SEPARATION OF DESIRED SPEECH FROM INTERFERING SPEECH REVERBERATING IN A ROOM

by

Hiroshi Sekiguchi

Bachelor of Engineering
The University of Tokyo
Tokyo, Japan
(1978)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS OF THE
DEGREES OF

MASTER OF SCIENCE
IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February, 1984

Hiroshi Sekiguchi

Signature of Author Signature redacted
Department of Electrical Engineering and Computer Science
Feb 7, 1984

Certified by Signature redacted
Thesis Supervisor

Accepted by Signature redacted
Chairman, Departmental Committee on Graduate Students

SEPARATION OF DESIRED SPEECH FROM INTERFERING SPEECH REVERBERATING IN A ROOM

by

Hiroshi Sekiguchi

Submitted to the Department of Electrical Engineering and Computer Science, February, 1984, in partial fulfillment of the requirements for the degree of Master of Science

ABSTRACT

A new algorithm has been developed to achieve the separation of desired speech from interfering speech reverberating in a room. The algorithm is developed by viewing this problem as a two-data channel adaptive noise cancellation problem.

Using a least squares error criteria in the adaptive noise canceler approach leads to classical system identification problem. The solution estimates the room reverberant system and restores the desired speech by filtering the interfering speech with the estimated filter. The system identification problem is a well-known research field; however, this thesis project is characterized by two major features. One is the fact that it deals with a system involving a large number (more than 1000) of parameters to be estimated. The other is that it handles non-stationary speech data signals, rather than just stationary noise signal.

Of the many available techniques for system identification, a spectral analysis estimation method is best suited to this particular problem by virtue of its efficient computation, using the FFT, and the flexibility for applying it to any type of room.

The proposed new algorithm is a spectral analysis estimation method, a modified technique for dealing with non-stationary speech signal. The Maximum Likelihood estimation technique directly derives this new algorithm by assuming a Gaussian colored noise for the desired speech, and by allowing the colored noise power spectral density to change frame by frame. The filtering process is realized by the overlap-save method.

The algorithm uses finite length frames and is implemented in a recursive fashion, providing the capability of real-time processing and adaptability to possible changes of the room acoustic environment. In each frame, the filter estimate is updated on a frequency-by-frequency basis.

The algorithm was implemented in a computer software program. Experiments using synthetic speech data were performed with two different types of room transfer functions, a 32 point delayed delta function, and a 1024 point room response experimentally measured in a room. The algorithm achieved above 20 DB signal-to-noise ratio in the output speech, when original speech signals with three different signal-to-noise ratio (6 DB, 0 DB, -12 DB) were input into the algorithm. For actually recorded speech data, however, the result is too obscure to present in this thesis. More work needs to be done for this case.

Thesis Supervisor: Bruce R. Musicus

Title: Assistant Professor of Electrical Engineering & Computer Science

To Tomoko

Acknowledgements

Many people have made this thesis possible and the following is dedicated to my thanks for these people.

First I am most indebted to Professor Bruce R. Musicus for his continuous encouragement and support of my work on this thesis. His great insights and keen intuitive guidance during the work helped me stay on the right track. Also, his patience and generosity made my first research experience enjoyable.

I sincerely thank Professor Alan V. Oppenheim for his warm support and precious advice in getting along with the difficult but rewarding rigors of research. His constructive suggestions provided me with the correct cues for the solution to the personal troubles I encountered.

I would like to thank Professor Louis Braida for providing the computer facilities on which most of the simulation programs in this thesis work were made.

I owe a variety of support and help in learning technical matters and using computers to the people in two groups: the Digital Signal processing Group and the Communications Biophysics Group. Especially, Evangelos Milios helped me get accustomed to the computer environment. Michael Wengrovitz provided some speech examples for the experiments. Dr. Greg Duckworth kindly gave me advice on the experiments performed in the project.

Special thanks to professor Hiroya Fujisaki in the University of Tokyo, who introduced me to Professor Oppenheim in Japan for the first time. By virtue of his help I started my academic work smoothly.

I also express my gratitude to people in Toshiba Corporation in Japan, for the financial support and encouragement during the entire time period of this work. Especially, I

sincerely thank Ichizo Takimoto for accommodating and encouraging my desire to pursue further study at MIT.

Personal thanks to my wife Tomoko, who always supported me mentally and managed to endure frequent attacks of anxiety. She also provided some speech examples for the experiment. Without her help, this work would have been impossible. Other personal gratitudes go to my parents Tadashi & Sumiko Sekiguchi for their continuing love, support, and encouragement.

Table of Contents

List of Figures

Chapter 1 Introduction

1.1. Motivation

The objective of speech enhancement is ultimately to improve the intelligibility of the desired speech or to separate the desired speech from interfering signals, such as noise, other speech, or echoes.

The need for separating a desired signal from an interfering signal is often encountered in the real world. One example is the problem of jet fighter cockpit noise cancellation.1 LSI technology has allowed the use of digital communication systems in a variety of equipment. Digital narrow-band vocoders are sometimes used when the application can tolerate the quality of the synthesized speech. However the presence of high levels of acoustic noise in the speech signal will degrade the quality of the resynthesized speech. This is the case with the fighter cockpit problem. A pilot voice transmitted over the communication line is seriously affected by the ambient noise created by engines or the vibration of the cockpit itself.

Another example is a broadcast news room problem. When a news anchor man is on the air in a broadcast studio, studio engineers running the broadcast often transmit to the news anchor man specific cues or messages to inform him of what is to be done next. If these messages are sent out over an audio monitor placed near the news anchor man, the microphone into which the newscaster speaks would also pick up the messages played out by the audio monitor. It would then be necessary to subtract out the monitor signal together with any acoustic reflections from the microphone signal.

This thesis will concentrate on seeking a practical and effective algorithm for canceling interfering speech in a real broadcast room environment.

1.2. Background

Generally, there are two approaches differing in the number of available observation signals for the problem solution. One approach is the one-data channel method. It utilizes only a single signal, which combines the desired speech and the interfering speech. The other approach is the two-data channel method, which primarily employs an adaptive noise canceling scheme.[2]

Several research papers have been written about the one-data channel method. Unfortunately they have only marginally succeeded in reducing the interfering speech. U.C. Shields[3] dealt with two added speech signals. He exploited the difference in the fundamental frequencies, or pitch periods between the interfering speech signal and desired speech signal to selectively eliminate the interfering speech from the microphone signal, using a comb filter. This method offers the acceptable noise reduction, unless there is unvoiced speech in it. Because his method is based on the fundamental frequency identification for two signals the method fails in unvoiced sections, which do not contain the fundamental frequency structure.

Other work has been done by Parsons.4 Parsons made use of a short time Fourier transform of windowed speech to select the harmonics of the desired voice. He identifies the peaks of the two speech signals and somehow assigns those peaks to each speaker. After leaving out undesired speech peaks, inverse short time Fourier transforming produces only the desired speech.

Steven Boll [5] performed spectral subtraction method. He first estimates the power spectrum of the background noise during intervals when the desired speech is not present. Then that estimate was subtracted from the spectrum of the noisy speech. It assumes the stationary signal for the noise. It fails when the noise is speech, because of that

stationary-assumption,

The two-data channel method has often achieved better performance. If additional information about only the interfering speech can be monitored at the second sensor, it is preferable to use the two-data channel method.

That second sensor is often called the reference signal. If the reference signal contains only a correlated version of the noise in the primary speech and moreover if the relation between these two can be approximated by a linear transfer function, then the two channel noise canceler scheme leads to an estimation of the transfer function. The estimation is performed by calculating the transfer function which minimizes the mean square error of the primary speech and the filtered version of the reference speech. After estimation, the desired signal is restored by subtracting the version of the reference signal filtered with that estimated filter, from the primary signal.

S. F. Boll [6] has done some work with the two channel adaptive noise cancellation scheme, using 1500 taps LMS method for the filter estimation. He attempts to put two microphones in the difference places in a room so that only one microphone collects the desired signal and the interfering signal, whereas the other microphone obtains the only interfering signal. He achieves -20 DB noise reduction in the output signal, when the interfering noise signal is so large that it masks the desired speech in the primary completely.

More related work with this thesis has been done by M. Paulik,[7] also using two channel data method in the broadcasting room problem. The desired signal degraded by the reverberant interfering signal would be put into one channel, on the other hand only the interfering signal would be fed to the other. Although his work was restricted to exploring the theoretical feasibility of the adaptive noise cancellation approach based on modeling

the room reverberant impulse response with a 21 tap filter with known coefficients, his results suggest that the approach may be quite practical. He examined three finite length filter techniques, such as the covariance method, LMS method (Least Mean Squares) and autocorrelation method. He utilized his own method to find the speech section where the desired speech is substantially small, and at the same time the reference speech is dominant in the primary. He then constructed good estimates by using that frame's data.

He achieved more than 40 DB SNR in the output after 5 sec when the algorithm starts with -7 DB SNR primary speech in the covariance method. Also it achieves 40 DB SNR in the output after about 5 sec when it starts with for the 10 DB SNR primary speech in the same method.

1.3. The scope of the thesis

The aim of this thesis is to expand the two-data channel adaptive noise canceler scheme to a general type of room transfer function in the broadcast room problem. It then leads to the issue of transfer function estimation technique. The most crucial features in the problem are a (i) long impulse response of the system (more than 0.1 sec) and (ii) the data is non-stationary speech. When the interesting frequency range is from 100 Hz to 5000 Hz, the sampling frequency is chosen as 10 KHz. The 0.1 sec long impulse response requires the estimation of more than 1000 sample points in a time domain. Our purpose is to establish and present a new algorithm, fast enough for real-time processing implementation, sufficiently robust for restoration of the desired speech with a broadcasting quality, and universal enough to apply to any kind of room environment. In chapter 2, we will define this problem precisely and clarify the importance and features of the problem.

The adaptive cancellation scheme leads to the well-known typical problem of the linear transfer function identification. The estimation process for such a long room

reverberation impulse response with more than 1000 points usually requires extraordinarily large amounts of the computation and higherly robust computation. Also the non-stationarity of the signal makes the problem much more difficult. Thus we will study possible techniques to deal with such a large number of estimates in chapter 3. The spectral analysis method, the pole-zero technique, and the finite length response technique (FIR) will be discussed as a candidate. Both advantages and disadvantages of each candidate will be compared and finally the most suitable method will be determined. The spectral analysis estimation method is selected because it can deal with a large numbers of estimates by using the FFT.

In chapter 4, the spectral analysis technique will be intensively discussed from the theoretical point of view. Although techniques are well-known for the stationary stochastic input signal case, the methods for the non-stationary input signal case are still poor. We will discuss two methods for non-stationary cases. The first one is the average of the transfer function estimate over frames (ATFE), the other is the method of averaging the weighted spectrum with the value of the reciprocal of the desired speech power spectrum estimate (AWSE). The first method will prove to be a poor estimation method, due to the high variance of the filter estimate. The second method will be found useful, this method uses a model with a Gaussian colored noise assumption for the desired speech, allowing the colored noise power spectral density to change frame by frame. This modeling is reasonable for the speech signal, even though it is not perfect.

In chapter 5, the performance of the second method will be examined by both synthetic speech data and actually recorded speech data in a room. The methodology of measuring a room impulse response will be given. In the synthetic speech data case, two kinds of room system functions are used. One is a mere 32 point delay delta function, the other is a 1024 point long impulse response actually measured in a room.

For the synthetic data cases, the proposed algorithms attain 20 DB signal-to-noise ratio in the output speech. It is not perfect, but acceptable. The subjective listening test shows that the interfering speech is considerably reduced at the end of the 14 sec performance.

For the actually recorded speech data case, however, the result is too obscure to clarify the performance of the algorithm. We suspect that the non-essential issues such as program bugs or the mistakes in the recording procedure caused these unexpected results.

References

1. W. A. Harrison , J. S. Lim, and E. Singer, "Adaptive Noise Cancellation in a Fighter Cockpit Environment," *will be published in IEEE Trans. on Acoust., Speech, Signal Processing* .

2. Bernard Widrow, "Adaptive Filters 1: Fundamentals," SU-SEL-66-126, Stanford Electronics Lab., Stanford University, Dec. 1966.

3. U. S. Shields Jr. , "The Subtraction of Added Speech by Digital Comb Filtering," *S.M. Thesis* , Massachusetts Institute of Technology, Dept. of Elec. Eng. and Comp. Science, 1970.

4. Thomas W. Parsons , "Separation of Speech from interfering speech by means of Harmonic Selection ," *Journal of the Acoustic Society of America*, vol. 60, no. 4, Oct. 1976.

5. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, April 1979.

6. S. F. Boll and D. C. Pulsipher, "Suppression of Acoustic Noise in Speech Using Two Microphone Adaptive Noise Cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, Dec. 1980.

7. Mark Paulik, , "The Separation of Acoustically Added Speech Signals by Adaptive Noise Canceling ," *M. S. Thesis*, Massachusetts Institute of Technology, Dept of Elec. Engi. and Comp. Science , 1983.

Chapter 2 The goals of the project

2.1. The problem statement

This thesis project is concerned with the separation of two acoustically added speech signals, particularly the restoration of one desired speech signal additively degraded by other speech in a broadcasting news studio. Fig 2.1 illustrates this problem. A news anchor man speaks into a microphone on the air in a studio, whose voice is denoted as $s(n)$, while an audio loud speaker placed some distance away from the microphone reproduces a studio engineer's voice, expressed as $x(n)$, in the studio room. Therefore, the latter voice is also collected in the microphone, forming an interfering signal to the desired newscaster signal. The electrical signal drives the audio loudspeaker (possibly with some distortion), then the acoustic waves propagate through the air to the microphone by a direct path as well as by indirect paths, reflecting many times off the studio walls, furniture and occupants. Due to the approximate linearity of acoustic propagation and reflection, this reverberant process can be represented as a linear system characterized by a system impulse response $h(n)$. The signal at microphone $y(n)$ is thus formed from the reverberated loudspeaker signal, $g(n) = x(n) * h(n)$, plus the newscaster's voice, $s(n)$;

$$y(n) = s(n)+g(n) = s(n)+x(n) * h(n) \tag{2.1.1}$$

The question is how to subtract the interfering version signal $x(n) * h(n)$ from the microphone signal $y(n)$ so as to restore and broadcast the newscaster's voice $s(n)$ by knowing only $x(n)$ and $y(n)$, but not knowing $h(n)$. (Note that $x(n)$ is available because the engineer's voice is monitored so that it may be sent out from the loud speaker.)

One effective solution is offered by the two-channel noise cancellation scheme advocated by Widrow.1 The scheme obtains two inputs, the reference signal $x(n)$, and the primary signal $y(n)$ which consists of the desired signal additively degraded from the

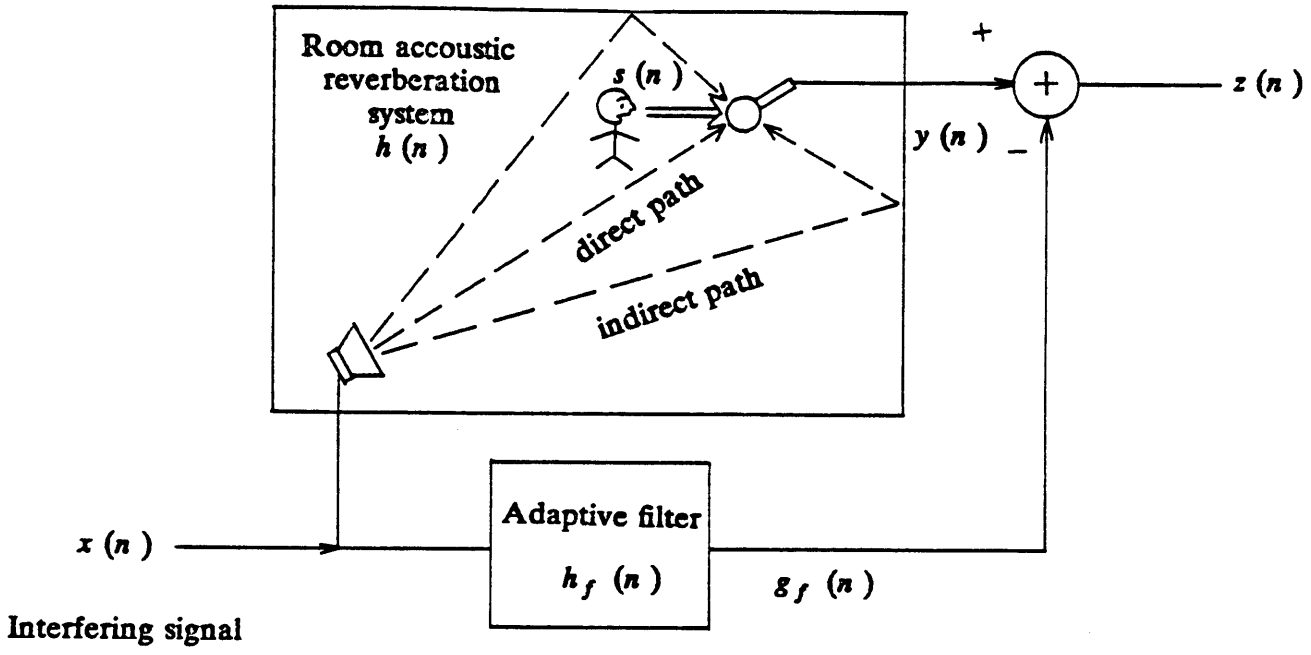


Fig 2.1 The broadcast room problem

reference signal. It also sets up an another filter $h_f(n)$ driven by the interfering signal $x(n)$, then estimates and adjusts the filter so that the output of the filter $g_f(n)$ may give a "good" approximation to $y(n)$. Moreover it subtracts the filter output $g_f(n)$ from the other input $y(n)$ to produce the canceler output $z(n)$. Since the output of the canceler is expressed in the form of

$$z(n) = y(n) - g_f(n) \tag{2.1.2}$$

where

$$y(n) = s(n) + x(n) * h(n) \tag{2.1.3}$$

$$g_f(n) = x(n) * h_f(n) \tag{2.1.4}$$

combining the above equations leads to

$$z(n) = s(n) + x(n) * (h(n) - h_f(n)) \tag{2.1.5}$$

Intuitively, if $h_f(n)$ can be adjusted to match $h(n)$ exactly, then the output of the canceler $z(n)$ would be the desired signal $s(n)$. This adjustment would be achieved in an elegant fashion by minimizing the least squares energy in $z(n)$ with respect to all possible filters $h_f(n)$'s. That is,

$$\hat{h}_f(n) \leftarrow\!\!- \min_{h_f(n)} \sum_n \left| y(n) - h_f(n) * x(n) \right|^2 \tag{2.1.6}$$

offers an optimal estimate of the room reverberant system $\hat{h}_f(n)$. By doing this, the adjusted filter $\hat{h}_f(n)$ would hopefully correspond to a good estimate for the room reverberant system $h(n)$. Consequently, the output of the canceler - the remainder after subtracting the filter output $x(n) * \hat{h}_f(n)$ from $y(n)$ would be a good estimate of desired signal $s(n)$.

Thus the noise cancellation approach leads to two major steps to follow for solving the acoustic separation problem. One is to estimate the room transfer function $h(n)$. The second is to filter the interfering speech $x(n)$ with the estimated $\hat{h}(n)$ and then to subtract the filter output $x(n) * \hat{h}(n)$ from the microphone signal $y(n)$.

The first issue, estimating the room transfer function $h(n)$ is in the field of identification of the transfer function. We will discuss several promising techniques for modeling which differ in how the transfer function is parameterized and how the estimated parameters are estimated. These methods include the spectral analysis technique, pole-zero modeling, and several finite impulse response (FIR) techniques.

The second computational issue is how to efficiently convolve the input $x(n)$ through the estimated $h(n)$ to cancel the loudspeaker interference. Several alternatives considered include discrete Fourier transform methods (overlap-add and overlap-save) or direct convolution of the filter and the input signal. Which method is most efficient usually depends

on which estimation technique is used for transfer function estimation. For instance, if the short-time spectral analysis technique is used for the estimation process, then the filter is estimated in the frequency domain, and it is natural to use FFT based filtering algorithms such as the overlap add or the overlap save methods. Polo-zero modeling as well as FIR modeling estimates the filter transfer function in the time domain, and therefore direct time domain convolution for implementing the filter will be more efficient.

The superiority of one technique over another in terms of performance will depend on the particular application. Critical aspects of the algorithm include its speed and accuracy. Behavior of the algorithm may be changed drastically by the characteristics of the signals which the problem deals with, and the characteristics of the transfer function such as the length of the impulse response.

In the broadcast room problem, important characteristics of the algorithm are

[1] Capability of canceling the interfering speech in real-time. The noise canceler should utilize a fast algorithm for canceling the interference so as to broadcast the restored newscaster's speech more or less immediately. It means the algorithm should be simple enough to process the data fast.

[2] Broadcast quality performance in canceling the interfering loudspeaker signal. This requires the algorithm to estimate $h(n)$ as accurately as possible. It should not only remove as much of the interfering speech as possible, but should also restore the newscaster's speech free of the artifacts which are often present in processed speech.

[3] Applicability to any type of room environment. No matter how different the size or shape the room would have, no matter what kind of the broadcasting equipment is used, the algorithm should work.

[4] Fast adaptation to a changing room response (transfer function) due to moderate changes in the microphone and the monitor speaker locations, as well as changes in the positions of people and other reflecting objects in the room. The algorithm should apply an adaptive scheme which watches possible changes of the $h(n)$ and adapts itself to these changes.

The type of signals this broadcast room problem deals with is speech. The ordinary estimation literature has often dealt with much easier cases such as white noise or color noise. A speech signal, a non-stationary signal, is a most complicated object to cope with. Consequently the algorithm will more likely require special techniques besides those developed for white noise, or color noise cases.

The characteristics of the actual transfer function may offer important and useful insights into the solution of the problem, because the performance of different estimation techniques may drastically vary depending on those characteristics of the actual transfer function. To choose appropriate techniques for this problem, it is essential to investigate their characteristics and to derive the insights into which technique is appropriate.

To examine a real room reverberant impulse response may provide some insight into what would be a suitable technique to solve the broadcast room problem. The measurement of the room impulse response shows a long time duration of the response. It may vary from 0.1 sec to 0.5 sec, dependent on a variety of room environments. If $h(n)$ is modeled as an FIR filter, then at a 10 KHz sampling rate, it would be necessary to estimate the 1000-5000 coefficients of $h(n)$. Such a large number of estimates requires an algorithm to perform the tremendous amount of computation. Although the techniques of transfer function estimation are a well-known subject, this requirement of the large number of estimates forces us to consider a practical technique for the problem. Another feature is that

because of the separation of the loudspeaker and the microphone, the function $h(n)$ is zero for the amount of time corresponding to the delay for the acoustic signal to propagate by the direct path from the loud speaker to the microphone. This may change from 0 sec to 0.1 sec in different rooms. Estimating the response $h(n)$ during this time would waste computation since it would only end up setting up to the first 1000 points of $h(n)$ to zero. A better idea might be to adaptively estimate the delay of the direct path, and set $h_f(n)$ to zero during this interval.

References

1. Bernard Widrow et al., "Adaptive Noise Canceling:Principles and Applications," *IEEE Proceeding,* , vol. 63, no. 12, Dec. 1975.

Chapter 3 Theoretical Background

3.1. Introduction

In this chapter some theoretical background related with the thesis project is developed. The adaptive noise cancellation scheme is described first. Then the discussion on the system identification problem is followed. Several estimation techniques to identify the system function are introduced, such as the spectral analysis method (the Fourier transform of the system function), pole-zero modeling, and finite impulse response (FIR) techniques. Advantages and disadvantages of each technique are discussed and the appropriate technique is selected for the broadcasting problem by examining the suitability of each technique in terms of the required specifications in the problem.

3.2. The adaptive noise canceling principles (The two-channel noise canceler)

We consider the following two-channel adaptive noise cancellation scheme illustrated in Fig 3.1. Assume that the signal source s and the interfering signal n are uncorrelated , zero mean, and statistically stationary. (Note that this assumption in not valid for speech.) Let n_1 be the interfering signal, and let n_0 be another signal which is correlated in some unknown way with the noise n_1. In the newscaster's problem, n_0 is the convolution of n_1 with a loudspeaker and a room reverberant impulse response h, which is unknown , and may be changing with time.

The signal s is transmitted over a channel to a sensor that also receives the interfering signal n_0. The combined signal $s+n_0$ forms the primary input to the canceler. A second sensor receives the interfering signal n_1. This sensor provides the reference signal to the canceler.

The interfering signal n_1 is filtered to produced an output g that is as close as possible to n_0 in some "good" sense. This output is then subtracted from the primary input $s+n_0$ to produce the system output $z = s+n_0-g$, which is hopefully the restored version of the signal s.

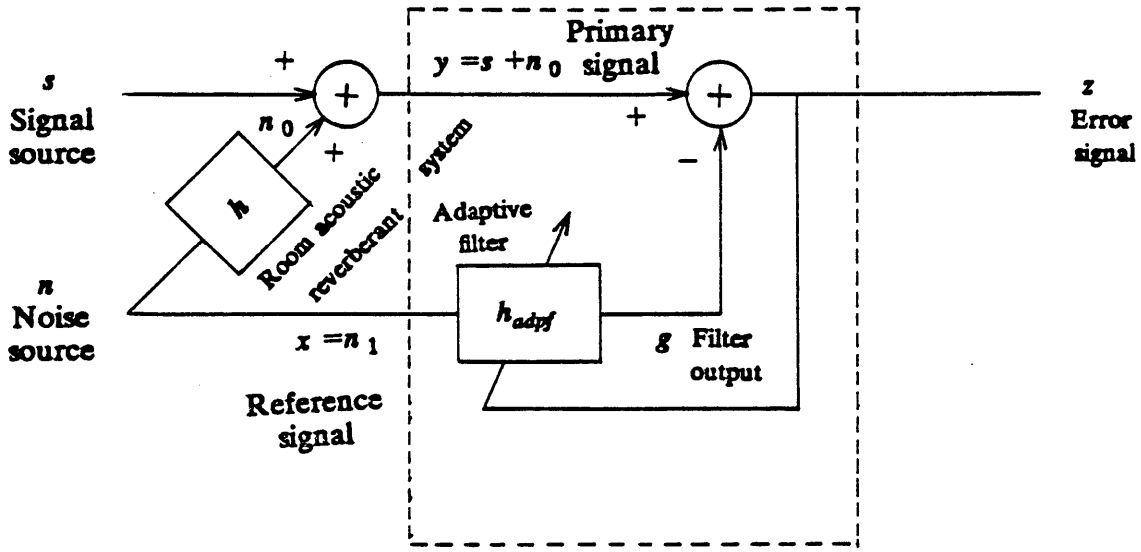


Fig 3.1

The block diagram of the adaptive noise canceler

Adaptive noise canceling schemes often use the least squares error criteria to choose a "good" match between the primary signal y and the filter output g.

$$\min E\ [(y-g)^2]$$

The output error z is

$$z = (s+n_0)-g \tag{3.2.1}$$

Squaring, one obtains

$$z^2 = s^2+(n_0-g)^2+2s(n_0-g) \tag{3.2.2}$$

Taking the expectation of both sides of (3.2.2) and taking into account that s is

uncorrelated with n_0 and g,

$$E[z^2] = E[s^2]+E[(n_0-g)^2]+2E[s(n_0-g)]$$
$$= E[s^2]+E[(n_0-g)^2] \tag{3.2.3}$$

Since the signal power $E[s^2]$ is independent of the filter adjustment, optimizing the filter to minimize $E[z^2]$ is equivalent to minimizing $E[(n_0-g)^2]$. Thus the optimal filter output is the best least squares estimate of n_0.

In particular, in the newscaster problem, the reference signal is

$$n_1 = x, \tag{3.2.4}$$

the interfering signal in the primary signal is

$$n_0 = h * x, \tag{3.2.5}$$

and the primary signal is

$$y = s+h * x. \tag{3.2.6}$$

* : convolution operator

where h is the impulse response of the room including the reverberant effect. Since the adaptive filter output g is

$$g = h_{adpf} * x \tag{3.2.7}$$

minimizing $E[(n_0-g)^2]$ is equivalent to

$$\min E[(n_0-g)^2] = E[(h * x-h_{adpf} * x)^2]$$
$$= E[((h-h_{adpf}) * x)^2] \tag{3.2.8}$$

Consequently, if the adaptive filter h_{adpf} exactly matches the actual room reverberant impulse response h, then the interfering signal is perfectly canceled. Thus minimizing the expected energy in z with respect to h_{adpf}

$$\hat{h}_{adpf} \longleftarrow \min_{h_{adpf}} E[|y-h_{adpf} * x|^2] \tag{3.2.9}$$

where E is the ensemble average, yields the optimal filter $\hat{h}_{adpf} = h$.

This minimization problem can also be applied to more general system identification problems, where $x(n)$, $y(n)$ are not necessarily stationary stochastic signals. The system function $h(n)$ is to be estimated from the observation signals $x(n)$ and $y(n)$, and the residue sequence $s(n)$ is to be found. The literature of system identification provides several practical methods for solving the underlying problem. Which method is the most appropriate for solving the problem is dependent on the details of the application, such as what additional information about $h(n)$ is available, or what kind of input signals are present.

3.3. System identification

System identification is one of the most important areas in engineering and economics because of its wide applicability. A great deal of research efforts has focused on studying the properties of a wide variety of algorithms for performing system identification. Consequently system identification techniques are generally well known and understood for a wide variety of applications. 1

In this section several promising techniques will be studied for use of the estimation of the room transfer function. Fig 3.2 shows a conventional system identification model. This is exactly equivalent to the problem statements in the preceding section, where $h(n)$ is a room transfer function; $x(n)$ is the interfering speech, $s(n)$ is the desired speech (the newscaster's voice) to be estimated and $y(n)$ is the microphone signal. Our problem is to estimate $h(n)$, then to estimate $s(n)$ from $x(n)$ and $y(n)$.

The system output (microphone signal) can be written in the form ;

$$\begin{aligned} y(n) &= x(n) * h(n) + s(n) \\ &= \sum_{m=0}^{M-1} h(m)\, x(n-m) + s(n) \end{aligned} \tag{3.3.1}$$

To solve for $h(n)$, we will minimize the least squares error signal $y(n) - x(n) * h(n)$.

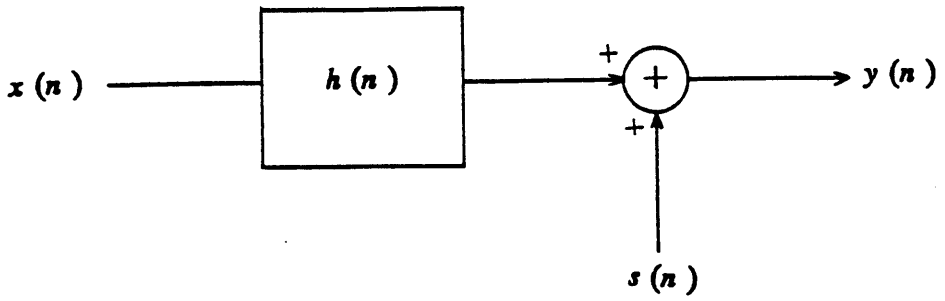


Fig 3.2

The block diagram of the conventional system identification problem.

That is,

$$\hat{h}(n) \leftarrow\!\!- \min_{h(n)} \sum_{n} \left| y(n) - h(n) * x(n) \right|^2 \qquad (3.3.2)$$

This is the objective function minimized by the two-channel noise canceler.

We assume that the system function $h(n)$ can be effectively modeled with a finite, although long, impulse response of length M.

The input signals which arise in the original broadcast room problem have the following properties:

[1] Both $x(n)$ and $s(n)$ are band-limited speech signals. They are not stationary, and have large variations in signal energy for the long period of time.

[2] The system function $h(n)$ is generally time-varying.

[3] The duration of the impulse response, M, is generally unknown and could be rela-

tively long. (A few thousand points)

We will look over three fundamentally different system identification methods. These include the spectral analysis method, the pole-zero model method, and the finite impulse response technique (FIR). The spectral analysis method attempts to identify the linear system from short-time spectral information. Such a method potentially has the capability of estimating systems with long impulse responses, without the need for recursive matrix inversions, when the analysis is implemented using FFT.

The pole-zero modeling allows the impulse response to be infinitely long, because the model has poles. Although pole-zero modeling methods have not been well studied mostly due to the need for solving non-linear equations, the number of the parameters to be estimated in the system function is relatively small, leading to a possibly simpler and more robust procedure.

The finite impulse response technique (FIR) is a method for estimating impulse responses which are relatively short. The covariance method is one FIR technique, utilizing efficient recursion methods to solve a least squares minimization problem exactly. Methods similar to this have found wide use in speech processing [2] The Least Mean Squares (LMS) adaptive algorithm is another FIR method. It takes a sample-by-sample adaptive approach for recursively updating linear system estimates. It is especially useful for efficiently estimating linear , slowly time-varying systems of long impulse responses (up to 500).

3.4. The spectral analysis technique (Fourier transform technique)

One estimation method for the transfer function is based on spectral analysis. An estimate of the filter transfer function $\hat{H}(\omega)$ is derived from the optimal solution to the minimization problem in the two-channel noise cancellation. Since we will deal with a

non-stationary signal in later chapters, here we just concentrate on the case where $x(n), y(n)$ are stationary stochastic signals for the simplicity of the argument. The least squares estimates of the filter is given as the function which minimizes the least squares error of

$$Err = E[|y(n)-h(n) * x(n)|^2] \tag{3.4.1}$$

Here E is the ensemble average. If we assume the signals $x(n)$ and $y(n)$ are ergodic, then we can take an average of $\bar{Err}$ over N different values of n,

$$\bar{Err} = E\left[\frac{1}{N}\sum_n |y(n)-h(n) * x(n)|^2\right] \tag{3.4.2}$$

Approximating N as infinite and applying Parseval's theorem, Err' will be

$$\bar{Err} = E\left[\frac{1}{N}\frac{1}{2\pi}\int |Y(\omega)-H(\omega)X(\omega)|^2 d\omega\right] \tag{3.4.3}$$

Taking the derivatives with respect to both the real part and the imaginary of $H(\omega)$, and setting these derivatives to zero, gives the optimal filter $\hat{H}(\omega)$

$$\hat{H}(\omega) = \frac{E\left[\frac{1}{N}X^*(\omega)Y(\omega)\right]}{E\left[\frac{1}{N}X^*(\omega)X(\omega)\right]} = \frac{S_{xy}(\omega)}{S_{xx}(\omega)} \tag{3.4.4}$$

where $S_{xx}(\omega)$ is the auto power spectral density of $x(n)$, and $S_{xy}(\omega)$ is the cross spectral density of $x(n)$ and $y(n)$.

Unfortunately, in practice we do not know these power spectral densities in advance, and we must estimate them from observations of $x(n)$ and $y(n)$. Traditional power spectral density estimation in the stochastic literature includes several methods for estimating the power or cross spectral density from an available data segment. One common method is the periodogram. Although there are various versions of the periodogram, a basic outline is given here. Suppose that K frames of a data sequence with L points in each frame

are available. Then using an N point DFT for the Fourier transform of each L point sequence ($L \leq N$), (3.4.4) would be expressed in terms of periodograms $P_{xy}(\omega)$, $P_{xx}(\omega)$ as

$$\hat{H}(k)=\frac{\hat{P}_{xy}(k)}{\hat{P}_{xx}(k)} \tag{3.4.5}$$

where

$$\hat{P}_{xy}(k) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{N}X_i^*(k)Y_i(k) \tag{3.4.6}$$

$$\hat{P}_{xx}(k) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{N}X_i^*(k)X_i(k) \tag{3.4.7}$$

$$X_i(k) = \sum_{n=1}^{N-1} x_i(n)exp(\frac{-j2\pi k\ n}{N}) \tag{3.4.8}$$

$$Y_i(k) = \sum_{n=1}^{N-1} y_i(n)exp(\frac{-j2\pi k\ n}{N}) \tag{3.4.9}$$

$$x_i(n) = x(n)w_r(iL-n) \tag{3.4.10}$$

$$y_i(n) = y(n)w_r(iL-n) \tag{3.4.11}$$

$$w_r(n) = \begin{cases} 1 & 0 \leq n \leq L-1 \\ 0 & n<0\ ,\ n \geq L \end{cases} \tag{3.4.12}$$

In these equations, the power spectrum density estimate is formed by taking the DFT of the available segments of data, multipling by the complex conjugate of the DFT $X^*_i(k)$, and then averaging over many segments.[3,4]

The actual processing of input data $x(n)$ and $y(n)$ is achieved in the overlap-save method as follows. Assume that the length of the true filter is M. After N ($\geq$ L+M-1) point DFT of i th frame L point inputs $x_i(n)$ and $y_i(n)$ are calculated, denoted as $X_i(k)$ and $Y_i(k)$ respectively, The output of the filter $\hat{H}(k)$ with the input signal $X_i(k)$ is denoted as $G_i(k)$ in the form of

$$G_i(k) = X_i(k)\hat{H}(k) \tag{3.4.13}$$

then take the inverse N point DFT of $G_i(k)$,

$$g_i(n) = \frac{1}{N}\sum_{k=1}^{N-1} G_i(k)exp(\frac{j2\pi k\ n}{N}) \tag{3.4.14}$$

Although the outcome $g_i(n)$ is L+M-1 point long, the first and last M-1 points are contaminated by the time aliasing. Then the middle points from $n=M-1$ to $n=L-1$ are subtracted from $y(n)$, the residue is the estimated desired speech $\hat{s}_i(n)$. The window for sectioning input data is shifted by L-M+1 points, the next frame (i+1 th) input data are read and same procedure is taken to obtain another L-M+1 point $\hat{s}_{i+1}(n)$.

One main advantage of the periodogram approach is the computational efficiency if the FFT is used in the periodogram. The number of multiplications required to estimate $h(n)$ is $3N\log_2 N$ for each frame. Here N is the FFT buffer size.(sample points) Another advantage is the applicability to whatever the room reverberant system function would be, if it is assumed to be smooth in the frequency domain.

An essential problem in the estimate (3.4.5) is that the estimate $\hat{H}(k)$ is never the exact value of the true filter, even if it might be close. One major reason is the end effect of the finite length window on each frame data. This end effect shows up in a variety of fashions. Time aliasing in multiplication or division of two DFT is one phenomenon. Another major reason is a smearing effect of the window due to the sideslobe leakage. Large sideslobe height in the frequency response of the window causes degradation in the filter estimate at these sideslobe frequencies.

The need for the average of power spectra over multiple frames is claimed in the following way. If power spectra from only one frame, say the i th frame, are used for obtaining $\hat{H}(k)$ in the form of

$$\hat{H}(k) = \frac{\frac{1}{N}X_i^*(k)Y_i(k)}{\frac{1}{N}X^*{}_i(k)X_i(k)} = \frac{Y_i(k)}{X_i(k)} \tag{3.4.15}$$

The filter output $G_i(k)$ is

$$G_i(k) = X_i(k)\hat{H}(k) = Y_i(k) \tag{3.4.16}$$

$$g_i(n) = y_i(n) \tag{3.4.17}$$

then the output of the noise canceler $\hat{S}_i(k)$ is

$$\hat{s}_i(n) = y_i(n) - g_i(n) = y_i(n) - y_i(n) = 0 \tag{3.4.18}$$

Undesirablly, the desired speech estimate $\hat{s}(n)$ becomes zero, although the true desired speech should be present. The problem is that L points data were used in one Fourier transform culculation, but not averaged.

