulse response. In case of multiple point equalization, it equalizes the common and unique parts of the room impulse responses.

1.4. REVERBERATION REDUCTION

In this method the room impulse response is separated into desirable and undesirable components. The components defined as undesirable in the room transfer function are removed. The process is divided into three steps: separation of the undesired transfer function components from the original room transfer function; de-reverberation

of the undesirable components and the addition of desired components.

Psycho acoustically derived criteria is developed which is used to influence the de-reverberation process. This incorporates the temporal masking properties of the human ear. A simplified set of rules for determining the audibility of components can be formulated using forward and backward masking concepts[11]. This idea is used to design windows which are used in separating the desirable and undesirable components in the room impulse responses.

Based on the perceptual approach, the human listener does not perceive all the detailed information contained in the room impulse response, since many room reflections are masked by the direct sound and other reflections and thus rendering these room reflections inaudible. This approach gives a basic idea of sound perception in a room. Therefore, to achieve de-reverberation, the direct path component can be maximized while minimizing the other components of the room impulse response thus removing most of the effects of the room. Section 3 discusses this approach in more detail.

2. EQUALIZATION FOR MULTIPLE POSITIONS

2.1. BACKGROUND

To understand the effects of single location equalization on other locations, consider a simple first order room reflection model as follows. Let $h_1(n)$ and $h_2(n)$ be the impulse responses from a single source to two positions 1 and 2, respectively. They are represented as,

$$\begin{aligned} h_1(n) &= \delta(n) + \alpha_2\delta(n-1); |\alpha_2| < 1 \\ h_2(n) &= \delta(n) + \beta_2\delta(n-1) \\ \beta_2 &\neq \alpha_2 \end{aligned} \tag{8}$$

This first order reflection model is valid. Consider two positions located along the same radius from a source, and each position has neighboring walls which absorb sound differently and negligible higher-order reflections from each wall. For simplicity, the absorption due to air and the propagation delay is ignored in this model. Ideal equalization at position 1 is achieved if the equalizing filter, $h_{eq}(n)$, is

$$h_{eq}(n) = (-\alpha_2)^n u(n) \tag{9}$$

Where, $u(n) = 1,$ for $n \geq 0$ is a discrete unit step function. Therefore, $h_{eq}(n) \otimes h_1(n) = \delta(n)$. However, the equalized response at position 2 can be shown to be,

$$h_{eq}(n) \otimes h_2(n) = \delta(n) - (\alpha_2 - \beta_2)(-\alpha_2)^{n-1} u(n-1) \tag{10}$$

There are two objective measures of the equalization performance for position 2, (i) frequency domain error function and (ii) time domain error function. The time domain error function can be computed easily which represents the deviation from the ideal equalized response (delta function) and can be defined as,

$$\begin{aligned} \varepsilon &= \frac{1}{I}\sum_{n=0}^{I-1} e^2(n) = \frac{1}{I}\sum_{n=0}^{I-1}\left(\delta(n) - h_{eq}(n) \otimes h_2(n)\right)^2 \\ &= \frac{(\alpha_2 - \beta_2)^2}{I}\sum_{n=1}^{I-1}(-\alpha_2)^{(2n-2)} \end{aligned}$$

The response at position 2 is clearly not equalized because, $\varepsilon > 0$. Thus to achieve good equalization, equalizers have to be designed such that it accounts for the changes in the room response due to variations in the source and listening position.

2.2. SPATIAL AVERAGING EQUALIZATION

One of the goals of equalization is to minimize the spectral deviations (peak and dips) in the magnitude frequency response through an equalization filter. There sound played through the loudspeaker system is therefore significantly improved through this correction. In essence, the system resulting from the equalization filter and the room response should have a perceptually flat frequency response.

The room impulse responses were generated using the image derived model [4]. There frequency domain responses were plotted in Figure 2.1. It can be seen from the plots that the room impulse responses have a lot of spectral deviations and they are different. If the spectral deviations are made flat by the use of a filter, the quality of sound played back through the loudspeaker system will be improved.

An equalization filter has to be designed such that the spectral deviations in the magnitude of the frequency response are minimized over a large space in the listening environment and simultaneously for multiple listeners. An example of performing single point equalization is shown in Figure 2.2. The top plot shows the equalization done for position 1. The bottom plot shows the equalization done for position 2 using the same equalizing filter. It can be clearly seen that the performance is degraded.

One method for providing equalization simultaneously is by spatially averaging the measured room responses at different positions for a given loudspeaker and stably inverting the result [1]. The microphones are positioned such that they correspond to the center of the listener's head.

The RMS (Root Mean Square) method is used widely due to its simplicity for computing equalization filter and the spatial average of the measured responses is given by,

$$H_{avg}(e^{j\omega}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|H_i(e^{j\omega})\right|^2} \tag{12}$$

$$H_{eq}(e^{j\omega}) = H_{avg}^{-1}(e^{j\omega})$$

Where, N is the number of listening positions, with responses $H_i(e^{j\omega})$ that are to be equalized. It is aimed at achieving uniform frequency response coverage for all listeners. The performance of the spectral average equalization is shown in the Figure 2.3. It can be

seen that the spectral deviations are minimized for both positions using the spatial average equalization filter.

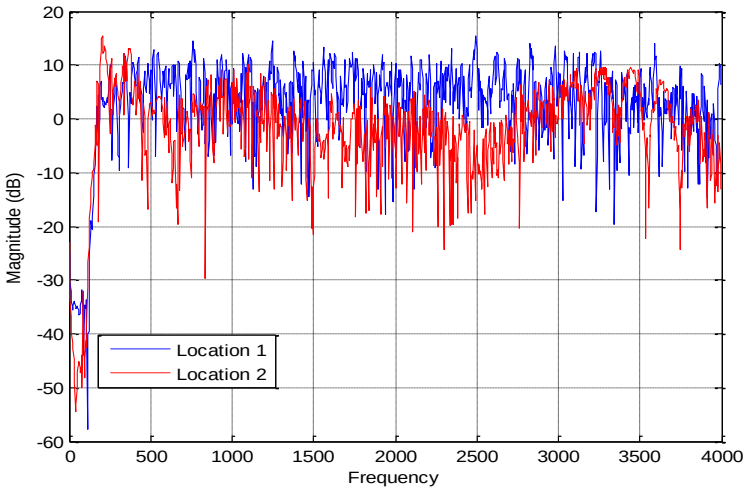

Figure 2.1. Frequency Domain Plots of the Room Impulse Responses

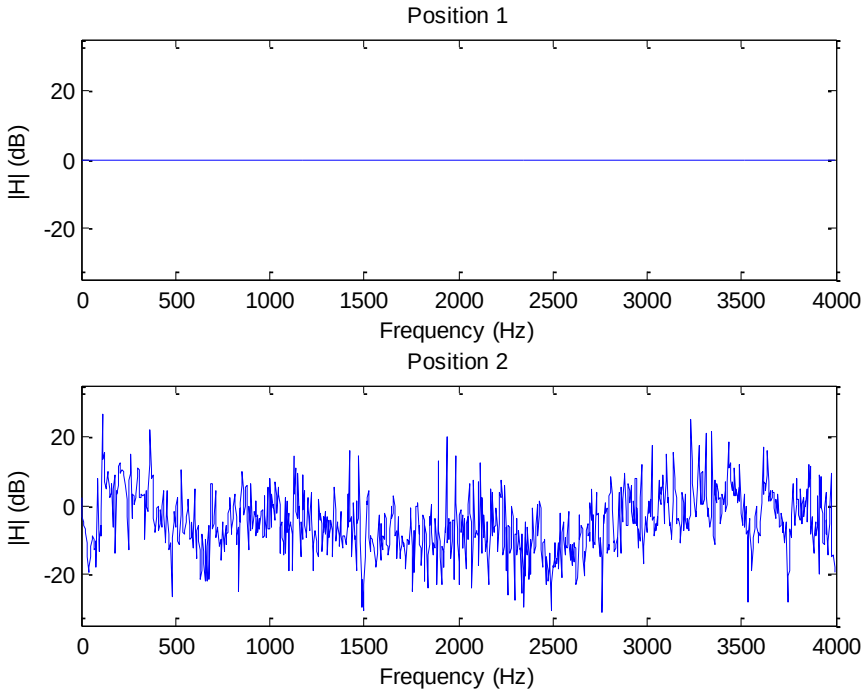

Figure 2.2. Frequency Domain Plot with only one Room Impulse Response Equalized

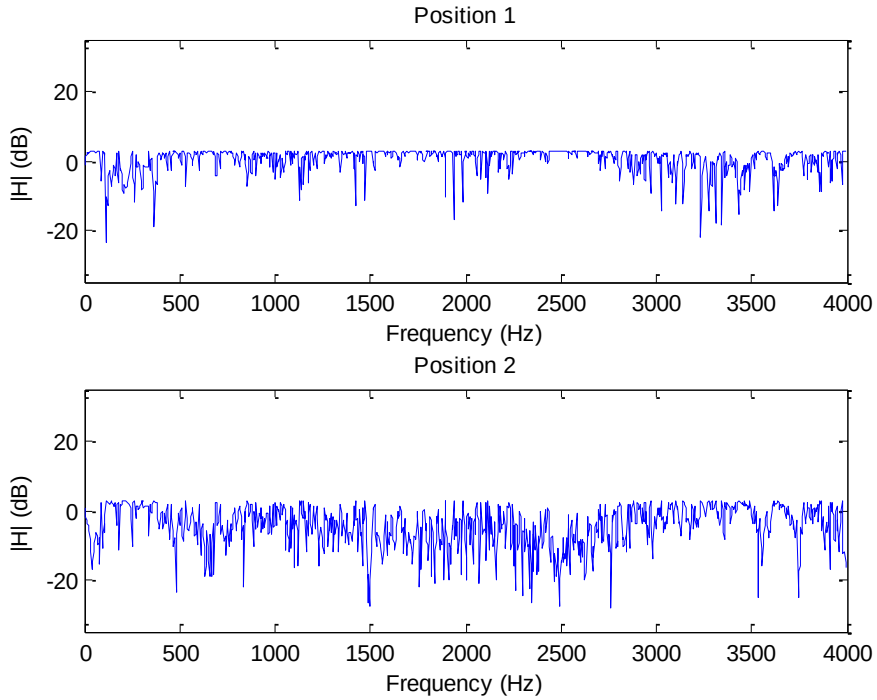

Figure 2.3. Magnitude Responses after Spatial Average Equalization of Responses at the Two Positions

However, the performance of spatial averaging can be limited by (i) a mismatch between the microphone measurement location and actual location for the center of the human head and (ii) variations in the listener's position. It also equalizes both audible and inaudible frequencies, since it equalizes all frequencies. In the time domain this causes certain unnecessary components in the room impulse response also to be heard.

2.3. MULTIPLE INPUT/OUTPUT INVERSE THEOREM (MINT)

Since all impulse responses do not have stable inverses, it is difficult to realize their exact inverses. This method realizes the exact inverses by constructing inverse from multiple FIR filters by adding extra signal transmission channels produced by multiple loudspeakers or microphones [7]. The coefficients of these FIR filters are computed using well known concepts of matrix algebra.

2.3.1. The Principle. Figure 2.5 shows a two input single output FIR filter. This system is obtained by adding an extra signal transmitting channel to the linear system shown in Figure 2.4. The two signal transmission channels are denoted as $C_1(z^{-1})$ and $C_2(z^{-1})$ and the two FIR filters $H_1(z^{-1})$ and $H_2(z^{-1})$ are connected to the inputs of $C_1(z^{-1})$ and $C_2(z^{-1})$, respectively. To realize the inverse filtering, $H_1(z^{-1})$ and $H_2(z^{-1})$ must satisfy the expression,

$$D\left(z^{-1}\right)=1=C_1\left(z^{-1}\right)C_1\left(z^{-1}\right)+C_2\left(z^{-1}\right)C_2\left(z^{-1}\right) \tag{13}$$

where $D(z^{-1})$ is the z-transform of $d(k)$ given by $d(k)=c(k)*h(k)$ according to Figure 2.4. Since $C_1(z^{-1})$, $C_2(z^{-1})$, $H_1(z^{-1})$ and $H_2(z^{-1})$ are polynomials in z^{-1}, a solution set of (13) has the following properties,

- Solution for (13) exist if and only if $C_1(z^{-1})$ and $C_2(z^{-1})$ do not have any common zeros in the z-plane.
- If (13) has a solution, it is unique if the orders of $H_1(z^{-1})$ and $H_2(z^{-1})$ are less than those of $C_2(z^{-1})$ and $C_1(z^{-1})$, respectively.

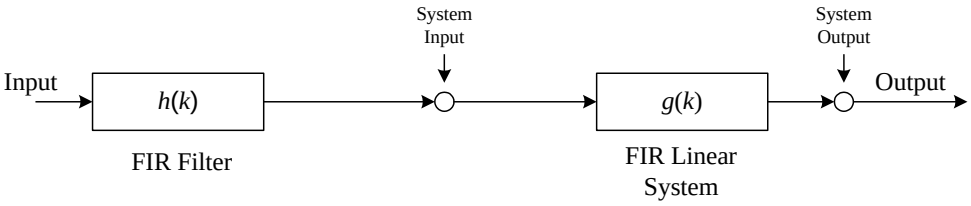

Figure 2.4. Conventional Inverse Filtering Method

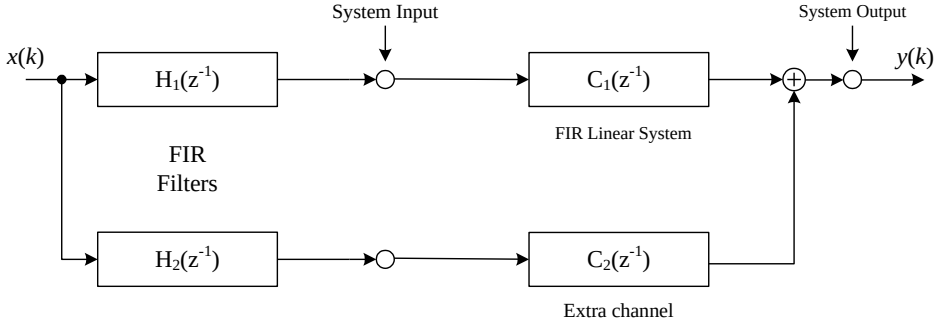

Figure 2.5. Inverse Filtering Method Based on MINT

This concept is useful to de-reverberate the acoustic signals transmitted in a room which involves two microphones. The system is shown in Figure 2.6. The transmission channels from source S to microphones M_1 and M_2 are denoted as $C_1(z^{-1})$ and $C_2(z^{-1})$, respectively. This system is equivalent to a single input two output linear FIR system. The output signals after the microphones and FIR filters $H_1(z^{-1})$ and $H_2(z^{-1})$ are added to satisfy (13).

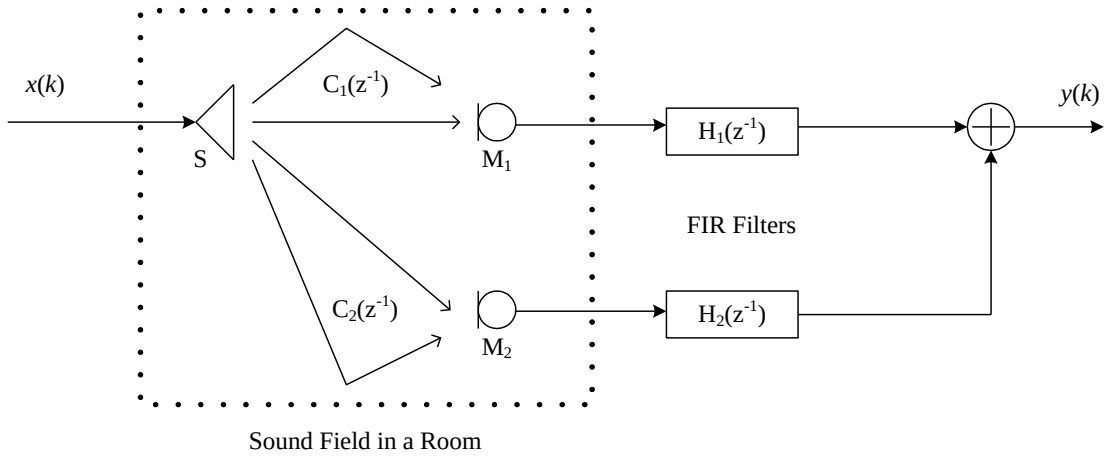

Figure 2.6. De-reverberation using MINT

2.3.2. Computation of FIR Filters for Exact Inversion. To simplify the explanation consider Figure 2.4. Equation (13) can be rewritten in the time domain as,

$$d(k) = c_1(k) * h_1(k) + c_2(k) * h_2(k) \tag{14}$$

where,

$$d(k) = \begin{cases} 1 & \text{when k=0} \\ 0 & \text{when k=1,2,....} \end{cases}$$

This can be expressed in matrix form as,

$$\mathbf{d} = \mathbf{C}_1\mathbf{h}_1 + \mathbf{C}_2\mathbf{h}_2 = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} \tag{15}$$

or,

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ \\ \\ \\ \\ 0 \end{bmatrix} = \begin{bmatrix} c_1(0) & & 0 & c_2(0) & & 0 \\ c_1(1) & \ddots & \vdots & c_2(1) & \ddots & \vdots \\ \vdots & & & \vdots & & \\ & & c_1(0) & c_2(n) & & \\ c_1(m) & & c_1(1) & 0 & \ddots & c_2(0) \\ 0 & \ddots & \vdots & \vdots & & c_2(1) \\ \vdots & & & & & \vdots \\ & & c_1(m) & & & c_2(n) \end{bmatrix} \begin{bmatrix} h_1(0) \\ h_2(1) \\ \vdots \\ \\ h_1(i) \\ h_2(0) \\ \vdots \\ h_2(j) \end{bmatrix} \tag{16}$$

Vector **d** is of size L+1 where, $L=m+i=n+j$. Where, m+1 and n+1 are the durations of g_1 and g_2, respectively and i and j are the orders of h_1 and h_2, respectively. The coefficients of the FIR filters $h_1(k)$ and $h_2(k)$ can now be computed using the relationship,

$$\begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} = \left[\mathbf{C}_1\ \mathbf{C}_2\right]^{-1} \mathbf{d} \tag{17}$$

2.3.3. Multiple-Input Multiple-Output System. The above mentioned concept can be extended to invert a multiple-input multiple-output FIR system. This can be used to cancel the effects of the room impulse response at multiple points in a room. The block diagram is shown in Figure 2.7.