Finally, a notable point is that the periodogram estimation is also interpreted from the system modeling point of view as an attempt to perform a least squares fit of the available time data to a harmonic model, namely discrete Fourier series, extrapolating outside of the available time data by a periodic sequence.

3.5. Pole-zero modeling method

Assume that a room is exactly rectangular without any furniture in it, and that the reflection coefficients of the walls are independent of angle. Also we suppose that an omnidirectional loudspeaker and microphone are used. Then the response of the room reverberation could be accurately modeled by a total of fifteen parameters consisting of three room size parameters, six location parameters for the microphone and the loudspeaker, and six reflection coefficients of six walls.[5] Because this model is unwieldy, and the parameters are hard to estimate, we might try to approximate the true room response by a pole-zero model with M zeros and N poles,

$$H(z) = \frac{\sum_{k=0}^{M-1} b_k z^{-k}}{\sum_{k=0}^{N-1} a_k z^{-k}} \tag{3.5.1}$$

Major motivations for the current interest in the modeling approach to the problem are

[1] The number of the parameters to be estimated is less than that in the periodogram or the FIR technique.

The drawbacks to which attention should be paid are

[1] The performance of the modeling will fully depend on how well it can match the actual room response, which in tern depends on the model order.

[2] The determination of the order of poles and zeros is usually difficult.

[3] Since equations to be solved for determining the values of parameters are nonlinear, estimating the parameters is computationally intensive, compared with solving linear equations for an FIR filter model.[6] Furthermore the convergence properties of most pole-zero algorithms and their robustness are not guaranteed.

3.6. FIR Filter Technique (zero modeling method)

The finite length filter approach is one of the most straightforward representations for a room reverberant system, since it is intended for estimating the impulse response of the room in the time domain. This approach leads to linear equations for the impulse response samples as follows.

Suppose that an impulse response is represented as an M point finite length filter,then equation (3.2.9) leads to

$$\hat{h}(n) \leftarrow\!\!- \min_{h(n)} \sum_{n} \left[(y(n) - \sum_{k=0}^{M-1} h(k)x(n-k))^2 \right] \tag{3.6.1}$$

The projection theorem for least squares approximation tells us that the minimum least squares error is obtained when the error is orthogonal to the data set $x(n)$, i.e., when

$$\sum_{n} [(y(n) - \sum_{k=0}^{M-1} h(k)x(n-k))x(n-j)] = 0 \tag{3.6.2}$$

Now $r_{xx}(k,j) = \sum_n [x(n-k)x(n-j)]$ and $r_{xy}(j) = \sum_n [y(n)x(n-j)]$. Then the equations to solve in terms of $h(n)$ for n=0,1...M-1 are

$$\sum_{k=0}^{M-1} h(k) r_{xx}(k,j) = r_{xy}(j) \qquad j=0,1,\ldots,M-1. \tag{3.6.3}$$

A variety of algorithms are offered for estimating the r_{xx} and r_{xy} from finite data, and for solving the resulting equations.[7]

The covariance method is a well-known technique for estimating the r_{xx} and r_{xy} from a block of data $x_i(n)$, $y_i(n)$ (i is a block number for the data); Straightforward Gaussian elimination method requires $O(M^3)$ multiplications and $O(M^2)$ storages. The auto correlation method is another method. The Levinson's recursion is famous for efficient recursive algorithm for the resultant matrix equation in the auto correlation method. It requires $O(M^2)$ multiplications, and $O(M)$ storages. Unfortunately, model orders of $M = 1000$ or more are necessary to properly estimate the filter in the broadcast problem. This means that enormous sets of equations must be solved, leading to potentially large numerical errors and huge computation times. These observations imply that the covariance technique may hardly satisfy either the real-time processing requirement or the robustness requirement in the newscaster's problem.

However, an advantage is that if an algorithm could be found that could solve these equations fast and robustly, even for model orders as high as M=1000, then an exact solution for the least squares estimate $\hat{h}(n)$ could be obtained by the covariance method. This is because unlike the periodogram technique, this method does not use windowing, and therefore avoids introducing bias into the estimates.

Another common method for FIR filter technique is the Least Mean Squares

(LMS) technique. It is an iterative, minimum-seeking method for determining the least squares solution. It recursively estimates the filter on a sample-by-sample basis, rather than on a block-by-block basis like the covariance method or the auto correlation method.

Denote the L point observation error criteria as E_L,

$$E_L = \sum_{n=0}^{L-1} e(n)^2 \tag{3.6.4}$$

$$= \sum_{n=0}^{L-1} \left[y(n) - \sum_{m=0}^{M-1} h(m)x(n-m) \right]^2 \tag{3.6.5}$$

Minimizing E_L with $h(n)$ leads to the estimate $\hat{h}(n)$. Let $\underline{\hat{h}_i}$ be a vector containing an estimate of the true filter $h(n)$ at time i th , defined as

$$\underline{\hat{h}_i} = \begin{bmatrix} \hat{h}_i(0) \\ \hat{h}_i(1) \\ . \\ . \\ \hat{h}_i(M-1) \end{bmatrix} \tag{3.6.6}$$

The LMS algorithm determines the new estimate $\underline{\hat{h}_{i+1}}$ by the update formula ;

$$\underline{\hat{h}_{i+1}} = \underline{\hat{h}_i} - \mu \underline{\Delta_i} \tag{3.6.7}$$

Where $\underline{\Delta_i}$ is the gradient of E_L with respect to $\underline{\hat{h}}$ in the i+1 th iteration and μ is a constant.

Basically $-\underline{\Delta_i}$ determines the direction in which the correction is made for the $i+1$ iteration and μ is a constant which controls the size of the step taken in that direction. Since E_L is a quadratic function of $\underline{\hat{h}}$, a single minimum exists in the error surface and it can be shown that the LMS algorithm converges to this minimum on average. The LMS algorithm offers an approximate estimate but not the exact solution, to the least squares error estimate of the actual filter. Although the resultant estimate is not as exact as that in the covariance method, the need for $O(M)$

computation and $O(M)$ storages makes LMS easy to implement in cases with large system order M.

3.7. Comparison of the three methods

The covariance method is the most exact solution of the three, providing excellent estimates for $h(n)$. But the one possible disadvantage of covariance method is that the implementation must solve an M th order matrix equation at every time step. Since the algorithm uses a recursion procedure, inaccuracies in both the computation of the covariance matrix and the estimate $\hat{h}(n)$ start to become measurable for M on the order of 50 with double precision arithmetic. As such, estimation of systems with large values of M would generally not be practical using the covariance method. In addition, the storage for the covariance matrix grows as M^2, and also the computational time increases in proportion to the order of M^3. These again makes the method impractical for large values of M.

The LMS can cope with system functions with a larger M, by virtue of the small computational cost of $O(M)$ and storage requirement of $O(M)$, though it does not offer a exact filter in any case. However its adaptive speed is usually slow and requires a larger amount of observation data than other methods, since it is affected by band-limited input and by high signal to noise ratio.

The pole-zero model is not the method of choice, since the appropriateness of the model to the room reverberant system is unknown, and there is no guarantee for its validity. Furthermore efficient robust solutions for the parameters are not available.

The spectral analysis method has the major advantage that the implementation is simple, and can be practically used for large M values (at least up to 2000) by virtue

of the fast computation of an FFT. The computational time is $O(M \log_2 M)$ and the storage is $O(M)$. Also the ability to model any smooth transfer function makes this spectral analysis method attractive. The method can be made adaptive by updating a filter estimate with each new block of data. In this sense, the spectral analysis method satisfies three requirements out of four mentioned in the section 2.1 in the broadcast room problem. Theses advantages seem to suggest that the spectral analysis method is superior to other methods.

Only one anxiety is the robustness. Since windowing is used in estimating the power and cross spectra, then the data misadjustment near window ends causes error in the filter estimation. Due to the use of DFT, time aliasing (circular aliasing) may also contaminate the filter estimation.

References

1. L. R. Rabiner, R. E. Crochiere, and J. B. Allen, "FIR System Modeling and Identification in the Presence of Noise and with Band-Limited Inputs," *IEEE Trans. on Acoust., Speech, Signal Processing,* vol. ASSP-26, no. 4, August 1978.

2. Makhoul, J., "Linear Prediction: A Tutorial Review," *Proc. IEEE,* vol. 63, no. 4, pp. 561-580, 1975.

3. R. B. Blackman and J. W. Tukey, , *The Measurement of Power Spectra form the Point of View of Communications Engineering,* Dover, 1959., New York: .

4. P.D. Welch, "Spectra: A Method Based on Time Averaging over Short Modified Periodograms," *IEEE Transactions on Audio and Electroacoustics,* vol. 15, pp. 70-73, June 1967.

5. J. B. Allen and D. A. Berkely, "Image Method for Efficiently Simulating Small-Room Acoustics ," *J. A. Acoust. Soc. Am.,* , vol. 65, no. 4 , April 1979.

6. M. Morf , D. T. Lee, J. R. Nickolls , and A. Vieira, "A classification of algorithms for ARMA models and ladder realizations," *IEEE Int. Conf. Acoust. Speech, Signal Processing,* pp. 13-19, Hartford CT: , 1977.

7. M. Morf, B. Dickinson , T. Kailath , and A. Vieira , "Efficient solution of covariance equations for linear prediction," *IEEE Trans. Acoust. Speech, Signal Processing,* vol. ASSP-25, pp. 424-433 , Oct. 1977.

Chapter 4 Spectral Analysis Estimation

4.1. Introduction

In chapter 3 several candidate algorithms for estimating linear transfer functions were summarized, and their advantages and disadvantages for this broadcast studio problem were discussed. Careful analysis of the suitability of each candidate for this broadcasting problem suggests that the best approach would be spectral estimation, because of its superiority over the other techniques in terms of the crucial requirements for this problem.

In this chapter, the spectral analysis estimation technique is intensively discussed from the theoretical point of view. The discussion begins with deriving the classical frequency domain estimate of a transfer function via a least squares procedure, assuming stationary noise signals. This leads to estimating the filter as a ratio of a cross spectrum and a power spectrum. To form robust averaged estimates of these spectra, the concept of data frames is introduced. This frame concept views an infinite time sequence as a set of sequences presented on a frame by frame basis. This idea is commonly used for dealing with an infinite time sequence like noise or speech. The approach can then be made recursive by incorporating frames of data one at a time, updating the estimate with each new frame to track changes of the system function in time.

Next we introduce the Maximum Likelihood interpretation of the least squares error problem and conclude that the same minimization problem results if we only assume that the newscaster speech is Gaussian White noise. Stationarity or a careful stochastic description of the input reference signal are not needed.

We consider two approaches for speech case, each a variation of the spectral analysis

estimation method. In the first approach the estimate of the transfer function is the average of the transfer function estimates over frames (ATFE), unlike the averaging of the spectral in the periodogram technique. This approach was suggested by the observation that for non-stationary signals such as speech, the room transfer function estimate should be more stable from frame to frame than any spectral estimate derived from different frames of data.

The second method averages the weighted spectrum with the value of the reciprocal of the desired speech power spectral density estimate (AWSE). This second method relies on a Gaussian colored noise assumption for the newscaster's speech, allowing the colored noise power spectral density to change frame by frame. This modeling is acceptable for speech signals, though it is still not perfect for modeling the non-stationary of speech.

4.2. Spectral analysis solution for the least squares error criteria in the case of stationary signals.

In this section, we study the transfer function estimation technique in the case of stationary signals. We model the reference signal $x(n)$ as a stationary signal, whereas we model the newscaster's speech $s(n)$ as Gaussian white noise. $y(n)$ will then be stationary which is a combination of two stochastic signals. First an ideally theoretical solution will be given. The ideal theoretical solution estimates $H(w)$ as the ratio of the expectation of the spectral densities. Next, a practical solution using observed data will be given, which uses a periodogram technique to estimate the spectral densities.

4.2.1. The ideal theoretical solution for the least squares error.

The newscaster's solution for the noise canceler in the frequency domain can be found by minimizing the least squares error between $y(n)$ and the filtered reference

$h(n) * x(n)$

$$Err = E[|y(n) - h(n) * x(n)|^2] \tag{4.2.1}$$

Here E is the ensemble average. If we assume the signals $x(n)$ and $y(n)$ are ergodic, then we can take an average of Err over L different values of n,

$$\bar{Err} = E[\frac{1}{L}\sum_{n} |y(n) - h(n) * x(n)|^2] \tag{4.2.2}$$

Applying Parseval's theorem, $\bar{Err}$ will be

$$\bar{Err} = E[\frac{1}{L}\frac{1}{2\pi}\int_{-\pi}^{\pi} |Y(\omega) - H(\omega)X(\omega)|^2 d\omega] \tag{4.2.3}$$

Taking the derivatives with respect to both the real and imaginary parts of $H(\omega)$, and setting these derivatives to zero, gives the optimal filter $\hat{H}(\omega)$

$$\hat{H}(\omega) = \frac{E[\frac{1}{L}X^*(\omega)Y(\omega)]}{E[\frac{1}{L}X^*(\omega)X(\omega)]} = \frac{S_{xy}(\omega)}{S_{xx}(\omega)} \tag{4.2.4}$$

where $S_{xx}(\omega)$ is the auto power spectral density of $x(n)$, $S_{xy}(\omega)$ is the cross spectral density of $x(n)$ and $y(n)$.

4.2.2. The periodogram technique

Unfortunately, in practice we do not know these power spectral densities in advance, and we must estimate them from observations of $x(n)$ and $y(n)$. Traditional power spectral density estimation literature includes several methods for estimating the power or cross spectral density from an available data segment.[1] One common method is the periodogram. Although there are various versions of the periodogram, the basic approach can be described easily.

Assume that we need to estimate for spectral densities given only one finite length set of observation data $x(n)$, $y(n)$. Let us consider how to estimate the power spectral

density from this finite data record. (The argument for cross spectral density estimation is similar.) Suppose that one has a finite N point time sequence observation of a stationary stochastic process. (colored noise) Set $x(n) = 0$ for $n \geq N$, and $n < 0$. One primitive estimate of the power spectrum is then to calculate the average magnitude squared of the Fourier transform of the available sequence $x(n)$;

$$I_N(\omega) = \frac{1}{N} |X(\omega)|^2 \tag{4.2.5}$$

Here $X(\omega)$ is the discrete time Fourier transform (DTFT) of the available finite N point time sequence. That is,

$$X(\omega) = \sum_{n=0}^{N-1} x(n) exp(-j\,\omega n) \tag{4.2.6}$$

The expected value of $I_N(\omega)$ is

$$E[I_N(\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{xx}(\theta) W_B(\omega - \theta) d\theta \tag{4.2.7}$$

where $S_{xx}(\omega)$ is the true power spectral density, namely the Fourier transform of the true covariance

$$S_{xx}(\omega) = \sum_{m=-\infty}^{\infty} R_{xx}(m)\, exp(-j\,\omega m) \tag{4.2.8}$$

and $W_B(\omega)$ is the Fourier transform of the the triangular window $w_B(n)$;

$$w_B(n) = \begin{cases} \dfrac{N - |m|}{N} & |m| \leq N-1 \\ 0 & |m| > N \end{cases} \tag{4.2.9}$$

$$W_B(\omega) = \frac{1}{N} \left[\frac{\sin\left(\frac{\omega N}{2}\right)}{\sin\left(\frac{\omega}{2}\right)} \right]^2 \tag{4.2.10}$$

It suggests that $I_N(\omega)$ is a biased estimate due to the smearing effect of the window. The covariance of the estimate $I_N(\omega_1)$, and $I_N(\omega_2)$ is expressed in the form of

$$Cov[I_N(\omega_1)I_N(\omega_2)] \approx S_{xx}(\omega_1)S_{xx}(\omega_2)\left\{\left(\frac{\sin\left[\frac{(\omega_1+\omega_2)N}{2}\right]}{N\sin\left[\frac{(\omega_1+\omega_2)}{2}\right]}\right)^2+\left(\frac{\sin\left[\frac{(\omega_1-\omega_2)N}{2}\right]}{N\sin\left[\frac{(\omega_1-\omega_2)}{2}\right]}\right)^2\right\} \tag{4.2.11}$$

and the variance is

$$Var[I_N(\omega)] = (S_{xx}(\omega))^2\left\{1+\left(\frac{\sin(\omega N)}{N\sin\omega}\right)^2\right\} \tag{4.2.12}$$

These estimates can be calculated by discrete Fourier transforms (DFT). The DFT of $I_N(\omega)$ (4.2.5) is just values of $I_N(\omega_k)$ evaluated at equally spaced frequency samples $\omega_k = \frac{2\pi k}{N}$. If the variance between two arbitrary frequency samples spaced a multiple of $\frac{2\pi}{N}$ in the equation (4.2.11) is evaluated, one notices that the covariance (4.2.11) between those DFT coefficients is approximately zero, Therefore samples of $I_N(\omega)$ separated by $\frac{2\pi}{N}$ are approximately uncorrelated. The problem with this periodogram estimate is that the variance (4.2.12) at any frequency sample is the square of the true spectral density value $S_{xx}(\omega)$ and does not go to zero as the data length $N \to \infty$. Therefore, no matter how much data is available, the estimate $I_N(\omega_k)$ does not get close to the true power spectral density. (the estimate is not consistent.) Increasing the data length simply increases the number of independent frequency samples in $I_N(\omega)$ and thus it increases the rapidity of fluctuations of the estimate.

One common approach for reducing the variance of estimates is to take the average of estimates over several independent data sets. Let's assume that an available data segment $x(n)$ for $n = 0, \ldots, N-1$ is divided into K frames of L samples each e.g. $N = KL$.

$$x^{(i)}(n) = x(n+(i-1)L) \qquad 0 \le n \le L-1, 1 \le i \le K \tag{4.2.13}$$

and compute the magnitude squared of the Fourier transform for each frame.

$$I_L{}^{(i)}(\omega)=\frac{1}{L}\left|\sum_{n=0}^{L-1}x^{(i)}(n)\exp(-j\,\omega n)\right|^2 \qquad 1\le i\le K \tag{4.2.14}$$

The Bartlett periodogram estimate is then defined as the average of $I_L(\omega)$ over several frames

$$B_{xx}(\omega)=\frac{1}{K}\sum_{i=1}^{K}I_L{}^{(i)}(\omega) \tag{4.2.15}$$

Though $B_{xx}(\omega)$ is still a biased estimate, it has less variance than $I_N(\omega)$, because $I_L{}^{(i)}(\omega)$ $i=1,\ldots,K$ are approximately independent.

$$Var[B_{xx}(\omega)] \approx \frac{1}{K}Var[I_L(\omega)] \tag{4.2.16}$$

$$\approx \frac{1}{K}(S_{xx}(\omega))^2\left\{1+\left(\frac{\sin[\omega L]}{L\sin[\omega]}\right)2\right\} \tag{4.2.17}$$

Now $B_{xx}(\omega)$ is a consistent estimate, since the variance decreases to zero with the growth of K. However if only a limited amount of data is available, so that N is fixed, then increasing K means decreasing the length of each frame L, thus reducing the frequency resolution of the estimate. Therefore there is a tradeoff between spectral resolution and the variance, when we choose K and L given a fixed length N. In practice, values of L and K can often be determined by prior information about the underlying signal. For example, if the power spectral density is anticipated to have a sharp peak, then L should be large enough to offer sufficient frequency resolution.

Another approach for reducing the variance of the estimate is to multiply each frame of data by a smoothing window with a broader bandwidth and lower sidelobes than a rectangular window. Successive overlapping frames are then averaged to smooth the power spectrum estimate. Welch[2] offers a technique using a Hanning instead of a rectangular window for shaping each frame. Since a Hanning window has a broader bandwidth than a

rectangular window, the estimate is smoothed over neighbor frequency samples to form a more smoothed frame. The data frames $x^{(i)}(n)$ are usually overlapped in time. Welch method is expressed as

$$J_L{}^{(i)}(\omega) = \frac{1}{LU} \left| \sum_{n=0}^{L-1} x^{(i)}(n) w(n) \exp(-j\,\omega n) \right|^2 \qquad i = 1, \ldots, K \tag{4.2.18}$$

$$U = \frac{1}{L} \sum_{n=0}^{L-1} w^2(n) \tag{4.2.19}$$

Then, the power spectrum estimate is

$$B_{WW}(\omega) = \frac{1}{K} \sum_{i=1}^{K} J_L{}^{(i)}(\omega)$$

The expectation is still biased, but the variance goes down in proportion to $\frac{1}{K}$.

$$Var[B_{WW}(\omega)] \approx \frac{1}{K} S_{xx}{}^2(\omega) \tag{4.2.20}$$

4.2.3. Transfer function estimation using the periodogram technique

Now suppose that K frames of input data $x(n)$ and $y(n)$ are available with L points in each frame. Then using the DFT for the Fourier transform of each sequence, (4.2.4) would be expressed in terms of periodograms $P_{xy}(\omega), P_{xx}(\omega)$ as

$$\hat{H}(k) = \frac{\hat{P}_{xy}(k)}{\hat{P}_{xx}(k)} \tag{4.2.21}$$

where

$$\hat{P}_{xy}(k) = \frac{1}{K} \sum_{i=1}^{K-1} \frac{1}{L} X_i{}^*(k) Y_i(k) \tag{4.2.22}$$

$$\hat{P}_{xx}(k) = \frac{1}{K} \sum_{i=1}^{K-1} \frac{1}{L} X_i{}^*(k) X_i(k) \tag{4.2.23}$$

$$X_i(k) = \sum_{n=0}^{N-1} x_i(n) exp\left(\frac{-j\,2\pi k\ n}{L}\right) \tag{4.2.24}$$

$$Y_i(k) = \sum_{n=0}^{N-1} y_i(n) exp\left(\frac{-j\,2\pi k\ n}{L}\right) \tag{4.2.25}$$

$$x_i(n) = x(n) w_r(iL - n) \tag{4.2.26}$$

$$y_i(n) = y(n)w_r(iL - n) \tag{4.2.27}$$

$$w_r(n) = \begin{cases} 1 & 0 \le n \le L-1 \\ 0 & n < 0, n \ge L \end{cases} \tag{4.2.28}$$

It shows that the power spectrum density estimation is accomplished by taking the DFT of the available segments of data, multiplying by the complex conjugate of the DFT of $X^*_i(k)$, and then averaging over many segments.[3,4,5]

4.2.4. Recursive frame-by-frame solution

The idea of dividing an infinitely long signal into a set of frames, processing each frame, and then recombining a set of frames to generate an infinitely long output sequence, is a common method in digital signal processing. Even if a signal is non-stationary over long time intervals, this frame concept may often deal with a portion of the signal as stationary when the signal is assumed to be quasi-stationary. Also, if a system varies slowly, one may still view the system as time-invariant in a short time period.

We now utilize a frame approach to recursively estimate the transfer function. The recursive method integrates well with real-time estimation and filtering, and also provides the capability of adapting to slowly time-varying transfer functions. Assume that a filter estimate from the previous frame is available. When a current frame of data is acquired, the algorithm updates the estimate of the transfer function, then estimates the current frame of newscaster speech. We then iterate, using the next frame of data to further improve the transfer function.

The recursive method is:

Given the i-1 th stage accumulated estimates for the power spectral densities $\hat{P}_{xx}^{(i-1)}(k)$, $\hat{P}_{xy}^{(i-1)}(k)$, and the transfer function

$$\hat{H}_{i-1}(k) = \frac{\hat{P}_{xy}^{(i-1)}(k)}{\hat{P}_{xx}^{(i-1)}(k)} \tag{4.2.29}$$

Then the i th accumulated estimates are updated from the i th frame data by:

$$\hat{P}_{xx}^{(i)}(k)=(1-\frac{1}{i})\hat{P}_{xx}^{(i-1)}(k)+\frac{1}{i}X_i^*(k)X_i(k) \tag{4.2.30}$$

$$\hat{P}_{xy}^{(i)}(k)=(1-\frac{1}{i})\hat{P}_{xy}^{(i-1)}(k)+\frac{1}{i}X_i^*(k)Y_i(k) \tag{4.2.31}$$

$$\hat{H}_i(k)=\frac{\hat{P}_{xy}^{(i)}(k)}{\hat{P}_{xx}^{(i)}(k)} \tag{4.2.32}$$

To restore the newscaster's speech $s(n)$, it is reasonable to use a Fourier transform technique, such as the overlap-save method, since the filter estimate is available in the frequency domain. Assume that the filter length is M. After N ($\geq$ L+M-1) point DFT's of the i th frame L point inputs $x_i(n)$, $y_i(n)$ are calculated, the output of the filter $\hat{H}_i(k)$ is computed as;

$$G_i(k)=X_i(k)\hat{H}_i(k) \tag{4.2.33}$$

then taking the inverse DFT of $G_i(k)$,

$$g_i(n)=\frac{1}{N}\sum_{k=1}^{N-1}G_i(k)\exp(\frac{j\,2\pi k\,n}{N}) \tag{4.2.34}$$

Only the middle portion, sample points from n=M-1 to n=L-1 are valid linear convolution terms. These valid sample points are subtracted from the corresponding $y_i(n)$ and L-M+1 point valid $s_i(n)$ are obtained. The window for sectioning input data is shifted by L-M+1 points, the next frame (i+1 th) input data are read and same procedure is used to obtained another L-M+1 points $s_{i+1}(n)$.

4.2.5. The effect of added newscaster's speech on transfer function estimation

This section deals with the effect of the presence of newscaster speech on estimation performance. In the broadcast room problem, the interfering speech (the studio engineer's voice $x(n)$) is regarded as an 'undesired signal', thus to remove that undesired speech from the microphone is the ultimate goal in the noise canceler. Ordinarily the presence of

a loud newscaster speech is favorable, because the interfering speech tends to be concealed by the loud speech. However from the point of view of the transfer function estimation problem, the newscaster's voice (the newscaster speech) plays the role of a disturbance signal. When the newscaster talks loudly, accurate transfer function identification becomes difficult. The reason will be given below.

When the true transfer function is denoted as $H(k)$, and if the frame length is approximately infinite, then the primary speech is written in the DFT expression as

$$Y_i(k) = S_i(k) + H(k) X_i(k) \tag{4.2.35}$$

Substituting the above equation into equation (4.2.21), the i th filter estimate $\hat{H}_i(k)$ is expressed in the form of

$$\hat{H}_i(k) \approx H(k) + \frac{\frac{1}{i}\sum_{l=1}^{i} X^*_l(k) S_l(k)}{\frac{1}{i}\sum_{l=1}^{i} X^*_l(k) X_l(k)} \tag{4.2.36}$$

The second term of the above equation behaves as the deviation from the true value of the filter. This deviation term becomes large either when the denominator is small, or when the numerator is large.

Suppose that the newscaster's speech $s(n)$ is not present at all during the entire period of the time. Then $S_l(k)$ is always zero for all frame l. As a result, the second term in (4.2.36). will be zero, and within our approximation accuracy, the estimate $H_i(k)$ for any frame i is equal to the true filter $H(k)$. Thus the algorithm will adapt to the true filter after the first frame.

If the newscaster's speech $s(n)$ is non-zero, however, then even if $x(n)$ and $s(n)$ are uncorrelated, the numerator in the second term does not necessarily vanish perfectly for the frame i. The reason is that the expectation of the second term is zero, but its variance is in a crude sense proportional to the power in $S_l(k)$, and inversely proportional to the

power in $X_l(k)$. Thus, the large energy in the newscaster's speech, or the small energy in the reference speech, equally makes the variance worse. Therefore the newscaster's speech causes the deviation in the filter estimate and the extent to which it degrades the filter estimate is roughly determined by the ratio of the power of the newscaster's speech to that of the reference speech.

4.3. Maximum Likelihood (ML) interpretation to the least squares criteria

We introduce Maximum Likelihood Estimation (ML) in order to interpret least squares error estimation of the transfer function from another point of view. This alternative approach provides new insights into the underlying system identification problem. The aim of this section is to derive the ML technique assuming that the newscaster signal is Gaussian white noise. We will not need to assume that reference signal (the loudspeaker) is stationary - in particular, since $x(n)$ and $y(n)$ may be non-stationary, we will not be able to refer to the "power spectrum" or "cross power spectrum" of $x(n)$ and $y(n)$. Surprisingly, despite this difficulty, we will show that the ML method, when $s(n)$ is stationary white noise, is identical to the least squares minimization problem in the previous section.

Suppose that the system identification problem in Fig 3.2 be restated as ;

Assume that $s(n)$ be a Gaussian white noise with an unknown constant power spectral density σ^2. Let $x(n)$ be a deterministic known signal, which is processed through a fixed though unknown filter $h(n)$, then corrupted by $s(n)$ to form $y(n)$. Given the observations, the random signal $y(n)$ and the known deterministic signal $x(n)$, estimate the unknown deterministic filter coefficients $h(n)$ and the unknown constant power spectral density σ^2.

A well-known technique for estimating unknown deterministic parameters when the probability density function of the observations $y(n)$ is known, is Maximum Likelihood (ML) estimation. ML in general works as follows. Given known $x(n)$ and $y(n)$, set parameters $h(n)$ $n=0,\ldots,M-1$ and σ^2 to arbitrary values, and then calculate the probability density function of $y(n)$. Repeat this calculation for all possible sets of parameters values for $h(n)$ and σ^2. Among all possible sets of values, choose the set of parameter values that provide the maximum probability density of $y(n)$.

An actual procedure can be realized in the following manner. Since $s(n)$ is zero-mean Gaussian white noise, $y(n)$ is also Gaussian white noise with mean $x(n)*h(n)$, and covariance $R_{yy}(n,n+m)=\sigma^2\delta(m)$. Then the probability density of the point $y(n)$ is :

$$\log p\ (y\ |\underline{x},\underline{h},\sigma^2)=-\frac{1}{2}\left[\frac{|y-\underline{x}*\underline{h}\,|^2}{\sigma^2}+log\,(2\pi\sigma^2)\right] \tag{4.3.1}$$

Now we suppose L observation points $y(n)$ and $x(n)$ for $n=0,\ldots,L-1$ are available. Then the joint probability density of

$$\underline{y}=\begin{bmatrix} y(0)\\ y(1)\\ .\\ .\\ .\\ y(L-2)\\ y(L-1)\end{bmatrix} \tag{4.3.2}$$

is a Gaussian density with ;

$$\text{mean}\ ;E[\underline{y}]=\underline{h}\underline{x}$$
$$\text{covariance matrix}\ ;Cov[\underline{y}]=\Lambda_{yy}$$

where

$$\underline{x}=\begin{bmatrix} x(0)\\ x(1)\\ .\\ .\\ .\\ x(L-2)\\ x(L-1)\end{bmatrix} \tag{4.3.3}$$

$$\mathbf{h} = \begin{bmatrix} h(0) & 0 & 0 & . & . & . & 0 \\ h(1) & h(0) & 0 & . & . & . & 0 \\ . & h(1) & h(0) & . & . & . & 0 \\ h(M-1) & . & h(1) & . & . & . & 0 \\ 0 & h(M-1) & . & . & . & . & 0 \\ . & 0 & h(M-1) & . & . & . & 0 \\ 0 & 0 & 0 & h(M-1) & h(M-2) & . & h(0) \end{bmatrix} \tag{4.3.4}$$

$$\Lambda_{yy} = \begin{bmatrix} \sigma^2 & 0 & . & . & . & 0 \\ 0 & \sigma^2 & . & . & . & 0 \\ 0 & 0 & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \sigma^2 \end{bmatrix} \tag{4.3.5}$$

Thus it follows that the average log joint probability density has the form :

$$\frac{1}{L}\log p\,(\underline{y}\,|\underline{x},\mathbf{h},\sigma^2) = -\frac{1}{2L}\left[(\underline{y}-\mathbf{h}\underline{x})^T \Lambda_{yy}^{-1}(\underline{y}-\mathbf{h}\underline{x}) + \log(|2\pi\Lambda_{yy}|)\right] \tag{4.3.6}$$

The reason for considering the average of the log probability density, instead of merely the probability, is that we will want to observe the asymptotic behavior of the probability density $L \to \infty$. Substituting Λ_{yy} into the above equation ;

$$\frac{1}{L}\log p\,(\underline{y}\,|\underline{x},\mathbf{h},\sigma^2) = -\frac{1}{2}\left[\frac{1}{L}\sum_{n=0}^{L-1}\frac{|y(n)-h(n)*x(n)|^2}{\sigma^2} + \log(2\pi\sigma^2)\right] \tag{4.3.7}$$

Note that the first term inside the summation corresponds to the least squares error between the primary signal and the filtered version of the reference signal normalized with σ^2 This normalized error simply represents a Gaussian white noise with variance 1. The second term is a gain factor controlled by only σ^2. ML suggests we estimate $\hat{h}(n)$ by maximizing the average L point log joint probability with respect to $h(n)$ for $n = 0, \ldots, L-1$ and σ^2. Then estimates $\hat{h}(n)$ and $\hat{\sigma}^2$ are given by ;

$$\hat{h}(n),\hat{\sigma}^2 \leftarrow \max_{h(n),\sigma^2} -\frac{1}{2}\left[\frac{1}{L}\sum_{n=0}^{L-1}\frac{|y(n)-h(n)*x(n)|^2}{\sigma^2} + \log(2\pi\sigma^2)\right] \tag{4.3.8}$$

First, maximizing the average log probability function with respect to $h(n)$ leads to minimizing the numerator of the first term ;

$$\hat{h}(n) \leftarrow \min_{h(n)} \sum_{n=0}^{L-1} |y(n) - h(n) * x(n)|^2 \tag{4.3.9}$$

Taking the partial derivative of the average log probability function with respect to σ^2, and setting it to zero, one obtains ;

$$\hat{\sigma}^2 = \frac{1}{L} \sum_{n=0}^{L-1} |y(n) - \hat{h}(n) * x(n)|^2 \tag{4.3.10}$$

Interestingly, the equation (4.3.9) is exactly the same minimization problem as presented in the least squares criteria in the two channel noise canceler scheme, and the equation (4.3.10) is the minimum error energy in that minimization problem. These facts suggest that the commonly-used least squares criteria is equivalent to the ML criteria with a Gaussian white noise assumption for the newscaster's speech. In other words, the approach of building the model in the fig 3.2 and solving the equation (3.3.2) to obtain estimate $\hat{h}(n)$ is based on the Gaussian white noise assumption for the optimal speech, whereas the reference signal is regarded as a known deterministic sequence. As discussed later, one may conjuncture that this approach may lack the capability to deal with non-stationary signal and highly colored newscaster's speech signal.

4.4. Method 1 for non-stationary speech signals

4.4.1. Motivation

The Fourier transform estimate $\hat{H}(k)$ for the room reverberant transfer function and the estimate of the newscaster's speech have been derived for stationary signals in the preceding section. $\hat{H}(k)$ is estimated as the ratio of the cross power spectral density estimate to the auto power spectral density estimate. These spectral density estimates are

found by averaging $\frac{1}{L}X_i^*(k)X_i(k)$, or $\frac{1}{L}X_i^*(k)Y_i(k)$ over frames as in (4.2.22) and (4.2.23). Beware that for speech signals, however, because of its non-stationary property, (4.2.22) and (4.2.23) do not represent spectral density estimates. Non-stationary signals do not have spectral densities in the sense that these spectral densities do not correspond to an expectation of the magnitude of signals at a certain frequency . In other words, the averaging frames $\frac{1}{L}X_i^*(k)X_i(k)$, or $\frac{1}{L}X_i^*(k)Y_i(k)$ will not necessarily converge to a particular estimate, since such limit values do not exist for speech signals. In practice, the shape of these terms $\frac{1}{L}X_i^*(k)X_i(k)$, or $\frac{1}{L}X_i^*(k)Y_i(k)$ vary drastically frame by frame.

In this section, one of our estimation methods for coping with speech signals is given.

4.4.2. Averaging the transfer function estimates

Since power spectral density estimates do not make sense for non-stationary data, one might try estimating the transfer function by the fellowing alternative procedure;

$$\hat{F}_i(k)=\frac{\frac{1}{L}X^*_i(k)Y_i(k)}{\frac{1}{L}X^*_i(k)X_i(k)} \tag{4.4.1}$$

$$\hat{H}_i(k) = \frac{1}{K}\sum_{i=1}^{K}\frac{\frac{1}{L}X^*_i(k)Y_i(k)}{\frac{1}{L}X^*_i(k)X_i(k)} \tag{4.4.2}$$

$$= \frac{1}{K}\sum_{i=1}^{K}\hat{F}_i(k) \tag{4.4.3}$$

This indicates that the current transfer function estimate $\hat{F}_i(k)$ is computed, from the i th frame data, and then the average is taken of the frame filter estimates $\hat{F}_i(k)$ in order to obtain the smoothed estimate $\hat{H}_i(k)$. To calculate this adaptively, a geometrically weighted average is calculated, which can be calculated recursively in the following way:

$$\hat{H}_i(k) = (1-\mu)\hat{H}_{i-1}(k) + \mu\hat{F}_i(k) \tag{4.4.4}$$

where μ is a parameter. (μ may be a function of k)

Suppose that $H(k)$ represents the DFT of the true transfer function. For sufficiently long frame length, we can approximately express $Y_i(k)$ as

$$Y_i(k) \approx S_i(k)+H(k)X_i(k) \tag{4.4.5}$$

The reason why this is just the approximation is that owing to the finite frame length, the circular aliasing and windowing effects causes inequality. We will discuss this issue in detail in a later section.

Substituting this into the expression of the current estimate from the i th frame data, one obtains

$$\hat{F}_i(k) = \frac{Y_i(k)}{X_i(k)} \tag{4.4.6}$$

$$\approx \frac{S_i(k)+H(k)X_i(k)}{X_i(k)} \tag{4.4.7}$$

$$= H(k)+\frac{S_i(k)}{X_i(k)} \tag{4.4.8}$$

Thus the averaged estimate $\hat{H}_i(k)$ becomes

$$\hat{H}_i(k) \approx H(k)+\sum_{l=-\infty}^{i} (1-\mu)^{i-l}\mu \frac{S_i(k)}{X_i(k)} \tag{4.4.9}$$

The justification for using the estimate (4.4.9) is that if $X_i(k)$ and $S_i(k)$ are uncorrelated and zero mean, then the second term will average to 0, leaving the first term which is the true transfer function.

Unfortunately, this estimate is subject to extremely large errors (tremendously large variance), since the second term introduces a very large variance in $\hat{H}_i(k)$, particularly because $\frac{S_i(k)}{X_i(k)}$ the second term becomes enormously if $X_i(\omega)$ is small.

One technique to avoid this large variance is to try to distinguish frequency components which will have low versus high variance, and let only the low variance frequency

components of the estimates $\hat{F}_i(k)$ contribute to the accumulated estimate $\hat{H}_i(k)$. The term $\frac{S_i(k)}{X_i(k)}$ becomes large, either if $S_i(k)$ is substantial at a certain frequency k or if $X_i(k)$ is too small at a certain frequency k. At this frequency, the current estimate $\hat{F}_i(k)$ gives a poor estimate, and one should avoid adding it into $\hat{H}_i(k)$. On the contrary, if $S_i(k)$ is nearly zero or $X_i(k)$ is reasonably large, then the current estimate $\hat{F}_i(k)$ at that frequency component offers a good estimate.

In order to know whether $\hat{F}_i(k)$ offers a good estimate at each frequency component k, one should have at least rough estimate of $S_i(k)$. Unfortunately knowing the optimal speech $S_i(k)$, is the ultimate goal in the problem. One reasonable way is to roughly "pre-estimate" $S_i(k)$ using the previous frame's averaged estimate $\hat{H}_{i-1}(k)$ and the i th frame's input data $x(n)$, $y(n)$. That is,

$$\hat{S}_i(k) = Y_i(k) - \hat{H}_i(k)X_i(k) \tag{4.4.10}$$

To judge whether $\hat{F}_i(k)$ obtained in the current frame is a good estimate at frequency component k, one evaluates $\frac{\hat{S}_i(k)}{X_i(k)}$ and if this value is significantly larger than $H(k)$ at frequency k, one claims that $\hat{F}_i(k)$ is poor at that frequency, and don't update the accumulated estimate $\hat{H}_i(k)$ at frequency k. On the contrary, if the value happens to be small at k, then one supposes that $\hat{F}_i(k)$ is good at k, and updates $\hat{H}_i(k)$ at this frequency.

The next question is how to determine the degree of the contribution of the current frame estimate to the averaged filter estimate at each frequency k.

We introduce a rough scheme to answer this question. The aim of the following discussion is, however, not to present a legitimate algorithm, but to bring an insight into the encountered question. This discussion leads to another approach to the same question, which will be discussed in the next section.

The rough scheme for implementing the frequency selective contribution is as follows. Generally the extent to which $\hat{F}_i(k)$ contributes to $\hat{H}_i(k)$ may be controlled by the step size parameter μ. To realize the relation of μ with the validity of the current filter estimate at frequency k, we study a parameter, the ratio of of the estimated filtered reference speech to the estimated optimal speech, defined as;

$$Rt(k) = \left| \frac{X_i(k)\hat{H}_{i-1}(k)}{\hat{S}_i(k)} \right| \tag{4.4.11}$$

$$\approx \left| \frac{X_i(k)\hat{H}_{i-1}(k)}{Y_i(k) - \hat{H}_{i-1}(k)X_i(k)} \right| \tag{4.4.12}$$

Note that $Rt(k)$ is a positive parameter. Suppose that $Rt(k)$ is small, then the disturbance signal (the newscaster's speech) is large, in comparison with the reference speech. In the case, μ should be small to prevent the contribution of the current estimate $\hat{F}_i(k)$ to the averaged filter estimate $\hat{H}_i(k)$. Whereas if $Rt(k)$ is large, then the reference speech is dominant over the disturbance, which should lead to the significant contribution to the averaged filter estimate. Thus, μ should be large. Therefore the desirable characteristics of a function μ of the parameter $Rt(k)$ is a monotonically increasing function.

For a simple example, suppose that we use a threshold test. If $Rt(k)$ at the frequency k is smaller than a certain threshold value, μ is zero, that is, the filter estimate is not updated. If $Rt(k)$ at the frequency k larger than the threshold value, μ is non zero constant μ_0. This threshold case is equivalent to choosing a step function shown in Fig 4.1 Therefore at each frequency component k, $Rt(k)$ is evaluated and μ for each k is determined separately.

If the estimated filter length is M, the FFT buffer is N, and the length of data to be read for the frame L, then the whole algorithm is

[1] Given the previous frame's accumulated filter estimate $\hat{H}_i(k)$ (an M point sequence), observe the current frame L point data $x_i(n), y_i(n)$. calculate the pre-estimate $\tilde{S}_i(k)$ by

$$\tilde{S}_i(k) = Y_i(k) - \hat{H}_i(k)X_i(k) \tag{4.4.13}$$

where $X_i(k), Y_i(k)$ are the N point DFT of the $x_i(n)$ and $y_i(n)$.

[2] Compute $Rt(k)$ in the form of

$$Rt(k) = \left| \frac{X_i(k)\hat{H}_{i-1}(k)}{\tilde{S}_i(k)} \right| \tag{4.4.14}$$

[3] Refer to a chosen function $\mu(Rt(k))$ so as to obtain the value of the μ. For the step function case, this is equivalent to perform the threshold test.

[4] Calculate the current transfer function estimate $\hat{F}_i(k)$ by

$$\hat{F}_i(k) = \frac{\frac{1}{L}X_i^*(k)Y_i(k)}{\frac{1}{L}X_i^*(k)X_i(k)} \tag{4.4.15}$$

[5] Compute the current frame accumulated filter estimate by

$$\hat{H}_i(k) = (1-\mu)\hat{H}_{i-1}(k) + \mu\hat{F}_i(k) \tag{4.4.16}$$

[6] Make the sequence $\hat{h}_i(n)$ M points long. This is done by, take the inverse DFT of $\hat{H}_i(k)$, truncating the last N-M points of the resultant time sequence of $\hat{h}_i(n)$, to get the M point filter estimate. Then take the N point DFT of truncated filter estimate.

[7] Filter $X_i(k)$ with $\hat{H}_i(k)$ by the overlap-save method.

For the particular function $\mu(Rt(k))$ shown in Fig 4.2. however, the above idea fails badly. Suppose one applies this algorithm to the case with no newscaster speech in the primary signal. We will usually set the initial estimate $\hat{H}_0(k)$ to 0. Start the algorithm. For the first frame, $Rt(k)$ is zero for all k, since $\hat{H}_0(k) = 0$. Then $\mu(Rt(k)) = \mu(0) = 0$, leading

to no update in the filter estimate, leaving $\hat{H}_1(k)$ zero. Surprisingly, the algorithm never updates the filter estimate in any frame.

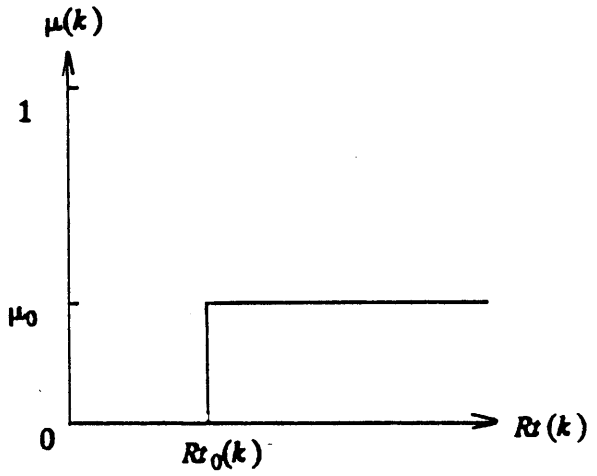


Fig 4.1 An example of a function for $\mu(k)$ vs Rt(k) (a threshold test.)

This example suggests that once the previous estimate $\hat{H}_{i-1}(k)$ becames zero at a frequency k, the estimates for the rest of frames do not adapt to the true filter at all at this frequency, but remain zero. This happens because we arbitrarily chose a function μ which happens to have value 0 at $Rt(k) = 0$. To come up with a better function μ, requires a more careful theoretical derivation. We will not continue the discussion on the selection of the correct function. Instead, we will develop another approach for the non-stationary signal case from a different point of view in the next section.

4.5. Method 2 for non-stationary speech signals

4.5.1. A new system identification model

We consider a new approach which allows input signals $x(n)$, $y(n)$ to be non-stationary signals in estimating the transfer function. It also allows $s(n)$ to be time-varying colored white noise. To illustrate the new approach, consider the linear system shown in the Fig. 4.2.

An observation, a random signal $y(n)$, is assumed to be comprised of two signals; (1) an output of a filter with unknown deterministic filter coefficients $h(n)$, excited by a known deterministic signal $x(n)$; (2) $s(n)$, a zero-mean colored noise with unknown deterministic power spectral density $\sigma^2(\omega)$. In a stochastic processing context, $s(n)$ can be viewed as an output of a 'shaping filter' $\sigma(n)$ excited by a normalized white Gaussian noise $v(n)$, where the magnitude squared of the Fourier transform of $\sigma(n)$, is equal to $\sigma^2(\omega)$. Given signals $x(n)$ and (y), we will estimate the unknown deterministic parameters $h(n)$ and $\sigma^2(\omega)$.

Modeling speech as colored noise with power spectrum is reasonably valid for short periods of time. To deal with the non-stationarity of speech, it seems reasonable that, by sectioning the whole signal into frames with a finite length, modeling a desired speech (newscaster's speech) as a random signal with fixed power spectral density within each frame, but different densities in different frames, is likely to work well.

Again, a technique for estimating the parameters is the Maximum Likelihood (ML) estimation method. ML begins with calculating the log probability density of $y(n)$. The statement is as follows;

Suppose that K sets of infinitely long observations $y_i(n)$, $x_i(n)$ for $i=0,\ldots,K$, $n=-\infty,\ldots,\infty$ are obtained in K independent trials with the same system function $h(n)$ but different $\sigma^2_i(\omega)$ in the Fig 4.2. We will calculate the joint probability density of the K

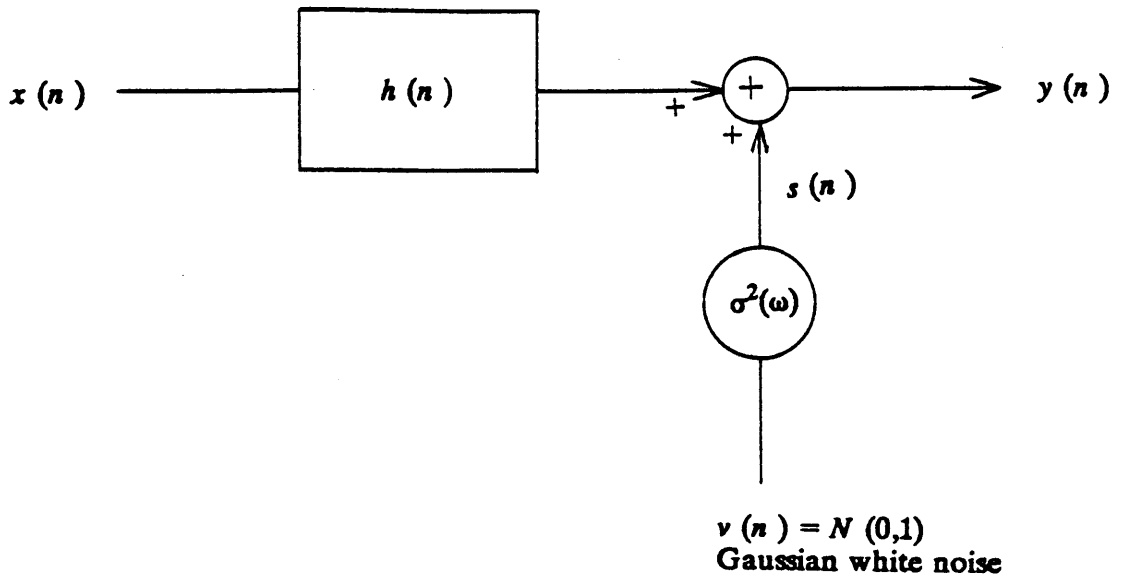


Fig 4.2 The block diagram of a new system identification problem.

independent experiment $p(y_1(n),\ldots,y_K(n) \mid x_1(n),\ldots,x_K(n), h, \sigma^2_1(\omega),\ldots,\sigma^2_K(\omega))$, then will find the values of $h(n)$ and $\sigma^2_1(\omega),\ldots,\sigma^2_K(\omega)$ which maximize that probability.

We divide the derivation into several stages for the sake of clear presentation.

[1] Compute the L point joint average probability density function for the i th experiment $p(\underline{y_i} \mid \underline{x_i}, h, R_{y_i y_i}(0))$

Suppose L (an enormous number) observation points are available for each experiment. (Later, L will go to infinity and the asymptotic behavior of a solution will be obtained.) These observations in the experiment i are denoted as $y_i(n)$ and $x_i(n)$ for $n = 0, 1, \ldots, L-1$. Since $v(n)$ is a zero-mean Gaussian white random process, $y_i(n)$ is a Gaussian colored random process with a mean $h(n) * x_i(n)$, and covariance

$$R_{y_i y_i}(n, n+m) = R_{y_i y_i}(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \sigma^2_i(\omega)\exp(j\,\omega m)\,d\omega \qquad (4.5.1)$$

Then the L point joint average log probability density of $y_i(n)\, n = 0, \ldots, L-1$ is also

Gaussian with ;

$$\text{mean } ; E[\underline{y_i}] = \mathbf{h}\,\underline{x_i} \tag{4.5.2}$$

$$\text{covariance matrix } ; Cov[\underline{y_i}] = \Lambda_{y_i y_i} \tag{4.5.3}$$

where

$$\underline{y_i} = \begin{bmatrix} y_i(0) \\ y_i(1) \\ . \\ . \\ . \\ y_i(L-2) \\ y_i(L-1) \end{bmatrix} \qquad \underline{x_i} = \begin{bmatrix} x_i(0) \\ x_i(1) \\ . \\ . \\ . \\ x_i(L-2) \\ x_i(L-1) \end{bmatrix} \tag{4.5.4}$$

$$\mathbf{h} = \begin{bmatrix} h(0) & 0 & 0 & . & . & . & 0 \\ h(1) & h(0) & 0 & . & . & . & 0 \\ . & h(1) & h(0) & . & . & . & 0 \\ h(M-1) & . & h(1) & . & . & . & 0 \\ 0 & h(M-1) & . & . & . & . & 0 \\ . & 0 & h(M-1) & . & . & . & 0 \\ 0 & 0 & 0 & h(M-1) & . & h(1) & h(0) \end{bmatrix} \tag{4.5.5}$$

$$\Lambda_{y_i y_i} = \begin{bmatrix} R_{y_i y_i}(0) & R_{y_i y_i}(1) & . \; . \; . & R_{y_i y_i}(L-1) \\ R_{y_i y_i}(1) & R_{y_i y_i}(0) & . \; . \; . & R_{y_i y_i}(L-2) \\ . & . & . \; . \; . & . \\ . & . & . \; . \; . & . \\ . & . & . \; . \; . & . \\ R_{y_i y_i}(L-1) & R_{y_i y_i}(L-2) & . \; . \; . & R_{y_i y_i}(0) \end{bmatrix} \tag{4.5.6}$$

Thus it follows that the average log joint probability density for the i th experiment is in the form of;

$$\frac{1}{L}\log p\,(\underline{y_i}\,|\underline{x_i}, \mathbf{h}, \Lambda_{y_i y_i}) = -\frac{1}{2L}\left[(\underline{y_i} - \mathbf{h}\underline{x_i})^T \Lambda_{y_i y_i}^{-1}(\underline{y_i} - \mathbf{h}\underline{x_i}) + \log\left\{(2\pi)^L\,|\Lambda_{y_i y_i}|\right\}\right] \tag{4.5.7}$$

The reason for considering the average of the log probability density, instead of merely the probability density, is that we will shortly want to observe the asymptotic behavior of the probability density for $L \to \infty$.

We present an asymptotic expression of this log probability density when the number of observations becomes large. The following discussion is originally offered by Lim[6] , and discussed by Musicus[7] We describe only the main stream of their discussion. The whole point here is to attempt to transform the expression into the frequency domain, because ML for stationary processes with long observation time can often be simplified dramatically and can be more easily handled.

We add one more assumption. The stability of the stationary process $y_{\underline{i}}$, that is, a covariance $R_{y_i y_i}(m)$ of the process $y_{\underline{i}}$ is

$$\sum_{m=-\infty}^{\infty} |R_{y_i y_i}(m)| < \infty \tag{4.5.8}$$

Also we assume that the inverse covariance function $R^{-1}{}_{y_i y_i}(m)$ is stable.

$$\sum_{m=-\infty}^{\infty} |R^{-1}{}_{y_i y_i}(m)| < \infty \tag{4.5.9}$$

Then for sufficiently large L, the eigenvalues of $\Lambda_{y_i y_i}$ will be approximately ;

$$\lambda^{(i)}{}_l = \sigma^2{}_i(\omega_l) \quad \text{for } l = 0, \ldots, L-1 \tag{4.5.10}$$

and the eigenvectors will be approximately ;

$$u^{(i)}{}_l = \frac{1}{\sqrt{L}} \begin{bmatrix} \exp(j\,0\omega^{(i)}{}_l) \\ \cdot \\ \cdot \\ \exp(j\,(L-1)\omega^{(i)}{}_l) \end{bmatrix} \qquad l = 0, \ldots, L-1 \tag{4.5.11}$$

Moreover, since $\Lambda_{y_i y_i}$ and $\Lambda^{-1}{}_{y_i y_i}$ are decomposed in the form of

$$\Lambda_{y_i y_i} = \sum_{l=1}^{L} \sigma^2{}_i(\omega_l)\, u^{(i)}{}_l \left\{u^{(i)}{}_l\right\}^{*T} \tag{4.5.12}$$

$$\Lambda^{-1}{}_{y_i y_i} = \sum_{l=1}^{L} \frac{1}{\sigma^2{}_i(\omega_l)} u^{(i)}{}_l \left\{u^{(i)}{}_l\right\}^{*T} \tag{4.5.13}$$

where $\left\{u^{(i)}{}_l\right\}^{*T}$ is a complex conjugate transpose vector of $u^{(i)}{}_l$. Substituting $\Lambda_{y_i y_i}$ into the equation (4.5.7)

$$\frac{1}{L}\log p\,(\underline{y_i}\,|\underline{x_i}, \mathbf{h}, \sigma^2{}_i) = -\frac{1}{2}\left[\frac{1}{L}\sum_{l=0}^{L-1}\left\{\frac{(\underline{y_i} - \mathbf{h}\underline{x_i})^T u^{(i)}{}_l \left\{u^{(i)}{}_l\right\}^{*T} (\underline{y_i} - \mathbf{h}\underline{x_i})}{\sigma_i{}^2(\omega_l)} + \log\left\{2\pi\sigma^2{}_i(\omega_l)\right\}\right\}\right] \tag{4.5.14}$$

It is easy to show that ;

$$\underline{y_i}^T u^{(i)}{}_l = \frac{1}{\sqrt{L}} Y_i{}^*(\omega_l) \tag{4.5.15}$$

$$(\mathbf{h}\underline{x_i})^T u^{(i)}{}_l = \frac{1}{\sqrt{L}} H^*(\omega_l) X_i{}^*(\omega_l) \tag{4.5.16}$$

$$\left\{u^{(i)}{}_l\right\}^{*T} \underline{y_i} = \frac{1}{\sqrt{L}} Y_i(\omega_l) \tag{4.5.17}$$

$$\left\{u^{(i)}{}_l\right\}^{*T} (\mathbf{h}\underline{x_i}) = \frac{1}{\sqrt{L}} H(\omega_l) X_i(\omega_l) \tag{4.5.18}$$

then the equation (4.5.14) becomes

$$\frac{1}{L}\log p\,(\underline{y_i}\,|\underline{x_i}, H(\omega), \sigma^2{}_i(\omega))$$

$$\approx -\frac{1}{2}\left[\int_{-\pi}^{\pi}\left\{\frac{\frac{1}{L}\left|Y_i(\omega) - X_i(\omega)H(\omega)\right|^2}{\sigma^2{}_i(\omega)} + \log(2\pi\sigma^2{}_i(\omega))\right\}\frac{d\omega}{2\pi}\right] \tag{4.5.19}$$

This is the L point average probability density of the i th experiment observation.

[2] Next we compute the joint log probability density of the K independent experiments. That is

$$p\,(y_1(n), \ldots, y_K(n)\,|\,x_1(n), \ldots, x_K(n), h, \sigma^2{}_1(\omega), \ldots, \sigma^2{}_K(\omega))$$

$$= \sum_{i=1}^{K} \frac{1}{L} \log p\,(\underline{y}_i \,|\underline{x}_i, H(\omega), \sigma^2{}_i(\omega)) \tag{4.5.20}$$

$$= \sum_{i=1}^{K} -\frac{1}{2} \left[\int_{-\pi}^{\pi} \left\{ \frac{\frac{1}{L} |Y_i(\omega) - X_i(\omega) H(\omega)|^2}{\sigma^2{}_i(\omega)} + \log(2\pi\sigma^2{}_i(\omega)) \right\} \frac{d\omega}{2\pi} \right] \tag{4.5.21}$$

[3] We find $H(\omega)$, $\hat{\sigma}^2{}_i(\omega)$ that maximize the K independent experiment joint log probability density.

$$\hat{H}(\omega), \sigma^2{}_1(\omega), \ldots, \sigma^2{}_K(\omega) \leftarrow\!\!- \max_{H(\omega), \sigma^2{}_i(\omega)} \sum_{i=1}^{K} \frac{1}{L} \log p\,(\underline{y}_i \,|\underline{x}_i, \mathbf{h}, \sigma^2{}_i(\omega)) \tag{4.5.22}$$

$$= \max_{H(\omega), \sigma^2{}_i(\omega)} \sum_{i=1}^{K} -\frac{1}{2} \left[\int_{-\pi}^{\pi} \left\{ \frac{\frac{1}{L} |Y_i(\omega) - X_i(\omega) H(\omega)|^2}{\sigma^2{}_i(\omega)} + \log(2\pi\sigma^2{}_i(\omega)) \right\} \frac{d\omega}{2\pi} \right] \tag{4.5.23}$$

It can be accomplished by taking the derivatives of the $\sum_{i=1}^{K} \frac{1}{L} \log p\,(\underline{y}_i \,|\underline{x}_i, H(\omega), \sigma^2{}_i(\omega))$ with respect to $H(\omega)$, $\sigma^2{}_i(\omega)$, $i = 1, \ldots, K$ respectively. Taking derivative with respect to $\sigma^2{}_i(\omega)$ $i = 1, \ldots, K$,

$$\hat{\sigma}^2{}_i(\omega) = \frac{1}{L} |Y_i(\omega) - X_i(\omega) H(\omega)|^2 \qquad i = 1, \ldots, K \tag{4.5.24}$$

Just substitute this equation into the objective function (4.5.21), again, giving

$$\hat{H}(\omega) \leftarrow\!\!- \max_{H(\omega)} \sum_{i=1}^{K} -\frac{1}{2} \int_{-\pi}^{\pi} \left[1 + \log\left(2\pi \frac{1}{L} |Y_i(\omega) - H(\omega) X_i(\omega)|^2\right) \right] \frac{d\omega}{2\pi} \tag{4.5.25}$$

To minimize with respect to $H(\omega)$, take the derivative to respect with $H(\omega)$, and set it to zero,

$$\sum_{i=1}^{K} \frac{-2X_i{}^*(\omega) Y_i(\omega) + 2X_i{}^*(\omega) X_i(\omega)}{|Y_i(\omega) - H(\omega) X_i(\omega)|^2} = 0 \tag{4.5.26}$$

Substituting the equation (4.5.24) into the above,

$$\sum_{i=1}^{K} \frac{-2X_i{}^*(\omega) Y_i(\omega) + 2X_i{}^*(\omega) X_i(\omega)}{\hat{\sigma}^2{}_i(\omega)} = 0 \tag{4.5.27}$$

Thus,

$$\hat{H}(\omega) = \frac{\sum_{i=1}^{K} X_i^*(\omega) Y_i(\omega) / \hat{\sigma}^2{}_i(\omega)}{\sum_{i=1}^{K} X_i^*(\omega) X_i(\omega) / \hat{\sigma}^2{}_i(\omega)} \tag{4.5.28}$$

Replacing $H(\omega)$ in the equation (4.5.24) by $\hat{H}(\omega)$, the $\hat{\sigma}^2{}_i(\omega)$ $i=1, \ldots, K$ is formed by;

$$\hat{\sigma}^2{}_i(\omega) = \frac{1}{L} |Y_i(\omega) - X_i(\omega) \hat{H}(\omega)|^2 \tag{4.5.29}$$

Therefore, the equations (4.5.28) and (4.5.29) are the solutions for the K independent experiment case. Notice that in the above that the two equations for $\hat{H}(\omega)$ and $\hat{\sigma}^2{}_i(\omega)$, for $i=1, \ldots, K$. depend on each other. We need to arrange the computation to estimate each separately. This arrangement will be made in the form of a recursive method in the next section.

4.5.2. Adaptive scheme

We consider a reasonable adaptive scheme based on the result in the preceding section, to provide an algorithm adaptable to changes of the system function $h(n)$. Again we perform the same experiments as in the preceding section in the system shown in Fig 4.2. We assume that we have performed an enormous number of experiments (the experiment number $k=-\infty, \ldots, i$) and obtained the large set of observations $x_{-\infty}(n)$, $y_{-\infty}(n)$, . . . , , $y_k(n)$, . . . , $x_i(n)$, $y_i(n)$ so far. Here i is the latest experiment number. Moreover, the length of each observation, L, is assumed to be very long.

Given these observations up to experiment number i, we try to find the $\hat{h}_i(n)$ and $\hat{\sigma}^2{}_k(\omega)$, for $k=-\infty, \ldots, i$ that maximize the joint probability density. Note that $\hat{h}_i(n)$ is the estimate of the system $h(n)$ on the basis of observations until the experiment number i. Let us assume that the system function $h(n)$ is changing slowly in time. As for the probability to be maximized, since the system is changing, it seems reasonable to take a

weighted average of the log probability density over the past experiments, rather than a straight arithmetic average. This allows more recent observations to contribute significantly to the error criteria, but suppresses older observations. As one choice, we introduce the exponential average of the averaged log probability density so that the latest experiment is weighted most, the previous one second most, and so on. Using an exponential average parameter α ($0 \le \alpha \le 1$), the objective function ϵ_i is formed by;

$$\epsilon_i(\mathbf{h}, \sigma^2_{-\infty}(\omega), \ldots, \sigma^2_i(\omega))$$

$$= \sum_{k=-\infty}^{i} (1-\alpha)^{i-k} \alpha \log p(\underline{y}_k | \underline{x}_k, \mathbf{h}, \sigma^2_k(\omega))$$

$$= \frac{1}{2} \sum_{k=-\infty}^{i} (1-\alpha)^{i-k} \alpha \int_{-\pi}^{\pi} \left[\frac{\frac{1}{L} |Y_k(\omega) - X_k(\omega) H(\omega)|^2}{\sigma^2_k(\omega)} + \log(2\pi\sigma^2_k(\omega)) \right] \frac{d\omega}{2\pi} \qquad (4.5.30)$$

Then maximizing the averaged log probability density with $H(\omega)$, and $\sigma^2_{-\infty}(\omega), \ldots, \sigma^2_i(\omega)$, leads to

$$\hat{\sigma}^2_k(\omega) = \frac{1}{L} |Y_k(\omega) - X_k(\omega) \hat{H}(\omega)|^2 \qquad (4.5.31)$$

for $k = -\infty, \ldots, i$.

$$\hat{H}_i(\omega) = \frac{\sum_{k=-\infty}^{i} (1-\alpha)^{i-k} \alpha \left\{ \frac{1}{L} X^*_k(\omega) Y_k(\omega) \right\} / \hat{\sigma}^2_k(\omega)}{\sum_{k=-\infty}^{i} (1-\alpha)^{i-k} \alpha \left\{ \frac{1}{L} X^*_k(\omega) X_k(\omega) \right\} / \hat{\sigma}^2_k(\omega)} \qquad (4.5.32)$$

Under a certain assumption, we can introduce a more efficient way to calculate $\hat{H}_i(\omega)$ and $\sigma^2_i(\omega)$, given the estimates from the previous experiment and new observations from the experiment i. It allows one to obtain the current estimates without recomputing $\hat{\sigma}^2_k(\omega)$, for $k = -\infty, \ldots, i-1$. The assumption is that both the filter estimate $\hat{H}_{i-1}(\omega)$ and power spectrum estimates $\hat{\sigma}^2_k(\omega)$, $k = -\infty, \ldots, i-1$ are so correctly obtained in the previous frame i-1 that the these power spectra $\hat{\sigma}^2_i(\omega)$ for $k \le i-1$ need not be reestimated in

the current frame operation, then only the current frame power spectrum $\hat{\sigma}^2_i(\omega)$ should be computed from the current frame data.

The efficient recursive method is described as follows: As for $\hat{\sigma}^2_i(\omega)$, $\hat{H}(\omega)$ in the equation (4.5.31) is replaced with the estimate in the previous experiment $\hat{H}_{i-1}(\omega)$, then the $\hat{\sigma}^2_i(\omega)$ is represented as

$$\hat{\sigma}^2_i(\omega) = \frac{1}{L}\left|Y_i(\omega) - X_i(\omega)\hat{H}_{i-1}(\omega)\right|^2 \tag{4.5.33}$$

Rewriting the equation (4.5.32), and substituting the above-obtained estimate $\hat{\sigma}^2_i(\omega)$, the current estimate $\hat{H}_i(\omega)$ is

$$\hat{H}_i(\omega) = \frac{(1-\alpha)N_{i-1}(\omega) + \alpha\left\{\frac{1}{L}X^*_i(\omega)Y_i(\omega)/\hat{\sigma}^2_i(\omega)\right\}}{(1-\alpha)D_{i-1}(\omega) + \alpha\left\{\frac{1}{L}X^*_i(\omega)X_i(\omega)/\hat{\sigma}^2_i(\omega)\right\}} \tag{4.5.34}$$

where $N_{i-1}(\omega), D_{i-1}(\omega)$ are the numerator, and denominator of $\hat{H}_{i-1}(\omega)$, respectively. That is

$$\hat{H}_{i-1}(\omega) = \frac{N_{i-1}(\omega)}{D_{i-1}(\omega)} \tag{4.5.35}$$

$$N_{i-1}(\omega) = \sum_{k=-\infty}^{i-1}(1-\alpha)^{i-1-k}\alpha\left\{\frac{1}{L}X^*_k(\omega)Y_k(\omega)\right\}/\hat{\sigma}^2_k(\omega) \tag{4.5.36}$$

$$D_{i-1}(\omega) = \sum_{k=-\infty}^{i-1}(1-\alpha)^{i-1-k}\alpha\left\{\frac{1}{L}X^*_k(\omega)X_k(\omega)\right\}/\hat{\sigma}^2_k(\omega) \tag{4.5.37}$$

Therefore the recursive algorithm to cancel the interfering speech is summarized as follows.

Given the numerator and the denominator of the filter estimate in the experiment i-1 $N_{i-1}(\omega)$, $D_{i-1}(\omega)$, compute the estimate of the power spectral density in the experiment i $\hat{\sigma}^2_i(\omega)$ by

$$\hat{\sigma}^2_i(\omega) = \frac{1}{L}\left|Y_i(\omega) - X_i(\omega)\hat{H}_{i-1}(\omega)\right|^2 \tag{4.5.38}$$

where

$$\hat{H}_{i-1}(\omega) = \frac{N_{i-1}(\omega)}{D_{i-1}(\omega)} \tag{4.5.39}$$

Then, calculate the auto spectrum and the cross spectrum,

$$P_{x_i y_i}(\omega) = \frac{1}{L} X^*_i(\omega) Y_i(\omega) / \hat{\sigma}^2_i(\omega) \tag{4.5.40}$$

$$P_{x_i x_i}(\omega) = \frac{1}{L} X^*_i(\omega) X_i(\omega) / \hat{\sigma}^2_i(\omega) \tag{4.5.41}$$

Then, compute the current numerator and denominator in the form of

$$N_i(\omega) = (1-\alpha) N_{i-1}(\omega) + \alpha \left\{ P_{x_i y_i}(\omega) / \hat{\sigma}^2_i(\omega) \right\} \tag{4.5.42}$$

$$D_i(\omega) = (1-\alpha) D_{i-1}(\omega) + \alpha \left\{ P_{x_i x_i}(\omega) / \hat{\sigma}^2_i(\omega) \right\} \tag{4.5.43}$$

Then the current filter estimate is

$$\begin{aligned} \hat{H}_i(\omega) &= \frac{N_i(\omega)}{D_i(\omega)} \\ &= \frac{(1-\alpha) N_{i-1}(\omega) + \alpha \left\{ P_{x_i y_i}(\omega) / \hat{\sigma}^2_i(\omega) \right\}}{(1-\alpha) D_{i-1}(\omega) + \alpha \left\{ P_{x_i x_i}(\omega) / \hat{\sigma}^2_i(\omega) \right\}} \end{aligned} \tag{4.5.44}$$

The desired speech in the current experiment i can be obtained by the filtering process.

$$\hat{s}_i(n) = y_i(n) - x_i(n) * \hat{h}_i(n) \tag{4.5.45}$$

Move to the experiment i+1.

The filtering process can be computed by any filtering method. We use the overlap-save method here, because it is easy to implement.(Note that this method assumes that the length of the filter M is finite.) We will present the overlap-save method in a later section.

4.5.3. Property of convergence and adaptation

In this section, we will first examine the nature of the exponential average and the role of its parameter α, and then the convergence behavior of the estimate $\hat{H}_i(\omega)$ is discussed.

If the L point averaged log probability density $p\,(\underline{y_k}\,|\underline{x_k}, h, \hat{\sigma}^2_k(\omega))$ in the experiment k is viewed as a sequence in time k, the exponential average of the joint log probability density in the equation (4.5.30) over the past frames is equivalent to filtering the sequence of densities through a single pole low pass filter.

$$\epsilon_i = \alpha \epsilon_{i-1} + \log p\,(y_i(n)\,|x_i(n), h, \sigma^2_i(\omega)) \tag{4.5.46}$$

This filtering tends to suppress rapid changes in the log probability density sequence. The value of $\frac{1}{\alpha}$ controls the number of frames effectively contributing to the current averaged log probability ϵ_i, It equivalently determines the cut-off frequency in the low-pass operation. Higher order averaging filters could also be used to achieve sharper cut-offs and less "jitter" in the error criterion and the resulting estimates, but these would require more computational effort.

Next, we show that on the average the recursive method will improve the estimate $\hat{H}_i(\omega)$ on each iteration. As $\hat{H}_i(\omega)$ improves, it is easy to show that the colored noise estimate $\hat{\sigma}^2_i(\omega)$ also becomes closer to the true power spectrum of the desired speech in experiment i. Now we show that on average the estimate in the experiment i, $H_i(\omega)$ is closer to the real filter than the i-1 th estimate $\hat{H}_{i-1}(\omega)$.

By assumption, $y_i(n)$ is generated in the form of

$$y_i(n) = s_i(n) + x_i(n) * h(n) \tag{4.5.47}$$

Since observations $x_i(n)$ $y_i(n)$ are approximately infinitely long, then the Fourier transform of the equation (4.5.47)

$$Y_i(\omega) = S_i(\omega) + X_i(\omega) * H(\omega) \tag{4.5.48}$$

can hold almost exactly. Substituting $Y_i(\omega)$ into (4.5.44) the current estimate $\hat{H}_i(\omega)$ can be rewritten in the form of

$$\hat{H}_i(\omega) = \frac{(1-\alpha)N_{i-1}(\omega)+\alpha\left\{P_{x_i y_i}(\omega)/\hat{\sigma}^2{}_i(\omega)\right\}}{(1-\alpha)D_{i-1}(\omega)+\alpha\left\{P_{x_i x_i}(\omega)/\hat{\sigma}^2{}_i(\omega)\right\}}$$

$$= \hat{H}_{i-1}(\omega) + \frac{\alpha P_{x_i x_i}(\omega)/\hat{\sigma}^2{}_i(\omega)}{(1-\alpha)D_{i-1}(\omega) + \alpha P_{x_i x_i}(\omega)/\hat{\sigma}^2{}_i(\omega)}(H(\omega) - \hat{H}_{i-1}(\omega))$$

$$+ \frac{\alpha P_{x_i s_i}(\omega)/\hat{\sigma}^2{}_i(\omega)}{(1-\alpha)D_{i-1}(\omega)+\alpha P_{x_i x_i}(\omega)/\hat{\sigma}^2{}_i(\omega)} \tag{4.5.49}$$

We define the coefficient of $H(\omega) - \hat{H}_{i-1}(\omega)$ as $q_i(\omega)$,

$$q_i(\omega) = \frac{\alpha P_{x_i x_i}(\omega)/\hat{\sigma}^2{}_i(\omega)}{(1-\alpha)D_{i-1}(\omega) + \alpha P_{x_i x_i}(\omega)/\hat{\sigma}^2{}_i(\omega)} \tag{4.5.50}$$

$$= \frac{\alpha P_{x_i x_i}(\omega)/\hat{\sigma}^2{}_i(\omega)}{D_i(\omega)} \tag{4.5.51}$$

Then

$$\hat{H}_i(\omega) = \hat{H}_{i-1}(\omega)+q_i(\omega)(H(\omega) - \hat{H}_{i-1}(\omega)) + q_i(\omega)\frac{P_{x_i s_i}(\omega)}{P_{x_i x_i}(\omega)} \tag{4.5.52}$$

Consider the parameter $q_i(\omega)$ first. Because both the numerator and the denominator of the $q_i(\omega)$ are positive and the numerator is always smaller than the denominator, $q_i(\omega)$ takes on real positive values between 0 and 1. Therefore the second term in the equation (4.5.52) is a legitimate correction term which updates the estimate from the previous stage filter estimate in the direction of the true filter. Thus $q_i(\omega)$ plays the role of an interpolation parameter. Note that if the system is not changing too quickly with time, then we would expect, on average, that

$$D_{i-1}(\omega) \approx \frac{P_{x_i x_i}(\omega)}{\hat{\sigma}^2{}_i(\omega)} \tag{4.5.53}$$

and thus $q_i(\omega)$ is almost α. Thus the speed of the convergence is approximately equal to the value α.

The third term in the equation, however, is a disturbance to the current estimate. Because $x_i(n)$ and $s_i(n)$ are uncorrelated, we would expect on average

$$E\left[X^*_i(\omega)S_i(\omega)\right]=0 \tag{4.5.54}$$

However, when estimating $P_{x_i s_i}(\omega)$, using only one pair of sample functions $x_i(n)$ and $s_i(n)$, we will generally find that $P_{x_i s_i}(\omega)$ is not zero. In a very crude sense, the variance of the third term in equation (4.5.52) will be approximately proportional to the power in $S_i(\omega)$ and inversely proportional to the power in $X_i(\omega)$. Thus random fluctuations in the third term may cause the filter estimate to fluctuate. This fluctuation does not go to zero, since the range of the average is only $\frac{1}{\alpha}$ experiments at any experiment stage i. Note, however, that if the desired speech is not present for the entire period of time, then the third term vanishes perfectly. Consequently, the update will always improve the filter estimate.

Now we study the behavior of updating the estimate in detail. First of all, one easily notices that the updating of the estimate is performed independently at each frequency ω. For simplicity, suppose that the previous filter estimate $\hat{H}_{i-1}(\omega)$ is close to the true filter so that $\hat{\sigma}^2_i(\omega)$ also represents the true power spectral density of the current desired speech $s(n)$ correctly. Then equation (4.5.52) shows that at frequencies where either the current desired speech has large energy or the current reference speech has small energy, the value of $q_i(\omega)$ becomes small, preventing the current new observation information from contributing significantly to the filter estimate. On the other hand, at frequencies where the current desired speech is small and the current reference speech is substantial, the $q_i(\omega)$ may be close to α and allows the current information to affect the current filter estimate

considerably. This consideration reflects the intuition that the presence of desired speech perturbs the filter estimate. For example, if the desired speech is not present at all at a certain frequency, then the speed of improvement is the fastest and would be somewhat close to α. As a result, the proposed algorithm tries to selectively update the filter estimate at frequencies where the reference is louder than the the desired speech in the primary speech $y_i(n)$. If we define the signal-to-noise ratio as the ratio of the desired speech energy to the correlated reference speech in the primary, then having the low signal-to-noise ratio in the primary speech allows for better filter estimation.