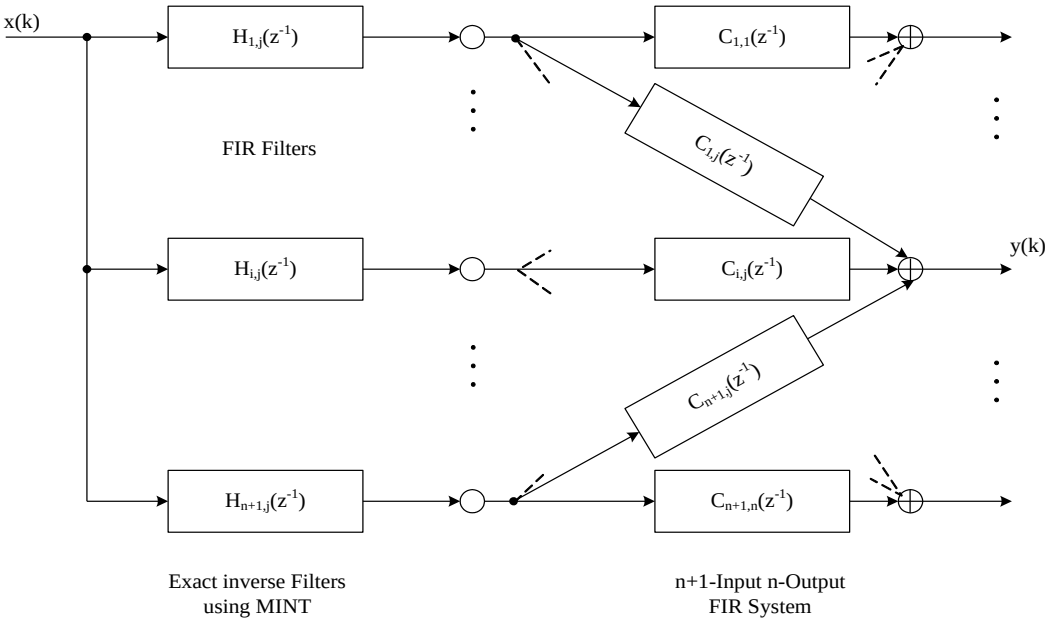

Figure 2.7. Inverse Filtering Method for Multiple Input Multiple Output System

The above system is an n+1-input, n-output system. In the above figure, $C_{i,j}(z^{-1})$ (i=1,2,…., n+1; j=1,2,…,n) is denoted as a signal transmission channel between the i^{th} input and the j^{th} output of the system. $H_{i,j}(z^{-1})$ denotes the FIR filter connected to the i^{th} input of the system. By using the principle of MINT the exact inverse of a multiple-input multiple-output linear FIR system can be realized.

2.3.4. Results. The algorithm was tested for a two channel case by taking a single source and two microphones at two different locations. The image derived model [4] was used to generate the impulse responses at the microphones. The impulse responses at the two channels are represented as c_1 and c_2, respectively. The size of the room chosen was 36 feet by 18 feet by 15 feet. The size of the impulse responses generated were 1024 at a sampling rate of 8 kHz. The impulse responses are shown in Figure 2.8. The equalized response is shown in Figure 2.9. The length of the filter was chosen to be 1024 taps. This is considered the true length, since it has the same length as that of the original room impulse response. It can be seen that the equalized response given by Equation (14) is a unit impulse response which is what the algorithm desires to achieve.

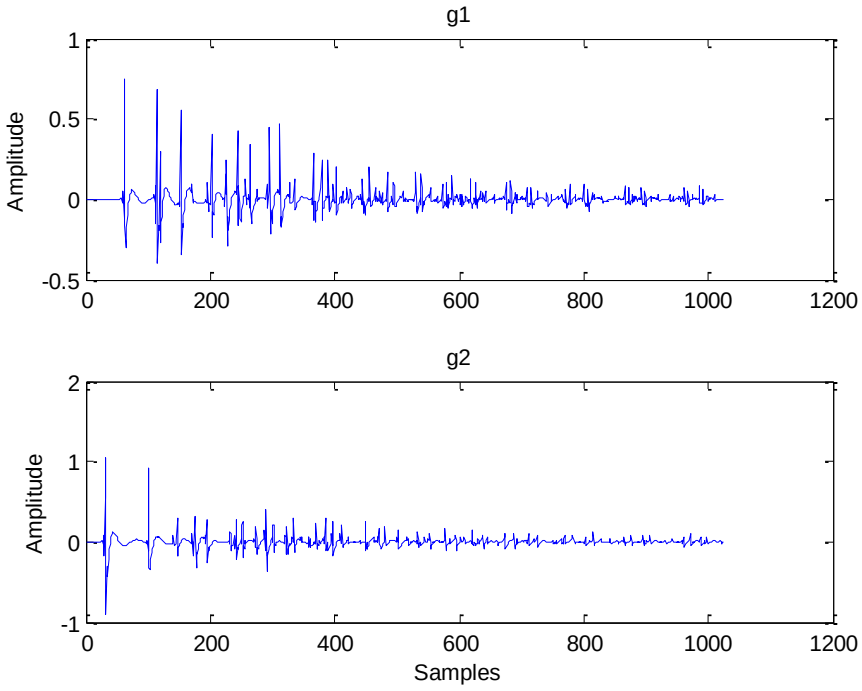

Figure 2.8. Original Two Channel Room Impulse Responses

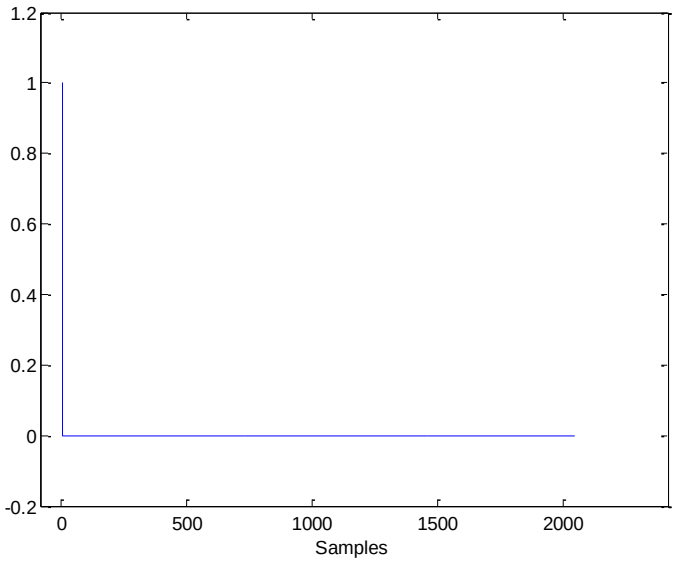

Figure 2.9. Equalized Response

The filters were now designed by choosing a shorter length of 600 taps. The equalized response was plotted and is shown in Figure 2.10. It can be seen that the equalized response has a lot of distortion throughout.

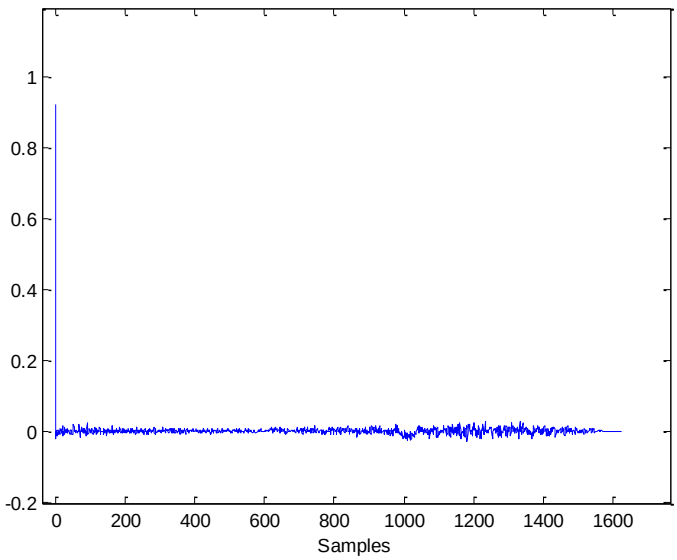

Figure 2.10. Equalized Response for Shorter Filter Lengths

The filter lengths were varied and the difference between the equalized responses for different filter lengths and the ideal equalized response were taken. If $d(n)$ represents the equalized response given by Equation (14) for different lengths of the filters $h_1(n)$ and $h_2(n)$ and if $d_{\text{true}}(n)$ represents the equalized response for true lengths of the filter $h_1(n)$ and $h_2(n)$, then the error is given as $\mathbf{e} = \mathbf{d}_{true} - \mathbf{d}$. True length is the actual length of the room impulse response or the length of the sound transmission channel. The error energy is computed using the equation, $\varepsilon = \mathbf{e}^T\mathbf{e}$. The variation of the error energies for different filter lengths was plotted and is shown in Figure 2.11. It can be seen that the results are better if the filter lengths are greater than the true length.

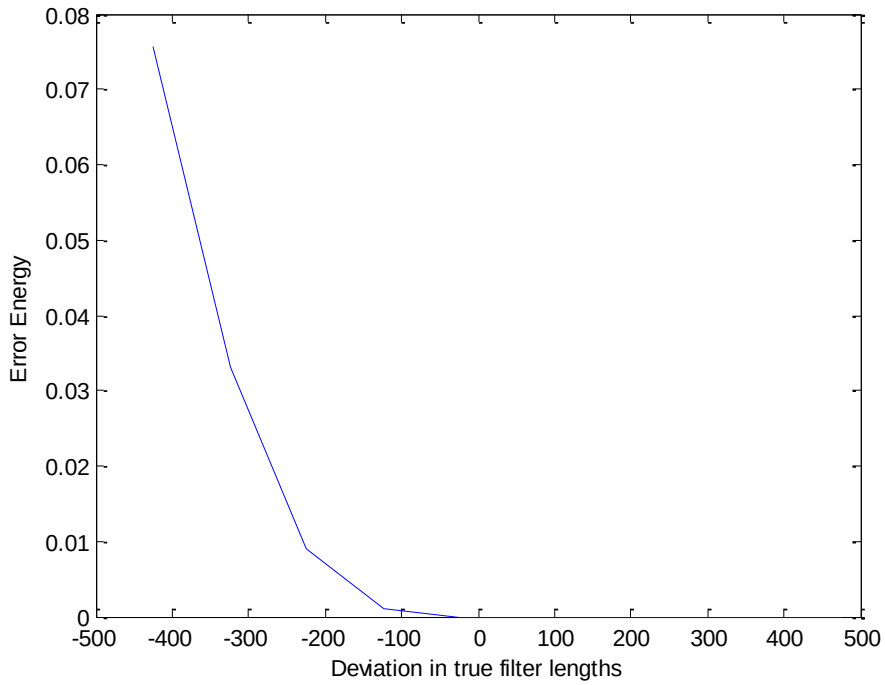

Figure 2.11. Variation of Error Energies for Different Filter Lengths

Thus, in this method it is required to know the true length to be able achieve perfect equalization. Since error is really small for lengths greater than the true length, larger length filters are required to achieve perfect equalization. However, in an actual

system it is difficult to accurately determine the true length of the impulse response. For larger length channels, even larger filter length is required which adds to the difficulty in computing the inverses given in Equation (17). Thus, methods which provide flexibility in designing filter of any length have to be looked at. The following section provides the development of a method which does not require this constraint of length and in addition explains the psychoacoustic properties of the human ear required for the development of the method.

3. SHORTENING/RESHAPING OF IMPULSE RESPONSES

3.1. ROOM-REVERBERATION COMPENSATION

For the enhancement of speech intelligibility in reverberant rooms, the loudspeaker signals need to be preprocessed to compensate for the reverberation. This approach is slightly different from the approach of channel equalization. In channel equalization, the objective is to try and recover the original signal from the received signal which is achieved by inverting the channel. In room-reverberation compensation, only the channel needs to be compensated in such a way that signal is perceived without reverberation. In other words, the room impulse response is only partially equalized such that all the audible echoes are removed and the inaudible echoes remain. This would ease the problem of trying to design a compensation system. This approach takes into consideration the psychoacoustic properties of the human auditory system and design pre-filters that are optimized to give best intelligibility. For better understanding of room-reverberation compensation, some psychoacoustic criteria or mainly the temporal masking effects of the human auditory system will be discussed next.

3.2. MASKING EFFECTS OF HUMAN AUDITORY SYSTEM

One way to determine the audibility of reflections is by considering its amplitude. Depending on a number of parameters, if a low level reflection can be masked by the direct sound component, the listener is unable to perceive the reflection. By increasing the amplitude of reflection, a Reflection Masking Threshold (RMT) is reached and the reflection becomes audible and its effect is observed as variation in timbre and loudness and still temporarily fused with the direct sound [12]. Further increase in reflection amplitude leads to the Echo Threshold (ET) being reached. The reflection then is heard as an echo. The Reflection Masked Threshold (RMT) can be defined as the amplitude threshold below which the human listener is unable to perceive single reflection, multiple reflections and reverberation. This effect of perceiving includes all possible sound attributes such as loudness, spatiality, localization, coloration, timbre, temporal structure, etc.

Consider a direct sound and a test reflection, the direct sound masks the test reflection. This is the concept of ordinary masking. Ordinary masking can be classified into simultaneous masking or post-masking and non-simultaneous masking or pre-masking. This is not the case with room masking, since the reflected sound will overlap, extend and succeed the direct sound. Thus both the effects of ordinary masking might appear simultaneously. For room masking, pre-masking effects are negligible especially for signals arriving from different directions [12].

The authors Bochholz et al. [12] derive a perceptual model for such room masking effects and propose that the Room Masking Function (RMT) can be describe by a functions of nine parameters. That is, $RMT = f\left\{\phi_d, \delta_d, \phi_r, \delta_r, \tau_r, p_d, f_{d,r}, n_r, s_d\right\}$. These parameters are described as follows:

ϕ_d, δ_d are the incidence angles of azimuth and elevation, respectively of the direct signal with respect to the orientation of the listener's head which is at $\phi_d = 0^\circ$ and $\delta_d = 0^\circ$ when the listener is looking directly towards the direct sound source. For single test reflection the direct sound angle of reflection may affect the RMT by 10dB as described in [12].

ϕ_r, δ_r are the incident angles of azimuth and elevation, respectively of the test reflection relative to the orientation of the listener's head. The RMT of a test reflection depends on the angle of the direct sound and the orientation of the listener's head. It is also mentioned in [12] that the masking effect was found to be strongest for equal directions of incidence of the direct sound and the test reflection, but the RMT was found to be 10dB lower for different directions. Also, changing elevation has the same effect as changing the azimuth.

τ_r is the time delay of the test reflection. For noise bursts RMT increases linearly with increasing delay time[12]. This decay can be described by a time constant which increases for increasing direct sound levels. By observing the curves of RMT for different direct sound levels, we can determine a maximum delay time τ_{max} which is of the order of a hundred milliseconds.

L_d is the sound level of the direct signal. For noise bursts the RMT decreases linearly with increasing sound level of the direct signal. This implies for louder sounds a

room reflection is easier to perceive than for the case of softer sounds [12]. For an absolute RMT, this increases linearly with increase in sound level.

$f_{d,r}$ describes the frequency content (spectrum) of the direct sound and of the test reflection. It is mentioned that the masking effect of the direct sound is strong if the spectral distribution of the direct sound and the test reflection coincide. In a realistic environment the frequency dependence of reflectivity on room surface or boundaries leads to attenuation of high frequency components of the reflected sounds.

n describes the combined effect of additional reflections and reverberation. By adding diffused reverberation to the anechoic signals increases the RMT of the single test reflection. In other references listed in [12] it is described that additional reflections can cause RMT to be raised or extended in time or in some cases replace the direct sound as the masker.

s_d is a signal dependent parameter which describes the effect of the type of the direct signal. The RMT for a test reflection has a strong signal dependency [12]. Based on this, the aspects of time overlap between the direct and the reflected signal and the effective duration of the signal has to be considered [12].

The above explanation for RMT suggests that masking effects of the human auditory system are signal dependent. This calls for signal dependent filtering to achieve ultimate performance. A good compromise between the masking curves obtained for various signals is the average masking curve. By using optimality criteria based on the average masking curve, linear signal independent filtering can be used.

Thus more emphasis can be laid on non-simultaneous masking in determining the audibility of time varying signals in a simpler manner to deal with the complex nature of loudspeaker-room transfer function and de-reverberation filters. Non simultaneous masking is divided in two types, backward masking and forward masking. In backward masking, the masked signal occurs before the louder masker and situation is reversed in forward masking.

Backward masking depends significantly on the training of the listeners. It is mentioned in [10] that untrained listeners experience substantial amount of backward masking whereas trained listeners experience little or none. It is also indicated that backward masking effects were completely gone if the masked signal preceded masker

by 20ms. Also, significant portion of backward masking disappears in approximately 5ms. Thus sound components occurring more that 15ms earlier will be audible only in isolation. Thus, it can be concluded that backward masking limit is 15ms.

Forward masking is dependent on the type of the masker and masked signal. The effect of forward masking is highly dependent on frequency relationship between the masker and masked signal [10]. It was determined in one of the references that forward masking effect begins as simultaneous masking and falls in a straight line on a linear-log scale of masking reduction in decibel versus time. Forward masking has been found to extend 100-200ms. It is also indicates the average forward masking criterion which is defined as having no reduction of masking compared to simultaneous masking for shorter time intervals of about 4ms and later falls at a rate of 35dB/decade. Thus, it can be concluded that forward masking acts like simultaneous masking for the first 4ms and then falls off at 35dB/decade.