In the case that $\hat{H}_{i-1}(\omega)$ is a poor estimate, even though desired speech is not present at all for the whole time, $\hat{\sigma}^2{}_i(\omega)$ will be non-zero due to the mismatch between the previous filter estimate and the true filter. This effect tends to underestimate the value of $q_i(\omega)$, and it leads to slow convergence. For example, if we start the algorithm with initial states of the filter estimate $\hat{H}_0(\omega) = 0$, $\hat{\sigma}^2{}_0(\omega) = 0$, then it does not lock on the true filter at the first frame, but instead it only gradually approaches the true filter.

4.5.4. Frame concept (Practical estimation method)

In the last few sections, we have assumed that multiple experiments with nearly infinitely long observation data in each are available. However, in practice, only finite length sections of one set of data sequences $x(n), y(n)$ are available in the broadcast room problem. In this section, we discuss the similarities and differences between the infinite data frame case and finite data frame case. The similarity is that one can use the result for the filter estimate mentioned in the previous section in almost the same fashion. A distinct advantage for the finite length frame data case is that the computation in the frequency domain can be computed using a DFT.

The recursive method can be re-stated in the following fashion by using the DFT. First break the input data stream $x(n), y(n)$ into multiple overlapping frames. As the i th frame of data, $x_i(n), y_i(n)$ becomes available, a power spectrum $\frac{1}{L}X^*_i(k)X_i(k)$ and a cross spectrum $\frac{1}{L}X^*_i(k)Y_i(k)$, are computed using DFT. After $\hat{\sigma}^2_i(k)$ is calculated by

$$\hat{\sigma}^2_i(k) = \frac{1}{L}\left|Y_i(k) - X_i(k)\hat{H}_{i-1}(k)\right|^2 \tag{4.5.55}$$

the numerator and denominator of the filter estimate are updated. The filter estimate $\hat{H}_{i-1}(k)$ is obtained by

$$\hat{H}_i(k) = \frac{N_i(k)}{D_i(k)} \tag{4.5.56}$$

$$= \frac{(1-\alpha)N_{i-1}(k) + \alpha\left\{\frac{1}{L}X^*_i(k)Y_i(k)/\hat{\sigma}^2_i(k)\right\}}{(1-\alpha)D_{i-1}(k) + \alpha\left\{\frac{1}{L}X^*_i(k)X_i(k)/\hat{\sigma}^2_i(k)\right\}} \tag{4.5.57}$$

The estimate of the desired speech for the current frame is then obtained by using an overlap-save method of filtering.

However, two crucial differences between the infinite frame case and the finite frame case are end effects due to the finite length of the data for each frame, and the bias effect of using a sectioning window. The estimated filter in the finite observation length case is unavoidably contaminated by time aliasing. This time aliasing is caused by the end effect of the frame data, when multiplication and division of DFT are performed. In order to consider the end effect, let us consider the following example with a rectangular windowed frame.

Assume that sequences $x(n)$ and $y(n)$ are related by convolution

$$y(n) = h(n) * x(n) \tag{4.5.58}$$

where $h(n)$ is an M point long filter. In the frequency domain, if the Fourier transform

exists for $x(n), y(n), h(n)$, then

$$Y(\omega) = H(\omega)X(\omega) \tag{4.5.59}$$

Now assume that we section sequences into multiple frames with an L point rectangular window $L \geq M$, giving the L point sequences in frame i $x_{i(L)}(n)$, $y_{i(L)}(n)$ $n = 0, \ldots, L-1$ $L > M$, such as

$$x_i(n) = x(iF + n) \tag{4.5.60}$$

$$y_i(n) = x(iF + n) \tag{4.5.61}$$

where F is the frame shift. Now the convolution relation does not hold strictly for these L point sections, $x_{i(L)}(n)$, $y_{i(L)}(n)$. This error occurs at the endpoints of the rectangular windowed sections. Denote the error in the frame i as $\Delta_{i(L)}(n)$;

$$\Delta_{i(L)}(n) \equiv y_{i(L)}(n) - x_{i(L)}(n) * h(n) \neq 0 \tag{4.5.62}$$

We examine the filter estimate problem as follows: When one has L point input and output observations of the system function $h(n)$, one wishes to estimate $h(n)$ by

$$\hat{H}(k) = \frac{\sum_i \frac{1}{L} X_{i(L)}^*(k) Y_{i(L)}(k)}{\sum_i \frac{1}{L} X_{i(L)}^*(k) X_{i(L)}(k)} \tag{4.5.63}$$

where

$$X_{i(L)}(k) = \sum_{n=0}^{L-1} x_i(n) \exp(-j\,\omega n) \tag{4.5.64}$$

$$Y_{i(L)}(k) = \sum_{n=0}^{L-1} y_i(n) \exp(-j\,\omega n) \tag{4.5.65}$$

We will evaluate the effect of the error $\Delta_{i(L)}(n)$ on the filter estimate $\hat{H}(k)$. If we denote the L point DFT of true filter as

$$H_{(L)}(k) = \sum_{n=0}^{L-1} h(n) \exp(-j\,\omega n) \tag{4.5.66}$$

the L point DFT of $\Delta_{i(L)}(n)$ is

$$\Delta_{i(L)}(k) \equiv Y_{i(L)}(k) - H_{(L)}(k)X_{i(L)}(k)$$
$$= \sum_{n=0}^{M-2}\left[y_i(n) - \sum_{m=0}^{M-1} h(m)x_i(n-m \text{ modulo } L)\right]\exp(-j\,2\pi k\, n/L) \qquad (4.5.67)$$

$$\Delta_{i(L)}(k) = \sum_{n=0}^{M-2}\sum_{m=n+1}^{M-1} h(m)(x_i(n-m) - x_i(L+n-m))\exp(-j\,2\pi k\, n/L) \qquad (4.5.68)$$

Note that $\Delta_{i(L)}(k)$ is an aliasing error, mixing signal points located L points apart. as L $\to\infty$, $\Delta_{i(L)}(k)$ stays approximately constant in the expected energy, because the number of terms in $\Delta_{i(L)}(k)$ is independent of L. Thus, multiply the equation (4.5.67) by $X_{i(L)}^{*}(k)$,

$$\frac{1}{L}X_{i(L)}^{*}(k)\Delta_{i(L)}(k) = \frac{1}{L}X_{i(L)}^{*}(k)Y_{i(L)}(k) - H_{(L)}(k)\frac{1}{L}|X_{i(L)}(k)|^2 \qquad (4.5.69)$$

Rewriting this equation,

$$\hat{H}(k) = H_{(L)}(k) + \frac{\sum_i \frac{1}{L}X^{*}_{i(L)}(k)\Delta_{i(L)}(k)}{\sum_i \frac{1}{L}X^{*}_{i(L)}(k)X_{i(L)}(k)} \qquad (4.5.70)$$

Thus the estimated filter $\hat{H}_{(k)}$ deviates from the true filter $H_{(L)}(k)$ because of the end effect errors $\Delta_{i(L)}(k)$. Note that the second term of the above equation, is ;

$$\left|\frac{\sum_i \frac{1}{L}X^{*}_{i(L)}(k)\Delta_{i(L)}(k)}{\sum_i \frac{1}{L}X^{*}_{i(L)}(k)X_{i(L)}(k)}\right|$$

$$= \left|\frac{\sum_i \frac{1}{L}X^{*}_{i(L)}(k)\Delta_{i(L)}(k)}{\left[\frac{1}{L}\sum_i |X_{i(L)}(k)|^2\right]^{\frac{1}{2}}\left[\frac{1}{L}\sum_i |\Delta_{i(L)}(k)|^2\right]^{\frac{1}{2}}}\right| \left|\frac{\frac{1}{L}\sum_i |\Delta_{i(L)}(k)|^2}{\frac{1}{L}\sum_i |X_{i(L)}(k)|^2}\right|^{\frac{1}{2}} \qquad (4.5.71)$$

$$\leq \left|\frac{\frac{1}{L}\sum_i |\Delta_{i(L)}(k)|^2}{\frac{1}{L}\sum_i |X_{i(L)}(k)|^2}\right|^{\frac{1}{2}} \to 0 \quad \textit{as frame length} \to\infty. \qquad (4.5.72)$$

it leads to

$$H_{(L)}(k) = \frac{\sum_i \frac{1}{L} X_{i(L)}(k) Y_{i(L)}(k)}{\sum_i \frac{1}{L} |X_{i(L)}(k)|^2} \quad \text{as} \quad L \to \infty \tag{4.5.73}$$

The error term has magnitude roughly proportional to $\frac{M}{L}$, since $\Delta_{i(L)}(k)$ has M terms, whereas $X_{i(L)}(k)$ has L terms. This argument, of course, is crude but it does illustrate the magnitude of the error due to the use of finite frame lengths. In particular, the smaller the ratio $\frac{M}{L}$, the better the filter estimate. It is desirable to choose L as at least about 10 times bigger than M to keep reasonable accuracy. However using such long frames causes a heavy computational burden, which makes it harder to implement the algorithm in real time. Furthermore, long frame lengths lead to long delays in a real-time system and slow down the adaptation rate. When short frames are used, we will expect that the estimated filter $\hat{H}(n)$ will contain circular aliasing, thus leading to reverberation artifacts in the speech.

If a Hanning window $h_D(n)$ is applied to the data instead of a rectangular window;

$$x_i(n) = h_D(n) x(iF + n) \tag{4.5.74}$$

$$y_i(n) = h_D(n) y(iF + n) \tag{4.5.75}$$

where is F is the shift of frame. Then the error;

$$\Delta_i(n) = y_i(n) - h(n) * x_i(n) \tag{4.5.76}$$

is more uniformly distributed over the interval $0 \le n < L$, rather than being bunched at the left edge. In particular, it is easy to show that aliasing error at the left edge has been tapered off by using a tapered windowed $h_D(n)$, and has effectively been traded for a smeared error across the entire data interval. This type of error is usually less objectionable because it does not cause such strong reverberation effects.

Windowing data to form finite length sections also causes a bias in the filter estimate.

The effect of sidelobe leakage is one example of this bias effect. If a rectangular window is used, this leakage is noticeably large, degrading the estimate $\hat{H}(k)$ when the signals are strongly harmonic. When a Hanning window is used, the effect of the leakage is less noticeable.

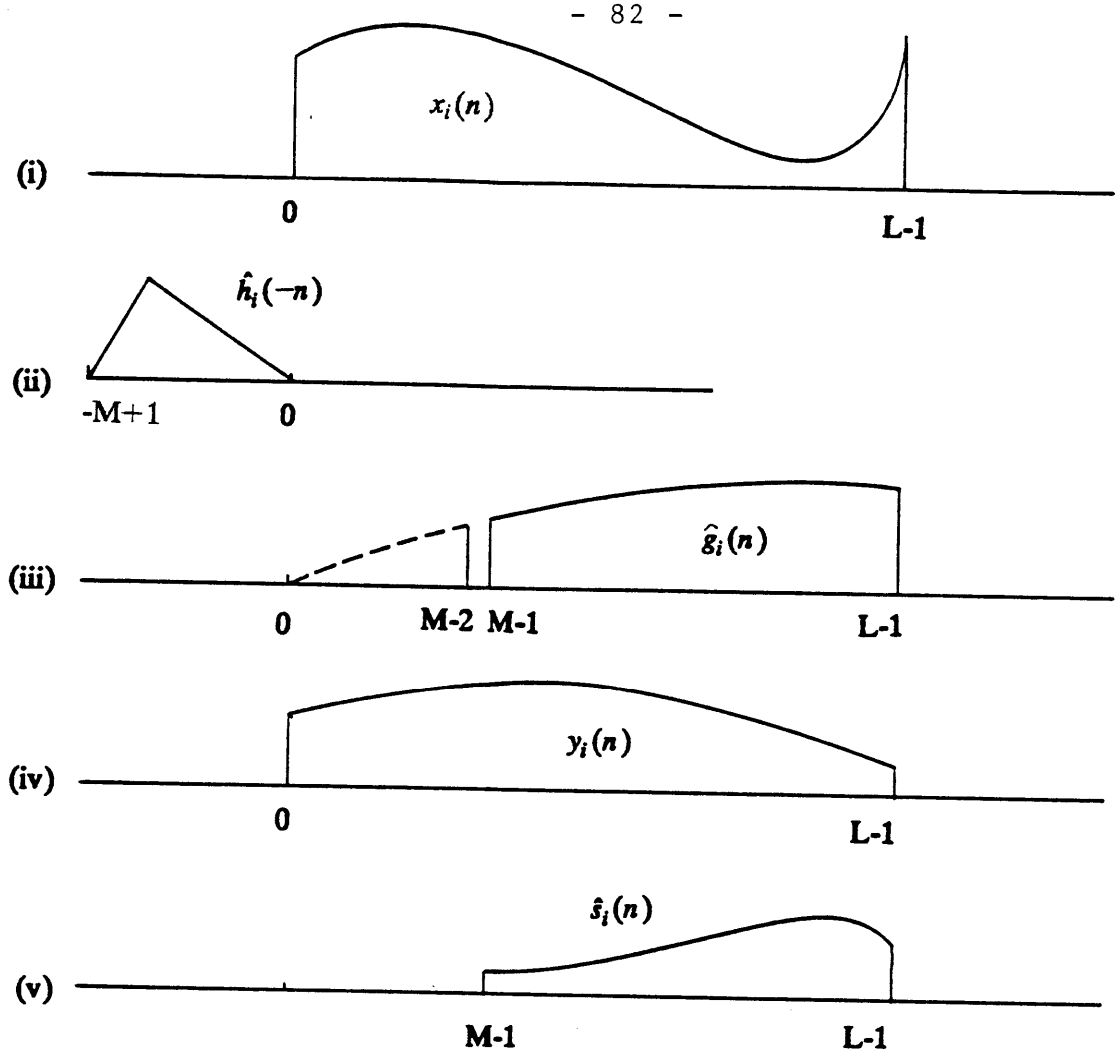
Next we consider the length of the window shift, F. It controls the rate of how often the filter must be reestimated. To avoid a sudden change of the filter estimate from frame to frame, a reasonable time shift between adjacent frames is necessary. The reason for preventing a sudden change is to make the estimated desired speech from different frames change smoothly across frame boundary. The rate of change in the room response is a main factor in determining the value of the shift. The adaptation rate is also controlled by the exponential parameter α, since $\frac{1}{\alpha}$ controls the number of frames effectively contributing to the current estimate. If α is close to 0, then data from many old frames contribute to the current filter estimate, leading to slow adaptation rate. This suggests then that the values of parameter α and the window shift F should be set appropriately at the same time to provide the desired adaptation rate.

4.5.5. Overlap-save method for the filtering process

When the current filter estimate is obtained, desired speech of the current frame $\hat{s}_i(n)$ will be obtained in a filtering process in the form of ;

$$\hat{s}_i(n) = y_i(n) - x_i(n) * \hat{h}_i(n) \qquad (4.5.77)$$

Although this filtering can be achieved by a variety of methods, we will use the overlap-save method because it is easily implemented for long filters. Assume that the filter $\hat{h}_i(n)$ is finite length with M points ($L \geq M$), and L points input data $x_i(n)$, $y_i(n)$ are available for each frame. First we think about the length of the FFT. Let the FFT size be N, then N is required to be bigger than L. Here we choose N = L. The overlap-save method



FFT size N $= L$

Fig 4.3 The overlap-save method

usually works as follows. (Fig 4.3)

Suppose that for one frame of processing, L points of input data $x_i(n)$ are read into an N point FFT buffer. Pad with zeros to the end of the buffer and take the N point FFT to obtain $X_i(k)$. Likewise feed M points of the current filter $\hat{h}_i(n)$ into the N point FFT buffer, pad with zeros, and then take the N point FFT to calculate $\hat{H}_i(k)$. Multiply $X_i(k)$ by $\hat{H}_i(k)$, to obtain the filtered output $\hat{G}_i(k)$,

$$\hat{G}_i(k) = X_i(k)\,\hat{H}_i(k) \tag{4.5.78}$$

Take the N point inverse FFT of $\hat{G}_i(k)$. The resultant sequence $\hat{g}_i(n)$ is N=L points long, but the first M-1 points are time aliased. The last N-M+1=L-M+1 points are valid data for the filtering. Therefore these correct N-M+1=L-M+1 points are subtracted from $y_i(n)$ to form the desired signal estimate $\hat{s}_i(n)$. For the next frame process, the window will be shifted by N-M+1=L-M+1 points, so that M-1 points are overlapped by adjacent windows. Thus, despite L points of the input data $x_i(n)$ and $y_i(n)$, the next filter process ought to read other L point input data portions, by shifting a window by N-M+1 points.

4.5.6. Improved recursive algorithm

This section deals with a estimation method which suffers from less time-aliasing than the original method in section 4.5.4.. To see the problem with our original algorithm, suppose that the filter estimate $\hat{H}_i(\omega)$ is exactly correct;

$$\hat{H}_{i-1}(k) = H(k) \tag{4.5.79}$$

Moreover suppose that no desired speech is present during the entire i th frame. We would therefore expect that our algorithm should not modify the filter estimate from the previous frame, since the old frame estimate can cancel the primary speech perfectly.

Unfortunately, the original algorithm proposed in section 4.5.4. will typically drive the filter estimate away from the true filter value, even in this ideal situation. First we analyize the reason for its failure, then propose alternative algorithm which improves drastically.

In the original algorithm, by assumption

$$\hat{H}_{i-1}(k) = H(k) \tag{4.5.80}$$

The power spectral density estimate $\hat{\sigma}^2{}_i(k)$ is calculated by the formula

$$\tilde{S}_i(k) = Y_i(k) - X_i(k)\hat{H}(k) \tag{4.5.81}$$

$$\hat{\sigma}^2{}_i(k) = \tilde{S}_i{}^*(k)\tilde{S}_i(k) \tag{4.5.82}$$

However, due to windowing effects,

$$Y_i(k) - X_i(k)H(k) \neq 0 \tag{4.5.83}$$

Thus equation (4.5.81) does not properly estimate the speech as $\tilde{S}_i(k) = 0$, but instead sets $\tilde{S}_i(k)$ to some non-zero value.

$$\tilde{S}_i(k) = Y_i(k) - X_i(k)H(k) \neq 0 \tag{4.5.84}$$

Rewriting this,

$$Y_i(k) = \tilde{S}_i(k) + X_i(k)H(k) \tag{4.5.85}$$

This causes the estimate $\hat{\sigma}^2{}_i(k)$ to be non zero. Substituting equation (4.5.85) into the formula for $\hat{H}_i(k)$ in equation (4.5.57) we obtain

$$\hat{H}_i(k) = \frac{(1-\alpha)H(k)D_{i-1}(k) + \alpha\left\{\frac{1}{L}X^*{}_i(k)\tilde{S}_i(k) + H(k)\frac{1}{L}X^*{}_i(k)X_i(k)\right\}/\hat{\sigma}^2{}_i(k)}{(1-\alpha)D_{i-1} + \alpha\frac{1}{L}X^*{}_i(k)X_i(k)/\hat{\sigma}^2{}_i(k)} \tag{4.5.86}$$

$$= H(k) + \frac{\alpha\frac{1}{L}X^*{}_i(k)\tilde{S}_i(k)/\hat{\sigma}^2{}_i(k)}{(1-\alpha)D_{i-1} + \alpha\frac{1}{L}X^*{}_i(k)X_i(k)/\hat{\sigma}^2{}_i(k)} \tag{4.5.87}$$

Because the second term does not vanish, the current estimate is driven away from the true filter value. The reason why the second term remains non-zero is that the windowing

effect caused by sectioning the input data $x_i(n)$ and $y_i(n)$, causes time-aliasing in $\tilde{S}_i(k)$ in equation (4.5.84).

To avoid this time-aliasing, we introduce an alternative method. Fig 4.4 is illustrative for this purpose. The old method uses $x_i(n)$ and $y_i(n)$ as input data for the filter estimation process. Since $x_i(n)$ and $y_i(n)$ do not strictly satisfy the convolution relationship, then the desired speech pre-estimate is contaminated by time aliasing. Instead, the new method first filters the reference speech $x_i(n)$ with the previous frame filter estimate $\hat{h}_{i-1}(n)$ by the overlap-save method and selects a portion free of time-aliasing as $\tilde{s}_i(n)$. The filter estimation process for frame i then uses $x_i(n)$ and $\tilde{s}_i(n)$ as inputs.

Now the detailed procedure in the new algorithm is presented.

[1] **The filter process for the i th frame pre-estimate of the desired speech by the overlap-save method.** (See Fig 4.5) The purpose is to obtain L points of the pre-estimate $\tilde{s}_i(n)$, $n=0,\ldots,L-1$ free of time-aliasing filtered with the i-1 th estimate $\hat{h}_{i-1}(n)$ by the overlap-save method. Note that we need L points of $\tilde{s}_i(n)$, since these points are used for the current filter estimate later. Assume that M points of the i-1 th (the previous) frame filter estimate $\hat{h}_{i-1}(n)$ are available and the size of the FFT is N ($N \geq L+2M-2$). The last M-1 points of the i-1 th frame plus L points of the i th (the current) frame of the input data $x(n)$ (the combined reference $\tilde{x}_i(n)$) should be available in order to compute L points of $\tilde{s}_i(n)$ free of time aliasing. L points of the i th frame of the primary $y(n)$ will also be needed. As a result, both the first and last M-1 points of the circular convolution are time aliased and the middle L point portion, $\tilde{s}_i(n)$ out of L+2M-2 points is valid.

[2] **Compute** $\hat{\sigma}^2{}_i(k)$ Denote the N point FFT of the correctly filtered L point pre-estimate $\tilde{s}_i(n)$ as $\tilde{S}_i(k)$. Compute the power spectral density estimate $\hat{\sigma}^2{}_i(k)$ by

$$\hat{\sigma}^2_i(k) = \bar{S}^*_i(k)\bar{S}_i(k) \tag{4.5.88}$$

[3] **Update $\hat{H}_i(k)$ and its numerator and denominator.** Substituting the obtained $\bar{S}_i(k)$ and $\hat{\sigma}^2_i(k)$ into equation (4.5.86), one obtains the current filter estimate

$$\hat{H}_i(k) = \frac{(1-\alpha)N_{i-1}(k) + \alpha\left\{\frac{1}{L}X^*_i(k)\bar{S}_i(k) + \hat{H}_{i-1}(k)\frac{1}{L}X^*_i(k)X_i(k)\right\}/\hat{\sigma}^2_i(k)}{(1-\alpha)D_{i-1} + \alpha\frac{1}{L}X^*_i(k)X_i(k)/\hat{\sigma}^2_i(k)} \tag{4.5.89}$$

The numerator $N_i(k)$ and the denominator $D_i(k)$ are presented as

$$N_i(k) = (1-\alpha)N_{i-1}(k) + \alpha\left\{\frac{1}{L}X^*_i(k)\bar{S}_i(k) + \hat{H}_{i-1}(k)\frac{1}{L}X^*_i(k)X_i(k)\right\}/\hat{\sigma}^2_i(k) \tag{4.5.90}$$

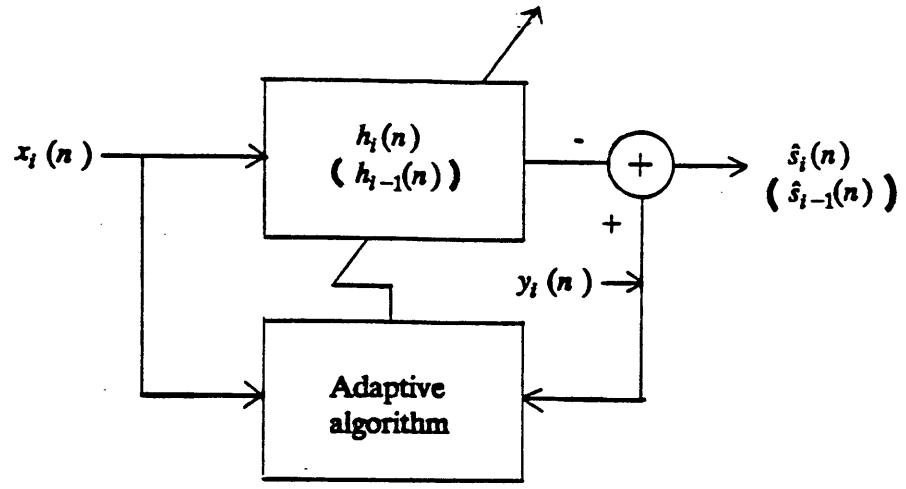
$$D_i(k) = (1-\alpha)D_{i-1} + \alpha\frac{1}{L}X^*_i(k)X_i(k)/\hat{\sigma}^2_i(k) \tag{4.5.91}$$

[4] **The filter process for the i th frame final desired speech estimate $\hat{\sigma}^2_i(k)$ by the overlap-save method.** After the current filter estimate $\hat{h}_i(n)$ is computed in equation (4.5.89), our goal is to calculate F points (F ; the shift of the window $F \leq L-M+1$) of non-circular aliased output $\hat{s}_i(n)$. (See Fig 4.6) Unlike the case for estimating $\bar{s}_i(n)$, only the current frame of L points of $x_i(n)$ and $y_i(n)$ are used to estimate $\hat{s}_i(n)$. After taking the N ($N \geq L+M-1$) point DFT's of $x_i(n)$ $y_i(n)$ and $\hat{h}_i(n)$, the filtered reference signal $g_i(n)$ is computed by

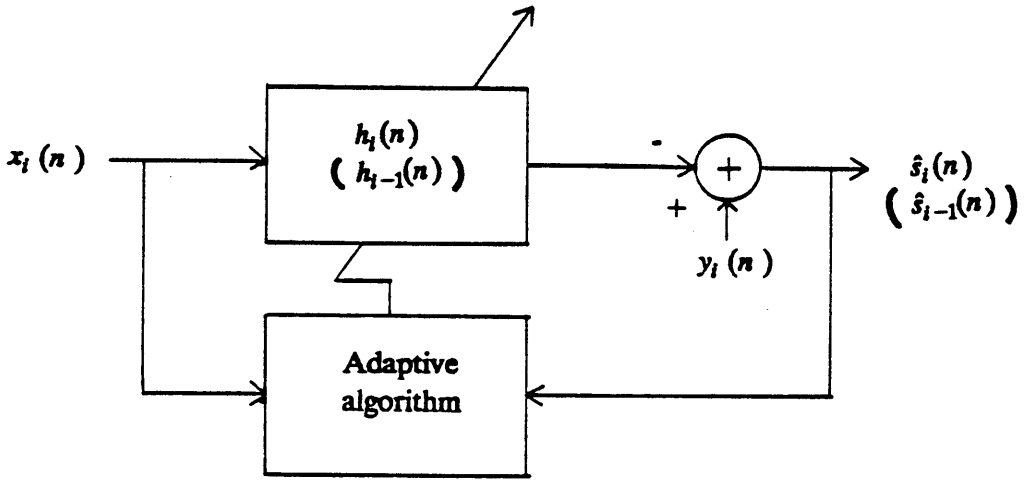
$$G_i(k) = \hat{H}_i(k)X_i(k) \tag{4.5.92}$$

$$g_i(n) = \frac{1}{N}\sum_{i=0}^{N-1} G_i(k)\exp\left(j\frac{2\pi}{N}k\,n\right) \tag{4.5.93}$$

Because $N \geq L+M-1$, $g_i(n)$ is $L+M-1$ points long. However $\hat{h}_i(n)$ may be up to M points long, the first M-1 points and the last M-1 points in the $L+M-1$ samples of $g_i(n)$ are not valid. Therefore the middle $L-M+1$ points are legitimate. The first F points of the valid middle portion is subtracted from $y_i(n)$. These F points are $\hat{s}_i(n)$.

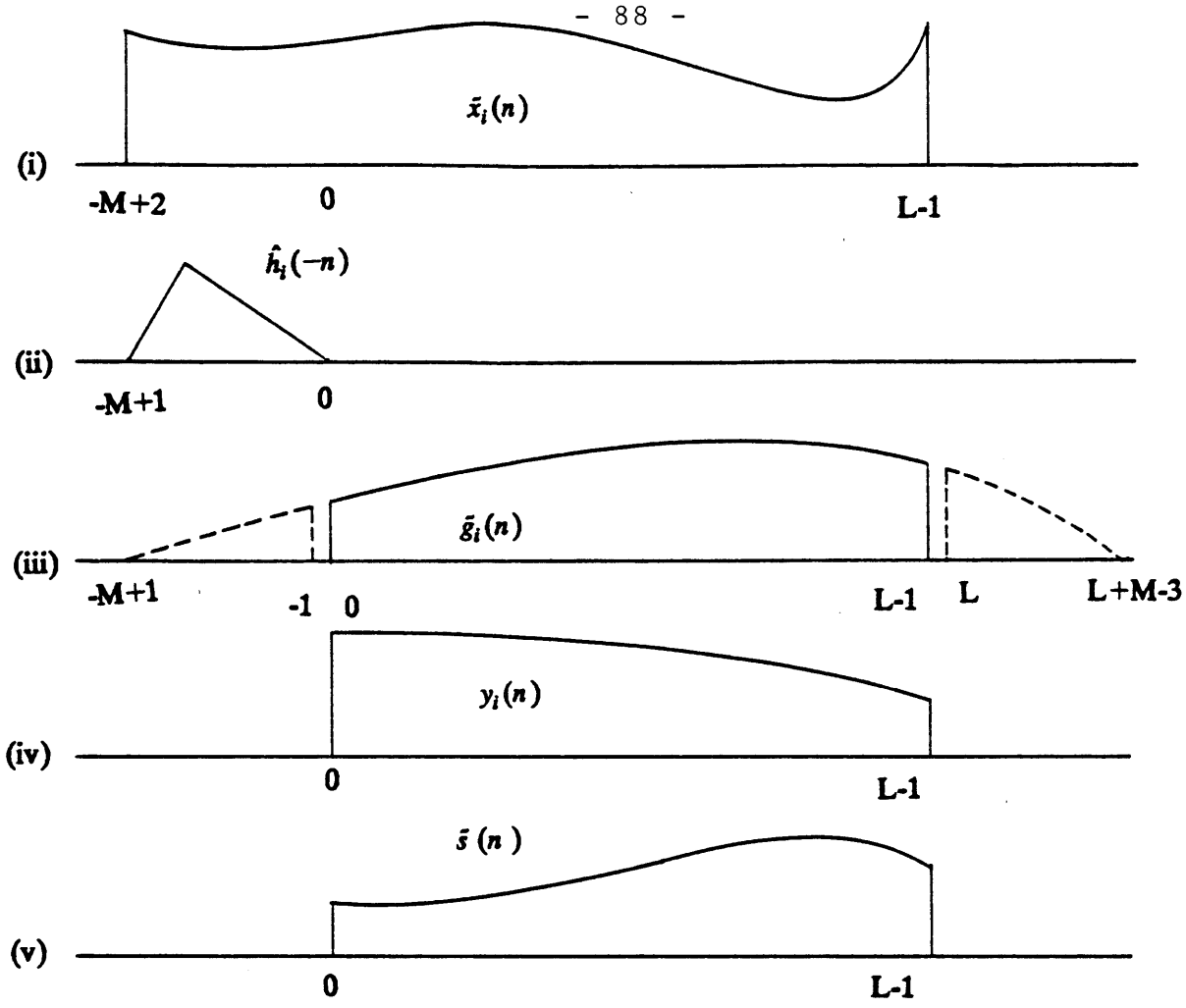


(a) The before-improvement system



(b) The after-improvement system

Fig 4.4 The estimation process and the filtering process



The FFT size ≥ L+2M-2

Fig 4.5 The filtering process for estimating $\tilde{s}_i(n)$

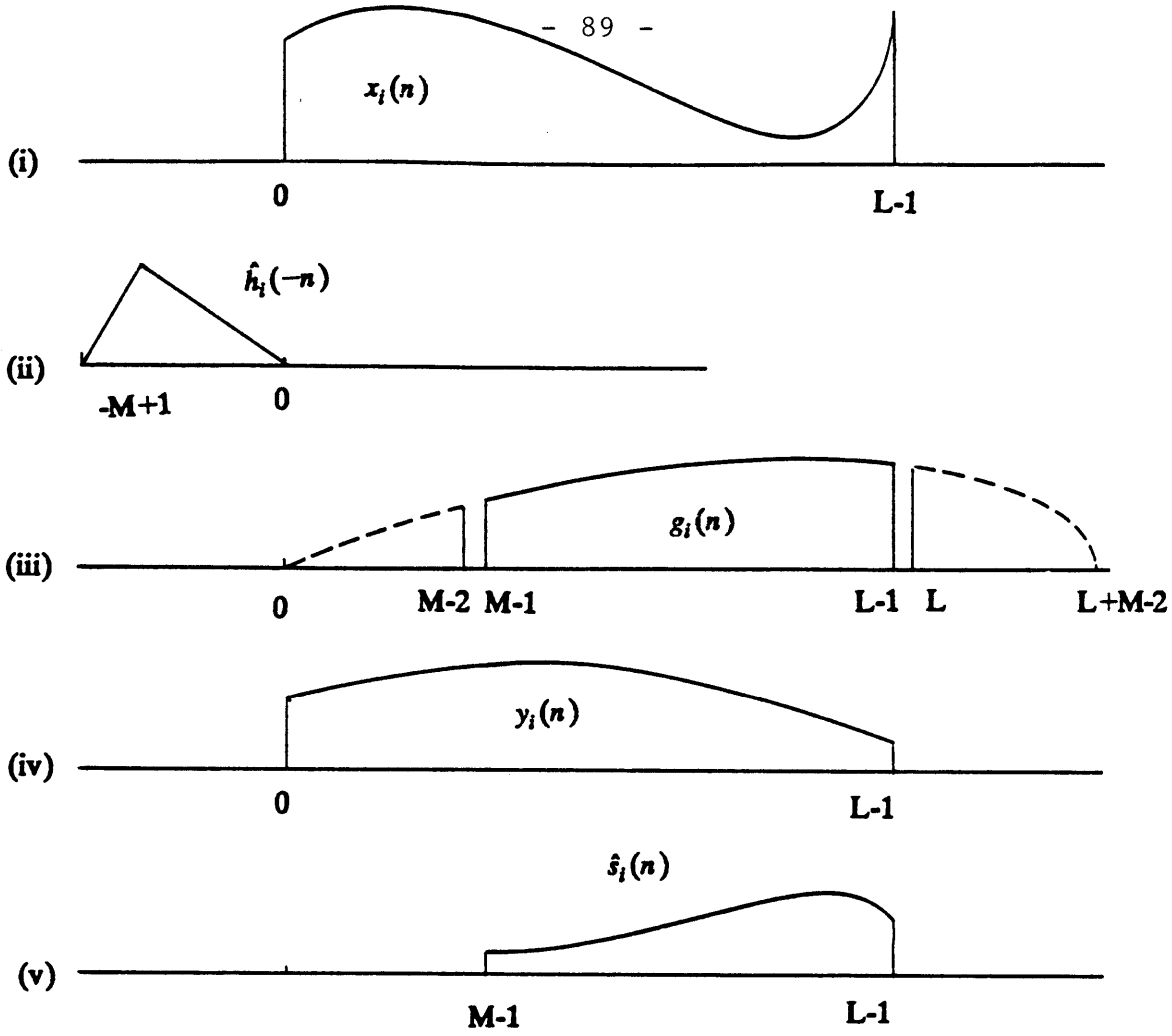


Fig 4.6 The filtering process for estimating $\hat{s}_i(n)$

4.5.7. Size of FFT

In practice the DFT is computed by the Fast Fourier Transform (FFT) algorithm. Determining the size of the DFT (FFT) means defining the number of the frequency samples that will be calculated in all functions such as $P_{x_i s_i}(k)$, $P_{x_i x_i}(k)$, $\hat{\sigma}^2{}_i(k)$, $\hat{H}_i(k)$, and $\hat{S}_i(k)$.

The size of FFT will be determined by two major factors. One is to avoid time aliasing in $P_{x_i \tilde{s}_i}(k)$, $P_{x_i x_i}(k)$, $\hat{\sigma}^2{}_i(k)$. Since $x_i(n)$, $\tilde{s}_i(n)$ for each frame are both L points sequences, all auto correlations $x_i(-n) * x_i(n)$, $\tilde{s}_i(-n) * \tilde{s}_i(n)$, and cross correlation $x_i(n) * \tilde{s}_i(n)$ are 2L-1 points long. Thus it is desirable to use FFT's of length greater than 2L-1 points in order to avoid time aliasing in these correlations.

The second factor arises from the need to use the estimated filter $\hat{H}_i(k)$ to generate both the preliminary estimate $\tilde{s}_i(n)$, and the final estimate $\hat{s}_i(n)$. The filtering process is most easily realized by the overlap-add method or the overlap-save method, because these approaches directly use the DFT to implement the filtering. Referring to section 4.5.5., we summarize the overlap-save method as follows ;

To filter the entire reference signal $x(n)$ with the filter $\hat{h}_i(n)$ and to estimate the final desired speech, $\hat{s}(n)$, $x(n)$ and $y(n)$ are sectioned into L point frames with M-1 points overlapping. Each $x_i(n)$ for the frame i is read into the N point FFT buffer, padding zeros at the end of the sequence. Then take the N point FFT, $X_i(k)$. Also read the filter $\hat{h}_i(n)$ into the N FFT point buffer, pad with zeros and take the N point FFT, $\hat{H}_i(k)$. Then the filter output $\hat{S}_i(k)$ is calculated by

$$\hat{S}_i(k) = Y_i(k) - \hat{H}_i(k) X_i(k) \tag{4.59-4}$$

Taking the N point inverse DFT of $\hat{S}_i(k)$,

$$\hat{s}_i(n) = \sum_{k=0}^{N-1} S_i(k) \exp\left(j \frac{2\pi}{N} kn\right) \tag{4.5.95}$$

where $X_i(k)$, $Y_i(k)$, $S_i(k)$, $\hat{H}_i(k)$ are N point DFT's of $x_i(n)$, $y_i(n)$, $s_i(n)$, $\hat{h}_i(n)$ respectively. Because the multiplying DFT coefficients corresponds to circular convolution, it is crucial to identify portions in $\hat{s}_i(n)$ free of circular aliasing. in order to get valid samples of $\hat{s}_i(n)$. Although the FFT size N in the optimal overlap-save method (the case where the least size for the same effect) is $N = L$ points long [1], we are allowed to use a bigger FFT size $N \geq L$, particularly $N = L + M - 1$ for simplicity here.