3.3. FORWARD MASKING LEVEL

The forward masking of a sinusoid signal by the same sinusoid was investigated for frequencies ranging from 125 and 4000 Hz in [13]. Forward masking in decibels is proportional to both masker level and log signal delay at each frequency. More forward masking occurs at low frequencies than at high frequencies with the maskers being at the same sensation levels. Masked thresholds are greater at low frequencies than at high frequencies with maskers having equal sound pressure level. Several experiments were conducted by [13] to estimate forward masking level as a function of masker level and signal level and to observe the effects of frequency. In all these experiments a sinusoid masker was presented with the same frequency as the sinusoid signal and a threshold of where a brief sinusoid was detected was determined. The masker signal frequency, masker intensity and signal delay were varied parametrically.

3.3.1. Masker Level and Signal Delay. To analyze the forward masking as a function of masker level and signal delay the data in one of the experiments conducted by [13] were plotted as a function of log signal delay with masker level as a parameter and

masker level with signal delay as a parameter. The data from the plots were to fitted to be straight lines. The data can be described by the following equation,

$$M = a\left(b - \log \Delta t\right)\left(L_m - c\right).$$

Where M is the amount of masking, Δt is the signal delay, L_m is the masker level, and *a*, *b* and *c* are constants. The slope of masking at a given signal delay is given by, $a\left(b - \log \Delta t\right)$. The three parameters, *a*, *b* and *c* allow the estimation of the amount of masking that will be produced by any combination of masker level and signal delay or estimate the masker level required for constant amount of masking at a given signal delay. For low level maskers ($L_m < c$) and long signal delays ($\Delta t > 10^b$) and for greater masking levels the above equation predicts too little masking. But, it summarizes data at a particular frequency for a range of signal delays and masker levels and is used as a tool for data reduction.

3.3.2. Frequency. Experiments were also done to determine whether forward masking varies as a function of frequency. From the analysis in [13] it was found that forward masking is greater at low frequencies regardless of how the masker levels are compared that is sound level versus sound pressure level or amount of masking versus masked thresholds.

3.4. FREQUENCY DOMAIN PSYCHOACOUSTICS

The above topics covered the temporal or time domain aspects of the human auditory system. The requirement for equalization is spectral flatness. It is mentioned in [11] spectral peaks are more audible than notches. The audibility of peaks depends on the audio stimulus. Since white noise was found to be to most sensitive stimulus, the values obtained during detection for white noise are used for spectral flatness criterion. Peak level versus the Q factors for different frequencies was observed in [11]. At high values of Q factor the sensitivity to peaks is decreased. It was also observed that wide bandwidth notches are also audible though lesser than the peaks. Thus notches at certain bandwidths are also audible. Thus, the approaches of trying to invert the room impulse response or flatten the spectral response do not take into account the auditory aspects of

the human ear. In further sections, algorithms considering the psychoacoustic aspects discussed above will be explained.

3.5. LISTENING ROOM COMPENSATION

A filter for listening room compensation (LRC) is placed in the path of the signal in front of the loudspeaker. The goal is to reduce the influence of the succeeding room impulse response so that the signal obtained $y[n]$ at the position of the reference microphone is hardly distinguishable from the original signal $s[n]$ by the human listener. The basic setup is depicted in the Figure 3.1. The block $c[n]$ is the finite length room impulse response and $h[n]$ denotes the finite length equalizer. The finite length equalizers are usually designed by minimizing the squared error between the concatenation of $c[n]$, $h[n]$ and the given target system. Usually the target system is a band pass filtered version of the delayed impulse.

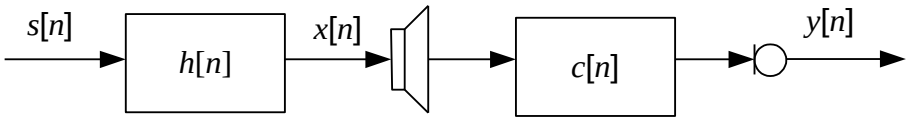

Figure 3.1. Single Channel Setup for Listening Room Compensation

Least Squares Method. In least squares equalization for LRC shown in Figure 3.2, a finite length $h[n]$ precedes the room impulse response $c[n]$. The equalizer is designed to minimize the square error between the concatenation $h[n]*c[n]$ and a target system $g[n]$ delayed by n_0 taps. The filter $g[n]$ is chosen as a band pass filter. The error signal $e[n]$ can be expressed as, $e[n]=\mathbf{s}^T[n]\mathbf{Ch}-\mathbf{s}^T[n]\mathbf{g}_{n0}$.

where,

$$\mathbf{s}[n]=\left[s\left[n\right],...,s\left[n-L_h-L_c+2\right]\right]^T$$

$$\mathbf{g}_{n0}=[\underbrace{0,...,0}_{L_g},g[0],...,g[L_g-1],\quad\underbrace{0,...,0}_{L_h+L_c-1-L_g-n_0}\quad]^T$$

L_h and L_c are the lengths of the equalizer and room impulse response. L_g represents the length of the target which is usually a band-pass filter. **C** is the convolution matrix of $c[n]$ whose dimension is, $L_h + L_c - 1 \times L_h$. The equalizer that minimizes the error signal's power $\mathrm{E}\{e^2[n]\}$ for a white noise input $s[n]$ is,

$$\mathbf{h} = \left(\mathbf{C}^H \mathbf{C}\right)^{-1} \mathbf{C}^H \mathbf{g}_{n0} \tag{18}$$

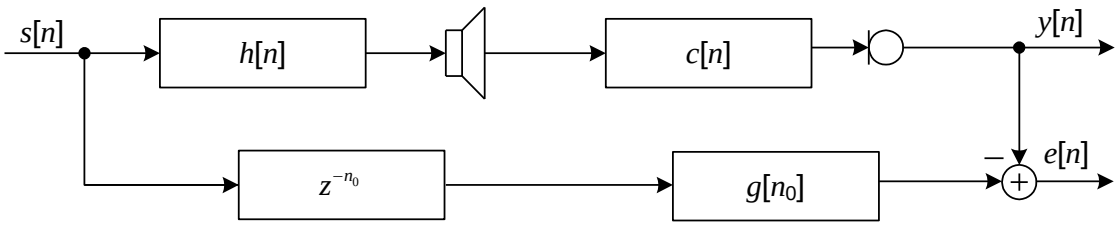

Figure 3.2. Setup for Listening Room Compensation using Least Squares

Instead of choosing a band-pass weighted function as the target system a more relaxed requirement is in psychoacoustics. One of them is the D50 measure for intelligibility of speech which is defined as the ratio of the energy within 50ms after the first peak of a room impulse response and the complete impulse response energy [15]. By choosing a target system with an optimized impulse response of 50ms, the D50 measure can be directly maximized. This idea is used in impulse response shortening.

3.6. CONCEPT OF IMPULSE RESPONSE RESHAPING/SHORTENING

A desired concatenated impulse response of the equalizer and the impulse response can be expressed by,

$$\mathbf{d}_d = \mathrm{diag}\{\mathbf{w}_d\}\mathbf{Ch} \tag{19}$$

in vector form. $\mathbf{w}_d$ is a vector that contains ones the desired region and zeros outside. **C** is the convolution matrix and **h** the equalizer as explained in the previous section. Accordingly,

$$\mathbf{d}_u = \mathrm{diag}\{\mathbf{w}_u\}\mathbf{Ch} \tag{20}$$

with

$$\mathbf{w}_u = \mathbf{1}_{[L_c+L_h-1]} - \mathbf{w}_d \tag{21}$$

represents the undesired part of the concatenated response. The energy of the unwanted part is kept constant while the energy of $\mathbf{d}_d$ is maximized. $\mathbf{1}_{[L_c+L_h-1]}$ is a vector of all ones of length as indicated. We can construct symmetric and positive semi-definite matrices **A** and **B** from (47) and (48) as given below.

$$\mathbf{d}_d^H\mathbf{d}_d = \mathbf{h}^H\mathbf{C}^H\mathrm{diag}\{\mathbf{w}_d\}^2\mathbf{Ch} = \mathbf{h}^H\mathbf{Ah} \tag{22}$$

$$\mathbf{d}_d^H\mathbf{d}_d = \mathbf{h}^H\mathbf{C}^H\mathrm{diag}\{\mathbf{w}_d\}^2\mathbf{Ch} = \mathbf{h}^H\mathbf{Bh} \tag{23}$$

Taking into account the loudspeakers limited playback capabilities at very low and very high frequencies, the maximization procedure is constrained to a broad band-pass area. Thus the bandpass $g[n]$ as described in the previous section is applied to the room impulse response. This can be written as,

$$c_{BP}[n] = c[n] * g[n] \tag{24}$$

Consequently, a convolution matrix $\mathbf{C}_{\mathrm{BP}}$ on the basis of c_{BP} can be assembled,

$$\mathbf{B}_{BP} = \mathbf{C}_{BP}^H\mathrm{diag}\{\mathbf{w}_{BP,d}\}^H\,\mathrm{diag}\{\mathbf{w}_{BP,d}\}\mathbf{C}_{BP} \tag{25}$$

The optimum equalizer $\mathbf{h}_{\mathrm{opt}}$ for maximizing the energy in a certain region is the solution of a generalized eigen value problem,

$$\mathbf{B}_{BP}\mathbf{h}_{\mathrm{opt}} = \mathbf{A}\mathbf{h}_{\mathrm{opt}}\lambda_{\max} \tag{26}$$

$\lambda_{\max}$ is the maximum eigen value and $\mathbf{h}_{\mathrm{opt}}$ is the corresponding eigen vector.

While designing the procedure for impulse response shortening goal is to avoid audible late echoes. The general shape of the room impulse response also has to be preserved which decays exponentially with time. Thus the temporal envelop should be such that it decays more quickly than the original impulse response thus yielding a shorter reverberation time. This can done by modifying the maximization window $\mathbf{w}_d$. Thus an exponentially decaying window with a reverberation time shorter than the original one can be used. One such window is used in [15] and can be treated as a design

rule. It is given by,

$$w_d[n] = \begin{cases} 0 & \text{for } 0 \le n \le n_0\text{-}1 \\ 10^{q(n-n_0)} & \text{for } n_0 \le n \end{cases}$$

As an example, q is chosen to be -3×10^{-5} and plotted against the original maximization window below in Figure 3.3.

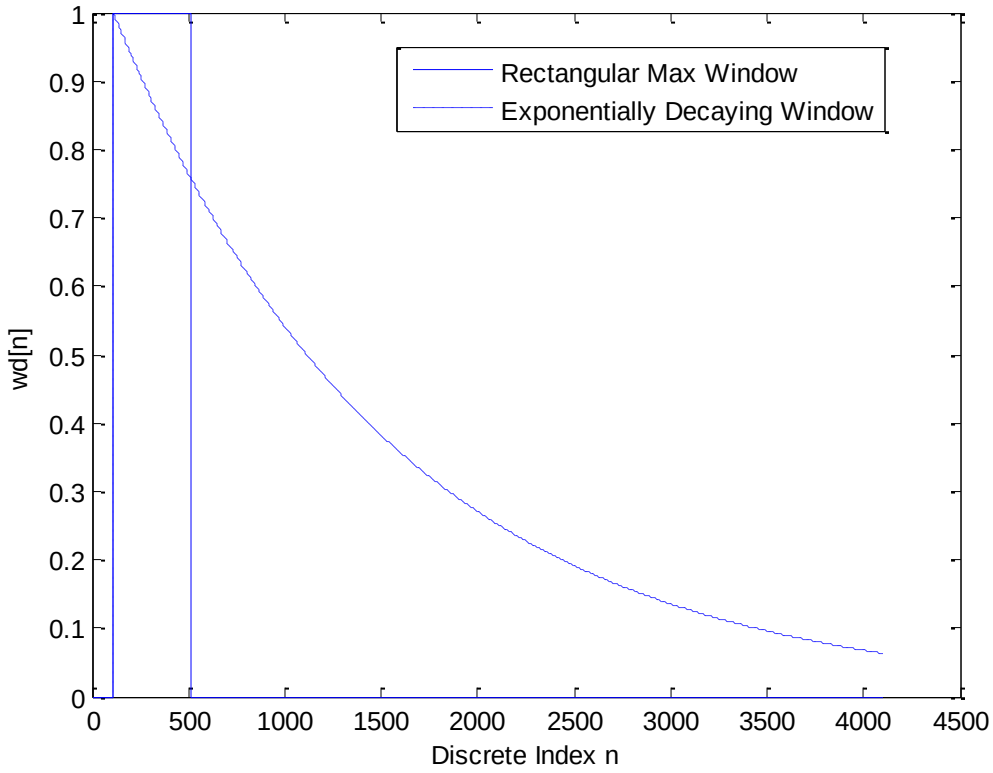

Figure 3.3. Maximization Windows as a Function of Time

If $c(n)$ is the impulse response of the room of length L_c and $h(n)$ the impulse response of the pre-filter of length L_h, then global impulse response of this pre-filter-loudspeaker-room is given by, $g(n) = h(n) * c(n) = \mathbf{Ch}$. **C** is the convolution matrix made up of c and is of size L_g-by-L_h as discussed earlier. The length of g is L_h+L_c-1. Main goal is to design a pre-filter $h(n)$ in such a way that the global response $g(n)$

attenuates faster than the impulse response of the room and also allow it to satisfy certain psychoacoustic conditions so that there is no audible echoes for a large class of signals.

For filter shortening and reshaping two windows $w_d(n)$ and $w_u(n)$ are used to derive a desired part $g_d(n) = w_d(n)g(n)$ and an unwanted part $g_u(n) = w_u(n)g(n)$ from the global impulse response $g(n)$. For shortening the windows $w_d(n)$ and $w_u(n)$ show no overlap whereas there may be significant overlap while doing reshaping. The purpose is to minimize some function of $g_u(n)$ while maximizing another function of $g_d(n)$ with respect to the pre-filter $h(n)$ without significantly affecting the magnitude frequency response of the global system. This means energy of $g_u(n)$ has to maximized while the energy of $g_d(n)$ is constant when not taking frequency responses into account for quadratic functions.

A conventional approach is to optimize $h(n)$ under the least squares criterion. That is,

$$\begin{cases} \text{MIN}_{\mathbf{h}} : f\left(\mathbf{h}\right) = \mathbf{g}_u^T \mathbf{g}_u \\ \text{S.T.}: \mathbf{g}_d^T \mathbf{g}_d = \text{constant} \end{cases}$$

This least squares problem is equivalent to the following eigen value decomposition

$$\mathbf{A}\mathbf{h}_{\text{opt}} = \lambda_{\min}\mathbf{B}\mathbf{h}_{\text{opt}}$$

with $$\mathbf{B} = \mathbf{C}^T \text{diag}\left[\mathbf{w}_d\right]^T \text{diag}\left[\mathbf{w}_d\right]\mathbf{C}$$

and $$\mathbf{A} = \mathbf{C}^T \text{diag}\left[\mathbf{w}_u\right]^T \text{diag}\left[\mathbf{w}_u\right]\mathbf{C}$$

The global impulse responses based on least squares is plotted in the time domain, log scale and in the frequency domain in Figures 3.4 to 3.6. The window based on the D50 measure (defined in Section 3.5) is used to design the filter in the figures plotted. Thus the window $w_d(n)$ is a rectangular window and its position is optimized to get an optimally shortened global impulse response $g(n)$. It can be seen that the pre-filter $\mathbf{h}_{\text{opt}}$ that is optimal in the least squares sense causes distortions in the frequency domain and late diffuse echoes in $g(n)$. Measures have been taken by using an exponentially decaying window as explained in the previous section. But, further improvements are needed in practice.

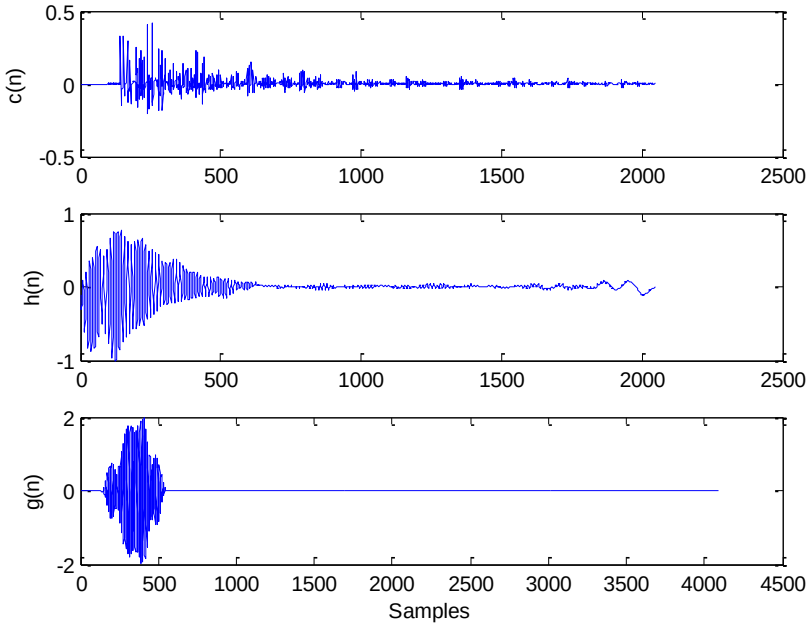

Figure 3.4. Original Response, Shortening Filter and Global Impulse Response (top to bottom)

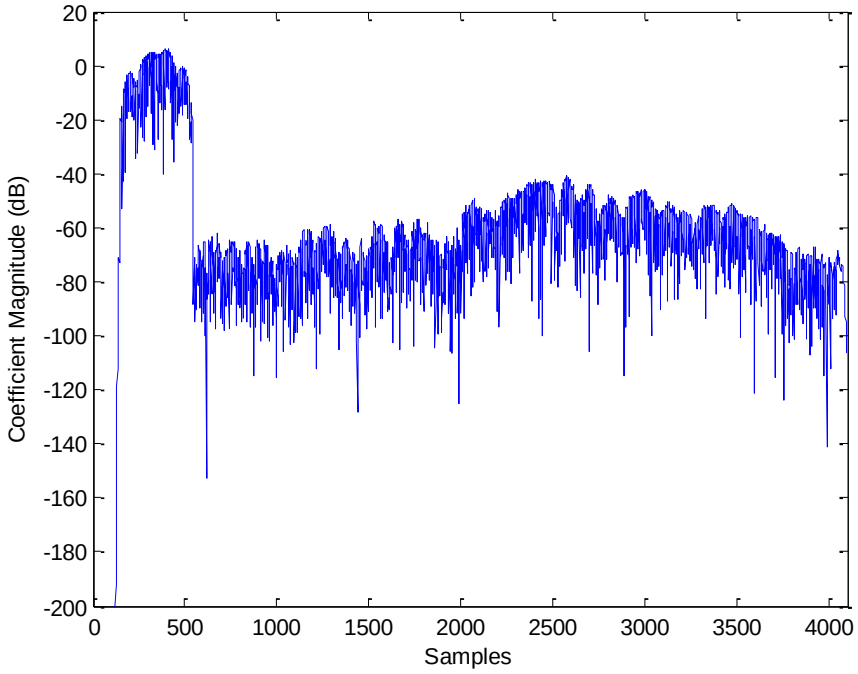

Figure 3.5. Decay of Global Impulse Response $g(n)$

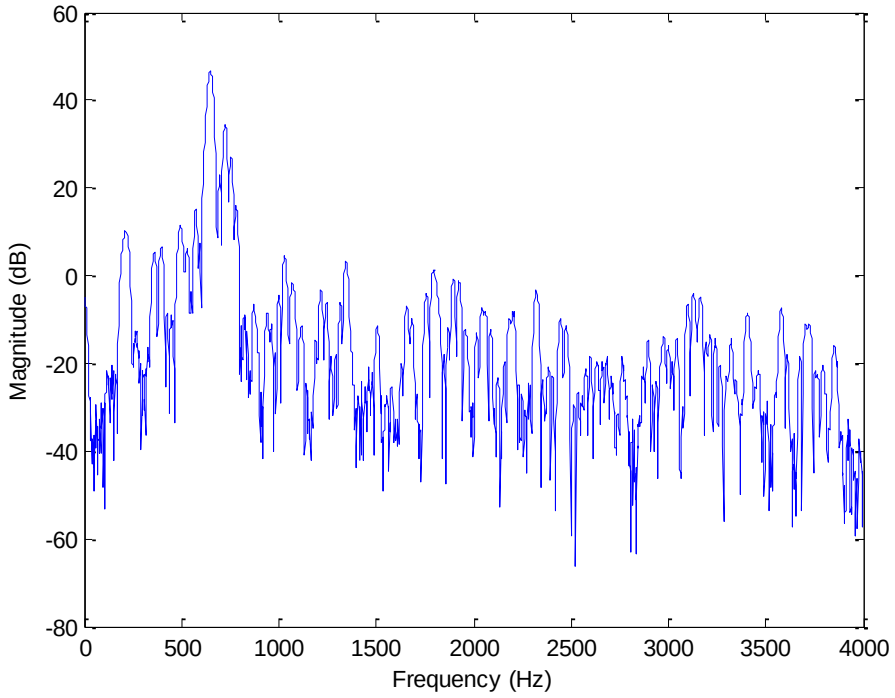

Figure 3.6. Magnitude Frequency Response of Shortened Global System Response

As an alternative to least-squares the infinity- and p- norm criteria, often used in robust estimation and control system design can more effectively influence the error behavior. Thus, in the next few sections, design of pre-filters based on the combining the infinity- and p- norm criteria and properties of the human auditory system in order control the perceived quality of sound will be explained.

For an optimal pre-filter the global impulse response $g(n)$ should have a quick and monotonically decaying characteristic so there are no noticeable echoes. This means that the attenuation characteristics of $g(n)$ has to be controlled. Properly selecting the windows $w_d(n)$ and $w_u(n)$ helps in achieving this requirement, but it is also important to use an optimization criteria that is suited to this requirement. For the optimization of pre-filters, the norm of the unwanted part $g_u(n)$ has to be minimized while keeping the norm of the desired part $g_d(u)$ as large as possible. The norm used is either the infinity-norm or the p-norm. With properly designed windows, it is possible to force the shortened or reshaped global impulse response to an approximately desired decaying

behavior.

3.7. INFINITY-NORM OPTIMIZATION

Since logarithm is a monotonic function, the optimization problem is as follows [16],

$$\text{MIN}_{\mathbf{h}}: f(\mathbf{h}) = \log\left(\frac{f_u(\mathbf{h})}{f_d(\mathbf{h})}\right) = \log\left(\frac{\text{Max}\left[\left|\mathbf{g}_u\right|\right]}{\text{Max}\left[\left|\mathbf{g}_d\right|\right]}\right) \tag{27}$$

Where, $f_d(\mathbf{h}) = \|\mathbf{g}_d\|_\infty = \text{Max}\left[\left|\mathbf{g}_d\right|\right]$ with $\mathbf{g}_d = \text{diag}\left[\mathbf{w}_d\right]\mathbf{Ch}$ is the infinity norm of the desired part, and $f_u(\mathbf{h}) = \|\mathbf{g}_u\|_\infty = \text{Max}\left[\left|\mathbf{g}_u\right|\right]$ with $\mathbf{g}_u = \text{diag}\left[\mathbf{w}_u\right]\mathbf{Ch}$ is the infinity norm of the unwanted part.

Minimizing $f(\mathbf{h})$ results in the minimization of $f_u(\mathbf{h})$ and at the same time results in the maximization of $f_d(\mathbf{h})$. Maximization of $f_d(\mathbf{h})$ leads to flat frequency domain characteristics of the global impulse response because, when one tap of $\mathbf{g}_d(n)$ is maximized, the other samples of $\mathbf{g}_d(n)$ becomes smaller and $\mathbf{g}_d(n)$ is dominated by a single tap. Minimization of $f_u(\mathbf{h})$ results in uniform distribution of errors across the time course of $g(n)$ because, all samples of the unwanted part $\mathbf{g}_u(n)$ would converge to almost the same value.

Learning rule based on taking the gradient of the optimality criterion is given by,

$$\mathbf{h}^{l+1} = \mathbf{h}^l - \mu\left(\frac{1}{f_u(\mathbf{h}^l)}\nabla_{\mathbf{h}} f_u(\mathbf{h}^l) - \frac{1}{f_d(\mathbf{h}^l)}\nabla_{\mathbf{h}} f_d(\mathbf{h}^l)\right) \tag{28}$$

where μ is a small step size.

If $\mathbf{g}_d(n)$ and $\mathbf{g}_u(n)$ have distinct maxima at positions I_d and I_u, respectively, then with $f_d(\mathbf{h}) = \left|g_d(I_d)\right|$ and $f_u(\mathbf{h}) = \left|g_u(I_u)\right|$ for a given $h(n)$ the corresponding gradients of $f_d(\mathbf{h})$ and $f_u(\mathbf{h})$ are as follows,

$$\nabla_{\mathbf{h}} f_d(\mathbf{h}) = \text{sign}\left[g_d(I_d)\right] w_d(I_d)\mathbf{C}_{I_d}^T \tag{29}$$

$$\nabla_{\mathbf{h}} f_u(\mathbf{h}) = \text{sign}\left[g_u(I_u)\right] w_u(I_u)\mathbf{C}_{I_u}^T \tag{30}$$

Where $\mathbf{C}_{I_d}$ and $\mathbf{C}_{I_u}$ are the I_dth and I_uth rows of the matrix $\mathbf{C}$, respectively. Thus the learning rule is given as follows,

$$\mathbf{h}^{l+1} = \mathbf{h}^{l} - \mu\left(\frac{1}{\left|g_u\left(I_u^l\right)\right|}\operatorname{sign}\left[g_u\left(I_u\right)\right]w_u\left(I_u\right)\mathbf{C}_{I_u^l}^T - \frac{1}{\left|g_d\left(I_d^l\right)\right|}\operatorname{sign}\left[g_d\left(I_d\right)\right]w_d\left(I_d\right)\mathbf{C}_{I_d^l}^T\right) \tag{31}$$

One of the advantages of infinity norm based algorithm is that the envelope of the unwanted part of the global response $g(n)$ is exactly determined by the inverse of the window function w_u. Thus, the attenuation behavior of $g(n)$ can be easily and exactly controlled enabling the removal of audible reverberation and echoes by exploiting the auditory masking property during the pre-filter design process. The implementation of the algorithm is as follows,

Algorithm

Step 1: Set the iteration index $l = 0$. Select a learning rate μ. Initialize the pre-filter $\mathbf{h}^l = \left[0.01, 0, ..., 0\right]^T$.

Step 2: Compute $\mathbf{g}^l = \mathbf{C}\mathbf{h}^l$, $\mathbf{g}_u^l = \operatorname{diag}\left(\mathbf{w}_u\right)\mathbf{g}^l$, $\mathbf{g}_d^l = \operatorname{diag}\left(\mathbf{w}_d\right)\mathbf{g}^l$; determine the positions of the maxima of $\left|\mathbf{g}_u^l\right|$ and $\left|\mathbf{g}_d^l\right|$, i.e. $f_u\left(\mathbf{h}\right) = \max\left(\mathbf{g}_u^l\right) = \left|g_u^l\left(I_u^l\right)\right|$ and $f_d\left(\mathbf{h}\right) = \max\left(\mathbf{g}_d^l\right) = \left|g_d^l\left(I_d^l\right)\right|$;

Step 3: Compute the gradients of $f_u\left(\mathbf{h}^l\right) = \left|g_u^l\left(I_u^l\right)\right|$ and $f_d\left(\mathbf{h}^l\right) = \left|g_d^l\left(I_d^l\right)\right|$ with respect to $\mathbf{h}^l$: $\nabla_{\mathbf{h}^l} f_u\left(\mathbf{h}^l\right) = \operatorname{sign}\left[g_u\left(I_u^l\right)\right]w_u\left(I_u^l\right)\mathbf{C}_{I_u^l}^T$ and

$\nabla_{\mathbf{h}^l} f_d\left(\mathbf{h}^l\right) = \operatorname{sign}\left[g_d\left(I_d^l\right)\right]w_d\left(I_d^l\right)\mathbf{C}_{I_d^l}^T$

Step 4: Update h: $\mathbf{h}^{l+1} = \mathbf{h}^l - \mu\left(\frac{1}{f_u\left(\mathbf{h}^l\right)}\nabla_{\mathbf{h}^l} f_u\left(\mathbf{h}^l\right) - \frac{1}{f_d\left(\mathbf{h}^l\right)}\nabla_{\mathbf{h}^l} f_d\left(\mathbf{h}^l\right)\right)$.

Step 5: Set $l := l + 1$ and go to Step 2.

3.8. P-NORM OPTIMIZATION

The corresponding optimization problem is given by [16],

$$f\left(\mathbf{h}\right) = \log\left(\frac{f_u\left(\mathbf{h}\right)}{f_d\left(\mathbf{h}\right)}\right) \tag{32}$$

$$f_d(\mathbf{h}) = \|g_d\|_{p_d} = \left(\sum_{k=0}^{L_g} |g_d(k)|^{p_d} \right)^{\frac{1}{p_d}}$$

with

$$f_u(\mathbf{h}) = \|g_u\|_{p_u} = \left(\sum_{k=0}^{L_g} |g_u(k)|^{p_u} \right)^{\frac{1}{p_u}}$$

and

where p_u and p_d are integers. The learning rule used is as given below,

$$\mathbf{h}^{l+1} = \mathbf{h}^l - \mu \nabla_h f(\mathbf{h}^l) \tag{33}$$

The gradients $\nabla_{\mathbf{h}} f_d(\mathbf{h})$ and $\nabla_{\mathbf{h}} f_u(\mathbf{h})$ are first calculated as,

$$\nabla_{\mathbf{h}} f_d(\mathbf{h}) = \left(\sum_{k=0}^{L_g-1} |g_d(k)|^{p_d} \right)^{\frac{1}{p_d}-1} \mathbf{C}^T \mathbf{b}_d \tag{34}$$

and

$$\nabla_{\mathbf{h}} f_u(\mathbf{h}) = \left(\sum_{k=0}^{L_g-1} |g_u(k)|^{p_u} \right)^{\frac{1}{p_u}-1} \mathbf{C}^T \mathbf{b}_u \tag{35}$$

with

$$\mathbf{b}_d = \text{diag}[\text{sign}[\mathbf{g}_d]]\text{diag}[\mathbf{w}_d]|\mathbf{g}_d|^{p_d-1}$$

$$\mathbf{b}_u = \text{diag}[\text{sign}[\mathbf{g}_u]]\text{diag}[\mathbf{w}_u]|\mathbf{g}_u|^{p_u-1}$$

Taking the gradient of $f(\mathbf{h})$ gives the following equation,

$$\begin{aligned} \nabla_{\mathbf{h}} f(\mathbf{h}) &= \frac{1}{f_u(h)} \nabla_{\mathbf{h}} f_u(\mathbf{h}) - \frac{1}{f_d(h)} \nabla_{\mathbf{h}} f_d(\mathbf{h}) \\ &= \frac{1}{\phi_{f_u}} \mathbf{C}^T \mathbf{b}_u - \frac{1}{\phi_{f_d}} \mathbf{C}^T \mathbf{b}_d \end{aligned} \tag{36}$$

where $\phi_{f_u} = \sum_{k=0}^{L_g-1} |g_u(k)|^{p_u}$ and $\phi_{f_u} = \sum_{k=0}^{L_g-1} |g_u(k)|^{p_u}$

In equation (64) the computational burden is due to $\mathbf{C}^T\mathrm{b}_u$ and $\mathbf{C}^T\mathrm{b}_d$. This burden can be eased by determining the convolution matrix **C** in the frequency domain by taking Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT). This can be done as follows, let $C(.) = \text{FFT}[c(.),\mathrm{L}_0]$ and $B_d(.) = \text{FFT}[b_\mathrm{d}(.),\mathrm{L}_0]$, where $L_0 \geq L_g$ is the

FFT size. Then $\mathbf{a}_d = \text{IFFT}\left[C^*(\bullet)B_d(\bullet)\right]$ which the same as the following result as given below,

$$\mathbf{C}^T\mathbf{b}_d = \left[a_d(0), a_d(1), \ldots, a_d(L_h - 1)\right]^T \tag{37}$$

similarly $\mathbf{C}^T\mathbf{b}_u$ is computed. Thus, the algorithm based on p-norm optimization can be summarized as follow.

Algorithm

Step 1: Set $l = 0$. Select a learning rate and FFT block size $L_0 = 2^\alpha \geq L_g$; compute $C(k) = \text{FFT}[c(n), L_0]$; initialize the pre-filter $\mathbf{h}^l = [0.01, 0, \ldots, 0]^T$.

Step 2: Compute $H^l(k) = \text{FFT}[h^l(n), L_0]$;

$\mathbf{g}_L = \text{IFFT}[C(k), H^l(k)]$, $\mathbf{g}_0 = [g_L(0), g_L(1), \ldots, g_L(L_g\text{-}1)]^T$; $\mathbf{g}_u = \text{diag}[\mathbf{w}_u]\mathbf{g}_0$, $\mathbf{g}_d = \text{diag}[\mathbf{w}_d]\mathbf{g}_0$.

Step 3: $\phi_{f_u} = \sum_{k=0}^{L_g-1}\left|g_u(k)\right|^{p_u}$, $\phi_{f_d} = \sum_{k=0}^{L_g-1}\left|g_d(k)\right|^{p_d}$

Step 4: $\mathbf{b}_u = \text{diag}[\text{sign}[\mathbf{g}_u]]\text{diag}[\mathbf{w}_u]\left|\mathbf{g}_u\right|^{p_u-1}$, $\mathbf{b}_d = \text{diag}[\text{sign}[\mathbf{g}_d]]\text{diag}[\mathbf{w}_d]\left|\mathbf{g}_d\right|^{p_d-1}$

$B_u(k) = \text{FFT}[\mathbf{b}_u, L_0]$, $B_d(k) = \text{FFT}[\mathbf{b}_d, L_0]$;

Step 5: $\mathbf{a}_u = \text{IFFT}[C^*(k)B_u(k)]$, $\mathbf{a}_d = \text{IFFT}[C^*(k)B_d(k)]$;

$\mathbf{a}_{u_0} = \left[a_u(0), a_u(1), \ldots, a_u(L_h - 1)\right]^T$, $\mathbf{a}_{d_0} = \left[a_d(0), a_d(1), \ldots, a_d(L_h - 1)\right]^T$.

Step 6: Update **h**:
$$\mathbf{h}^{l+1} = \mathbf{h}^l - \mu\left(\frac{1}{\phi_{f_u}}\mathbf{a}_{u_0} - \frac{1}{\phi_{f_d}}\mathbf{a}_{d_0}\right)$$

Step 7: Go to Step 2.

For values of $p_u = p_d = 2$ the problem reduces to that of a least mean squares solution. In the implementations of this algorithm in this thesis, the values of p_u and p_d is equal to 10 and 20, respectively for shortening and 20 and 10, respectively for reshaping as suggested in [16].

3.9. WINDOW FUNCTIONS

The energy decay properties and the frequency response of the global impulse response depend on the selecting the window functions. So, design of the window functions plays an important role in the entire shortening/reshaping filter design process. The global impulse response should decay in such a way that there are no audible echoes, which means the reverberation should be masked by the direct sound through the forward masking effect of the human auditory system. Similarly, the frequency domain characteristics should not change in such a way that the perceived timbre changes.