The first M-1 samples in $\hat{s}_i(n)$ must be discarded, because they are the result of the circular aliasing, so do not properly depend on the last M point input data $x_{i-1}(n)$, $y_{i-1}(n)$ from the i-1 th frame. Likewise, $\hat{s}_i(n) n = L+1 \cdots L+M-1$ is not valid either, since they should have further contribution from the next frame input data $x_{i-1}(n)$, $y_{i-1}(n)$. Consequently the middle portion of $\hat{s}_i(n)$, for $n = M, \ldots, L-1$ corresponds to the correct convolution of $x(n)$ and $\hat{h}_i(k)$.

The above discussion of the size of the FFT buffer suggests that the size of FFT should be large enough to make overlap-save processing convenient as well as avoid circular aliasing of the correlations. We will choose an FFT size $N \geq 2L$. (Remember that $L \geq M$ to achieve the proper frequency resolution.)

4.5.8. Smoothing effect

As shown in section 4.5.4., the size of the data length to be read for one frame, L must satisfy $L \gg M$. By choosing an FFT of length $N \geq 2L$, to calculate the filter estimate $\hat{H}_i(k)$, the time domain sequence obtained by the inverse FFT of $\hat{H}_i(k)$ is also N points long. Since we assumed that the room impulse response is M ($\leq N$) points long, we should make the length of estimated filter M points long by smoothing. An M points long filter

has frequency resolution $\frac{2\pi}{M}$. Letting the filter be N points long leaves too much resolution, and leads to estimates with large variance. Smoothing is needed to reduce the filter variance, and only retain a filter estimate of length M. We realize this smoothing in three different ways.

The correct choice of the parameter α in the exponential average is one way to control smoothing. It performs the average of elements in the numerator and the denominator of the filter estimate $\hat{H}_i(k)$ over the past $\frac{1}{\alpha}$ frames.

Secondly, we smooth the auto power spectrum $P_{x_i x_i}(k)$ and the cross spectrum $P_{x_i \tilde{s}_i}(k)$ by tapering the corresponding correlation sequences with a Hanning window. Alternatively, smoothing of these spectra can be achieved by multiplying a the data sections $x(n)$, $\tilde{s}(n)$ by a Hanning window. In both cases, the window effectively averages the spectra over adjacent frequency components. That window is called a smoothing window, since it smooths the spectrum over several frequencies. Here we study the case in which a rectangular or a Hanning window is applied to time correlation lags. Assume that the transfer function $h(n)$ is at most M points long. Thus, the necessary frequency resolution is at least $\frac{2\pi}{M}$. It means that within $\frac{2\pi}{M}$, the magnitude of $H(\omega)$ does not change much. To leave a margin for error, we will try to achieve twice the necessary frequency resolution, that is $\frac{\pi}{M}$. Let the length of a Hanning window be P (odd number). To achieve the frequency resolution of $\frac{\pi}{M}$ for $\hat{H}_i(k)$, one needs the same resolution in $P_{x_i x_i}(k)$ and $P_{x_i \tilde{s}_i}(k)$. A Hanning window with full length P has a frequency resolution $\frac{4\pi}{P}$, whereas a rectangular window with full length P has $\frac{2\pi}{P}$. Thus, P should be at least 4 times longer than M for a hanning window case, on the other hand, in a rectangular window case, P should be 2

times longer than M.

The final smoothing is to truncate $\hat{h}_i(n)$ in the time domain to exactly M points long after taking the inverse N point DFT of $\hat{H}_i(k)$. Even when the Hanning window is applied to the spectra, the division of two DFT's in $\hat{H}_i(k)$ causes circular aliasing. Thus the sequence after taking the inverse N point DFT of $\hat{H}_i(k)$ is not still M points long, but N points long. The truncation of sample points beyond M in $\hat{h}_i(n)$ is equivalent to applying a rectangular window to $\hat{h}_i(n)$. As mentioned in a later section, forcing the filter to be M points long helps suppress circular aliasing in the final filter processing for $\hat{s}_i(n)$.

Let us discuss the effect of the smoothing window for the spectra. A smoothing window is effective in reducing the variance of the filter estimate. The variance of the filter estimate is mainly dominated by the third term in equation (4.5.52). Since $x(n)$ and $s(n)$ are uncorrelated, the true value of the cross spectrum is zero. Therefore, the expectation of the cross spectrum $P_{x_i\hat{s}_i}(\omega)$ is zero, but the variance of our estimation of this quantity is not zero. As the number of frames increases, averaging $P_{x_i\hat{s}_i}(\omega)$ over multiple frames reduces this variance. In order to further reduce the variance in each iteration, a smoothing window is often used. It takes the average of $P_{x_i\hat{s}_i}(\omega)$ over adjacent frequency components within one frame filter estimate.

One problem with using a smoothing window for the estimation of the filter is that the estimated filter becomes biased. Although the bias problem already arises from sectioning input data into frames, the smoothing window makes the problem worse. Thus the estimate is not necessary updated toward the true filter in a strict sense, even if the disturbance signal (the newscaster's voice) is not present.

Another problem is that if the phase of $H(k)$ fluctuates rapidly, then smoothing $P_{x_i\hat{s}_i}(k)$ will average across frequency samples with quite different phase, thus averaging

the cross spectrum to zero regardless of its correct value. The phenomenon becomes worse especially when the frequency response of the filter contains a large linear phase shift caused by a large initial delay in the impulse response. For instance, suppose $h(n)$ is a causal symmetric triangular sequence in Fig 4.7, with values:

$$h(n) = \begin{cases} 0 & n \le n_0 - 1 \\ \dfrac{n - n_0}{n_1 - n_0} & n_0 \le n \le n_0 + n_1 - 1 \\ \dfrac{-(n - 2n_1 - n_0)}{n_1 - n_0} & n_0 + n_1 \le n \le n_0 + 2n_1 - 1 \\ 0 & n_0 + 2n_1 \le n \end{cases} \tag{4.5.96}$$

with a n_0 point delay and a big peak at n_0+n_1-1. Then the Fourier transform $H(\omega)$ is ;

$$H(\omega) = \exp(-j\,\omega(n_0 + n_1))H_1(\omega) \tag{4.5.97}$$

where

$$H_1(\omega) = \sum_{n=-n_1}^{n_1} h_1(n)\exp(-j\,\omega n) \tag{4.5.98}$$

and $h_1(n)$ is a zero-phase symmetric triangular waveform. The magnitude and phase of $H(\omega)$ are given by;

$$|H(\omega)| = |H_1(\omega)| \tag{4.5.99}$$

$$arg[H(\omega)] = -\omega(n_0 + n_1) \tag{4.5.100}$$

In the DFT phase, this linear phase term, $-\omega(n_0 + n_1)$, produces periodic linear phase, segments wrapped in 2π, with a period of $\dfrac{2\pi}{n_0 + n_1}$.

Consequently, the drastic variation of the phase in the original $h(n)$ comes from the overall linear phase $-\omega(n_0+n_1)$ module 2π. Thus, if windowing is used to smooth the filter in the frequency domain, and the width of the window in the frequency domain is comparable to $\dfrac{2\pi}{n_0 + n_1}$, then the averaging will give an estimate of $\hat{H}_i(k) \approx 0$. Thus the

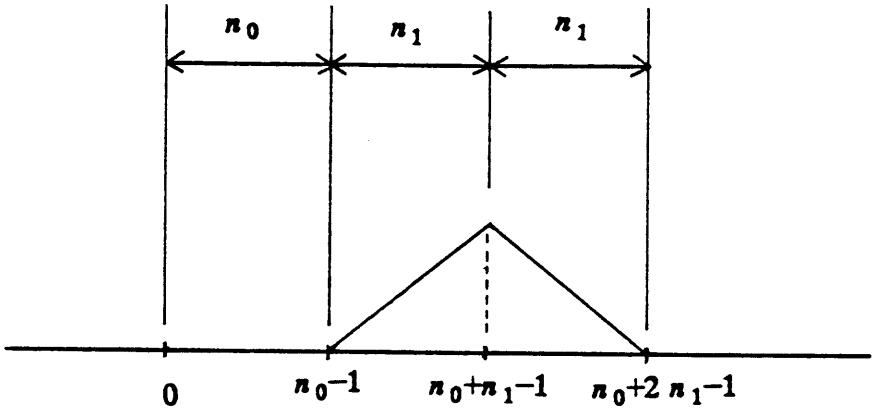


Fig 4.7 An example of a sequence containing delay time

averaging range (i.e. the frequency resolution) should be at least 10 times smaller than the period of the linear phase.

The room reverberation transfer function often contains an unknown long time delay at the beginning of its impulse response. The time delay corresponds to the traveling time of the acoustic signal over the distance from the loudspeaker to the microphone. For example, a distance of 4 feet provides about 40 points of delay at a 10 KHz sampling rate. The time delay increases in proportion to the distance. In the case of $h(n)$ with 40 points time delay , 60 points n_1, and a total length for the impulse response is 1024 points, the width of the main lobe of a smoothing window should be below $\frac{2\pi}{100}\times\frac{1}{10}=\frac{\pi}{500}$ radians in order to be 10 times narrower than one full linear phase cycle. It follows that a Hanning window, if used, will need a frequency resolution $\frac{4\pi}{P}<\frac{\pi}{500}$, which implies a window length $P>2000$.

4.5.9. Linear interpolation in the final estimate for the desired speech $\hat{s}_i(n)$.

There are some cases where the filter estimates $\hat{h}_i(n)$ from consecutive frames are quite different from one another. This is the case, for example, if the true filter $h(n)$ varies drastically the exponential averaging parameter α is set to 1, and little overlap between frames is used. Rapid changes in the filter estimate also occur during the initial transient. Due to this rapid change of $\hat{h}_i(n)$ from frame to frame, big mismatches in the waveform $\hat{s}_i(n)$ at frame boundaries occur, which may cause a clicking type of noise. To prevent this kind of noise, linear interpolation of the estimate $\hat{s}(n)$ is performed over a range of T ($<\frac{F}{2}$) points in the region where frames overlap.(See Fig 4.8)

When the middle portion of valid $L-M+1$ points is computed, the first $F+T$ points out of that valid portion are picked up and are multiplied with a equilateral trapezoid

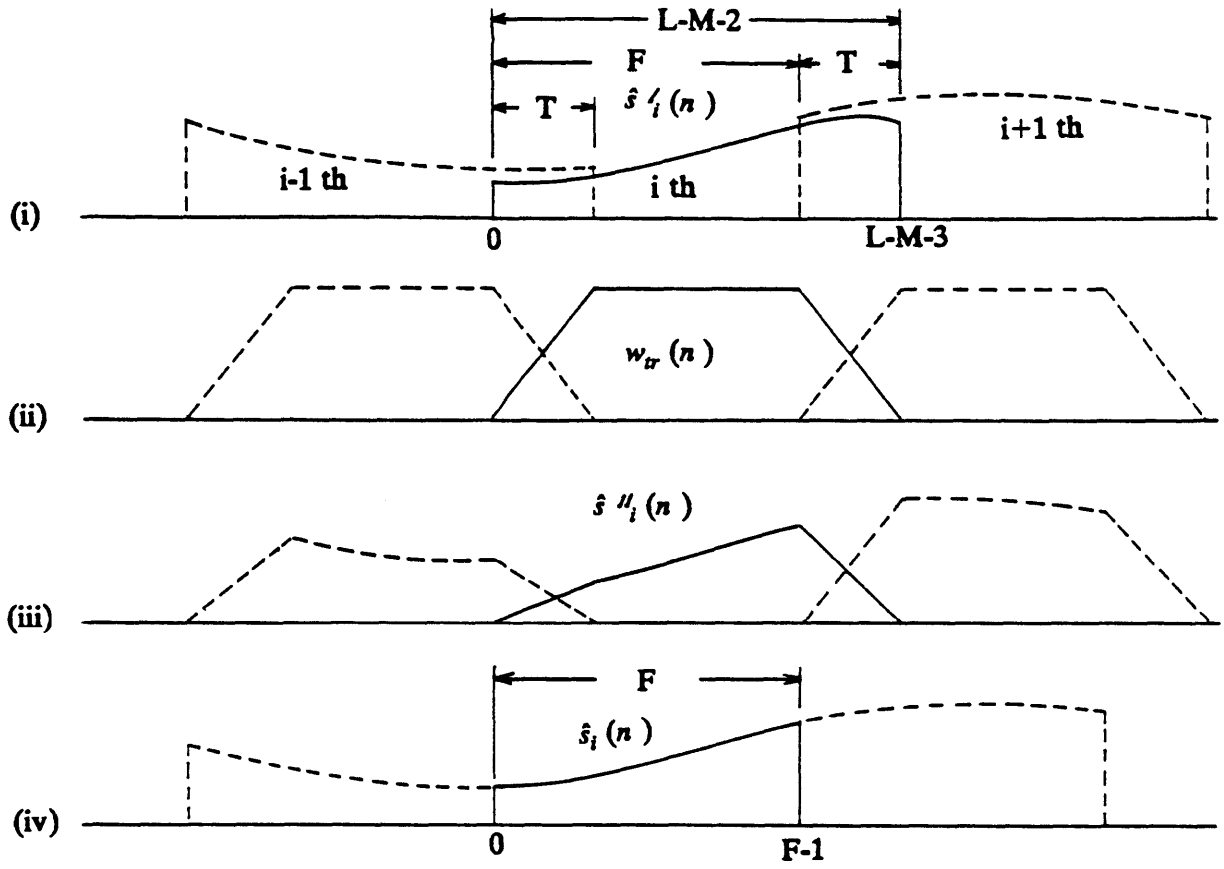


Fig 4.8 Linear Interpolation in the final estimate for $\hat{s}_i(n)$

(i) $\hat{s}'_i(n)$; A legitimate linear convolution portion of the filtering output, which is L-M-2 point long. (The solid line is the current frame portion, dotted lines are adjacent frame portions.)

(ii) $w_{tr}(n)$; A equilateral trapezoid window. Adjacent windows are overlapped in T points.

(iii) $\hat{s}''_i(n)$; An output windowed by the trapezoid.

(iv) $\hat{s}_i(n)$; The ultimate optimal speech estimate. The segments obtained in (iii) are added up at each sample point.

window. Each end region corresponding to the side of the trapezoid is overlapped with the end region of the windowed output in the adjacent frame. In those overlapped regions, the two linearly weighted adjacent frame outputs are interpolated. This interpolation provides a smooth transition in the filter output in the T point region at frame boundaries.

4.5.10. Implementation of the algorithm

In this section, the entire algorithm for both the estimation process and the filtering process will be presented in detail. We set several variables in the following way;

The length of the estimated filter (sample points) M.

The number of data points to be read for each frame L ; $L = 4M$.

The size of FFT buffer (sample points) N ; $N = 2L > L + 2M - 2$.

The length of a smoothing lag window P (should be odd).

The shift of the frame (sample points) F.

The number of points for interpolating the estimates $\hat{s}(n)$ between overlapping frames T.

[A] Set initial states $\hat{H}_0(k) = 0, N_0(k) = D_0(k) = 0$.

[B] Perform a threshold test for the reference energy $\sum_n x_i(n)^2$. If the energy is less than the threshold V_{th} then latch the value of the filter estimate from the previous frame for the current estimate, go to [E]. Otherwise, go to [C].

[C] Estimate $\hat{\sigma}^2_i(k)$. (See Fig 4.5)

[C-1] Read the i th frame L point data $x_i(n), y_i(n), n = 0, \ldots, L-1$. This is equivalent to picking them up with a rectangular window.

[C-2] Concatenate L point $x_i(n)$ to the last M-1 point data of the i-1 th frame $x_{i-1}(n)$, and make $L+M-1$ point $\tilde{x}_i(n), n=-M+2, \ldots, L-1$.

[C-3] Padding $N-(L+M-1)$ zeros to $\tilde{x}_i(n)$, take the N point DFT of $\tilde{x}_i(n)$, denoting the result as $\tilde{X}_i(k)$.

[C-4] Multiply $\tilde{X}_i(k)$ with the DFT of the previous frame filter estimate, $\hat{H}_{i-1}(k)$, then obtain $\tilde{G}_i(n)$ by

$$\tilde{G}_i(k) = \tilde{X}_i(k)\hat{H}_{i-1}(k) \tag{4.5.101}$$

[C-5] Take the N point IDFT of $\tilde{G}_i(k)$, denoting the result as $\tilde{g}_i(n)$. However, the first M-1 points and the last M-1 points in $\tilde{g}_i(n)$ do not represent the valid linear convolution of the signal $x(n)$ with $\hat{h}_{i-1}(n)$. The correct portion is from $n=0$ to $L-1$. Make $\tilde{g}'_i(n)$

$$\tilde{g}'_i(n) = \tilde{g}_i(n) \tag{4.5.102}$$

[C-6] Subtract $\tilde{g}'_i(n)$ from $y_i(n), n=0, \ldots, L-1$. The remainder is an L point sequence $\tilde{s}_i(n)$, a pre-estimate for the desired speech.

[C-7] Take N point DFT of the L point sequence $\tilde{s}_i(n)$, denoted as $\tilde{S}_i(k)$. Compute the squared magnitude of $\tilde{S}_i(k)$, $|\tilde{S}_i(k)|^2$.

[C-8] Take the N point IDFT of $|\tilde{S}_i(k)|^2$, obtaining the correlation sequence $R_{\tilde{s}_i\tilde{s}_i}(m)$. Multiply a P (an odd number) point Hanning window by the correlation sequence, and then take the N point DFT of the resultant sequence, to get a smooth power spectral density $\hat{\sigma}^2_i(k)$.

[D] Estimate the filter $\hat{H}_i(k)$

[D-1] Take the N point DFT of $x_i(n)$, padding with N-L point zeros, and denote it as $X_i(n)$.

[D-2]Calculate the current frame power spectral density estimate $P_{x_i x_i}(k)$,

$$P_{x_i x_i}(k) = \frac{1}{L} X^*_i(k) X_i(k) \tag{4.5.103}$$

also calculate the current frame cross power spectral density estimate $P_{x_i \bar{s}_i}(k)$,

$$P_{x_i \bar{s}_i}(k) = \frac{1}{L} X^*_i(k) \bar{S}_i(k) \tag{4.5.104}$$

[D-3]Smooth the current frame power spectral density estimates.(See Fig 4.9 Take the N point IDFT of $P_{x_i x_i}(k)$, $P_{x_i \bar{s}_i}(k)$, denoting the resultant correlation sequences as $R_{x_i x_i}(m)$, $R_{x_i \bar{s}_i}(m)$, respectively. They are both 2L-1 point sequences, so there is no circular aliasing in those N point sequences. Apply a P point Hanning window to those correlation sequences. Take the N point DFT of the windowed correlation sequences to get smooth power spectral density estimates, $\bar{P}_{x_i x_i}(k)$ $\bar{P}_{x_i \bar{s}_i}(k)$

[D-4]Calculate the smooth cross power spectral density estimate $\bar{P}_{x_i y_i}(k)$ by

$$\bar{P}_{x_i y_i}(k) = \bar{P}_{x_i \bar{s}_i}(k) + \hat{H}_{i-1}(k) \bar{P}_{x_i x_i}(k) \tag{4.5.105}$$

[D-5]Using $\hat{\sigma}^2{}_i(k)$ calculated in [C], the estimated filter is

$$\hat{H}_i(k) = \frac{(1-\alpha) N_{i-1}(k) + \alpha \bar{P}_{x_i y_i}(k) / \hat{\sigma}^2{}_i(k)}{(1-\alpha) D_{i-1}(k) + \alpha \bar{P}_{x_i x_i}(k) / \hat{\sigma}^2{}_i(k)} \tag{4.5.106}$$

[D-6]Smooth $\hat{H}_i(k)$. Take the N point IDFT of $\hat{H}_i(k)$, to get an N point filter $\hat{h}_i(n)$. Truncate the last N-M point of $\hat{h}_i(n)$, so that now $\hat{h}_i(n)$ is an M point filter. This truncation is equivalent to the smoothing in a frequency domain. Take the N point DFT of $\hat{h}_i(n)$, to get the ultimate estimate $\hat{H}_i(k)$.

[E] Filtering process to obtain the estimate $\hat{s}_i(n)$ for the desired speech.(See Fig 4.6)

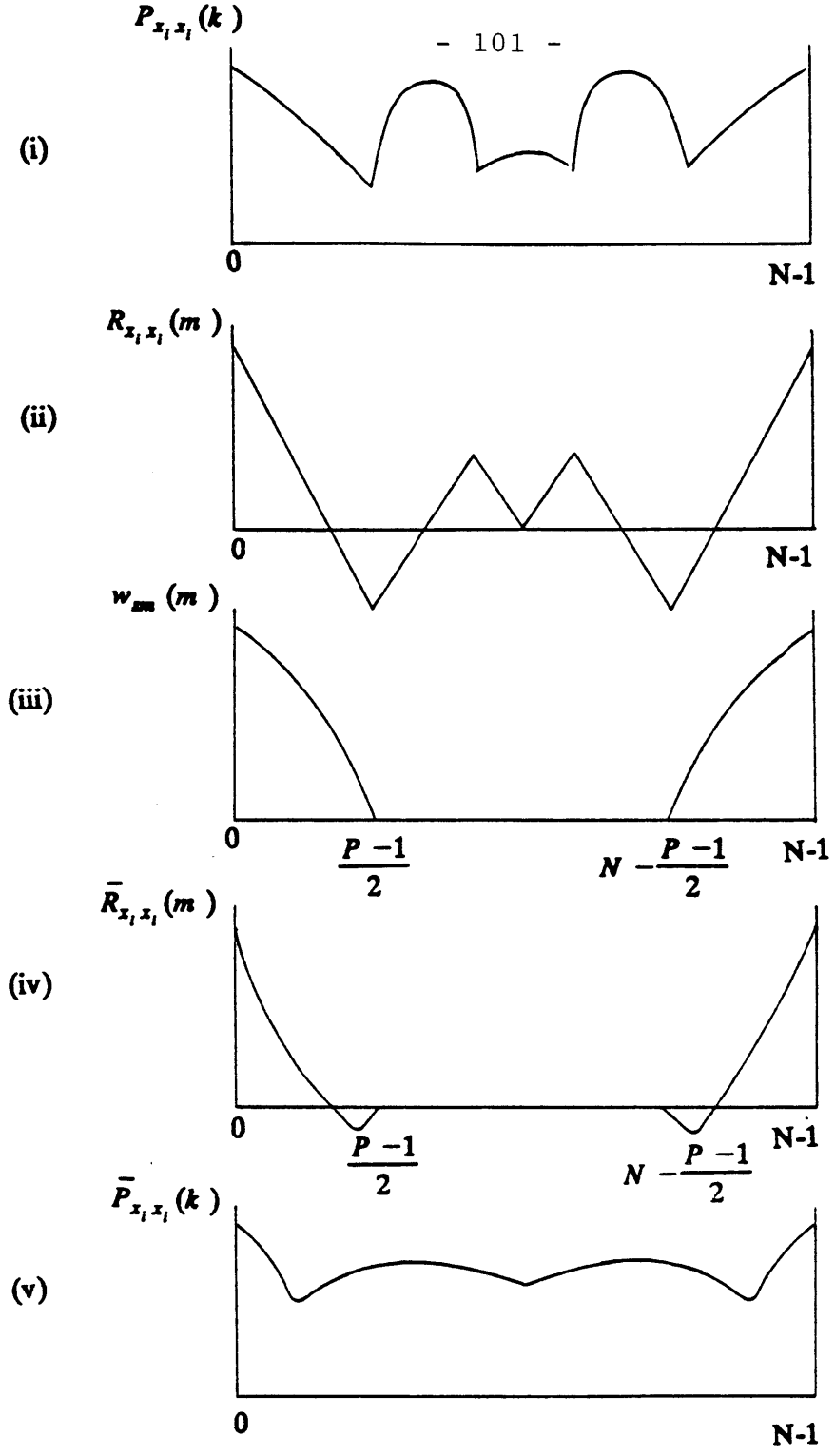


Fig 4.9 Smoothing a spectrum.

[E-1] Multiply the DFT of the current frame reference speech $X_i(k)$ by $\hat{H}_i(k)$ to obtain $G_i(k)$,

$$G_i(k) = X_i(k)\hat{H}_i(k) \tag{4.5.107}$$

[E-2] Take the N point IDFT of $G_i(k)$. The resultant sequence $g_i(n)$ has valid data corresponding to the linear convolution, from n=M-1 to n=L-1.

[E-3] Subtract $g_i(n)$ from $y_i(n)$ for n=M-1 to n=L-1, and obtain an L-M-1 point optimal speech estimate, $\hat{s}'_i(n)\, n = 0, \ldots, L-M$

$$\hat{s}'_i(n) = y_i(n-M+1) - g_i(n-M+1), n = 0, \ldots, L-M \tag{4.5.108}$$

[E-4] Smooth the desired speech estimate $s_i(n)$ across frame boundaries. Multiply $\hat{s}''_i(n)$ by an $F+T$ long trapezoidal window $w_{tr}(n)$, (See Fig 4.8)

$$w_{tr}(n) = \begin{cases} \dfrac{n}{T} & 0 \le n \le T \\ 1 & T+1 \le n \le F \\ -\left(\dfrac{n-F}{T}\right)+1 & F+1 \le n \le F+T \end{cases} \tag{4.5.109}$$

thus getting an $F+T$ point sequence $\hat{s}''_i(n)$,

$$\hat{s}''_i(n) = \hat{s}'_i(n)\, w_{tr}(n) \tag{4.5.110}$$

[E-5] Add the the previous frame windowed sequence $\hat{s}''_{i-1}(n)$ for the last T points and the current windowed sequence $\hat{s}''_i(n)$ for the first T points. Keep the next F points as is ;

$$\hat{s}''_i(n) = \begin{cases} \hat{s}''_i(n) + \hat{s}''_{i-1}(n+F) & 0 \le n \le T \\ \hat{s}''_i(n) & T+1 \le n \le F+T \end{cases} \tag{4.5.111}$$

[E-6] Denote the first F points in $\hat{s}''_i(n)$, as $s_i(n)$. The last T points in $\hat{s}''_i(n)$ are saved to use in step [E-5] for the next frame. Go to [B]

4.5.11. Performance measures for the algorithms

In order to understand both the limitations and advantages of these algorithms in the broadcast room problem, a study was performed on synthetic data with a known room reverberant linear system and a known newscaster's voice. In this section, we define and discuss several performance measures for analyzing the proposed estimation and filtering methods. They include the filter deviation γ, the noise reduction measure ρ, the signal-noise-ratio SNR_{out}, and the signal improvement measure β. γ, ρ, SNR_{out} can be monitored in synthetic data experiments, but not in real experiments, for in real experiments one does not know either the true transfer function or the newscaster's voice in advance. The signal improvement measure can be monitored in both cases.

Each measure indicates different aspects of the performance of the algorithms, although they are closely related with each other.

The filter deviation measure γ

The filter deviation measure γ is the ratio (expressed in DB) of the energy of the error in the filter coefficients to energy in the true filter coefficients ;

$$\gamma = 10 \log_{10} \left\{ \frac{\sum_{n=0}^{M-1} \Delta h^2(n)}{\sum_{n=0}^{M-1} h^2(n)} \right\} \tag{4.5.112}$$

where

$$\Delta h(n) = h(n) - \hat{h}(n), n = 0, \ldots, M-1 \tag{4.5.113}$$

Applying Parseval's theorem,

$$\gamma = 10 \log_{10} \left\{ \frac{\frac{1}{2\pi} \int_{-\pi}^{\pi} |\Delta H(\omega)|^2 d\omega}{\frac{1}{2\pi} \int_{-\pi}^{\pi} |H(\omega)|^2 d\omega} \right\} \tag{4.5.114}$$

where $H(\omega)$, $\Delta H(\omega)$ are the Fourier transforms of $h(n)$, $h(n)-\hat{h}(n)$. Using N point DFT's of $h(n)$, $h(n)-\hat{h}(n)$,

$$\gamma = 10\log_{10}\left\{\frac{\sum_{k=0}^{N-1}|\Delta H(k)|^2}{\sum_{k=0}^{N-1}|H(k)|^2}\right\} \tag{4.5.115}$$

γ is useful to characterize how well the estimate $\hat{h}(n)$ approximates the true transfer function. Also γ can be interpreted as the expected noise reduction when the reference signal and desired speech are white noise. If the reference signal $x(n)$ and the uncorrelated desired speech $s(n)$ are assumed to be white noise, then γ is on average a function of three system parameters, the total number of the observed data points N_{total}, the number of points in the impulse response M, and the signal-to-noise ratio in the primary signal,[8]

$$\gamma|_{white\ noise\ input} = 10\log_{10}\left[\frac{M}{N_{total}}\right] + SNR_{primary} \tag{4.5.116}$$

For the case of speech signal, where $x(n)$ and $s(n)$ are not white, the relationship of γ to the signal-to-noise ratio and the filter parameters can not derived as easily.

The noise reduction measure ρ

The noise reduction measure is defined as

$$\rho = 10\log_{10}\left\{\frac{\frac{1}{2\pi}\int_{-\pi}^{\pi}|\Delta H(\omega)X(\omega)|^2 d\omega}{\frac{1}{2\pi}\int_{-\pi}^{\pi}|H(\omega)X(\omega)|^2 d\omega}\right\} \tag{4.5.117}$$

In the DFT expression, it is

$$\rho = 10\log_{10}\left\{\frac{\sum_{k=0}^{N-1}|\Delta H(k)X(k)|^2}{\sum_{k=0}^{N-1}|H(k)X(k)|^2}\right\} \tag{4.5.118}$$

ρ is interpreted as the ratio of the energy of the remaining interfering signal in the output, to the energy of the filtered version of the interfering signal in the primary signal. For white noise signal $x(n)$, since $|X(k)|^2$ is constant on average, the measure γ is equivalent to ρ on average. For non-white stationary or non-stationary signals like speech, ρ is a much superior indication of the performance in the interfering signal cancellation. Because the ultimate purpose of the noise cancellation for the broadcast room problem is not to obtain equally good filter estimates $\hat{H}_i(k)$ at every frequency, but to cancel the interfering speech as much as possible to obtain a good desired signal estimate $\hat{s}(n)$, it is sufficient to have a good estimate of the system $H(k)$ only at frequencies where the input signal energy $X(\omega)$ is large. In a frequency region where the reference signal $X(k)$ is greatly attenuated, there is no information in the primary signal about the behavior of the system $h(n)$. In that frequency region, no system identification technique can offer reliable identification for the system function $h(n)$. Fortunately, in such a frequency region, it is not as important to have very good estimates of $H(\omega)$, since any error will not greatly affect the estimate of $S(k)$ frequencies. Consequently a frequency weighted measure of performance such as ρ is desirable, since ρ emphasizes the importance of these frequencies where $X(\omega)$ is large and deemphasizes the importance of those frequencies where $X(\omega)$ is small.

One caution is that neither γ nor ρ accurately measure the subjective quality of the algorithm. The reason is that even though γ in the white noise input case, or ρ in the speech signal case, may appear very good, the output signal may still sound unacceptable if the signal-to-noise ratio in the output is poor.

The signal-to-noise ratio in output SNR_{out}

Since the subjective quality will correspond to the signal-to-noise ratio, it is useful to

monitor the signal-to-noise ratio in the output signal;

$$SNR_{out} = 10\log_{10}\left\{\frac{\sum_n |s(n)|^2}{\sum_n |s(n)-\hat{s}(n)|^2}\right\} \tag{4.5.119}$$

Here in the denominator $s(n)-\hat{s}(n)$ represents the remaining interfering signal in the output, namely, $(h(n)-\hat{h}(n))*x(n)$.

Now we explain the relation of the noise reduction ρ and SNR_{out}. Define the signal-to-noise ratio in the primary signal;

$$SNR_{pri} = 10\log_{10}\left\{\frac{\sum_n |s(n)|^2}{\sum_n |x(n)*h(n)|^2}\right\} \tag{4.5.120}$$

It follows that the noise reduction ρ is the difference between SNR_{pri} and SNR_{out}.

$$\rho = SNR_{pri} - SNR_{out} \tag{4.5.121}$$

The signal improvement measure β

The signal improvement measure β is one of the few measures available in real experiments with recorded speech in a real room. The definition is,

$$\beta = 10\log_{10}\left\{\frac{\sum_n |y(n)|^2}{\sum_n |\hat{s}(n)|^2}\right\} \tag{4.5.122}$$

$$= 10\log_{10}\left\{\frac{\sum_n |s(n)+x(n)*h(n)|^2}{\sum_n |s(n)+x(n)*(h(n)-h(n))|^2}\right\} \tag{4.5.123}$$

This is not a particularly good performance measure, since it can not distinguish the contribution of loud $s(n)$ from the contribution due to the filter estimation error.

References

1. A. V. Oppenheim and R. W. Shafer, in *Digital signal processing*, Prentice-Hall, Englewood Cliffs, 1975.

2. P.D. Welch, "Spectra: A Method Based on Time Averaging over Short Modified Periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70-73, June 1967.

3. R. B. Blackman and J. W. Tukey, , *The Measurement of Power Spectra form the Point of View of Communications Engineering*, Dover, 1959., New York: .

4. S. M. Kay and S. L. Marple Jr., "Spectrum Analysis-A Modern Perspective," *Proceeding of the IEEE*, vol. 69, no. 11 , Nov 1981.

5. L. R. Rabiner and Jont B. Allen, "Short-time Fourier Analysis Techniques for FIR System Identification and Power Spectrum Estimation," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, April 1979.

6. L. R. Rabiner, R. E. Crochiere, and J. B. Allen, "FIR System Modeling and Identification in the Presence of Noise and with Band-Limited Inputs," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 4, August 1978.

7. B. R. Musicus, "An Iterative Technique for Maximum Likelihood Estimation With Noisy Data," *SM. Thesis*, Massachusetts Institute of Technology, Dept. of Elec. Engi. and Comp. Science, 1979.

8. L. R. Rabiner, R. E. Crochiere, and J. B. Allen, "FIR System Modeling and Identification in the Presence of Noise and with Band-Limited Inputs," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 4, August 1978.

Chapter 5 Empirical Results

5.1. Introduction

In the preceding chapter, we have proposed two different estimation methods for non-stationary signal cases. As mentioned there, the first method (the method of averaging the transfer function estimate over frames (ATFE)) is not a desirable algorithm because it has the high variance. Thus we will concentrate on evaluating the second method (the method of averaging the spectrum weighted with the value of the reciprocal of the optimal speech power spectral density estimate (AWSE)) in this chapter.

To investigate the performance of the second algorithm, we study several experiments. The experiments are divided into two major categories. The first category is a synthetic case, where we implement the system identification model Fig. 4.2 and simulate the primary speech by filtering the known reference speech through the known transfer function $h(n)$ and adding speech $s(n)$. Then we process both reference speech and primary speech with the proposed algorithm and examine how well it estimates the $h(n)$, as well as how well it restores the newsman's speech.

The other category is a real case, where we recorded reference speech and primary speech in an experiment in an actual room environment. Therefore neither an exact system function nor the newsman's speech are known in advance. This second category tests the performance of the algorithm in reality.

In the case of synthetic data, two types of time-invariant system functions are used to simulate the room. The first is merely a 32 point delayed delta function, the second is a 1024 point room impulse response, which is experimentally measured in a particular room.

Here we must note that, a room reverberation system, we have called in this thesis is

not only a pure room reverberation, but the overall system function including equipment characteristics, because the actual relation of the reference signal and the primary signal depend on the overall system.

We first discuss how this room impulse response was measured. Next we analyze what we expected to observe in a series of experiments. The result and discussion of experiments are presented.

5.2. The methodology for measuring a room reverberation transfer function

A room configuration

The room used for acquiring experimental data was an ordinary office room as shown in Fig 5.1. The room shape is almost rectangular, and its size is approximately $10' \times 20' \times 10'$. All four walls are made of plaster board, the ceiling is composed of acoustic paneling tiles, and the floor consists of tiles on concrete. The office contained a typical collection of acoustically reflecting objects, including several desks, chairs, sofa and shelves.

Equipment setting and experimental procedure

One method for measuring the transfer function in a room from the reference loudspeaker to the microphone, is to generate an impulse with short pulse width, feeding it into the loudspeaker, and then collecting an impulse response measured by the microphone.

Since we are interested in a frequency region of up to 5 KHz for speech, the pulse width of the impulse should be within 200 μsec . We used a 100 μsec width pulse for all the experiment. The equipment used was placed in the room as illustrated in Fig 5.1. We chose the distance between the loud speaker and the microphone to be 4 feet.

The Fig 5.2 shows the whole procedure from recording to digitizing. After both input

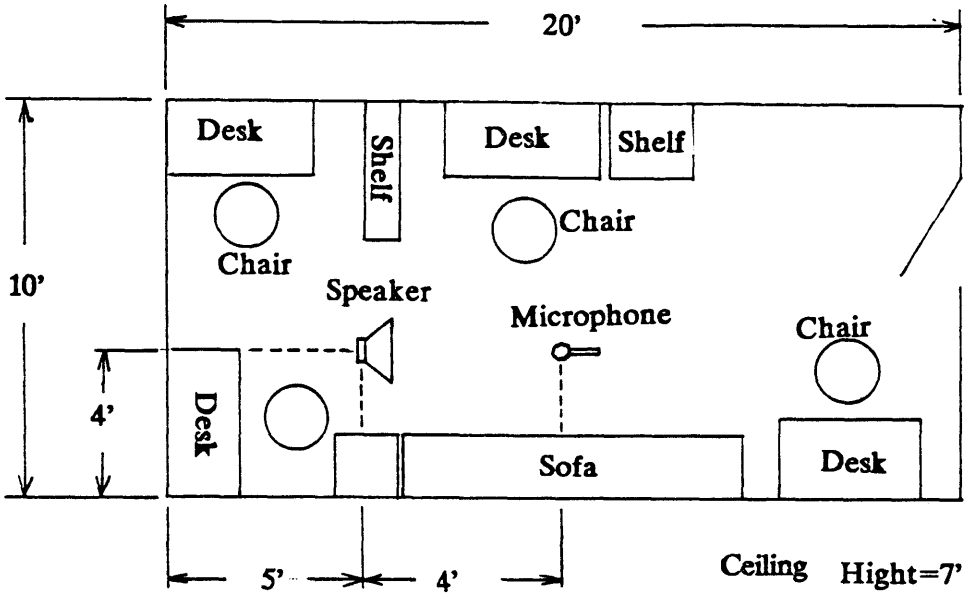


Fig 5.1 Room configuration

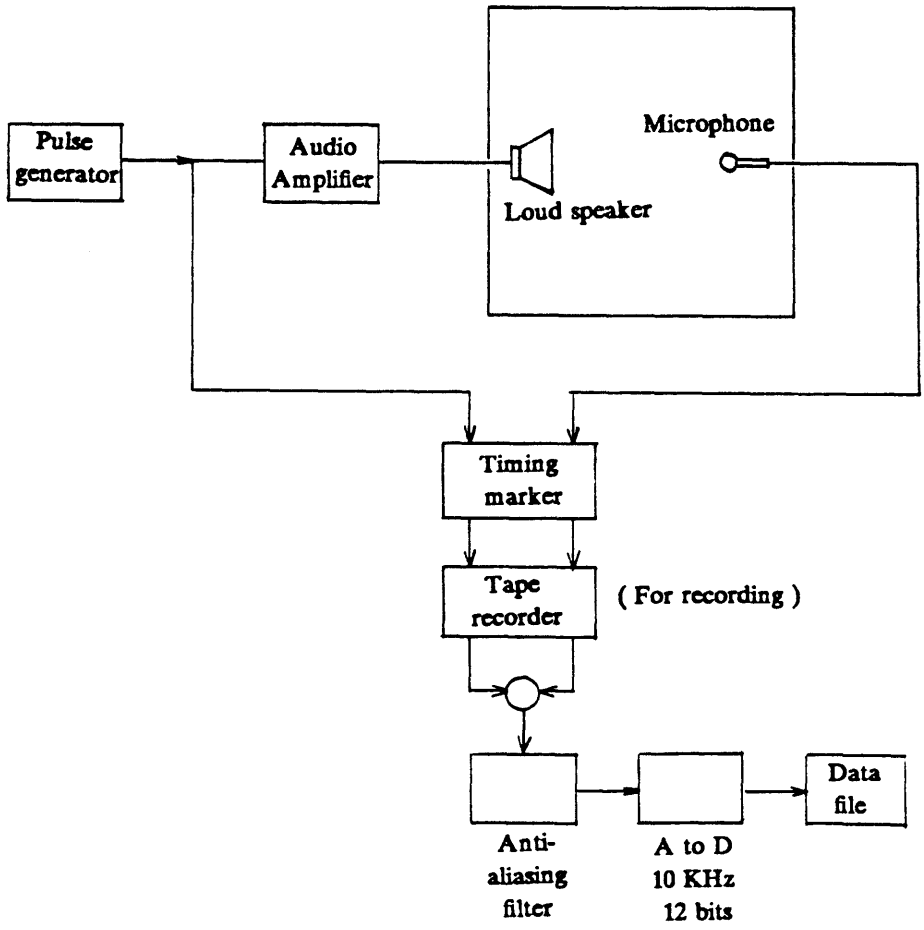


Fig 5.2 Equipment setting in measuring a room reverberant impulse response.

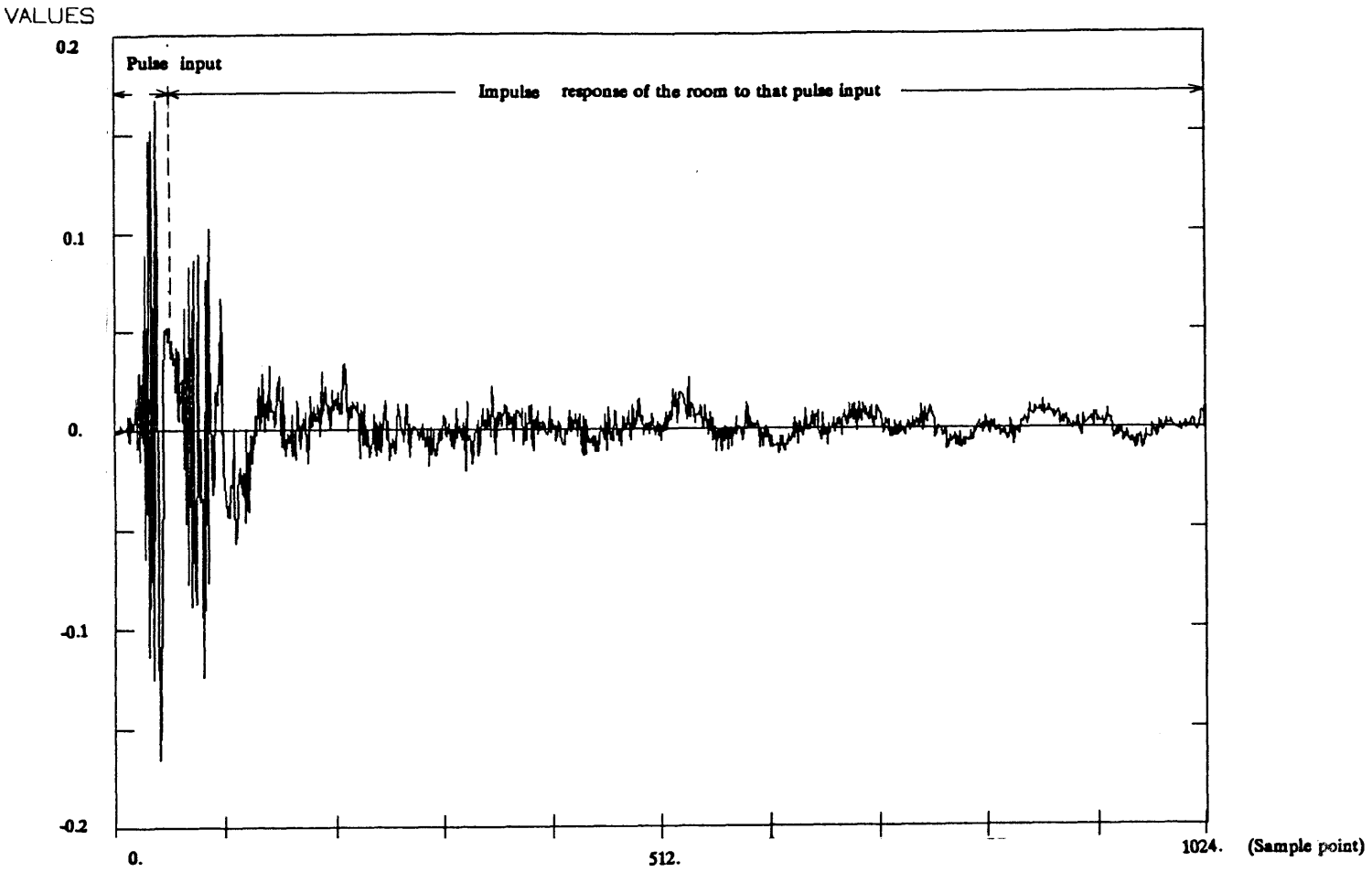


Fig 5.3 The input pulse and the room response to that pulse

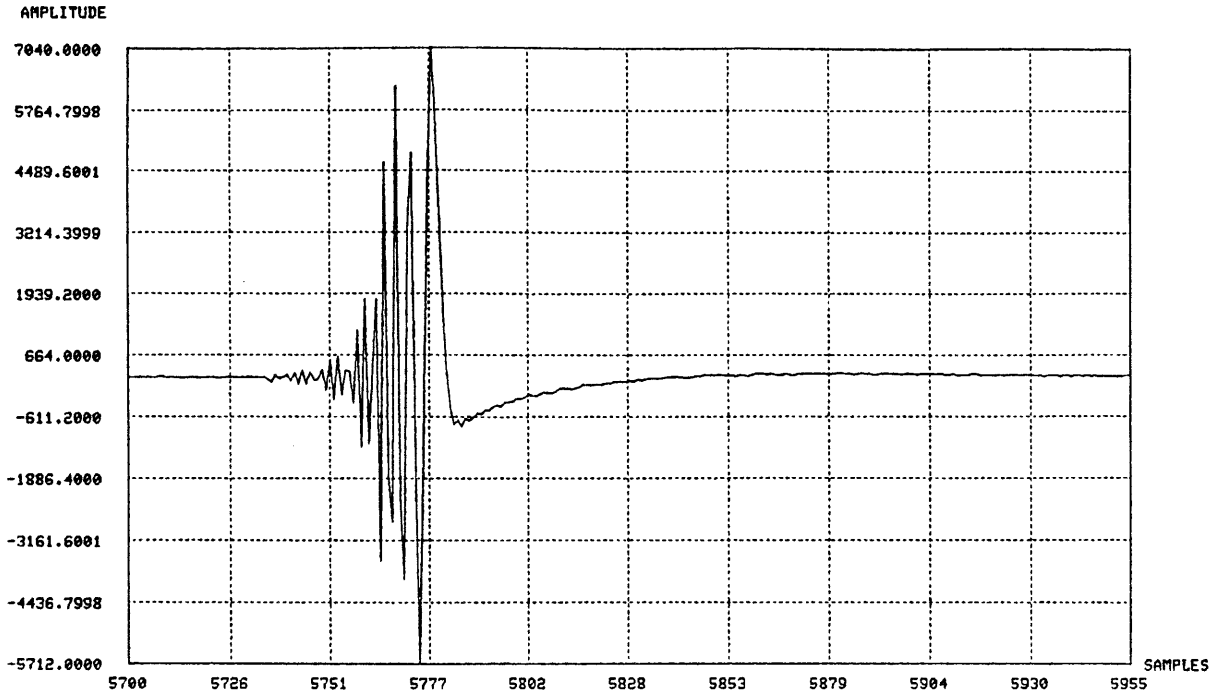
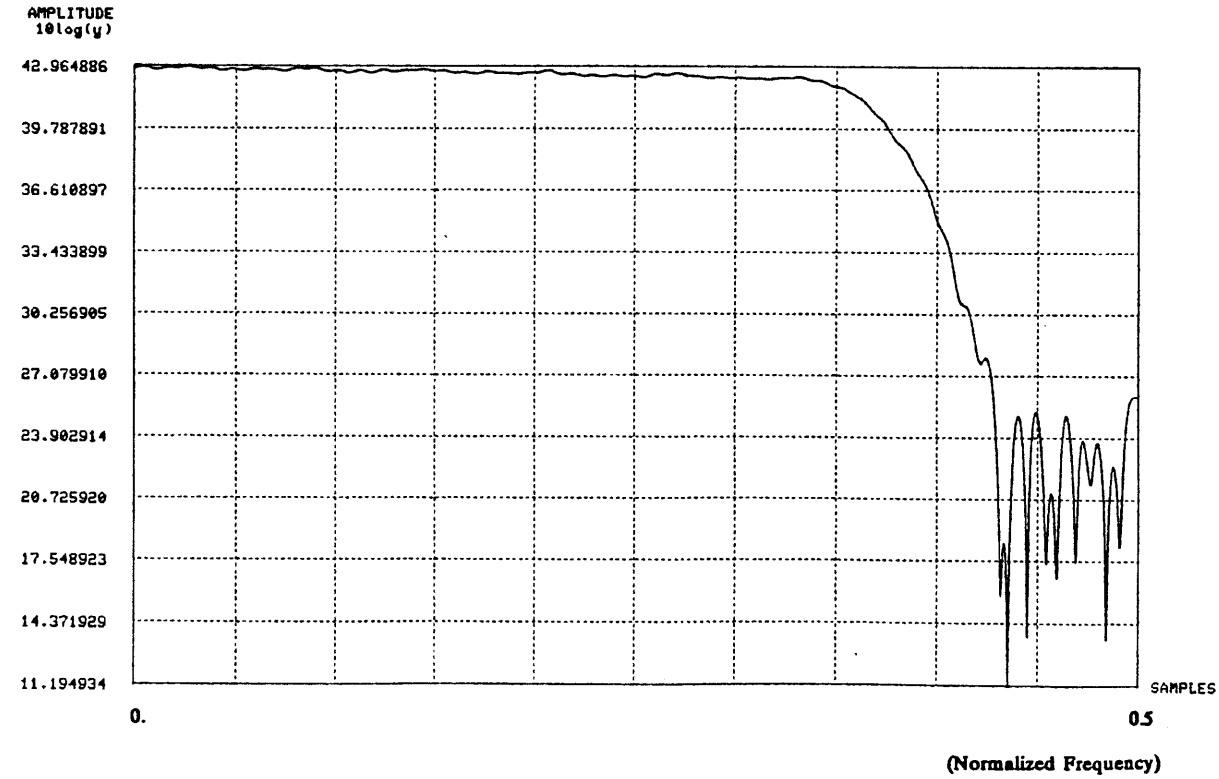


Fig 5.4 (a) The impulse response of the anti-aliasing filter The input is a 100 μ sec wide rectangular pulse, generated by a pulse generator



(b) The 4096 point Discrete Fourier Transform magnitude of the anti-aliasing filter impulse response

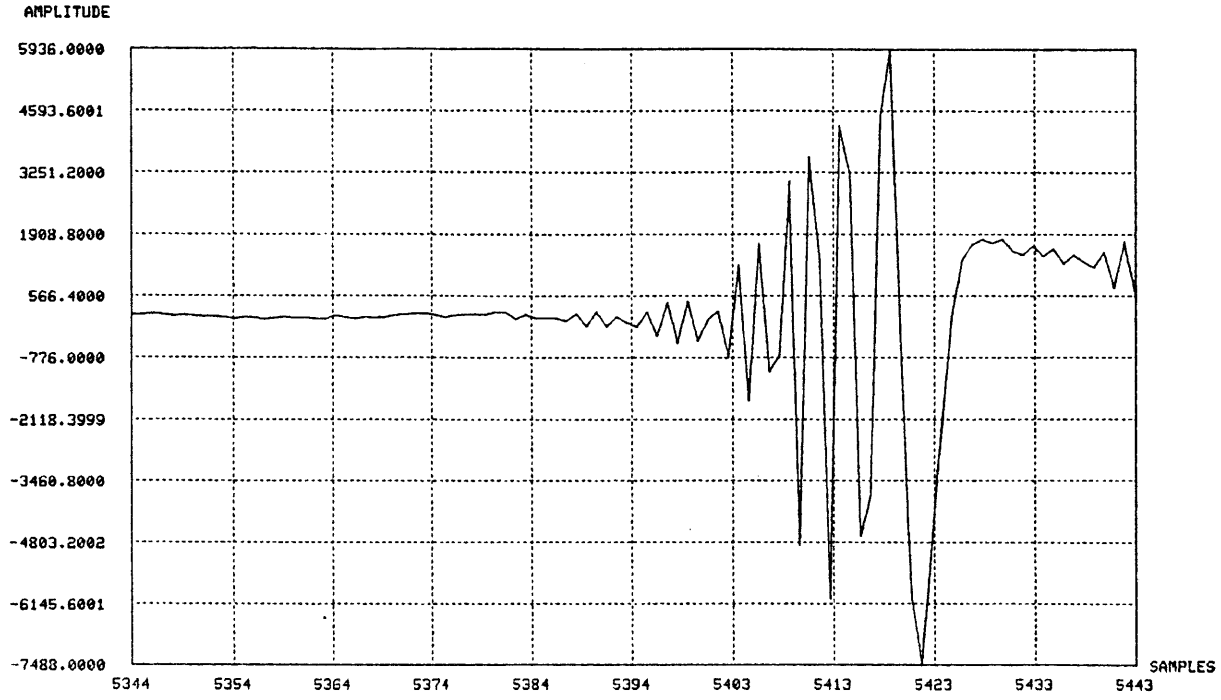
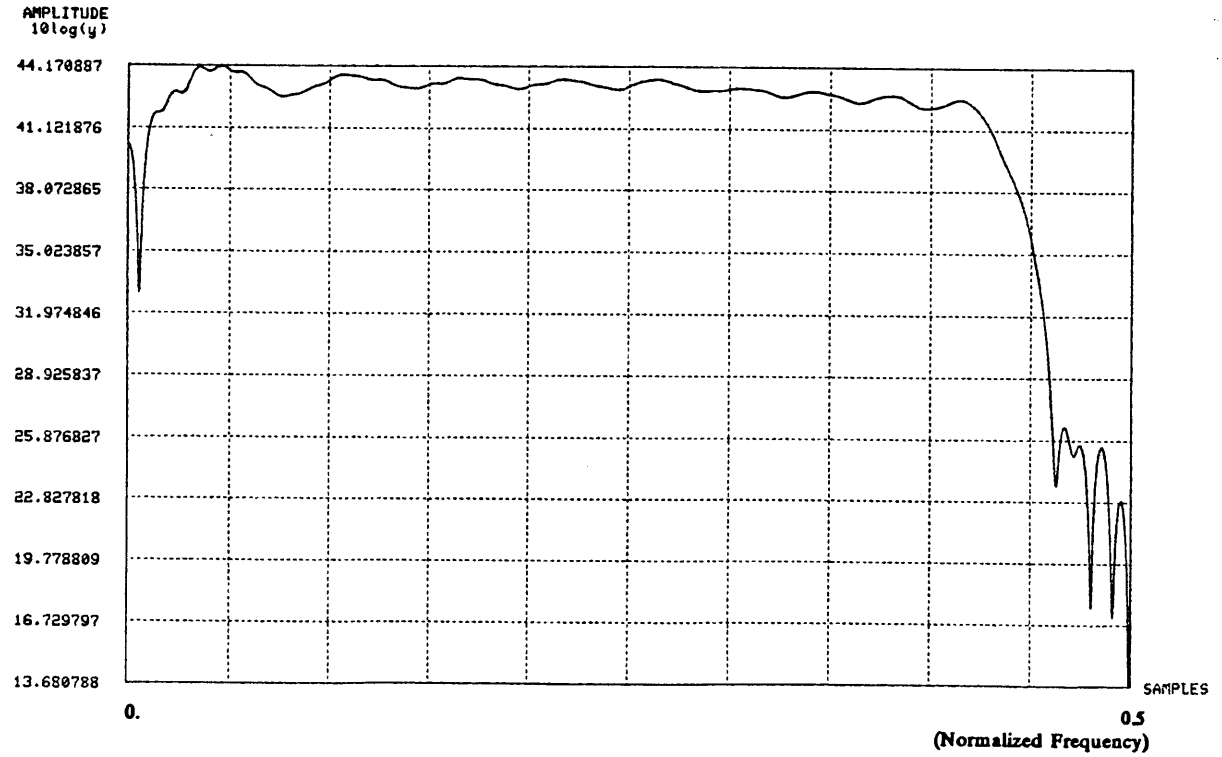


Fig 5.5 (a) The input pulse picked up from the Fig 5.3



(b) The 4096 point DFT magnitude of the input pulse

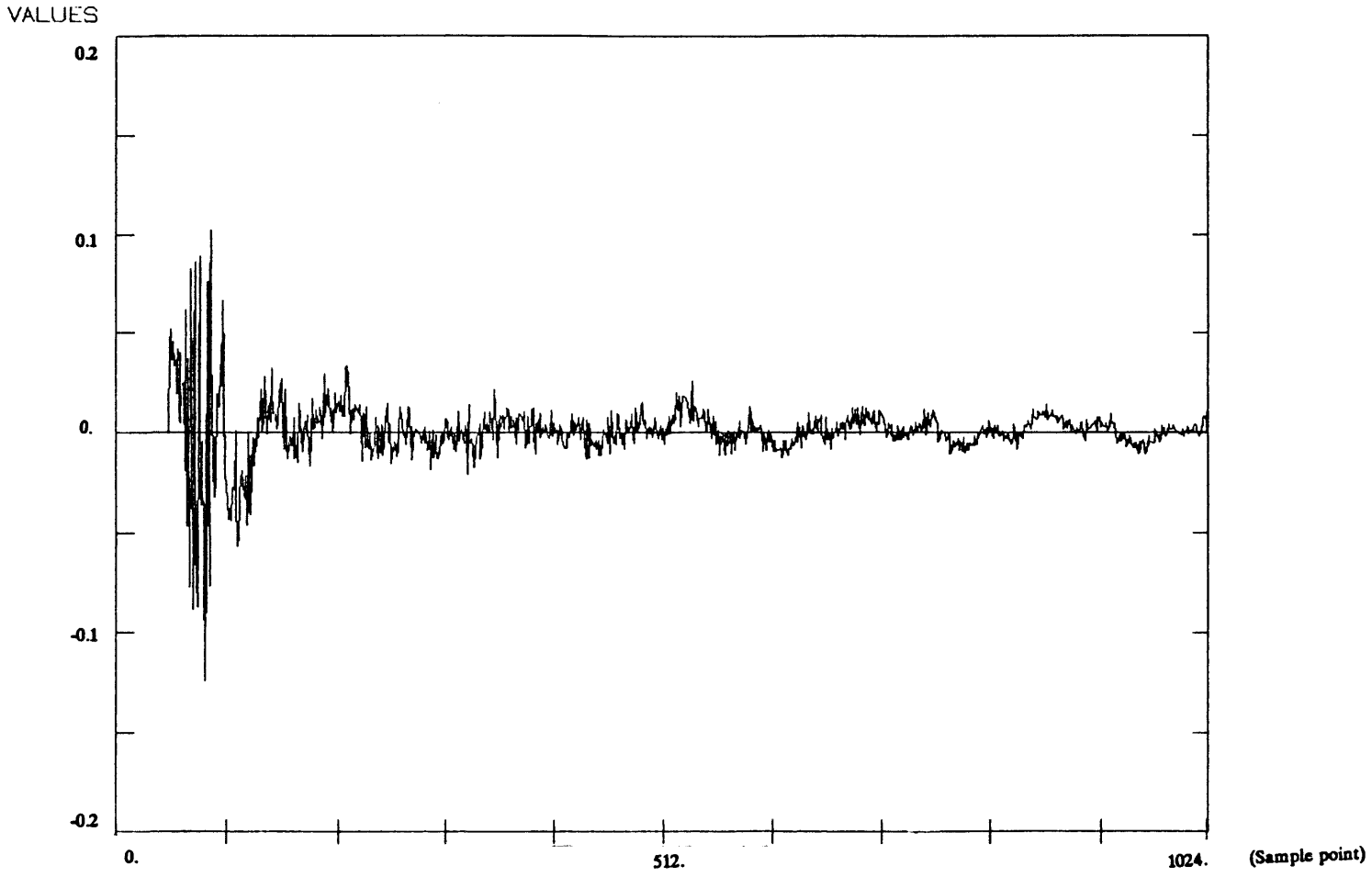
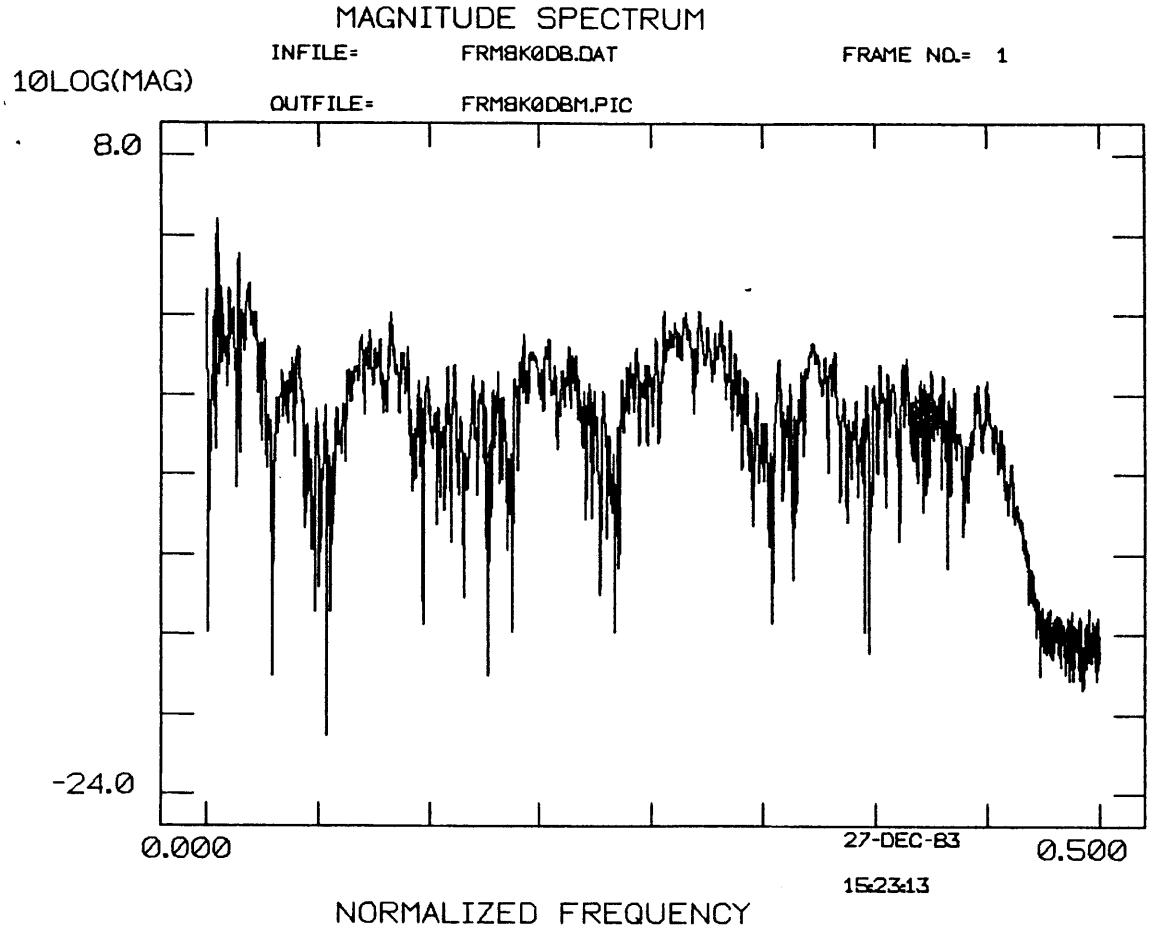
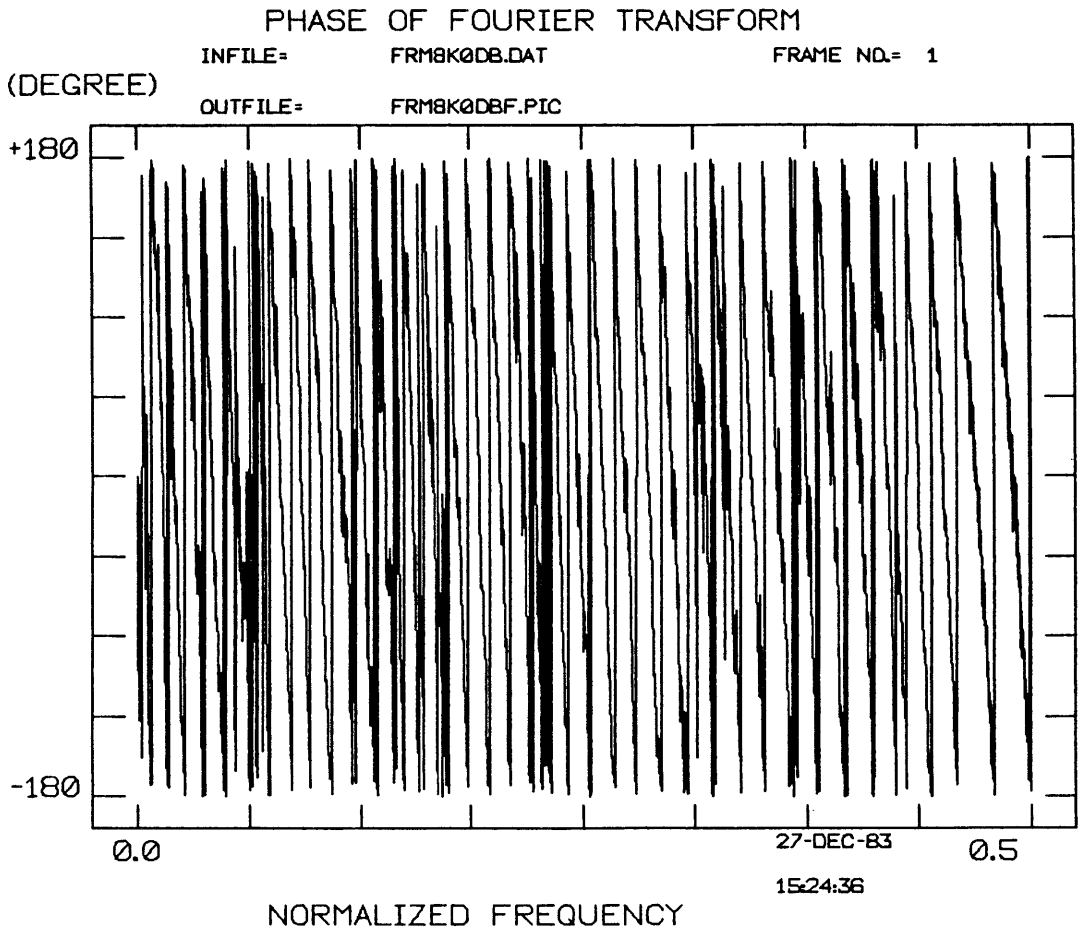


Fig 5.6 (a) The 1024 point room impulse response



(b) The 8192 point DFT magnitude of the 1024 point room impulse response



(c) The 8192 point DFT phase of the 1024 point room impulse response

and output of the reverberation system were recorded on each channel of a stereo audio tape, we mixed both signals into a single channel, processed with an anti-aliasing filter, and digitized it at a 10 KHz sampling frequency with 12 bits. The purpose of mixing the input pulse and the output response is to obtain information on the room delay time. The input impulse is so short that mixing does not corrupt the room impulse response and it is easy to calculate the initial delay time from the beginning of the impulse input to the beginning of the room response.

Results

Fig 5.3 shows the data sequence after digitizing the mixed signal, containing the input impulse and its response. The reason why the first signal, the input impulse, does not look like an impulse, seems to be due to the anti-aliasing filter frequency characteristics. In order to confirm this explanation, we fed a 100 μsec wide impulse, generated by a pulse generator, into the input of the anti-aliasing filter, and measured the response. Fig 5.4 shows the resultant 256 point impulse response of the anti-aliasing filter used in digitizing and the 10 log magnitude of its 4096 point DFT. On the other hand, Fig 5.5 represents the 50 point input pulse picked up from Fig. 5.3 and the 10 log magnitude of its 4096 point DFT. This "pulse" has a transform whose magnitude is almost flat up to 5 KHz, but whose phase is non-zero. It means that this input sequence is desirable as an impulse and that the resultant response to this pulse can be regarded as a correct impulse response except for phase. Comparing Fig 5.4 with Fig 5.5 , one concludes that both have similar time sequences, suggesting that the cause of the shape change in the pulse sequence is the characteristic of the anti-aliasing filter.

Fig 5.3 suggests that the impulse response in this particular room has significant energy in the first 0.1 sec (1000 points), and that beyond 0.1 sec no meaningful data are

present, except for 60 Hz hum noise contamination. Therefore we choose the first 1024 points as the impulse response, discarding the rest of data,

Fig 5.6 (a) shows only the dominant 1024 point room impulse response after carefully eliminating the input pulse portion and truncating beyond the 1025 th point. The initial time delay due to the direct path distance is 4.9 msec. The 4096 point DFT of this 1024 point sequence in Fig 5.6 (a) is illustrated in Fig 5.6 (b),(c). The fact that the impulse response is approximately 1024 points, implies that the required frequency resolution of the transfer function is at least 10 Hz.

Strictly speaking, the impulse response shown in Fig 5.6 (a) represents the combined impulse response of the entire audio chain, including the audio amplifier, the speaker, the microphone system, the timing marker, the pure room reverberation system, the anti-aliasing filter, and the transmitting cables. Since in the real broadcasting situation, $h(n)$ should represent the overall system function, we must also model the overall system as shown in Fig 5.2, instead of the pure room reverberation characteristics.

5.3. Speech data acquisition

5.3.1. Speech samples for synthetic experiments

The reference speech is 14 sec long with sentences in Japanese uttered by a female. The newsman's speech is another 14 sec long with sentences in English uttered by a male. These two sentences are completely different so that they may be considered uncorrelated signals. Two speech signals are recorded on each channel of a audio tape separately, then digitized into files on the computer system respectively. Each speech signal is first low-pass filtered to 4690 Hz to prevent aliasing, then sampled at a 10 KHz sampling frequency with a 12 bit analog-to-digital converter.

In the experiments, the primary speech signal (the microphone signal) is synthesized by adding the newsman's speech to a version of the reference speech filtered with a simulated room acoustic reverberant transfer function. Two types of synthetic transfer functions are used to systematically exercise the noise canceler. One is merely a 32 point delayed delta function $\delta(n-32)$, the other is a complicated 1024 point long impulse response, which was experimentally measured in a real room. A detailed description of how to measure the actual room transfer function was given in the previous section 5.2.

5.3.2. Speech samples for actual data

In this second category, the primary speech is collected in a real experiment. The experiment performed in a confined room.

The room used for the experiment is a office room of the size approximately $10' \times 20' \times 10'$, with plaster walls, a tile floor, and an acoustic paneling ceiling.(See section 5.2.) Because of the substantial sound absorption at the ceiling and walls, the reverberation time (the duration of an impulse response) of the room was rather short, around 0.1 sec.

Although $h(n)$ is not known exactly, it ought to be similar to the 1024 point transfer function used in the synthetic experiments. The reason is that the primary speech here was recorded in the same room with the exactly the same audio equipment setting, immediately after the experiment for the acquisition of the 1024 point room impulse response had been performed. Therefore the room environment in the experiment is almost same as described in the section 5.2. Two loud speakers are set in the room. One for the reference speech is placed around 4 feet away from a microphone, whereas the other for the newsman's speech sits at the microphone. (Although in the original broadcasting room problem, the newsman's speech is given by a real person, we substitute a loudspeaker for a real person for the convenience.)

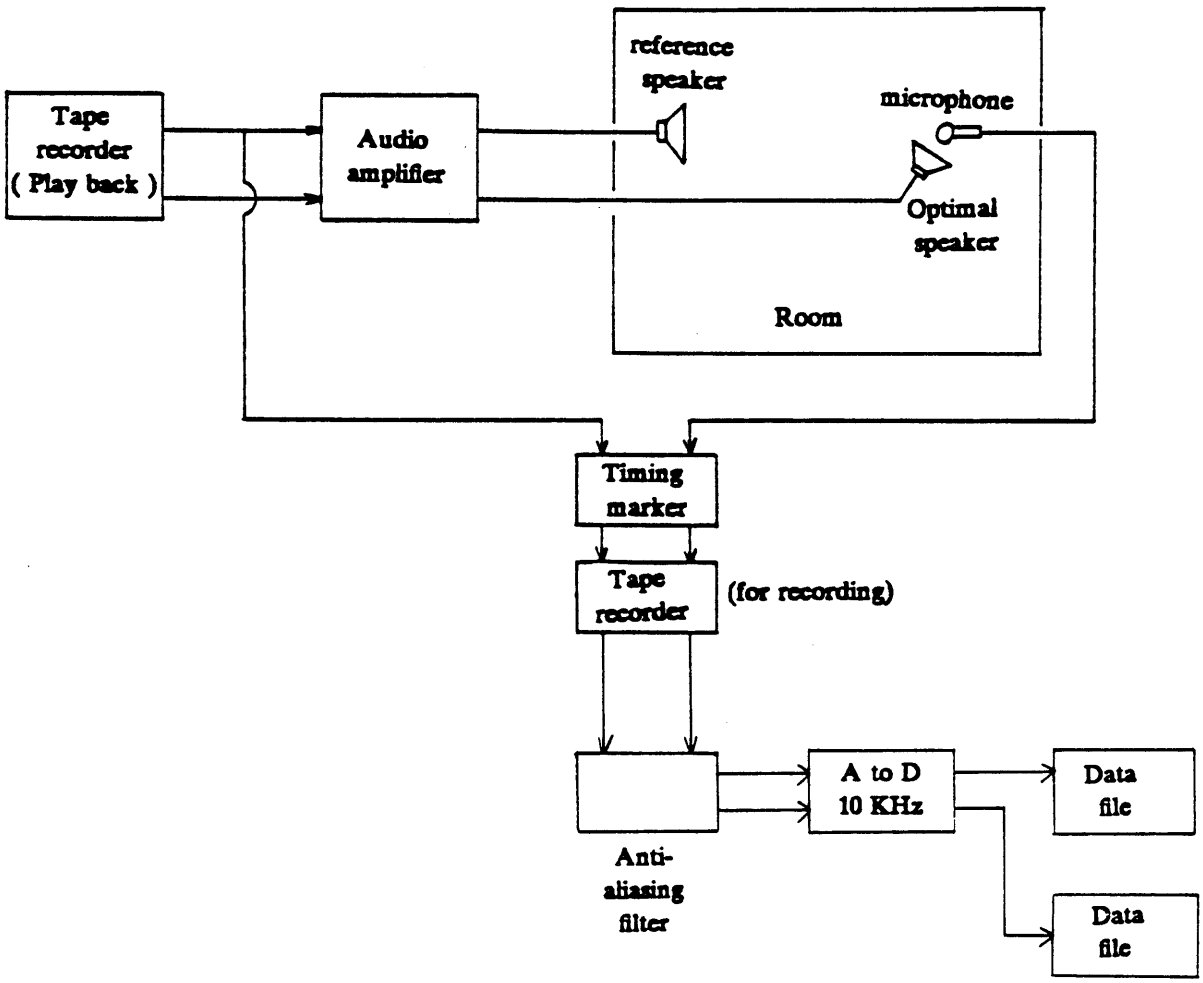


Fig 5.7 The procedure for acquiring speech data in a room

The 14 sec long Japanese sentence on the audio tape is played out from the reference speech loudspeaker at the same time as the 14 sec long English sentence by a male on another audio tape is delivered from the optimal speech loudspeaker. Both sentences are the same ones used for the synthetic data case in the synthetic experiments.

Both the collected microphone signal and the original reference signal are recorded on separate channels of tape. Both the microphone signal and the interference signal are digitized separately and stored in individual files. In order to make it easy to find the time origin for time alignment in both speech files, a timing marker is used to put pulse-shape marks in both channels simultaneously.

The Fig 5.7 shows the overall system for recording primary speech data acquisition. The algorithm in the computer reads the two digitized data files as input data, thus the transfer function $h(n)$ to be estimated in the problem represents the overall transfer function through the whole data transmission path, from the reference file to the primary file, that is, the loud speaker for the reference, the pure room reverberant system, the microphone, the timing marker, the tape recorder, and the anti-aliasing filter.

5.4. Theoretical analysis

Before the actual empirical data are presented, several issues expected to occur are discussed in this section. These issues are of importance especially for the interpretation of the synthetic experimental data. We discuss the behavior of the algorithm in terms of two aspects, the behavior of the filter estimate and the behavior of the estimated newsman's speech.

5.4.1. The behavior of the filter estimate

The effect of the finite length frame on the filter estimate.

In the section 4.5.3 we discussed the convergence property of the filter estimate for the ideal case with approximately infinite length frames. We showed that in this case, with all the data in the past frames contributing to the estimation process, with approximately infinite length frames, without any type of smoothing windows, and without any newsman's voice, the proposed algorithm converges to the true value immediately.

First we study the effect of the finite length frame data on the filter estimate in a practical case. The finite length frame causes two major types of degradation in the filter estimate. They include circular aliasing and the bias effect on the estimate. Using finite length DFT's is convenient for manipulating data, but multiplying or dividing such finite length transforms will cause circular aliasing. The bias effect is due to using a finite length window for sectioning input data into finite frames, a smoothing window for spectra $P_{x_i x_i}(k)$, $P_{x_i \tilde{s}_i}(k)$, and a truncation operation (multiplying by a rectangular window) on the filter estimate after the division of the spectra. The effects cause the filter estimate to approach some biased value, and not the value of the true filter, even if all data in the past are taken into account to estimate the current filter, that is $\alpha \approx 0$. In addition, since α is usually bigger than zero to allow adaptation to possible changes of the transfer function, only a finite amount of data in the past frames contributes to the estimate. Thus the estimate never converges to any value in a strict sense on the frequency point basis. The circular aliasing also causes degradation of the estimate, since it does not provide the exact filter value even if there is no newsman's voice all the time.