The forward masking effects of the human auditory system in real acoustic environments depends on both the signal under consideration and room characteristics described by the room impulse response as explained in Sections 3.2 and 3.3. Based on the explanations in Sections 3.1 to 3.4 a rough criterion for the assessment of average audibility of echo can be arrived at and reiterated as follows, the loudness of sound components is determined by convolution with a temporal integration function; the backward masking limit is set to 15ms; then the forward masking acts like simultaneous masking for the first 4ms after the initial direct sound and then falls of as 35dB/decade. Though the masking threshold computed according the above rules does not hold for every signal at hand, it is a compromise of masking thresholds for every signal at hand. Based on the above rules, the reshaping window can be defined.

3.9.1. Reshaping Window. The two window functions are defined as follows,

$$\mathbf{w}_u = [\underbrace{0, 0, ..., 0,}_{N_1+N_2} \underbrace{\mathbf{w}_0^T}_{N_3}]^T \tag{38}$$

$$\mathbf{w}_d = [\underbrace{0, 0, ..., 0,}_{N_1} \underbrace{1, 1, ..., 1,}_{N_2} \underbrace{0, 0, ..., 0}_{N_3}]^T \tag{39}$$

where $N_1 = t_0 f_s$, $N_2 = 0.004 f_s$, $N_3 = L_g - N_1 - N_2$, f_s is the sampling frequency and t_0 is the time taken by the direct sound. The window $\mathbf{w}_0$ is defined as follows [14],

$$w_0(n) = 10^{\frac{3}{\log(N_0/(N_1+N_2))}\log\left(\frac{n}{N_1+N_2}\right)+0.5} \tag{40}$$

with $N_0 = (0.2 + t_0)f_s$ and the time index ranges from $N_1 + N_2 + 1$ to $L_g - 1$.

The function $w_0(n)$ has a property that its reciprocal falls of approximately with 35dB/decade and hence represents the compromise masking limit of the human auditory

system. This is shown in Figure 3.7. It can be seen that its decay is -10dB at 4ms and decays exponentially to -70dB at 200ms.

3.9.2. Shortening Window. This is based on the D50 measure for intelligibility of speech. Therefore the two windows are defined as follows,

$$\mathbf{w}_u = [\underbrace{0,0,....,0}_{N_1+N_2},\underbrace{\mathbf{w}_0^T}_{N_3}]^T \tag{41}$$

$$\mathbf{w}_d = [\underbrace{0,0,....,0}_{N_1},\underbrace{1,1,....,1}_{N_2},\underbrace{0,0,....,0}_{N_3}]^T \tag{42}$$

where N_1 is same as defined for equation (67). The other parameters are $N_2 = 0.05f_s$ and $N_3 = N_1 + N_2 + 1$. The window w_0 is defined as,

$$w_0(n) = 1 + \frac{a-1}{N_3 - 1}n \tag{43}$$

where n = 0,1,2,…,$N_3 - 1$ and usually $a > 1$ for quick and uniform attenuation.

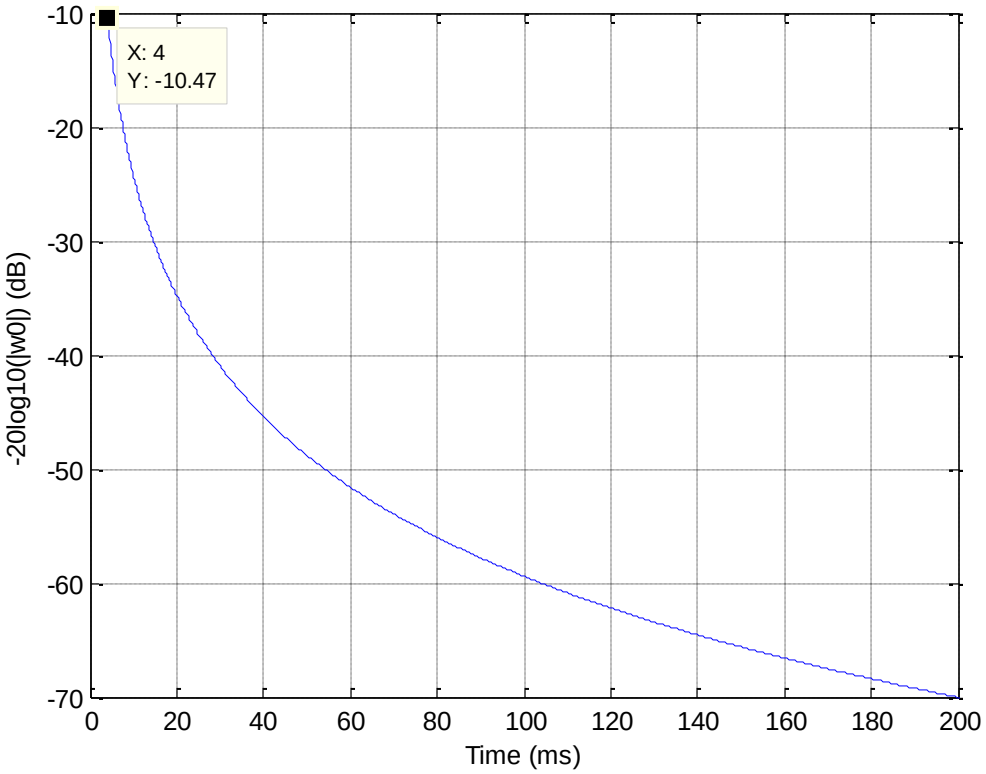

Figure 3.7. Logarithm Reciprocal of Window Function w_0(n) (Equation (40)).

3.10. SIMULATIONS

The room impulse response was simulated using the image derived model [4]. The length of the generated impulse responses was 2048 samples with a sampling frequency of 8000 Hz. This response was also designed to have frequencies greater than 50 Hz which are usually the frequencies generated by most loudspeakers which serve as mains on a multiple loudspeaker system. The dimensions of the room were chosen to be 36 ft by18 ft by 15 ft. The distance between the loudspeaker and microphone location in the simulated room response generation model was 18.72 ft.

For removing echoes, the pre-filters designed using the complex algorithms discussed above was inserted between the loudspeaker and microphone. The pre-filter performs either shortening or reshaping. Shortening satisfies the D50 measure and reshaping tries to make the global impulse response attenuate quickly to stay within the masking limit curve shown in Figure 3.7. Both the algorithms, p-norm and infinity- norm produce the same results with respect to reshaping and shortening. Hence, an extensive comparison is not made between these two algorithms. The infinity norm seems to converge slower than the p-norm algorithm. In the simulations shown below, the p-norm algorithm was used.

3.10.1. Reshaping. The parameters chosen to do the reshaping are as follows: p_u = 20, pd = 10, and learning rate $\mu = 10^{-6}$. The length of the pre-filter was 2048 and the algorithm was run for 2×10^6 iterations. The windows defined in equations (38) and (39) were used in the simulations for reshaping.

The room impulse response $c(n)$, the pre-filter $h(n)$ and the global impulse response $g(n)$ are shown in Figure 3.8. The Figure 3.9 shows the decay characteristics of $g(n)$ compared with the original room impulse response $c(n)$ and the masking limiting curve. The red line indicates the masking limiting curve which is the logarithmic reciprocal of $w_0(n)$. It can be seen that the envelope of the logarithm of the global impulse response $g(n)$ is completely controlled by the function $w_0(n)$. It can be seen that the reshaped version of $c(n)$ represented by $g(n)$ lies just under the masking limit.

3.10.2. Shortening. The parameters used in this case is as follows,

p_u = 10, p_d = 20 and learning rate $\mu = 10^{-6}$. The length of the pre-filter was 2048 again and the algorithm was run for 1×10^{6} iterations. The windows defined in equations (41) and (42) were used in this case. The simulations run for this method are represented in Figures 3.10 and 3.11. The Figure 3.10 again represents the plots of the different responses $c(n)$, $h(n)$ and $g(n)$ and this time $h(n)$ is the shortening filter. The Figure 3.11 shows the comparison of $c(n)$ and $g(n)$ in the logarithmic scale. It can be seen that $g(n)$ is the truncated version of $c(n)$.

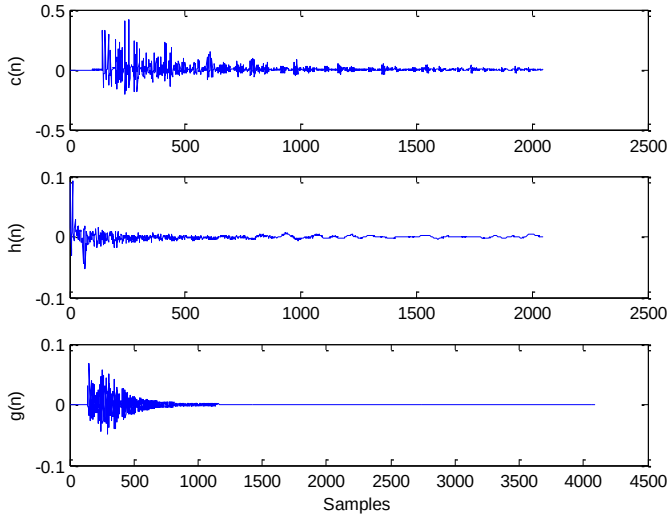

Figure 3.8. Original Filter, Reshaping Filter and Global Impulse Response (top to bottom)

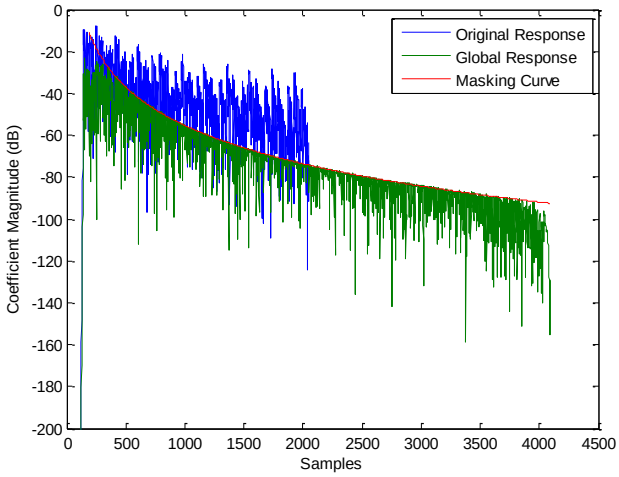

Figure 3.9. Decay of the Different Responses (reshaping)

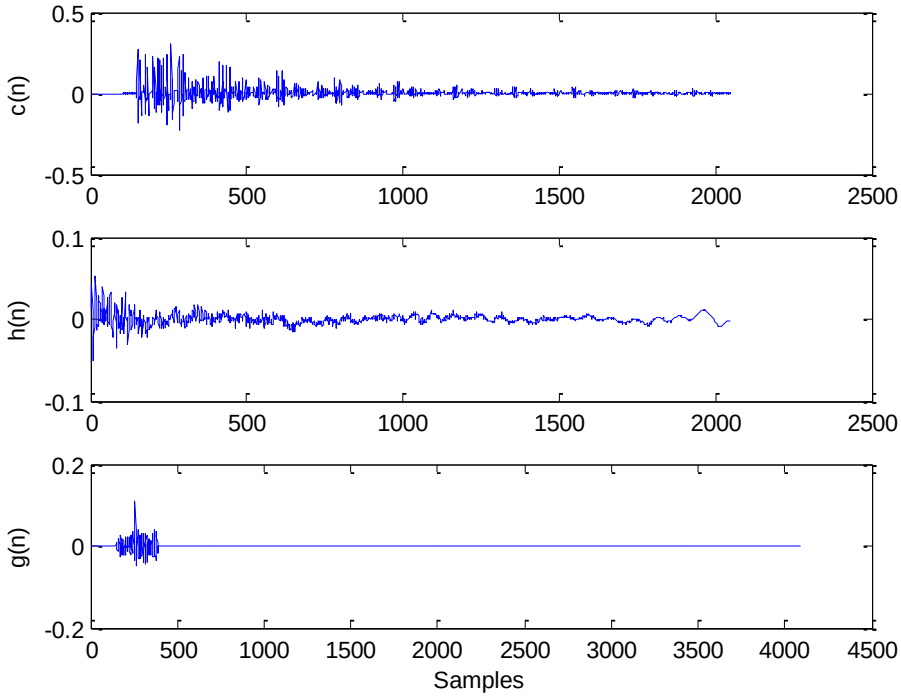

Figure 3.10. Original Filter, Shortening Filter, Global Impulse Response (top-bottom)

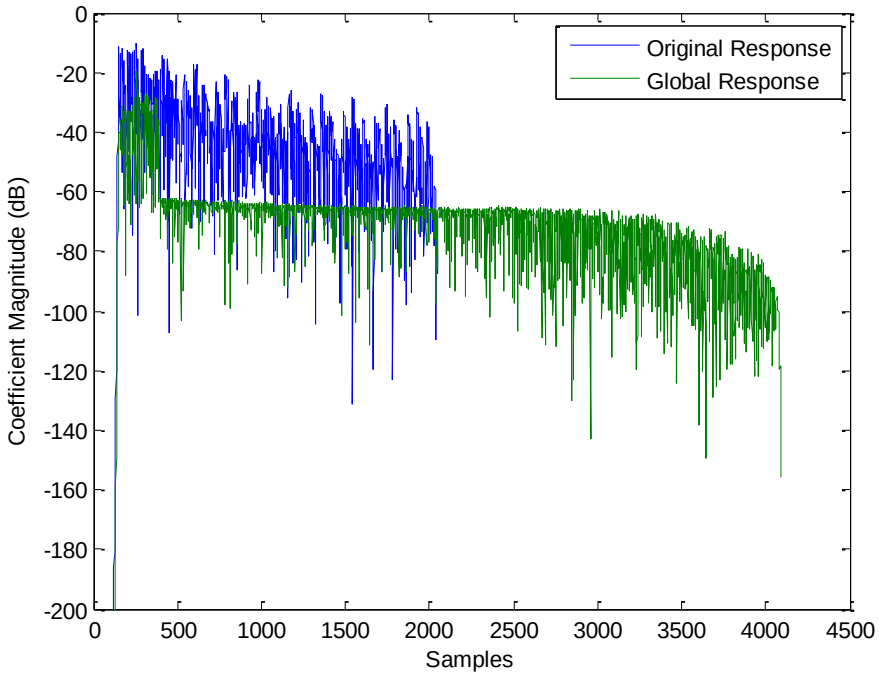

Figure 3.11. Decay of the Different Responses (shortening)

4. RESHAPING IMPULSE RESPONSES FOR MULTIPLE POSITIONS

4.1. PURPOSE

In practice, the room impulse response changes when the position of the source or receiver changes. The algorithms discussed in Section 3 did not consider the spatial robustness of the pre-filter. That is, verify whether the pre-filter performs in the same way with a room impulse response generated for a different location other than the location for which it was designed.

To validate the above discussion, a room impulse response was generated [4] for a particular loudspeaker and microphone location and reshaped using the methods discussed in Section 3. The reshaping filter obtained after performing this optimization was used to reshape another room impulse response generated for a different microphone location without changing the loudspeaker location. This was done to demonstrate that, when the microphone is moved to another location, the room impulse response changes and the filter designed for the original microphone location will not cause any reshaping to the new room impulse response. The results with the above setup were plotted and shown in Figures 4.1 to 4.4. Room Impulse Responses $c_{org}(n)$ and $c_{test}(n)$ were generated using the image derived model [4] at the original microphone location and a test microphone location, respectively, for a fixed loudspeaker location. The pre-filter $h(n)$ was designed to reshape the room impulse response $c_{org}(n)$ for location 1. The Figure 4.1 shows a plot of this room impulse response $c_{org}(n)$, the pre-filter $h(n)$ and the global response $g_{org}(n)$, obtained by reshaping $c_{org}(n)$ using the pre-filter $h(n)$. The same pre-filter $h(n)$ was used to reshape the room impulse response $c_{test}(n)$ to obtain the global response $g_{test}(n)$ which is shown in Figure 4.3. Comparing $g_{org}(n)$ and $g_{test}(n)$ it can be seen that the pre-filter $h(n)$ works well with the room impulse response at the original location but when the same pre-filter is applied to $c_{test}(n)$ there is not much reshaping occurring. To observe and compare the results better, Figures 4.2 and 4.4 depict the responses in the logarithmic scale along with the masking curves for the two cases discussed above. The Figure 4.3 shows that the global response $g_{org}(n)$ follows the masking curve and predominantly lies below the masking curve. The Figure 4.4 shows a large part of global response for the test location $g_{test}(n)$ lies above the masking curve.

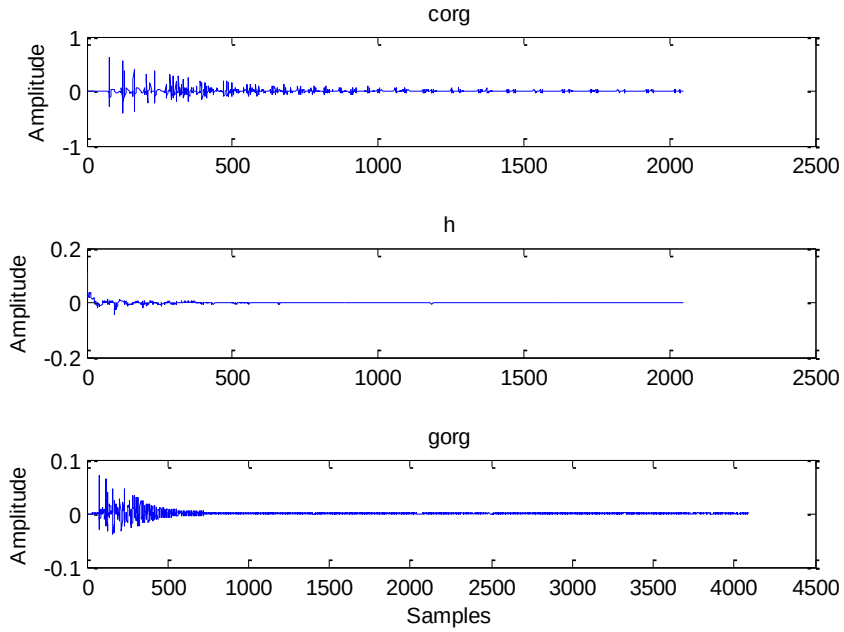

Figure 4.1. The Original Impulse Response, Reshaping Filter and Global Impulse Response for Location 1 (top to bottom).