However it is expected that the estimate goes toward the true filter on average to some extent, if parameters in algorithm are chosen appropriately so that the effect of the two major causes of degradation of the estimate are reduced. The minimization of time aliasing effect would be attained by setting the data length read for each frame L and the

length of the filter M to satisfy $\frac{M}{L} \ll 0.1$. The bias of the window is moderated by carefully choosing the window. For example, a window with high sidelobes but a narrow main lobe, like a rectangular window, produces a filter estimate with heavy sidelobe leakage near frequencies with large energy in the input signals. On the contrary, a window like a Hanning window has low sidelobes but a broad main lobe, which tends to uniformly distribute a small amount of bias.

The measures defined in the section 4.5.11 are indicators for the global behavior of the algorithm in a sense that they show only the physical quantity summed over all the frequency. Due to the on-average convergence property of the algorithm, they are one kind of data representation for the experiment data.

The rate of improvement in the filter estimate.

If we assume that the result of the convergence property in the section 4.5.3, can be approximately applied to the finite length frame case, then the current frame estimate is expressed in the form of

$$\hat{H}_i(k) = \hat{H}_{i-1}(k) + q_i(k)(H(k) - \hat{H}_{i-1}(k)) + q_i(k)\frac{P_{x_i s_i}(k)}{P_{x_i x_i}(k)} \tag{5.4.1}$$

When the newsman's voice is not present at all, the third term vanishes, then the rate of the improvement in the filter estimate is controlled by $q_i(k)$ where,

$$q_i(k) = \frac{\alpha P_{x_i x_i}(k)/\hat{\sigma}^2{}_i(k)}{(1-\alpha)D_{i-1}(k) + \alpha P_{x_i x_i}(k)/\hat{\sigma}^2{}_i(k)} \tag{5.4.2}$$

$$= \frac{\alpha P_{x_i x_i}(k)/\hat{\sigma}^2{}_i(k)}{D_i(k)} \tag{5.4.3}$$

In turn the $q_i(k)$ is determined by two basic parameters, the exponential average parameter α, and the energy of the spectrum of $x_i(n)$.

For a small value of α (≈ 0), the amount of data to be considered for the estimate becomes almost equivalent to all the data in the past frames, and the variance of the estimate becomes small. Thus the estimate converges to some value, thought not necessarily to the true filter owing to the bias factor. However a small value of α allows only very slow adaptation to possible changes of the true filter. Clearly, the trade off is necessary between the adaptive ability of the estimate and the low variance of the estimate, for the selection of the value α. Moreover, the formula for $q_i(k)$ in equation (5.4.2) suggests that the following conjuncture. Assume that the newsman's voice is not present all the time, and also assume that a fair amount of data is observed, so that $D_{i-1}(k)$ has reached to be a steady state. Then, $D_{i-1}(k)$ would be about the same order of magnitude as $P_{x_i x_i}(k)/\sigma^2{}_i(k)$. Then the denominator would become approximately $P_{x_i x_i}(k)/\sigma^2{}_i(k)$. This leads to $q_i(k) \approx \alpha$, which implies that α determines the speed of improvement.

The effect of the presence of the newsman's voice on the filter estimate.

Now we consider the effect of the newsman's speech to the rate of the improvement. In this case, the current estimate is given in equation (5.4.1). The speed of the improvement will be still controlled by the second term of the equation, and will be determined by α and ratio of the reference energy to the newsman's speech energy, $P_{x_i x_i}(k)/P_{s_i s_i}(k)$. Suppose that at a frequency k the current frame has large energy in the newsman's speech $S_i(k)$, or small energy in the reference speech $X_i(k)$. Then the denominator is dominated by the non zero first term, whereas the numerator goes near zero, leading to $q_i(k) \approx 0$. Therefore, the algorithm does not rely on the current information at the frequency k. On the other hand, if the current frame has small energy of the newsman's voice, but has large energy of the reference at a frequency k, then $q_i(k)$ becomes near 1, which is the best improvement rate. The third term is the fluctuation term to the filter estimate. Since

$x_i(n)$ is uncorrelated with $s_i(n)$, then the expectation of the third term is zero, whereas its variance depends on the number of the frames used for the current frame estimate. The variance of one frame is roughly $\frac{P_{s_i s_i}(k)}{P_{x_i x_i}(k)}$ 0, and for $\frac{1}{\alpha}$ frames, it goes down to zero in proportion to $\frac{1}{\alpha}$. Thus the third term never vanishs except when $\alpha = 0$

Measures γ, and ρ.

To visualize the rate of the improvement in the filter estimate, γ and ρ calculated for each frame will be plotted in figures with time on the horizontal axis. ρ is the noise reduction, and the filter estimation achieves smaller filter deviation γ at frequencies where $\frac{P_{x_i x_i}(k)}{P_{s_{s_i s_i}}(k)}$ has large energy than at frequencies where $\frac{P_{x_i x_i}(k)}{P_{s_i s_i}(k)}$ has small energy. If the reference has a flat spectrum, then γ is equivalent to ρ. But if the reference has non flat spectrum, then ρ achieves better values than γ.

5.4.2. The behavior of the estimated newsman's speech

Since the estimated filter never becomes the true filter in the practical case, then the estimated newsman's speech $\hat{s}_i(n)$ in the algorithm consists of the true newsman's signal $s_i(n)$ and the residue error due to filter mismatch, $(h(n) - \hat{h}_i(n)) * x_i(n)$, i.e.,

$$\hat{s}_i(n) = s_i(n) + (h(n) - \hat{h}_i(n)) * x_i(n) \qquad (5.4.4)$$

One possible problem caused by the residue error may happen when the reference speech to be removed is small, but a large mismatch in the filter estimate exists. In that case, although $x_i(n)$ is small, the residue error grows large. Consequently, the residue error may become larger than the original reference signal in the primary $\hat{h}_i(n) * x_i(n)$. This means that the algorithm does not subtract the reference signal (noise) but actually adds additional noise signal to the primary. It results in worse SNR_{out} than SNR_{pri}, and positive

values of noise reduction.

The effect of the residue signal on the sound quality of the estimated newsman's speech may result in artifacts in the newsman's speech, not just the superimposed noise to that speech. The estimated filtered $\hat{s}_i(n)$ is expressed as

$$\hat{s}_i(n) = y_i(n) - \hat{h}_i(n) * x_i(n) \tag{5.4.5}$$

it is not obvious that true newsman's signal $s_i(n)$ can be degraded by the estimation and the filtering process, since the above equation shows that $s_i(n)$ is never processed by the estimated filter. However the fact that the filter estimate $\hat{h}_i(n)$ depends on not only $x_i(n)$ but also $y_i(n)$, suggests that $s_i(n)$ in $y_i(n)$ affects the estimate $\hat{h}_i(n)$ implicitly. This argument is supported by the proposed algorithm which pre-estimates the power spectrum $\hat{\sigma}^2{}_i(k)$ first from $x_i(n)$ and $y_i(n)$, then uses the pre-estimate for the filter estimation. The extreme example is the case that only one frame of $x_i(n)$ and $y_i(n)$ are used for the filter estimate as follows :

$$\hat{H}_i(k) = \frac{P_{x_i y_i}(k)}{P_{x_i x_i}(k)} \tag{5.4.6}$$

$$= \frac{Y_i(k)}{X_i(k)} \tag{5.4.7}$$

Then the estimated newsman's voice is ;

$$\hat{S}_i(k) = Y_i(k) - \hat{H}_i(k) X_i(k) = 0 \tag{5.4.8}$$

It apparently means that all the newsman's voice has actually been incorporated into the filter estimate. Thus the resultant filter estimate completely degrades the newsman's speech to zero. This argument suggests that the newsman's voice may be affected even when averaging over multiple frames is used.

Another problem is that if the true filter contains initial delay time, the circular aliasing in the estimated filter degrades the output speech. In time delay portion, the estimated

values are not zeros, which causes reverberant effect in the residue signal.

5.5. Empirical Result and discussions

5.5.1. Setting values of parameters for experiments

We set the values of parameters in the algorithm to be suitable for a 1024 point room impulse response, even in the case of the 32 delayed delta function, since we are interested in using the 32 point delta function as an analytical simple example for more complicated filter cases. The parameter values chosen appropriate for a 1024 point room impulse response are called a "standard setting" in this whole chapter. They are; the FFT size is N = 8192 points, the data read in each frame is L = 4096 points sectioned by a 4096 point rectangular window, the filter length M = 1024, the frame shift F is 2048 points (50% overlapped). The correlation window is a 4097 points Hanning window, α=0.25, the interpolation range T is 1024. The reason for assigning those values to the standard setting parameters is as follows:

Since we assume that the true filter length could be up to 1024 points, M is 1024. To reduce the end effect of sectioning data, it is ideal to choose $L \gg 10 \times M$, leading to L=10240 points. However the most dominant part of the 1024 point impulse response is the first 400 points. This suggests that end effects will be caused primarily by the first 400 points. Short data frames are also desirable to make the assumption of stationarity of the transfer function in the frame valid. To compromise between these two factors, we chose L = 4096 (L = 4M). A rectangular window for sectioning data is used although a Hanning window might have led to more uniformly distributed bias.

To avoid the circular aliasing in the multiplication between two DFT's of 4096 point time sequences for the calculation of spectra $P_{x_i s_i}(k)$ and $P_{x_i x_i}(k)$, a 8192 FFT buffer is

appropriate. (In practice, smaller FFT buffer sizes may be used, since the 4000 th point correlation lag is assumed to be small enough to cause negligible circular aliasing.) However this 8192 FFT buffer is still not long enough to theoretically avoid the circular aliasing when the division of two 8192 DFT is computed to obtain $\hat{H}_i(k)$. In practice, this circular aliasing seems to be significant, and it never vanishes no matter how big the FFT size would be.

As for the smoothness of the filter, to estimate a 1024 point filter sequence, one needs to retain frequency resolution of at least 10 Hz. For extra margin, one may keep 5 Hz resolution during the calculation. Since the frame data length is 4096, containing 2.5 Hz frequency resolution, one need to smooth the filter estimate in several ways in order to reduce the variance of the estimate.

One way is to smooth spectra $P_{x_i s_i}(k)$ $P_{x_i x_i}(k)$. This is realized by applying the 4097 point Hanning window to the inverse DFT of these spectra. This Hanning window retains 5 Hz frequency resolution during the calculation. It corresponds to averaging adjacent 5 DFT coefficients in the 8 K FFT buffer. It equivalently offers 5 Hz frequency resolution. This 5 point cosine averaging is short enough to allow for one period of phase change within $\frac{\pi}{50}$ due to the linear phase factor of about 100 points delay. Thus 5 point smoothing would not cause degradation in phase information.

Another way to achieve smoothing is to apply a 1024 point rectangular window to the inverse DFT of the current filter estimate $\hat{H}_i(k)$, truncating the sample points beyond the 1025 th point. This cuts off a large amount of the circular aliasing, causing substantially reduced reverberation of the residue signal in the output.

The third smoothing method is controlled by the value of α, which also affects the adaptive rate of the algorithm. α controls the number of frames whose data the current

filter estimate is based on. In the experiments, α is chosen as 0.25, thus approximately the data of the last four frames are relied on by the estimation for each frame. This value is rather large for a time-invariant filter.

The size of shift of the rectangular window used for reading input data for the current frame , F, is 2048 (50 % overlap). The values of F and α determine the rate of the adaptation, the time which it takes to adapt itself from an old state to a new state, when the true filter changes instantaneously at a certain time. This is approximately $\frac{F}{\alpha}$ sample points. In this particular value setting, the time of the adaptation rate is about 0.8 sec.

The interpolation range T is determined to smooth connection of two adjacent frame outputs, since the estimated newsman's speech from the different filter estimates for two adjacent frames may not often connect with each other smoothly, causing a clicking type of noise. The value T ought to be chosen in order to suppress the noticeable clicking noise. In the experiments, T is defined as 1024. It means 512 points at both ends of 2048 point output for each frame are interpolated with corresponding data in adjacent frames.

In experiments shown here, the parameter values are a standard setting, unless it states otherwise explicitly.

5.5.2. The synthetic data Test 1

In test 1, we examine cases where no newsman's speech is present in the primary speech for the whole time duration. This study gives insight into how the proposed algorithm adapts to the ultimate transfer function from the initial state of $\hat{h}(n) = 0$. Also it shows the ultimate performance of the estimation process for the room transfer function, since no disturbance signal (newsman's speech signal) is present to confuse the estimator.

32-point delta function

A 32-point delayed delta function has a transform with uniform magnitude of 1 and linear phase. (See Fig 5.8 (a),(b),(c).) The important issues in analying the algorithm performance are to measure how windowing and aliasing degrade the estimates as well as examining the adaptation rate of the algorithm when $\hat{h}(n)$ stars from zero.

The result of the algorithm is given in the case with $h(n)=\delta(n-32)$ and no newsman's speech $s(n)=0$. The primary speech is the 32 point delayed version of reference speech. The parameters for the algorithm (N, L, etc.) are chosen as standard values. SNR in the primary is minus infinity, since there is no newsman's speech. The noise canceler should get rid of the 32 point version of the reference speech from the primary, leaving nothing in the output.

The noticeable phenomenon due to the end effect in the result of this section would be a slow speed of the improvement in the filter estimate. If the frame length is approximately infinite without any smoothing window, then the first frame estimate can hit on the values of true filter, and then the rest frame estimates locks on that true value for ever. Thus in this case it converges immediately.

The Fig 5.9 shows that the deviation of the filter estimate γ, decreases monotonously. According to the derivation of the filter estimate and ignoring end effects caused by finite frame length, the filter estimate error $\Delta H_i(k) = H(k) - \hat{H}_i(k)$ is

$$H(k) - \hat{H}_i(k) = H(k) - \left\{ \hat{H}_{i-1}(k) + q_i(k)(H(k) - \hat{H}_{i-1}(k)) \right\} \tag{5.5.1}$$

$$= (1 - q_i(k))(H(k) - \hat{H}_{i-1}(k)) \tag{5.5.2}$$

where

$$q_i(k) = \frac{\alpha P_{x_i x_i}(k)/\hat{\sigma}^2{}_i(k)}{(1-\alpha)D_{i-1}(k) + \alpha P_{x_i x_i}(k)/\hat{\sigma}^2{}_i(k)} \tag{5.5.3}$$

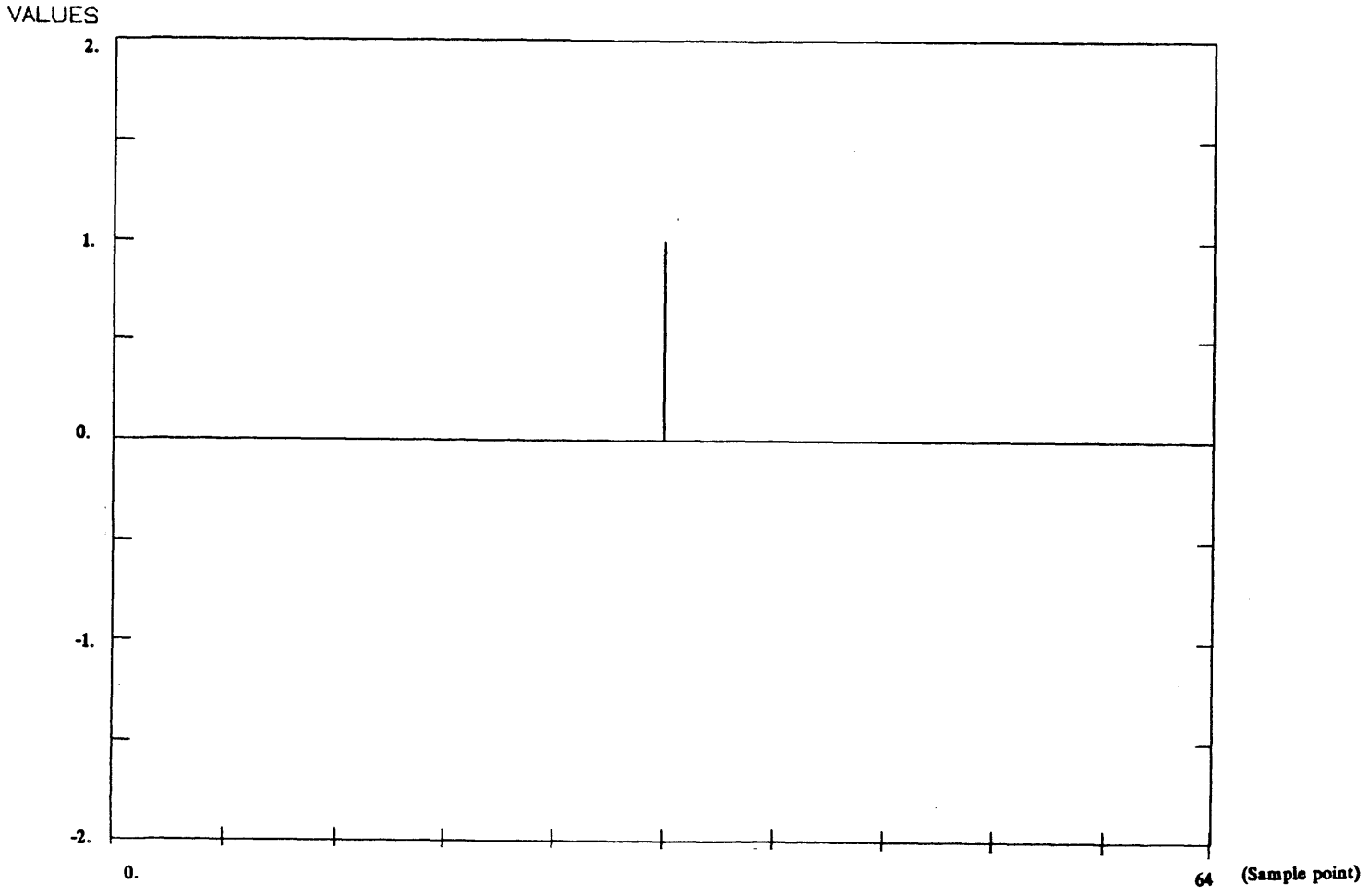
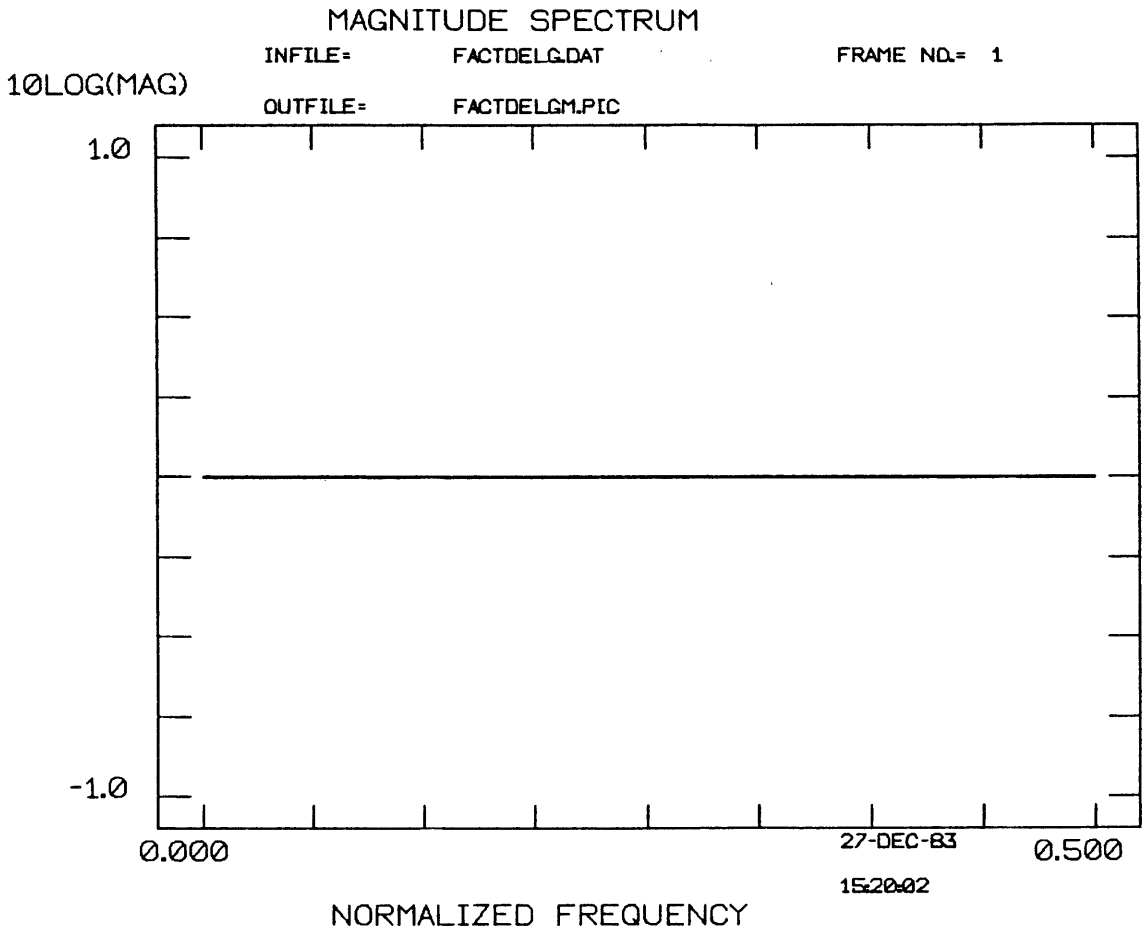
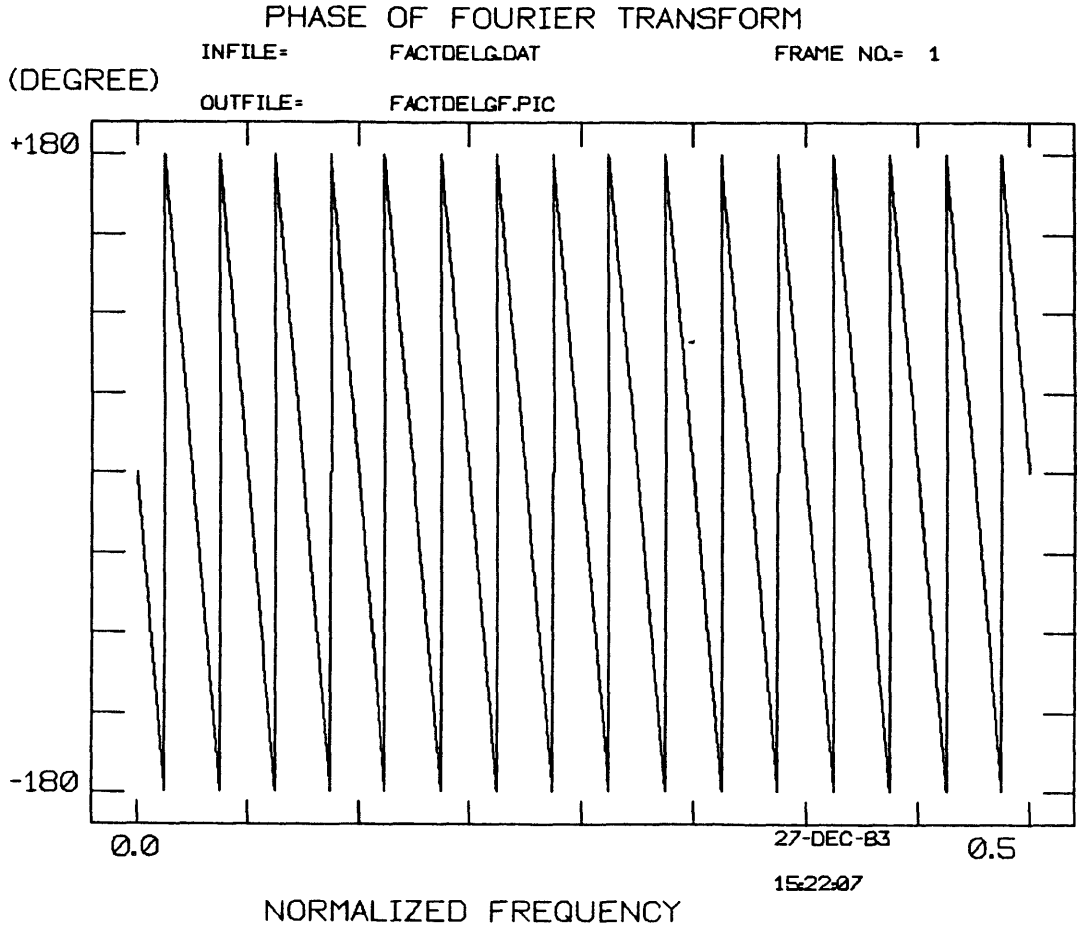


Fig 5.8 (a) The 32 point delayed delta function $h(n) = \delta(n-32)$



(b) The 8192 point DFT magnitude of the 32 point delayed delta function



(c) The 8192 point DFT phase of the 32 point delayed delta function $h(n) = \delta(n-32)$

Therefore the frame estimate error for the current frame is

$$\Delta H_i(k) = (1-q_i(k))\Delta H_{i-1}(k) \tag{5.5.4}$$

The measure γ_i for the current frame is expressed in terms of the $\Delta H_i(k)$

$$\gamma_i = 10\log_{10}\left\{\frac{\sum_{k=0}^{N-1}|\Delta H_i(k)|^2}{\sum_{k=0}^{N-1}|H(k)|^2}\right\} \tag{5.5.5}$$

$$= 10\log_{10}\left\{\frac{\sum_{k=0}^{N-1}(1-q_i(k))^2\,|\Delta H_{i-1}(k)|^2}{\sum_{k=0}^{N-1}|H(k)|^2}\right\} \tag{5.5.6}$$

Just assume both reference and desired speech are white noise, then on average

$$q_i(k) = \alpha = \text{constant} \tag{5.5.7}$$

Thus,

$$\gamma_i = 10\log_{10}(1-\alpha)^2 + \gamma_{i-1} \tag{5.5.8}$$

The improvement in the filter deviation for one step (from the i-1 th frame to the i th frame) is $10\log_{10}(1-\alpha)^2$. This represents the absolute value of the tangent in the filter deviation curve at the frame number i. Since α=0.25, the improvement is approximately -3 DB per frame. For speech case, however, the theoretical computation of the rate of improvement on γ, is impossible, because the filter deviation varies with how much energy the reference signal has for the entire period of observation, and that energy depends on utterances of speech. Nevertheless, the asymptotic behavior of the rate is expected be also a constant α. According to the result shown in the Fig 5.9, the filter deviation begins with -8 DB at the first frame, and achieves the largest improvement from the first frame to the second frame. Beyond the second frame the behavior of γ shows that the tangent of the curve decreases and seems to approach a constant slope of approximately -0.1 DB per

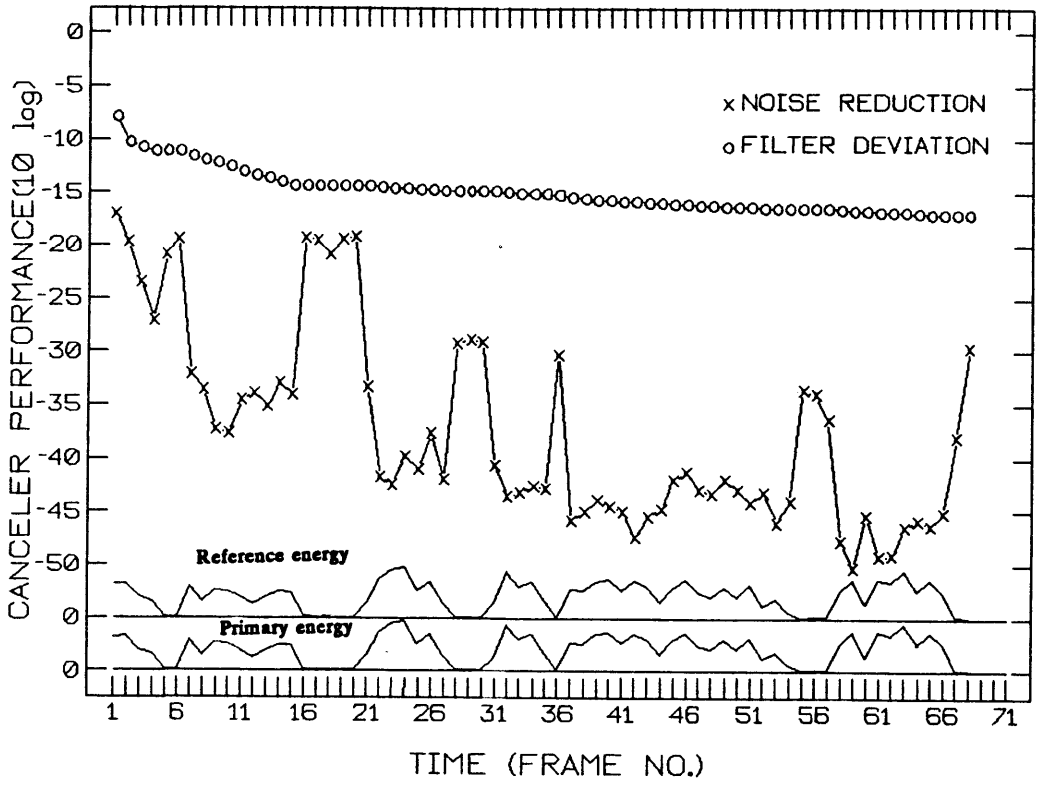
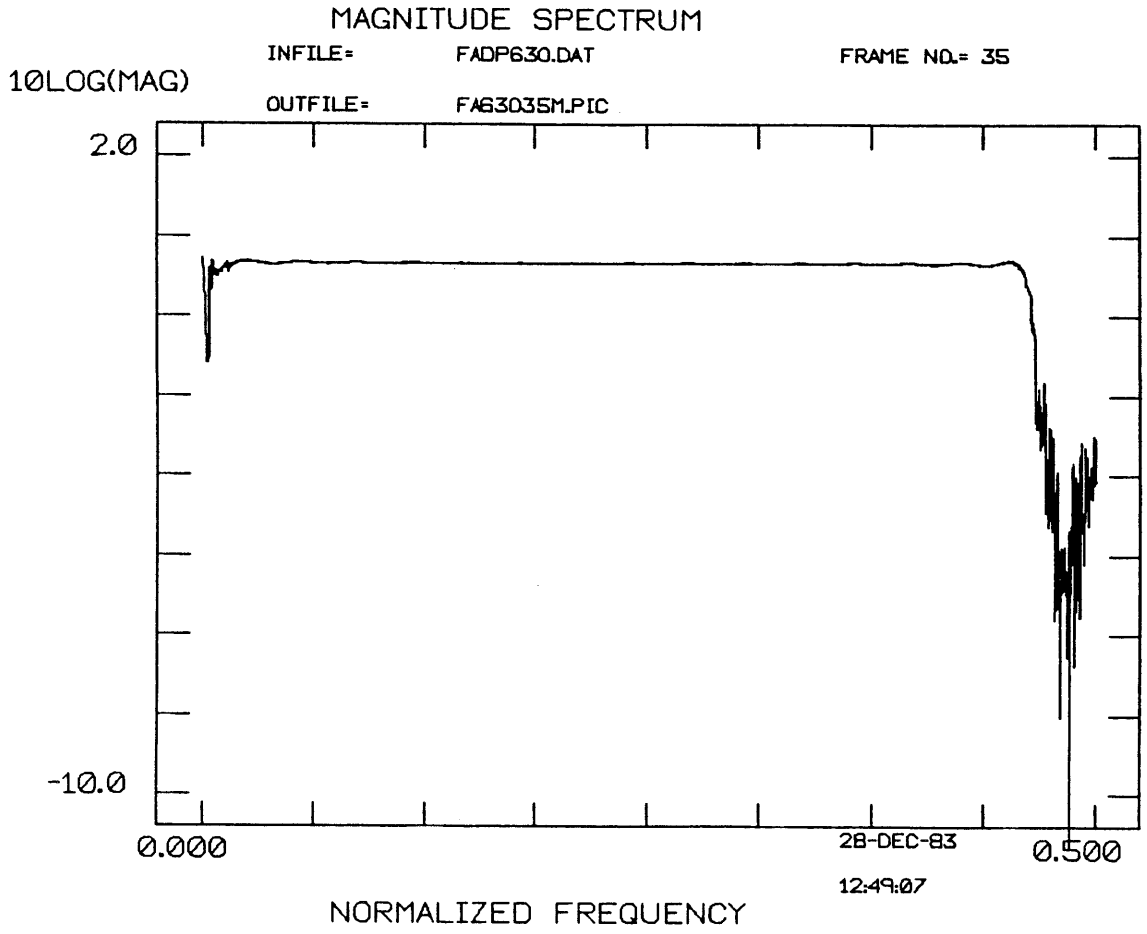
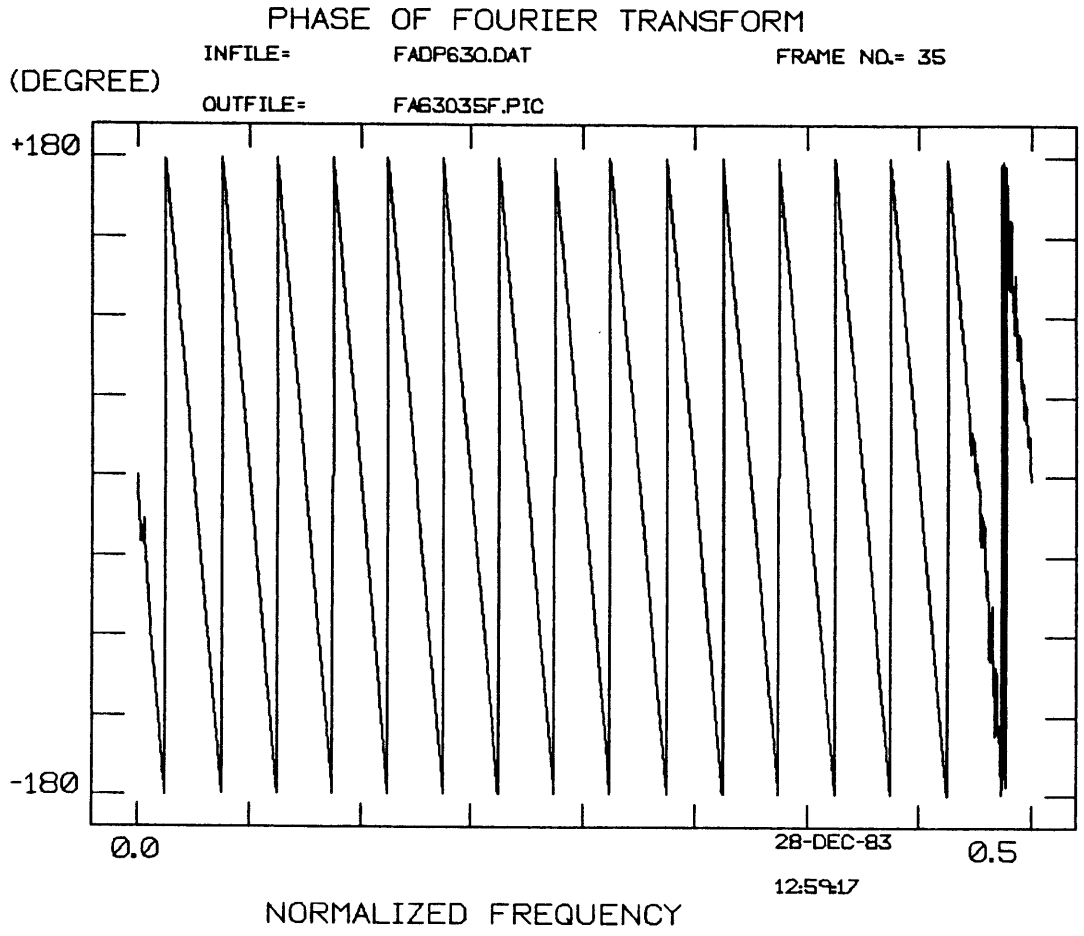


Fig 5.9 The filter deviation γ, the noise reduction ρ in a 32 point delayed delta function $\delta(n)$ with no newsman voice



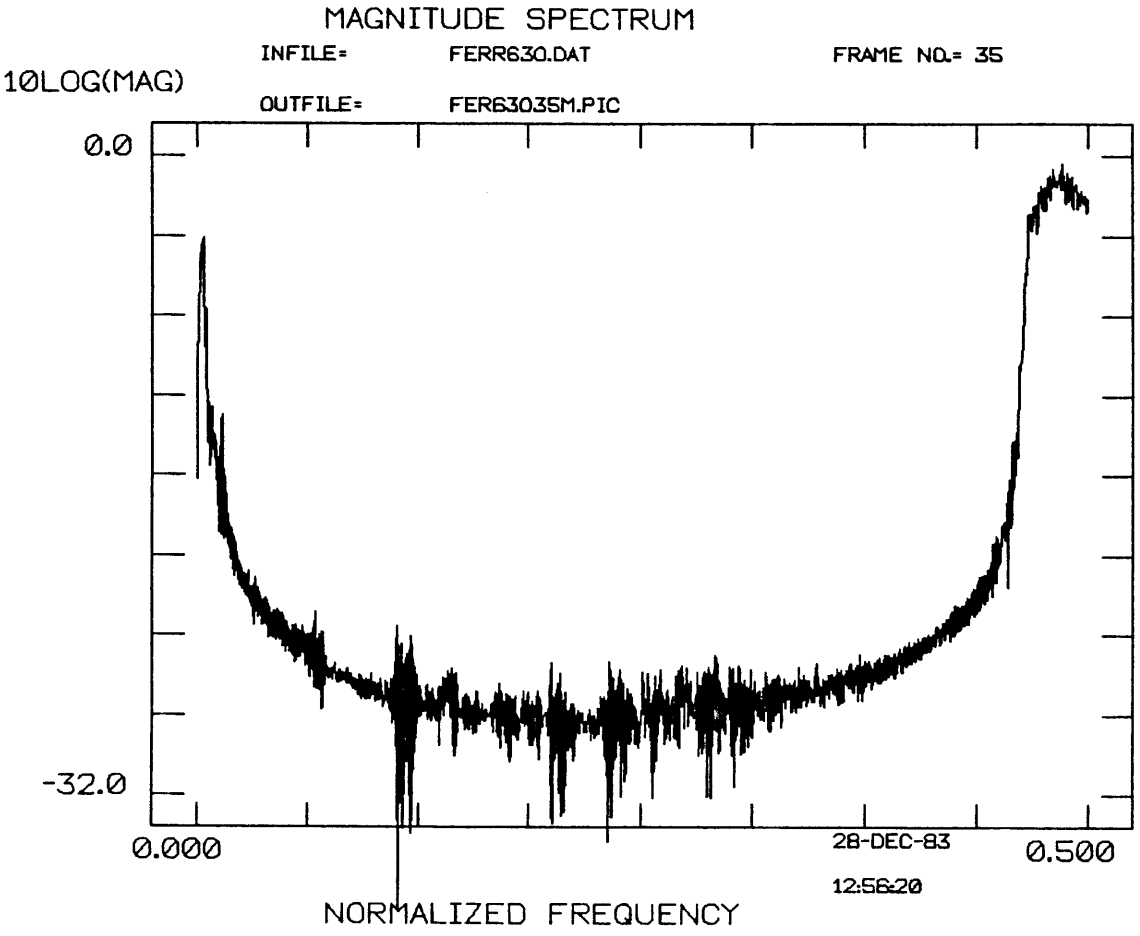
(a) The magnitude of the filter estimate in the 35 th frame

Fig 5.10 The behavior of the filter estimate for the true filter $h(n) = \delta(n-32)$ in the 35 th frame in the case of no newscaster's speech



(b) The phase of the filter estimate in the 35 th frame

Fig 5.10 The behavior of the filter estimate for $h(n) = \delta(n-32)$ in the 35 th frame in the case of no newscaster's speech



(c) The filter estimate error $\Delta H_i(k)$
$= H(k) - H_i(k)$ in the 35 th frame

Fig 5.10 The behavior of the filter estimate for $h(n) = \delta(n-32)$ in the 35 th frame in the case of no newscaster's speech

frame. Consequently it attains -17 DB at the end of the 14 sec. Apparently, the rate is much slower than the ideal noise case.