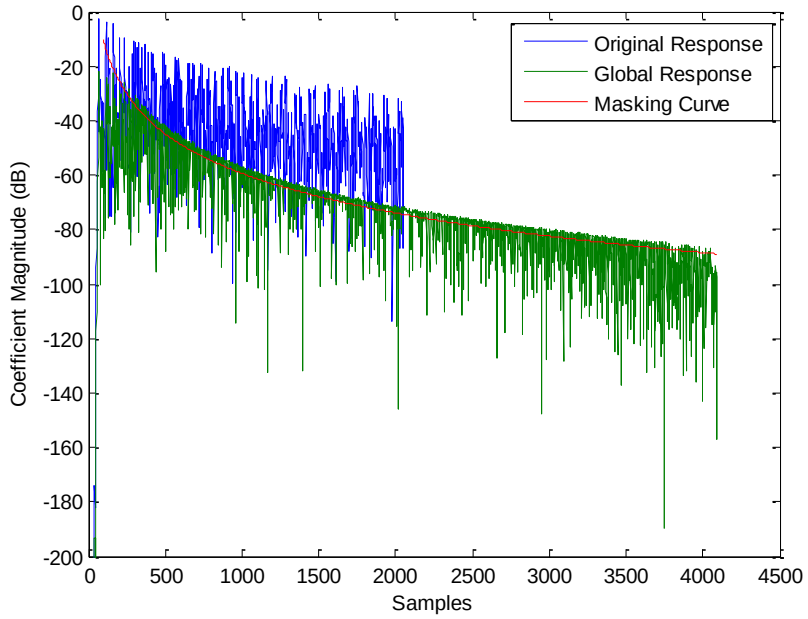

Figure 4.2. Comparison of the Responses in the Logarithmic Scale with the Masking Curve

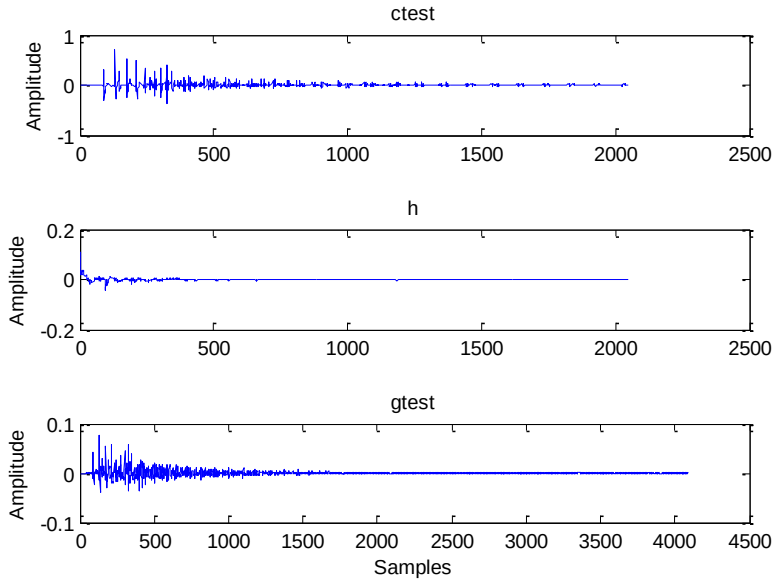

Figure 4.3. The Test Room Impulse Response, Pre-Filter and their Global Response

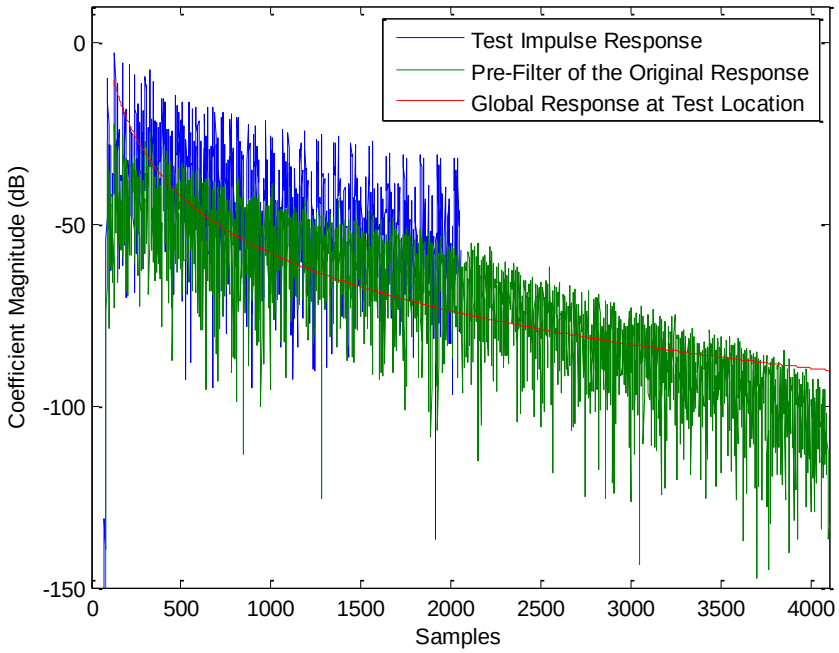

Figure 4.4. Illustration of the Global Response $g_{\text{test}}(n)$ lying above the Masking Curve

.

Thus, from the above the results the pre-filter designed for a given loudspeaker and microphone location is not spatially robust. In order for the pre-filter to be spatially robust, room impulse responses needs to be generated for different microphone locations and single pre-filter designed such that it would reshape the responses at their respective locations. The same pre-filter should also be able to shorten/reshape a room impulse response generated for any other location located within the vicinity of locations that were chosen to design the pre-filter.

4.2. METHOD I

In order to perform reshaping for multiple positions, a loudspeaker location is chosen and a set of random locations for microphone is chosen in front of the loudspeaker. The microphone locations are in close neighborhood with each other. The locations are chosen as a cluster of closely spaced points to perform reshaping over a given area in the room. Room impulse responses c_1(n), $c_2(n)$,…, $c_N(n)$ are generated for locations 1,2, …., N, respectively, where N total number of locations. Now the p-norm or infinity-norm optimization algorithms is used to perform reshaping for each of these room impulse responses c_1(n), $c_2(n)$,…, $c_N(n)$. Since the position of the direct path component changes for different room impulse responses, the desired w_d and undesired w_u windows used in the design of pre-filters are different for each of the room impulse responses $c_1(n)$, $c_2(n)$,…, $c_N(n)$. These windows are designed using equations (38), (39) and (40) with N_1 changing for each microphone position. After performing the optimization, each room impulse response $c_1(n)$, $c_2(n)$,…, $c_N(n)$ has its corresponding pre-filters h_1(n), $h_2(n)$,…, $h_N(n)$, respectively. To design a single pre-filter for each of these locations 1,2, …., N, it must have a representation of the shortened/reshaped results for all these locations 1,2, …., N. Averaging all these pre-filters is the best approach to satisfy this representation. Thus, the single averaged pre-filter can be written as,

$$\mathbf{h}_{avg} = \frac{1}{N}\left(\mathbf{h}_1 + \mathbf{h}_2 + ... + \mathbf{h}_N\right) \tag{44}$$

This pre-filter is applied in front of the loudspeaker to perform multiple position shortening/reshaping. The results for this method are discussed below.

Results. In this experiment the size of room used was 36feet by 18 feet by 15 feet. Five microphone locations were selected at (14.3,9,6) , (12.3,7.6,6) , (10.1,8.6,6), (10.8,7.9,6) and (12.3,9.4,6) denoted locations 1 to 5, respectively. These co-ordinates are specified in feet. The microphone and loudspeaker positions are shown in Figure 4.5. The image derived model [4] was used to generate room impulse responses $\{c_1(n), c_2(n), c_3(n), c_4(n), c_5(n)\}$ for all the fivr microphone locations. Each of these room impulse response $c_1(n)$, $c_2(n)$,….,$c_5(n)$ were reshaped using p-norm optimization. Again p-norm optimization was chosen for demonstrating the experimental results, since convergence in the case of p-norm optimization is faster compared to the infinity- norm optimization algorithms discussed in Section 3. For each room impulse responses five reshaping windows were used each of different sizes, since N_1, the length of the direct paths are different for each of these room impulse responses. The lengths of the direct paths were different for each room impulse response due to difference in distances of the microphone locations from the loudspeaker location. After reshaping, the pre-filters $h_1(n)$, $h_2(n)$,….,$h_5(n)$ are produced. There were averaged to produce $h_{avg}(n)$, using the method suggested in the previous section to realize a single pre-filter which reshapes the room impulse responses $\{c_1(n),c_2(n),c_3(n),c_4(n),c_5(n)\}$. After performing the reshaping, the global responses $g_1(n)$, $g_2(n)$,….,$g_5(n)$ are determined by convolving $h_{avg}(n)$ with $c_1(n)$, $c_2(n)$,….,$c_5(n)$, respectively. The Figures 4.6 to 4.10 show the plots of the logarithm of the room impulse responses $c_1(n)$, $c_2(n)$,….,$c_5(n)$ along with the logarithm of the global responses $g_1(n)$, $g_2(n)$,….,$g_5(n)$, respectively compared with the respective masking curves. From each of the Figures, it can be seen that a very small portion of the logarithm global impulse responses $g_1(n)$, $g_2(n)$,….,$g_5(n)$ lies above the masking curve. To test the robustness of the pre-filter h_{avg} a test location was chosen in the same neighborhood of the five locations chosen earlier. The image derived model [4] was then used to generate a room impulse response $c_{test}(n)$ at this reference location. Now the pre-filter $h_{avg}(n)$ is convolved with $c_{test}(n)$ to get the global response $g_{test}(n)$. Again the logarithm of $g_{test}(n)$, $c_{test}(n)$ along with the masking curve was plotted which is shown in Figure 4.11. It can be seen that the results are the same as in the case of the five locations that were chosen to do the reshaping by the average reshaping pre-filter h_{avg}. It can be seen that a very small portion of the global impulse response $g_{test}(n)$ lies above the masking curve.

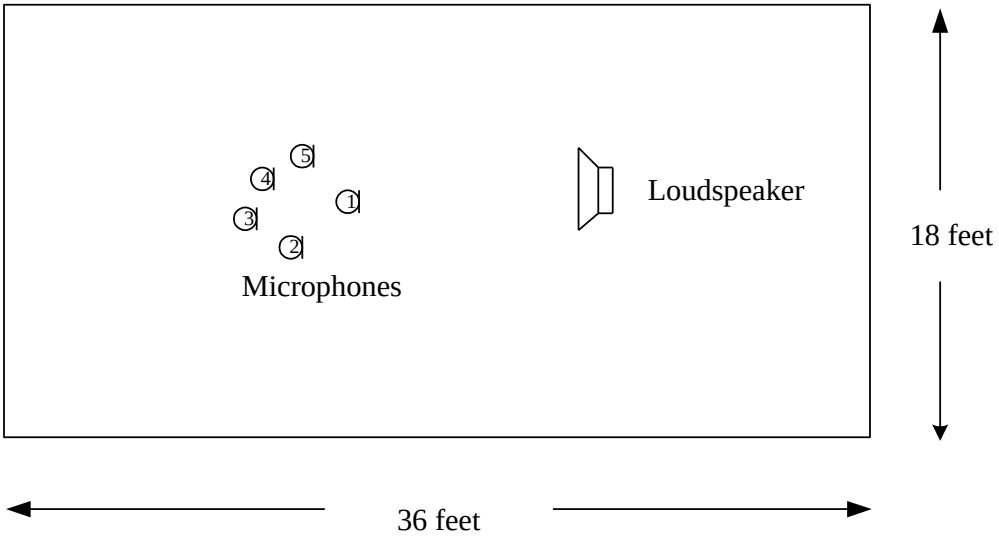

Figure 4.5. Experimental Setup

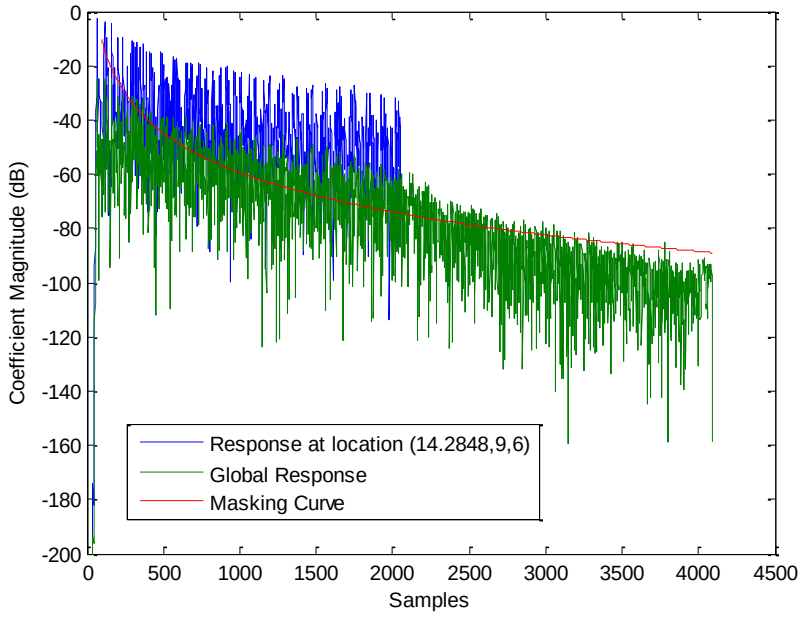

Figure 4.6. Logarithm Curves for Location 1

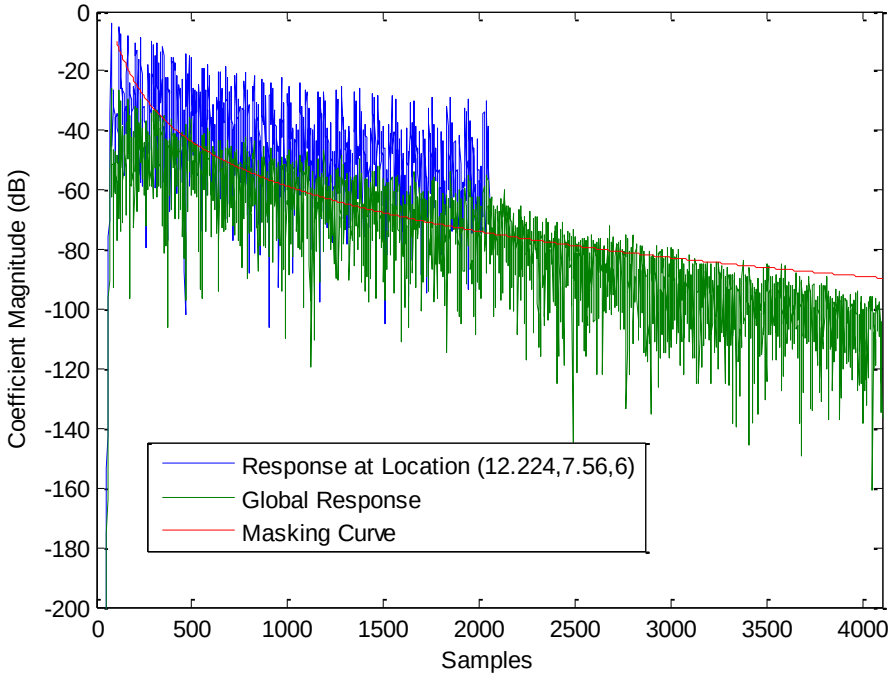

Figure 4.7. Logarithmic Curves for Location 2

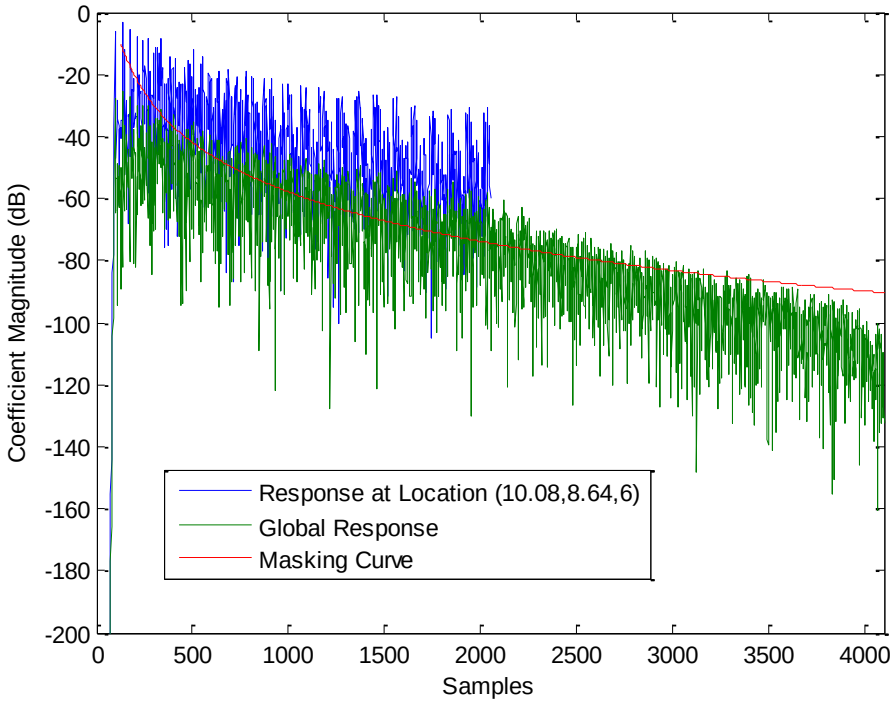

Figure 4.8. Logarithmic Curves for Location 3

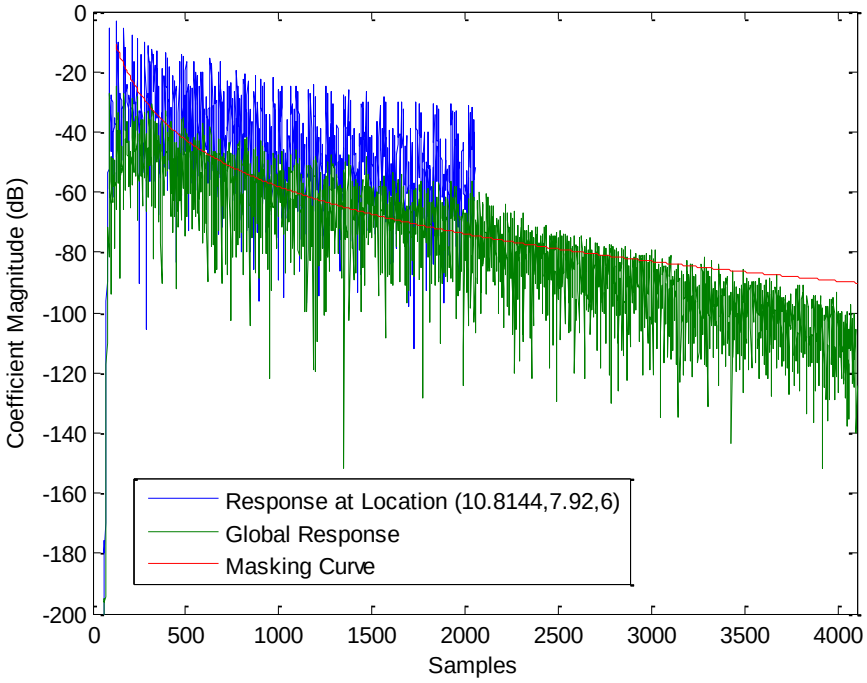

Figure 4.9. Logarithmic Curves for Location 4

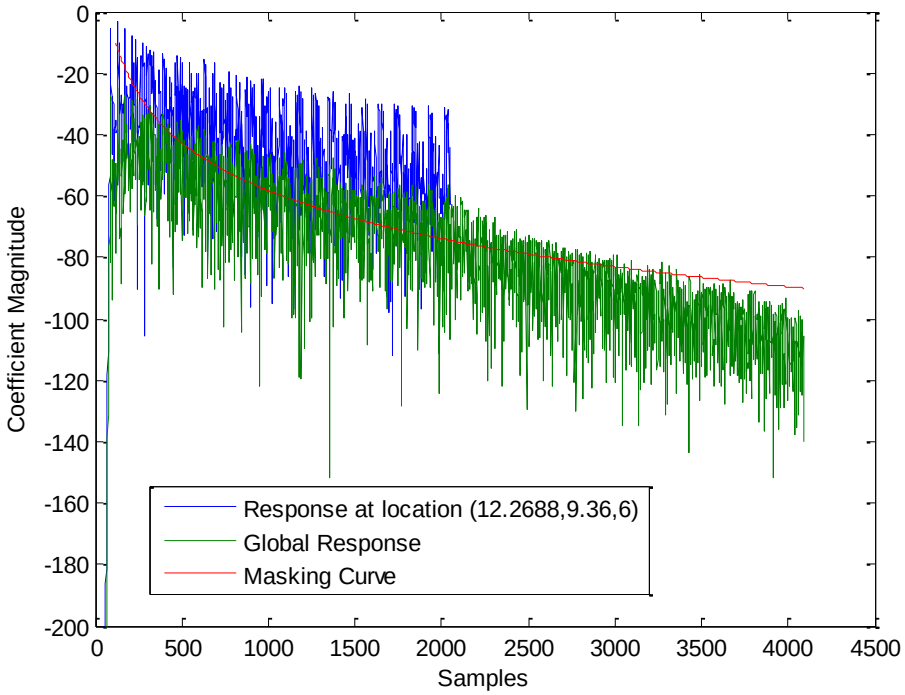

Figure 4.10. Logarithmic Curves for Location 5

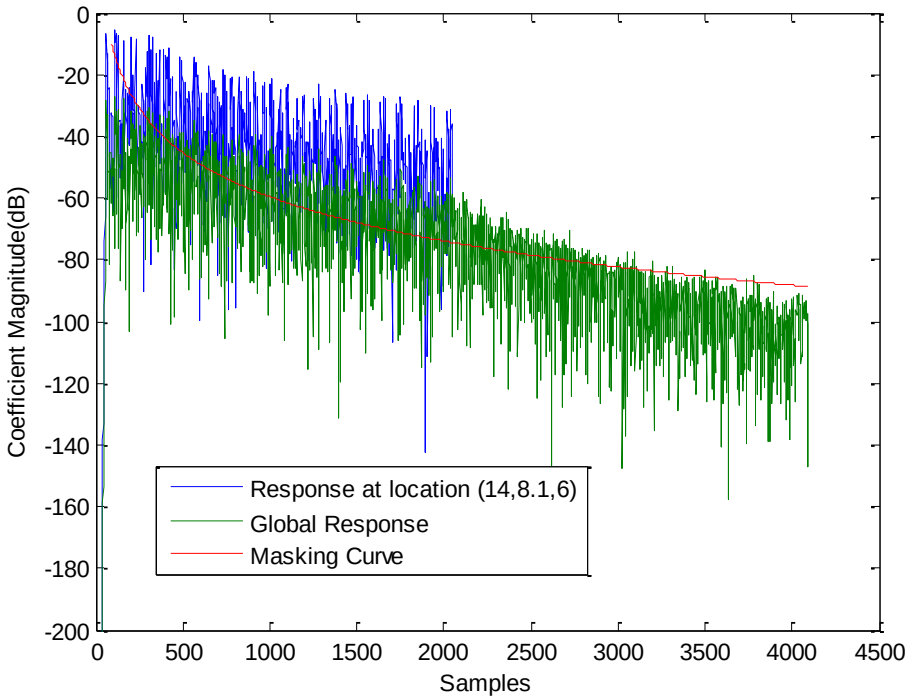

Figure 4.11. Logarithmic Curves for Test Location

The p-norm optimization algorithm was run for 1000000 iterations to design each of the pre-filters $h_1(n)$, $h_2(n)$,….,$h_5(n)$.

4.3. METHOD II

In another method to try and achieve reshaping for multiple positions, the single averaged pre-filter is calculated and used in each iteration of the optimization. The algorithm can be summarized as follows,

Step 1: The room impulse responses $\{c_1(n),c_2(n),\ldots,c_N(n)\}$ at different microphone positions 1,2,…,N, respectively, with a single loudspeaker are calculated.

Step 2: A single pre-filter $h_{avg}(n)$ is initialized to $\{0.01,0\ldots,0\}$ as was done while reshaping for one position.

Step 3: This single pre-filter is convolved with each of the room impulse responses $\{c_1(n),c_2(n),\ldots,c_N(n)\}$ to obtained the global responses $\{g_1(n),g_2(n),\ldots,g_N(n)\}$.

Step 4: The desired and undesired parts of each of the global responses $\{g_1(n),g_2(n),\ldots,g_N(n)\}$ are extracted using windows $\{w_{d1}(n),w_{d2}(n),\ldots,w_{dN}(n)\}$ and $\{w_{u1}(n),w_{u2}(n),\ldots,w_{uN}(n)\}$ each designed to perform reshaping for every microphone at positions 1,....,N, respectively.

Step 5: The desired and undesired global responses are then used to calculate pre-filters $\{h_1(n),h_2(n),\ldots,h_N(n)\}$ corresponding to each microphone location using p-norm or infinity norm optimization.

Step 6: The average of the pre-filters $h_{\text{avg}}(n)$ is calculated as,

$$\mathbf{h}_{avg}^{l} = \frac{1}{N}\left(\mathbf{h}_{1}^{l} + \mathbf{h}_{2}^{l} + \ldots + \mathbf{h}_{N}^{l}\right) \tag{45}$$

Where, l is the iteration number.

Step 7: Steps 3 to 5 are repeated until the optimization converges.

Results. For the sake of comparison, the same impulse responses $\{c_1(n), c_2(n),\ldots,c_5(n)\}$ used for Method I are used even for this method. The algorithm described above was applied to these impulse responses. The p-norm optimization algorithm was again used in this method in demonstrating the results. The algorithm was run for 2000000 iterations at which point the algorithm reached a satisfactory convergence point. The single pre-filter h_{avg} calculated from the above algorithm is convolved with each of the impulse responses $\{c_1(n), c_2(n),\ldots,c_5(n)\}$ to plot the global responses $\{g_1(n), g_2(n),\ldots,g_5(n)\}$ in the logarithm scale along with the masking curve and the impulse responses $\{c_1(n), c_2(n),\ldots,c_5(n)\}$. These plots are shown in Figures 4.12 to 4.16. It can be seen from the plots that the global responses follow the shape of the masking curve and only small portions are above the masking curve. Thus Method II works better than Method I for each of the locations chosen. To verify the results for a reference location, the same impulse response generated for the test location in the previous section was used even for this method. The global response at this reference location was calculated by convolving the pre-filter h_{avg} with this reference impulse response. This is plotted on the log scale along with the masking curve and logarithm of the test impulse response. Again a small portion lies above the masking curve compared to Figure 4.11 which corresponds to Method I.

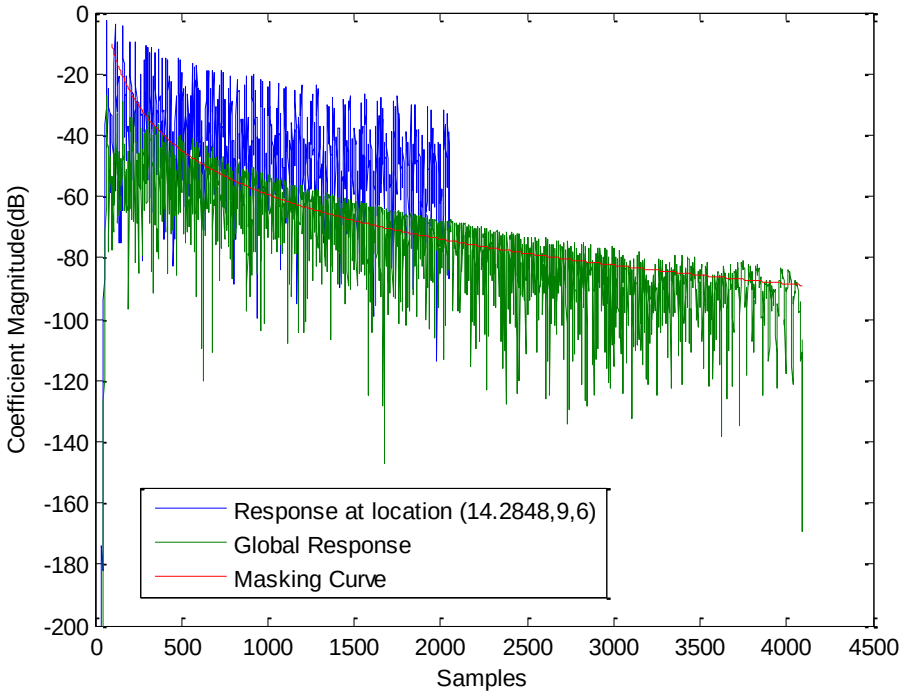

Figure 4.12. Logarithm Curves for Location 1

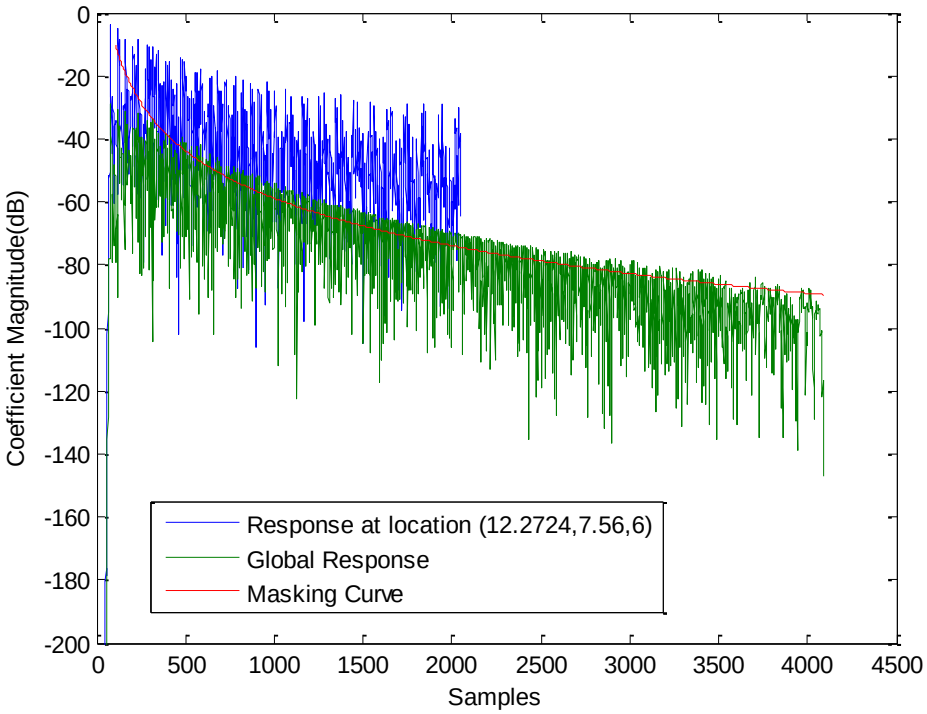

Figure 4.13. Logarithm Curves for Location 2

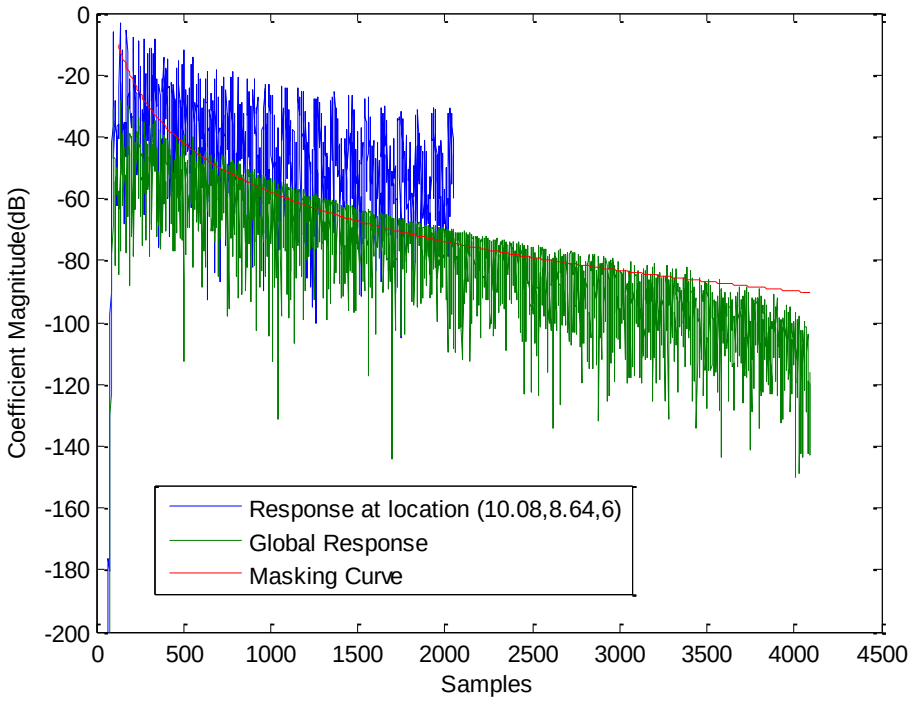

Figure 4.14. Logarithm Curves for Location 3

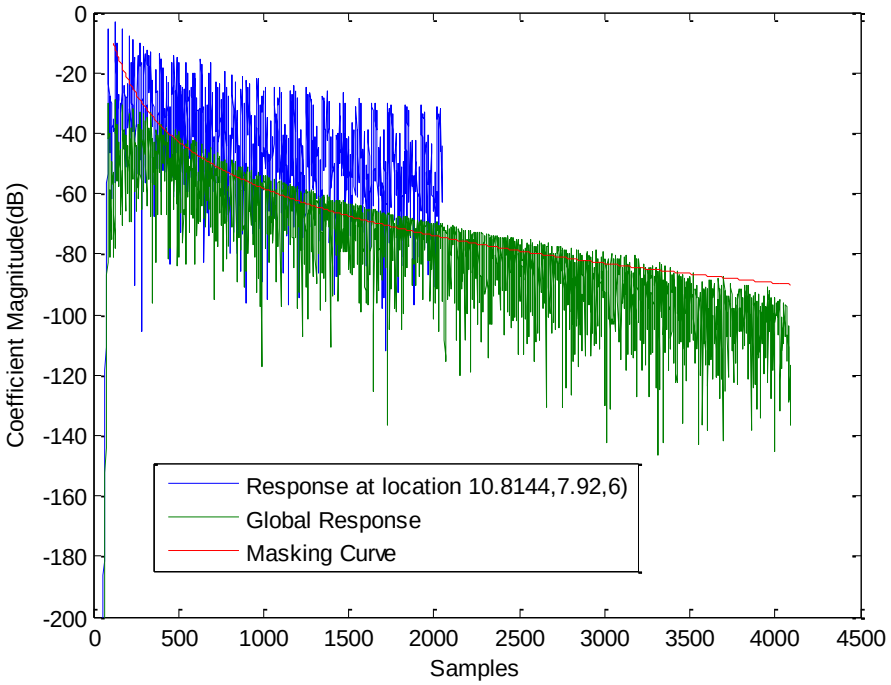

Figure 4.15. Logarithm Curves for Location 4

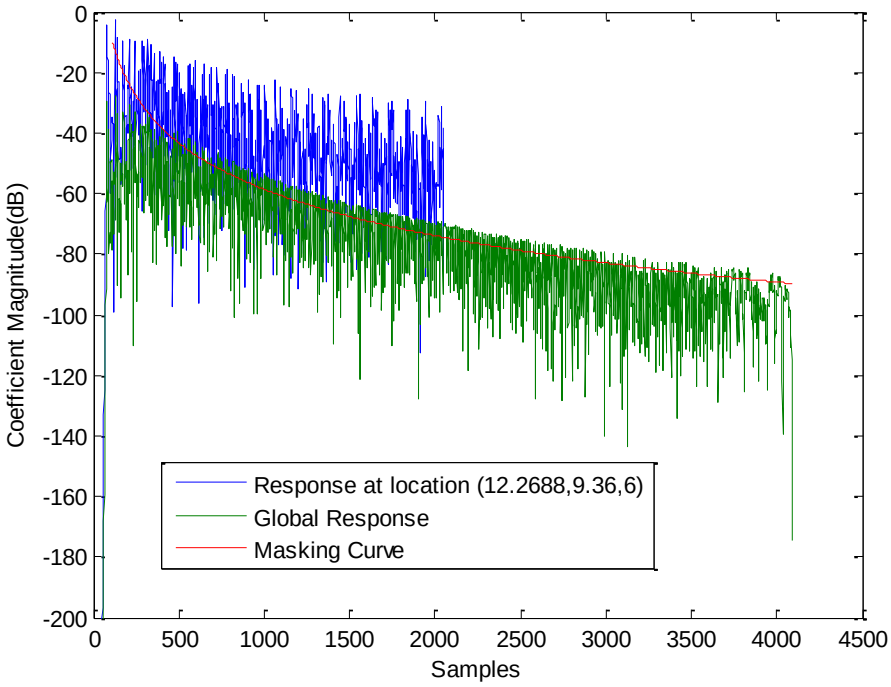

Figure 4.16. Logarithm Curves for Location 5

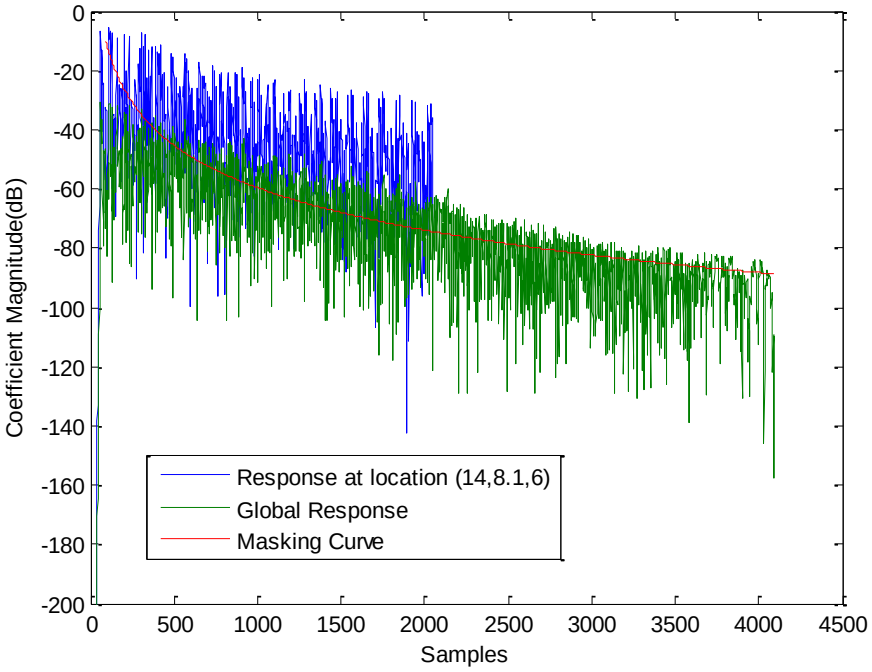

Figure 4.17. Logarithm Curves for Reference Location

4.4. COMPARISON OF METHOD I AND METHOD II

In order to compare the results of Method I and Method II, their global responses are plotted in the same figure along with the masking curve. This is shown in Figure 4.18 which is the plot for location 1. It can be seen that the global response of Method II is below the global response for Method I and much closer to the masking curve. Thus Method II is better than Method I for the trained location. The global responses of the two methods for the test location is also plotted which is shown in Figure 4.19. It can be seen than Method II is slightly better than Method I.

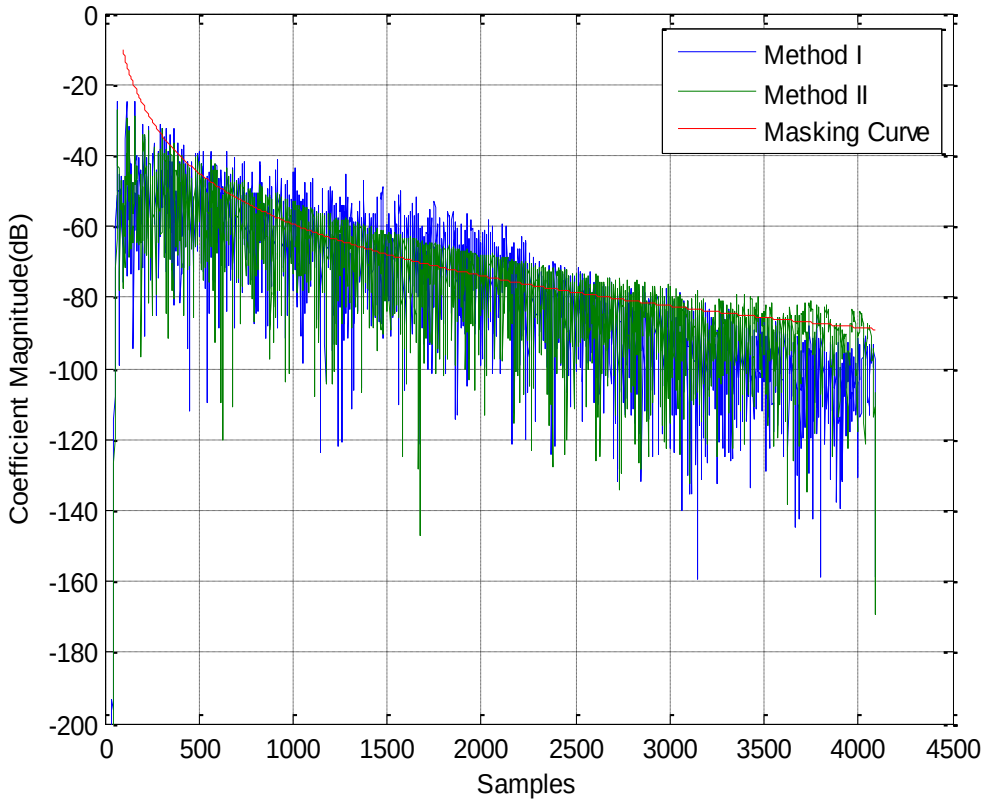

Figure 4.18. Comparison for Location 1

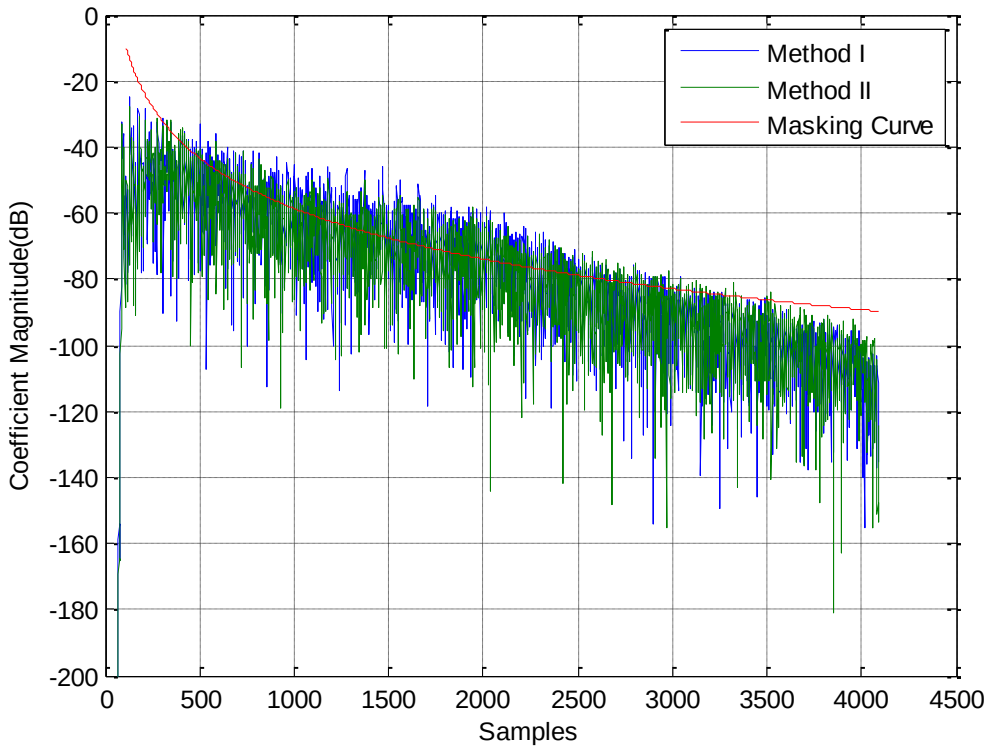

Figure 4.19. Comparison for Test Location

To validate the comparison better, a measure is needed which takes the difference in energy between the masking curve and the global response curve. In both the methods, the global response curves are above the masking curve due to the compensation done for multiple locations as opposed to a single location. Thus, the measure is taken by looking at the difference for those portions where the global response is above the masking curve. The energy difference measure is defined as,

$$EDM = \sum_{n} \left| 20\log_{10}\left(\frac{1}{w_u(n)}\right) - 20\log_{10}\left(g(n)\right) \right| \quad ;$$
$$\text{for } 20\log_{10}\left(g(n)\right) > 20\log_{10}\left(\frac{1}{w_u(n)}\right) \tag{46}$$

This measure gives the energy difference in decibels. It is computed for the five locations and test location selected earlier for both the methods. The values are shown in Table 4.1. The values denote the amount by which the global response is above the

masking curve.

Table 4.1. EDM Values of Method I and Method II

Location	Method I (dB)	Method II (dB)	Before Reshaping (dB)
1	1.4474	1.0573	6.9785
2	1.1421	0.6057	6.6431
3	1.2028	0.5372	6.1886
4	1.2362	0.5827	6.3497
5	1.1387	0.5975	6.5246
Test	1.3406	0.7537	6.1724

These values are normalized with respect to the length of the global response. The values in the table clearly indicate that the global response curve for Method II is closer to the masking curve indicated by the lower values of EDM in the column of Method II for all locations. Thus, Method II is a better method than Method I for performing reshaping at multiple positions. The table also shows the EDM values before doing the reshaping in the third column. From the higher values of EDM before reshaping, it can be concluded that there is a considerable improvement in the results after performing reshaping using the two methods.

5. CONCLUSION AND FUTURE WORK

In this thesis the standard methods of equalization for multiple positions were discussed but, could not meet the requirements of the human auditory system. The two methods discussed were Spatial Averaging and equalization based on MINT (Multiple input/output INverse Theorem). While Spatial Averaging was computationally simple, it neglected the temporal (time-domain) aspects of the Room Impulse Response. On the other hand, in case of MINT the processing was in the time-domain and produced exact inverses when the length and delay in the impulse responses were known. Also, computing inverses for long impulse responses is difficult. Thus, methods based on the requirements of human auditory system were discussed. The method used was impulse response shortening/reshaping with some optimization based on taking the infinity-norm or p-norm. Unlike MINT this method worked for filter lengths shorter than the room impulse response [14]. This was discussed for a single position and later extended for multiple positions using two methods which provided a compromised reshaping for all the training positions and test position. Unlike MINT which requires multiple filters corresponding to each channel these methods use a single preprocessed filter to achieve the reshaping at multiple positions. A measure was developed to compare the two methods and Method II was better than Method I.

The reshaping using the two methods was done for five room impulse responses and since it uses an iterative approach, it takes a long time to run. Thus in an audio application system, the system has to be trained for a long time before being used. Non-iterative methods of optimization have to be developed instead which takes much lesser time to run. The windows used in the optimization steps can be made adaptive such that it changes based on the changing error between the masking curve and the global response curve.

APPENDIX

REVERBERATION TIME OF ROOMS

If $I(t)$ is the sound intensity due to a loudspeaker transmitting with power $\Pi(t)$ at time t in a room of volume V and absorption of $a = \sum_i \alpha_i S_i$. Where, α_i and S_i are the absorption coefficient and surface area of wall I, respectively. The rate of change of total acoustic energy in the room can be expressed as,

$$\frac{d}{dt}\frac{4VI(t)}{c} = \Pi(t) - \alpha I(t) \tag{47}$$

where c is the speed of sound in the medium.

The solution to (72) can be written as,

$$I(t) = \frac{c}{4V} e^{-act/4V} \int_{-\infty}^{t} \Pi(r) e^{ac\tau/4V} d\tau$$

If the sound power $\Pi(t)$ fluctuates slowly relative to the time constant $4V/ac$, then the intensity will be approximately proportional to $\Pi(t)$ as

$$I(t) \approx \frac{\Pi(t)}{a}$$

When the sound power $\Pi(t)$ fluctuates faster than the time constant $4V/ac$, the intensity will not follow the fluctuations of $\Pi(t)$, and if the sound is shut off at time $t = 0$, the intensity can be expressed as,

$$I(t) = I_0 e^{-(act/4V)}$$

The reverberation time is defined as the time it takes for the intensity level to drop by 60dB after the source is switched off. Thus, if the dimensions of the room are measured in centimeters, the reverberation time is given by,

$$T_{60} = 60\frac{4V}{4.34ac}$$

The reverberation time computed through the above equation is based on geometrical room acoustics, where the wall is considered to be sufficiently irregular so that the sound energy distribution is uniform throughout the room.

The actual measurement of T_{60} can be done by the method of integrated impulse response proposed by Schroeder. This method uses an integration rule to determine an ensemble average of decay curves, $\left\langle g^2(t) \right\rangle$ from the square of the impulse response $h(t)^2$ using the following equation,

$$\left\langle g^2(t) \right\rangle = \int_t^\infty h^2(x)\,dx\,.$$

The result of the above equation is converted to dB scale to obtain the expression for computing T_{60},

$$T_{60} = 60\left(\frac{\Delta L}{\Delta t}\right)^{-1}$$

where $\Delta L/\Delta t$ is in dB/seconds.

BIBLIOGRAPHY

[1] S. Bharitkar and C. Kyriakakis, "Immersive Audio Signal Processing," Springer, 2006 pp. 54-56 and 98-104

[2] S. Bharitkar and C. Kyriakakis, "New Factors in Room Equalization Using a Fuzzy Logic Approach," Audio Engineering Society, 111th Convention, 2001

[3] S. Bharitkar and C. Kyriakakis, "A Cluster Centroid Method for Room Response Equalization at Multiple Positions," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 55-58, 2001

[4] http://www.2pi.us/rir.html

[5] S.T. Neely and J.B. Allen, "Invertibility of a Room Impulse Response," Journal of Acoustic Society of America, Volume 66, Issue 1, pp. 165-169, July 1979

[6] B.D. Radlovic and R.A. Kennedy, "Nonminimum-Phase Equalization and Its Subjective Importance in Room Acoustics," IEEE Transactions on Speech and Audio Processing, Volume 8, No. 6, pp. 728-737, November 2000

[7] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," IEEE Transactions on Acoustic, Speech and Signal Processing, Volume 36, No. 2, pp. 145-152, February 1988

[8] Y. Haneda, S. Makino and Y. Kaneda, "Muliple-Point Equalization of Room Transfer Functions by Using Common Acoustical Poles," IEEE Transactions on Speech and Audio Processing, Volume 5, No. 4, pp. 325-333, July 1997

[9] S.J. Elliott and P.A. Nelson, "Multiple-Point Equalization in a Room Using Adaptive Digital Filters," Journal of Audio Engineering Society, Volume 37, Issue 11, pp. 899-907, November 1989

[10] L.D. Fielder, "Analysis of Traditional and Reverberation-Reducing Methods of Room Equalization," Journal of Audio Engineering Society, Volume 51, Issue 1/2, pp. 3-261, February 2003

[11] L.D. Fielder, "Practical Limits for Room Equalization," Audio Engineering Society, 111th Convention, Paper Number 5481, November 2001

[12] J. M. Bucholtz, J. Mourjopoulos and J. Blauert, "Room Masking: Understanding and Masking of Room Reflections," Audio Engineering Society, 110th Convention, Paper Number 5312, May 2001

[13] W. Jesteadt, S.P. Bacon and J.R. Lehman, “Forward Masking Level as a function of frequency, masker level and signal delay,” Journal of Acoustic Society of America, Volume 71, Issue 4, pp. 950-962, April 1982

[14] A. Martins, T. Mei and M. Kallinger, “Room Impulse Response Shortening/Reshaping With Infinity- and *p*-Norm Optimization,” IEEE Transactions on Audio, Speech and Language Processing, Volume 18, No. 2, February 2010

[15] M. Kallinger, A. Mertins, “Room Impulse Response Shortening by Channel Shortening Concepts,” in Proceeding of Asilomar Conference Signals, Systems and Computers, 2005, pp. 898-902

[16] T. Mei, A. Mertins, M. Kallinger, “Room Impulse Response Shortening with Infinity-Norm Optimization,” IEE Conference on Acoustic, Speech and Signal Processing, pp. 3745-3748, April 2009

VITA

Raghavendra Ravikumar was born in Bangalore, Karnataka, India which is located in Southern India. He obtained his Bachelor's degree in Telecommunication Engineering on July 2008 from P.E.S. Institute of Technology, Bangalore which is affiliated to Visveswaraya Technological University (V.T.U.), Belgaum, Karnataka, India. He joined Missouri University of Science and Technology in fall of 2008 to pursue his Masters in Electrical Engineering specializing in Signal Processing and Communication. He worked under Dr. Steven L. Grant on a project funded by Leonard Wood Institute. The project requirement was to develop an Immersive Audio Environment (IAE) and an audio training scenario for soldiers. He will receive his Master of Science degree in Electrical Engineering in the month of December 2010.