The result shows that the noise reduction curve varies rapidly frame by frame. The reason is that the contribution of the reference speech in the primary, that is the denominator of ρ in the following equation

$$\rho_i = 10\log_{10}\left\{\frac{\sum_{k=0}^{N-1}|\Delta H_i(k)X(k)|^2}{\sum_{k=0}^{N-1}|H(k)X(k)|^2}\right\} \tag{5.5.9}$$

changes drastically, even though in the numerator the filter error $\Delta H_i(K)$ is not changing very rapidly. This effect is noticeable in frames with small energy of the reference speech. If there is no reference filtered signal in the primary to be eliminated, then the noise reduction is poor. Notice that in frame number 5, 6, from 16 through 20, from 28 through 30, 36, from frame 55 to 57, there happen to be no interfering speech (reference speech), and the noise reduction ρ is mediocre. Whereas in frames where interfering speech is present, the noise reduction improves, because the noise canceler tries to remove the substantial interfering speech from the primary. However, the overall trend of that curve goes down, and achieves -45 DB noise reduction at the end of 14 sec speech, because the filter estimate becomes better. The algorithm adapts the filter estimate to the true filter as the amount of the observation (input data) increases with time.

An informal subjective listening evaluation of the output speech indicates that the reference speech is considerably removed and that especially in the latter half of the 14 sec speech, it leaves almost nothing but the quantization noise from computation in the output speech. However, in the first 7 sec of speech, the remaining signal in the output (the residue signal) contains a small amount of reverberating sound. Probably the cause is

the circular aliasing produced by non-zero sample points beyond the 34 th point to the 1024 th point in the filter estimate, since the estimated filter length M is set to 1024. If one set the filter length in the algorithm to the correct value, M=32, then γ and ρ improved substantially.

One notices that during frames where the reference speech is almost zero, like frame number 5 and 6, etc, the filter deviation retains its previous values. This is because the algorithm performs a threshold test on the energy of the reference speech. If the energy of the reference speech is below the threshold value set up in advance, then the algorithm does not use the information from that frame for the filter update and just latches the previous frame estimate. Here the threshold is set to 125. (As reference, the usual energy of speech as compared with this algorithm is around 400.)

Fig 5.10 shows the filter estimate in the frequency domain in the 35 th frame. Fig 5.10 (a) is the magnitude of the filter estimate in a frequency domain, and Fig 5.10 (b) is the phase. Fig 5.10 (c), show the error of the estimated filter from the real filter. Fig 5.10 (c) is the magnitude of the error. Most of the error is caused at both extreme ends of the frequency region, below 200 Hz and above 4690 Hz, where no substantial speech energy exists. Therefore most of the error does not do harm to the noise canceler performance, represented in the measurement ρ.

The 1024 point room impulse response

The next example is a 1024 point room impulse response actually measured in a room. The transfer function is shown in Fig 5.6 It shows the impulse response of the filter in the time domain. Note that it is zero during the initial 49 points and its peak occurs around sample point 100. This impulse response was measured in the room reverberation experiment described in section 5.2

The phase and magnitude of the filter's frequency response are shown in Fig 5.6 (b),(c). In order to simplify the comparison with the 32 point delayed delta function, the overall gain of this 1024 point room impulse response is chosen so that the filtered reference signal has approximately same energy as the reference signal.

Fig 5.11 shows the noise reduction ρ and the deviation of the estimate filter γ. The filter deviation γ shows monotonic and gradual improvement, instead of an immediate lock on the true filter. It drastically improves for the first 11 frames, where about -17 DB is attained. After the frame 11, the rate of improvement is almost constant, except in frames where the reference speech is not present, such as frame 16 to 20. The value of the constant is about -0.1 per frame similar to that in the 32 point delayed delta function. Consequently it achieves -21 DB at the end of the 14 sec long speech.

Now we comment on the difference between the filter deviation γ of the 1024 point actual room impulse filter case and that of the 32 point delayed delta function case. If we compare both filter deviation curves, we notice that for frame 11 through 68, the deviation curve of a complicated case — the 1024 point actual room impulse filter — is about 5 DB better after 14 sec than that of a simple case — the 32 point delayed delta function case. This apparent contradiction is probably explained by the following argument. Due to the anti-aliasing low-pass filter in the conversion operation from analogue to digital, both the reference speech and the 1024 point impulse response are bandlimited up to 4.69 KHz. Since the primary speech is created by convolving the reference speech with a filter in the computer, it does not contain the information above 4.69 KHz to 5.0 KHz . Because the algorithm uses two speech signals, the reference and the primary, both having no information above 4.69 KHz to 5.0 KHz, the filter estimate by the algorithm does not have valid values in this high frequency region. Thus, values estimated in that region are totally random. However, when they are smoothed with a Hanning window, these random values

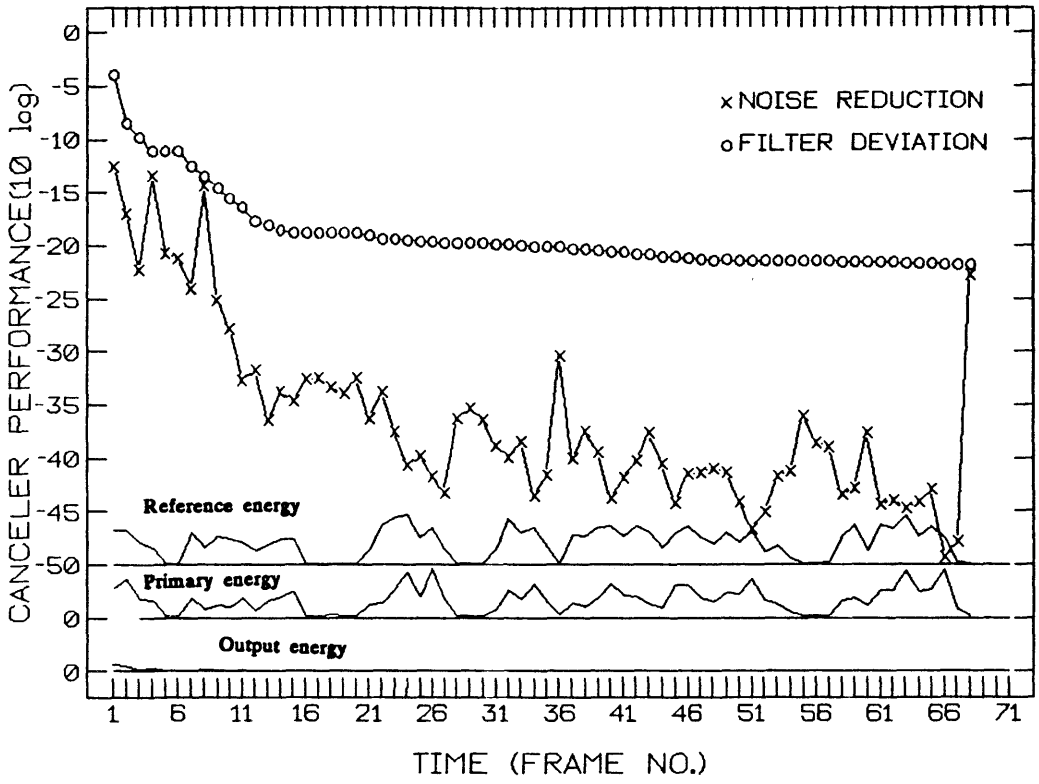
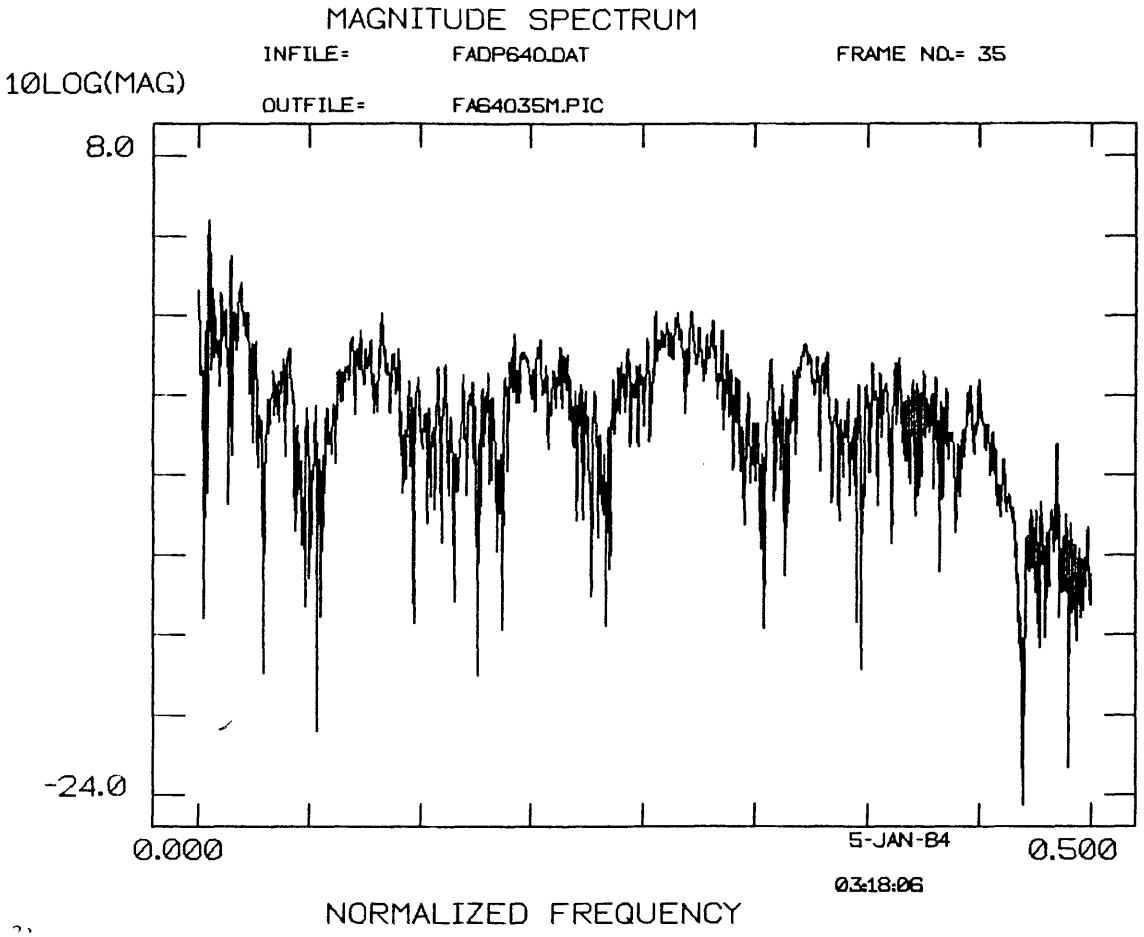
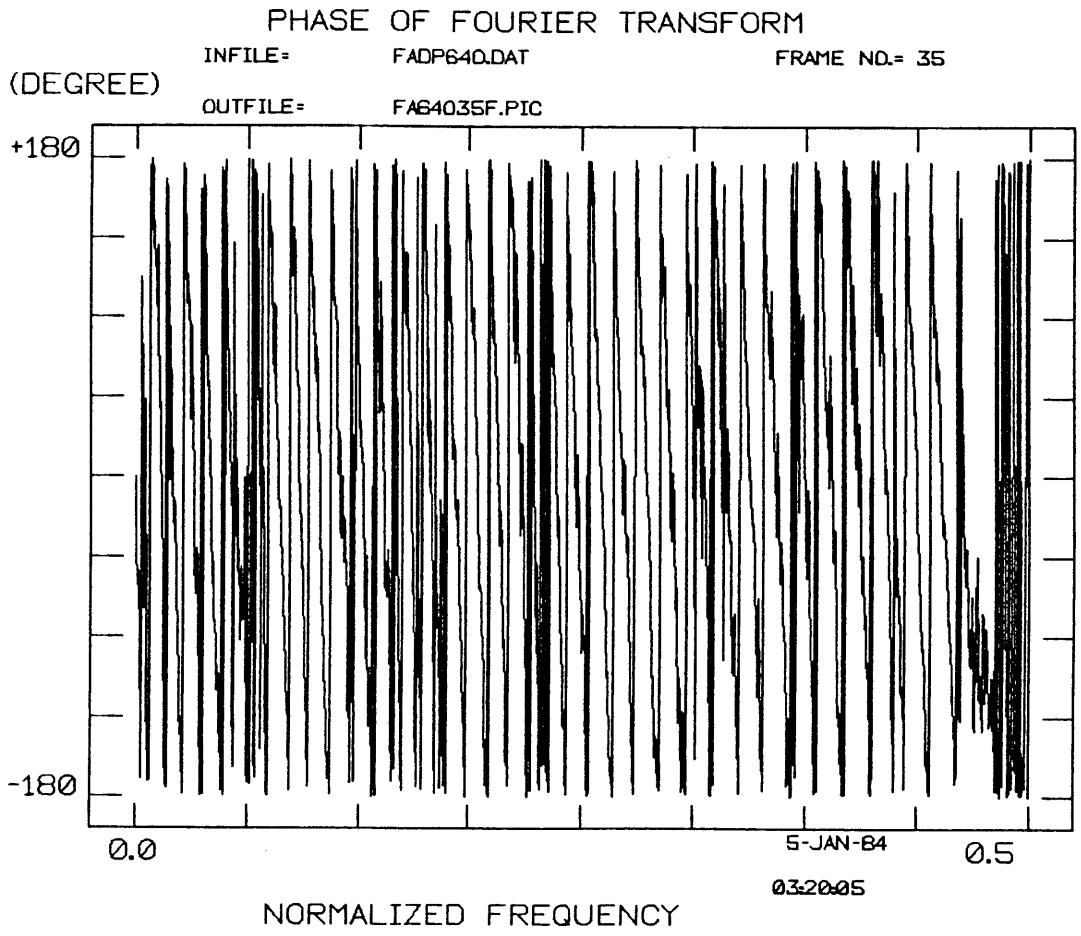


Fig 5.11 The filter deviation γ, the noise reduction ρ in the 1024 point room response with no newscaster's speech



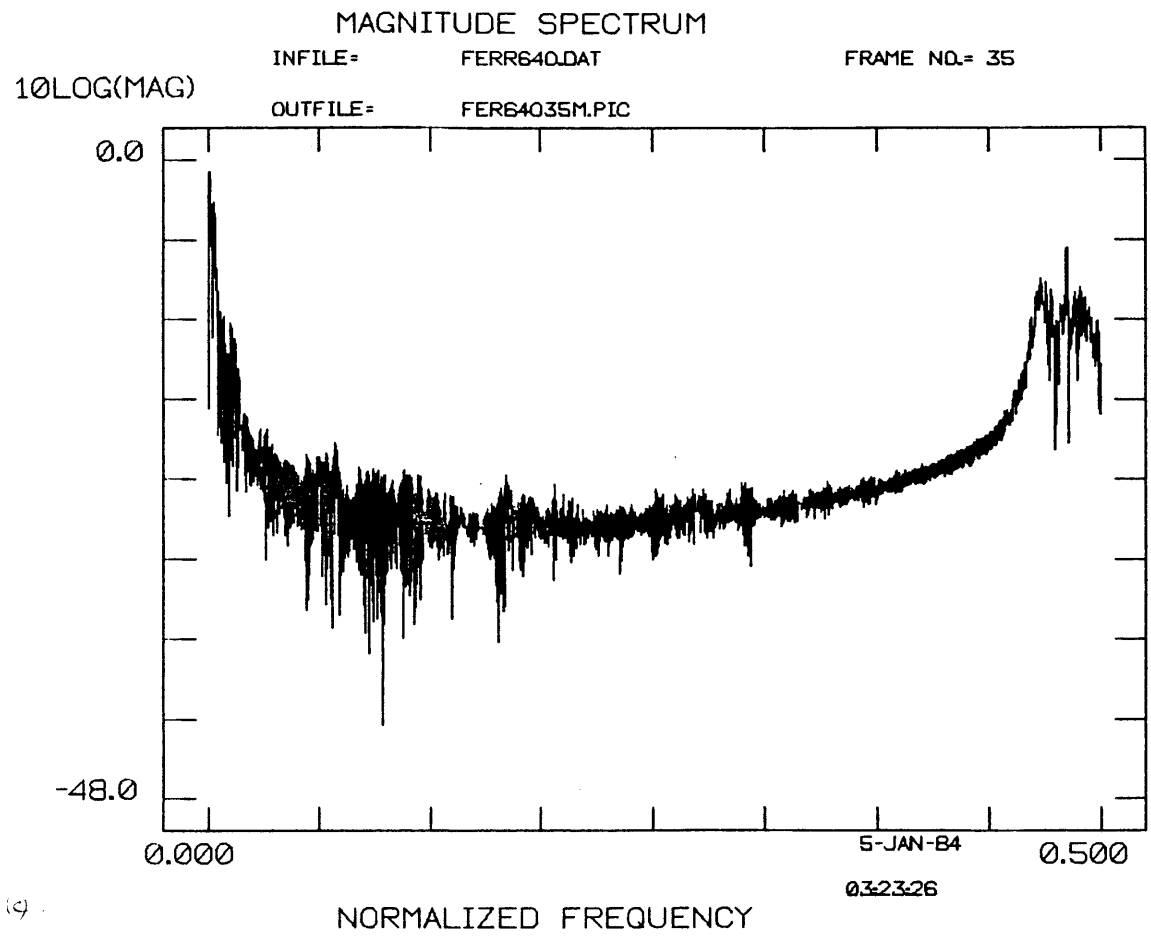
(a) The magnitude of the filter estimate in the 35 th frame

Fig 5.12 The behavior of the filter estimate for the 1024 point room response in the 35 th frame in the case of no newscaster's speech



(b) The phase of the filter estimate in the 35 th frame

Fig 5.12 The behavior of the filter estimate for the 1024 point room response in the 35 th frame in the case of no newscaster's speech



(c) The filter estimate error $\Delta H_i(k)$ in the 35 th frame

Fig 5.12 The behavior of the filter estimate for the 1024 point room response in the 35 th frame in the case of no newscaster's speech

average to nearly zero. Then the filter estimate has small value in that high frequency region.

On the other hand, the deviation of the filter estimate is calculated by

$$\gamma_i = 10\log_{10}\left\{\frac{\sum_{k=0}^{N-1}|H(k)-\hat{H}_i(k)|^2}{\sum_{k=0}^{N-1}|H(k)|^2}\right\} \tag{5.5.10}$$

In the case of the 32 point delayed delta function, the true filter

$$h(n) = \delta(n-32) \tag{5.5.11}$$

is generated in the computer, and this filter has substantial energy above 4.69 KHz. Thus the error factor $H(k)-\hat{H}_i(k)$ becomes large in the frequency region above 4.69 KHz , since $H(k)$ is still substantial but $\hat{H}_i(k)$ is small at this high frequency. Whereas in the case of the 1024 long actual room impulse, $H(k)$ is also band limited up to 4.69 KHz, and it has small value in the high frequency region since $H(k)$ is obtained through analogue to digital conversion from the experimental data. Consequently the error factor in the high frequency region in the 1024 point room response is still small. This difference makes the deviation of the simple case worse than that of the complicated case. This conjuncture proves true by computing in the case of the 32 point delayed delta function the filter estimate error from 0 Hz to 4.69 KHz and obtaining the result that only 5 % of filter estimate error comes from the frequency region below 4.69 KHz, whereas in the 1024 point filter case, 85 % of the total filter estimate error comes from that region. If we compare two different transfer function in terms of the filter deviation up to 4.69 KHz, the 32 delayed delta function achieves around -30 DB, while the 1024 point function remains -21 DB at the end of 14 sec speech.

Fig 5.12 illustrates the filter estimate and filter error in the frequency domain. Fig 5.12

(a), (b) show the magnitude, or phase of the filter estimate in the 35 th frame. It appears almost the same as the true filter Fig 5.6 (b),(c). The error between the estimate and the true filter in the 35 th frame is given in Fig 5.12 (c). One notices that the error occurs in the both frequency end regions mostly.

For the noise reduction, the Fig 5.11 illustrates that ρ achieves below -30 DB at frame number 11, where one can still hear the reference speech in the output, but that at the end of 14 sec speech it retains below -40 DB, where you hardly hear the reference speech. When ρ is compared with that in the 32 point delayed delta function case, there is no difference in performance which they achieve on average in frames where there are much of the reference speech. These frames with high energy reference speech are suitable for legitimate filter estimation. However, in frames where there are almost no reference speech, then the simpler system function gets worse than the complicated system. The reason is simple. We suspect that the values of parameters preferable to the complicated case causes this contradiction. The length of the filter estimate is 1024 in both cases, which does harm to the simple case, whereas it is the right choice in the complicated case. Since the estimate is truncated beyond the first 1024 points, it produces non zero values from the 33 th point to the 1024 th point in the filter estimate of the 32 point delayed delta function. It causes undesirable circular aliasing effect in the output. This circular aliasing effect appears clearly when the reference speech does not exist. If the M were chosen to be approximately 32, then this effect would disappear for the 32 point delta function case.

The informal subjective hearing test shows that the first half of the 14 sec speech still contains the smearing factor of the reference speech, whereas in the latter half latter half of the 14 sec, the smearing speech fades down to almost zero, leaving only quantization error. Although the noise reduction curve shows that most of improvement has been achieved before 10 frames, the smeared speech does not fade down until around frame

number 35. The first thing we should remind ourselves of is that the level of the speech signal in the hearing test is a relative parameter, thus how it sounds may vary with the volume of the audio amplifier playing back the output speech. Since the test 1 does not have the newsman's voice, the only way to evaluate the performance of the algorithm in the subjective listening test, is to listen to the primary speech and the output speech with the same volume on the audio amplifier at the same time. Therefore it is dangerous to determine the adaptive rate of the algorithm from the hearing test on only the output speech, since the volume is a parameter in that test. Consequently, we will determine the adaptive rate based on the noise reduction curve. Although it is hard due to the rapid fluctuation of the noise reduction, we define it as the elapse time from the beginning to the point at which the drastic rate of improvement stops. Therefore in this experiment the rate of adaptation is the duration of 11 frames (2.2 sec). The output of the first 30 frames (6.1 sec) or so, has some smeared reference speech. Its quality is largely attenuated and bandpassed by $H(k) - \hat{H}_i(k)$. The reverberant effect due to the circular aliasing is not clearly noticed. According to the results mentioned above, the algorithm works well in the complicated 1024 point filter case as well as the 32 point delta function case, when the parameters are chosen properly.

5.5.3. The Synthetic data Test 2

Test 2 includes cases with the newsman's speech present in the primary speech as well as the filtered version of the reference speech.

The presence of the optimal speech interferes with the correct estimation of the filter, thus the estimation will not be made as accurately as with no newsman's speech added (Test 1). The effect of the interference could be studied by changing the ratio of the newsman's speech to the filtered version of the reference speech in the 1024 point

room impulse response case. Again the standard values of parameters are chosen as the same values as in Test 1.

32 point delayed delta function

The simulated primary speech consists of 32 point delayed reference speech added to the newsman's speech. The ratio of the amplitude of the filtered version of the reference speech to the newsman's speech is almost 1 : 1. If we call the signal we wish to extract, namely the optimal speech, the "signal", whereas call the interfering signal, namely the filtered version of the reference speech, the "noise", then the signal-to-noise ratio is 0 DB in this example.

In the experiment, the standard values of parameters are chosen, as usual. Again note that this setting of parameters is not best for the $\delta(n-32)$, since this trial is made to compare with the 1024 point actual room impulse response case. The setting of parameters is chosen to be best for the 1024 point actual room impulse response.

Fig 5.13 shows the filter deviation γ, the noise reduction ρ. The curve of γ suggests that the noise canceler algorithm cannot exactly discriminate the correct innovation information for the filter estimate from the component of the interfering speech, because the γ does not monotonously decrease, but fluctuates. However the overall trend of the curve γ appears to be downward, and achieves around - 13 DB filter deviation at the end of the 14 sec speech. The failure of the discrimination is explained by the high variance of the third term in equation (5.4.1). The interesting observation about γ is that it improves in frames where only the reference signal is present in the primary. For instance, frame number 1, 21, and 36 have only the reference speech present with no newsman's voice, and these frames achieve good filter estimates. This sudden improvement is mainly provided by the fast adaptation rate, the big exponential average coefficient α=0.25, and so the adaptation

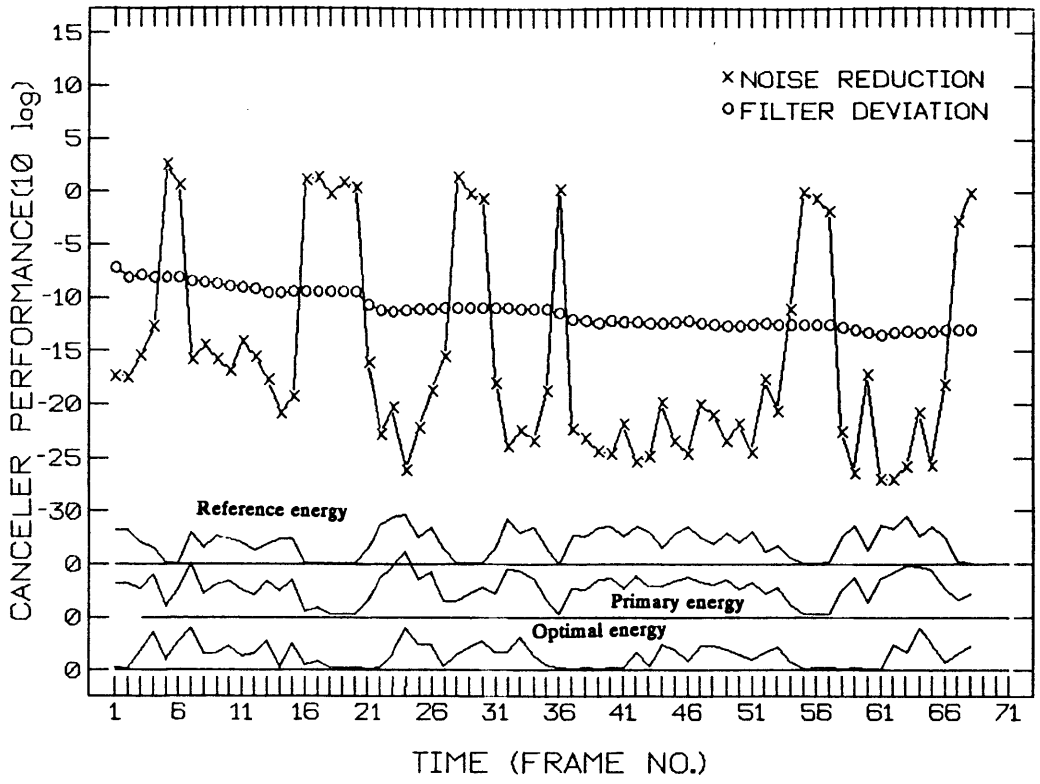


Fig 5.13 The filter deviation γ, the noise reduction ρ in $h(n) = \delta(n-32)$ with 0 DB SNR newscaster's speech

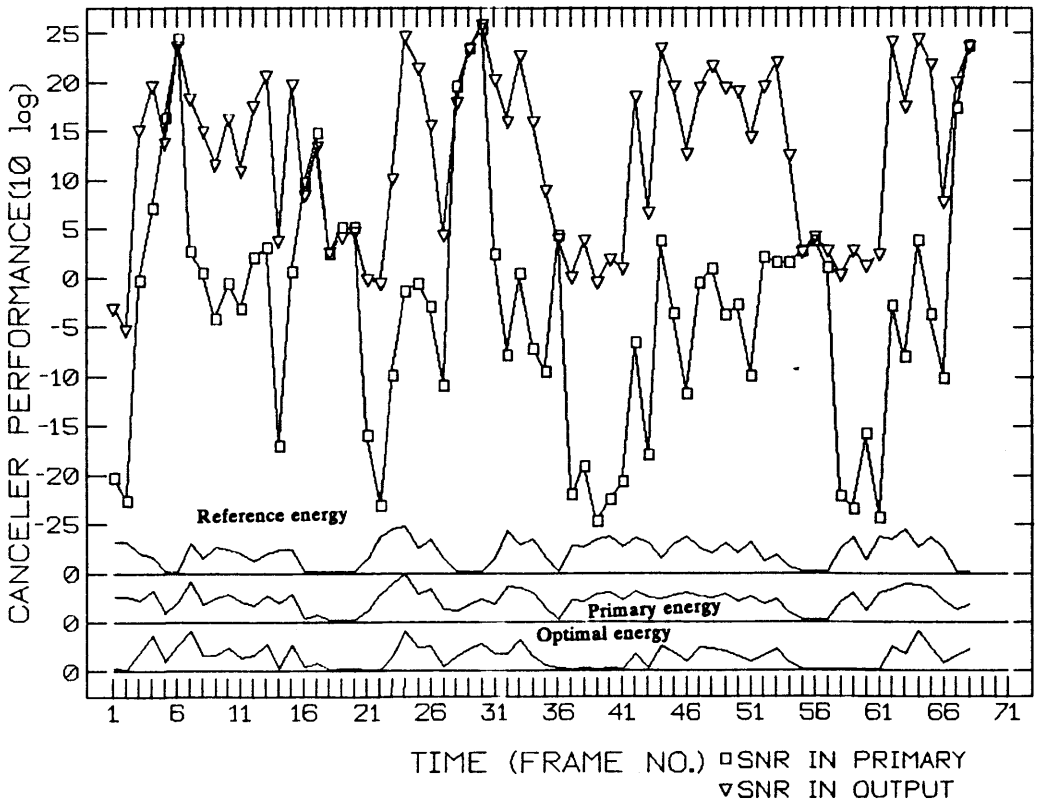


Fig 5.14 The signal-to-noise ratio (SNR) in the primary speech, the signal-to-noise ratio (SNR) in the output speech in $h(n) = \delta(n-32)$ with 0 DB SNR newscaster's speech

rate may vary with the value of α.

The noise reduction in Fig 5.13 achieves 20 DB at the end of 14 sec speech, which is rather worse than the case with the non newsman's speech due to the larger filter estimate deviation. Note that in speech frames with no reference, the noise reduction shows positive values instead of negative values. This suggests that in these frames, the noise canceler cannot remove the noise but instead adds noise. The reason is the same as the previous section, which is that the parameter setting for M is too big for the 32 point delayed delta function.

The subjective listening evaluation will be closely related with the signal-to-noise ratio in the output. In the Fig 5.14, both the signal-to-noise ratio in the primary signal, which is the input signal before processed, and the noise to the signal ratio in the processed output are plotted.

When there is no reference speech in the primary speech to be removed, such as frame 5 and 6, both lines are close. Whereas when there is substantial filtered reference speech in the primary speech, the difference between lines are large. This difference of both lines corresponds to noise reduction showed in Fig 5.13

The signal-to-noise ratio on the output is around 20 DB at the end of 14 sec speech. It is not perfect, but is acceptable for the noise canceler. The listening test suggests that, especially during the latter half of the speech, the interfering speech in the output is mostly eliminated. The quality of the estimated newscaster's voice is as same as the original without apparent artifacts.

A 1024 point actual room impulse response

The primary speech is simulated by convolving the reference speech with a known 1024 point actual room impulse response and then adding the newsman's speech to the

reference speech. The 1024 point actual room impulse response is exactly the same as that in Test 1, the second filter case. (See Fig 5.6) The gain of the impulse response is chosen so that the overall energy of the transfer function will be approximately 1, namely $\sum_{n=-\infty}^{\infty} h^2(n) = 1.$

Fig 5.15 shows the result in the case where the primary speech has 0 DB signal-to-noise ratio. Here we call the ratio of the newsman's speech to the filtered reference speech as the signal-to-noise ratio. The estimated filter does not improve as well as in the case with no newsman's speech. The existence of the newsman's speech interferes with the estimation process. The extent to which it interferes depends on how much newsman's speech is present in the primary, compared with the filtered reference signal.

In this 0 DB signal-to-noise case, the achieved estimated filter deviation is around -10 DB at the end of the 14 sec sample speech. The noise reduction results in around 20 DB.

The signal-to-noise ratio in the output of the noise canceler as well as the signal-to-noise ratio in the primary speech are plotted in Fig 5.16.

It suggests that 20 DB noise reduction from the 0 DB signal-to-noise ratio primary speech results in the 20 DB signal-to-noise ratio output speech of the canceler.

Fig 5.17 5.19 shows the deviation of $\hat{h}(n)$ and the noise reduction for the cases of -12 DB SNR in the primary speech, and 6 DB SNR in the primary respectively. Also the corresponding SNR in the output for both different primary SNR cases are shown in Fig 5.18 and Fig 5.20. In these cases, the SNR was adjusted by changing the gain on the reference signal $x(n)$, and the amplitude of the newsman's speech. In practice, the higher the SNR in the primary file, the worse the performance of the canceler becomes. The error of $\hat{h}(n)$ gets larger and the noise reduction becomes smaller. However, the higher the SNR in the primary speech, the less interfering speech needs to be removed. Thus at high SNR

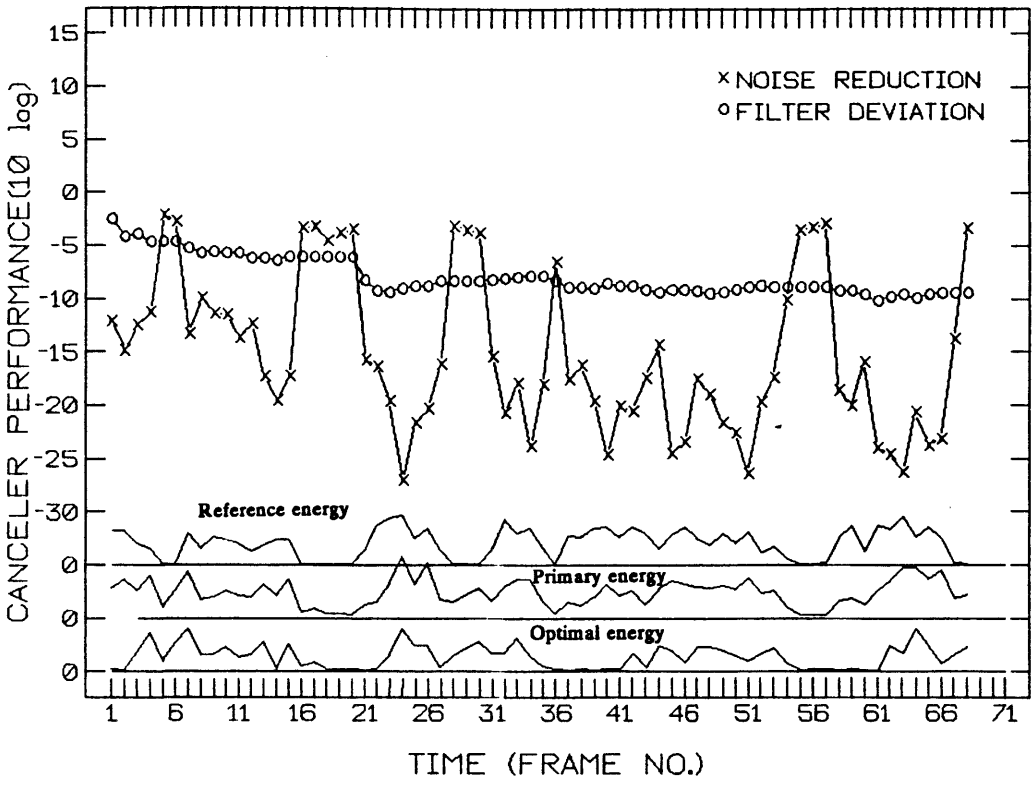


Fig 5.15 The filter deviation γ and the noise reduction ρ in the 1024 point room response with 0 DB SNR newscaster's speech

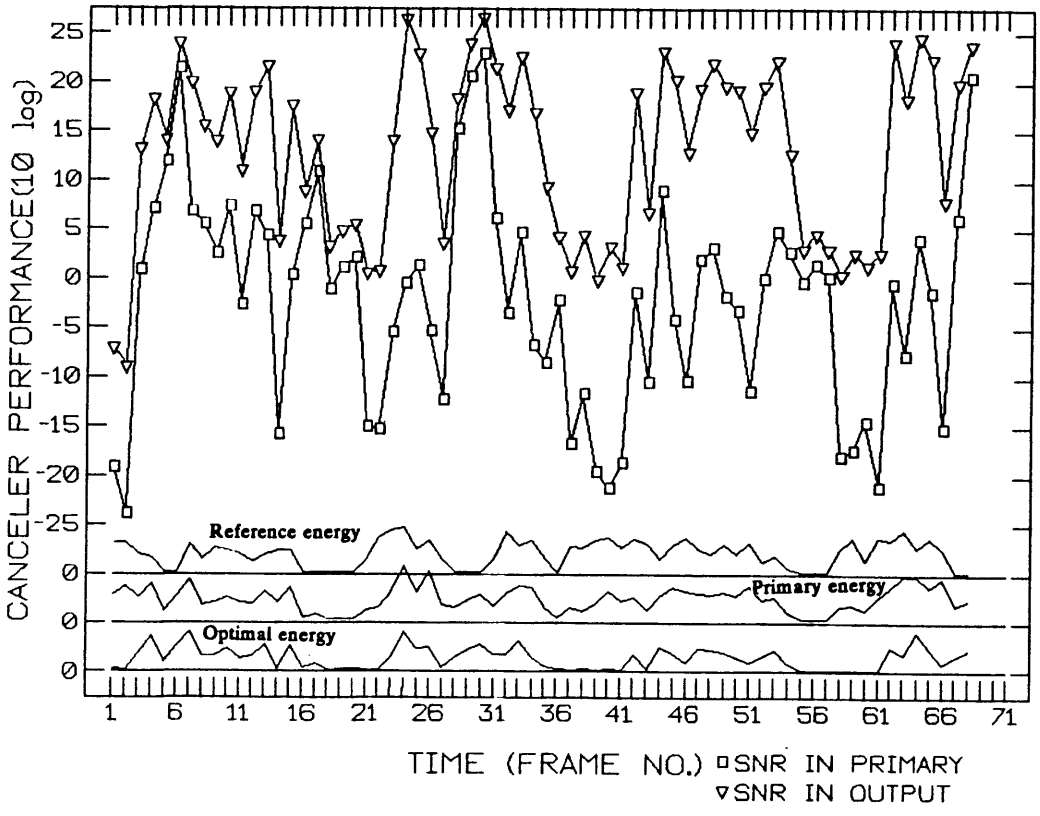


Fig 5.16 The SNR in the primary and the SNR in the output in the 1024 point room response with 0 DB SNR newscaster's speech

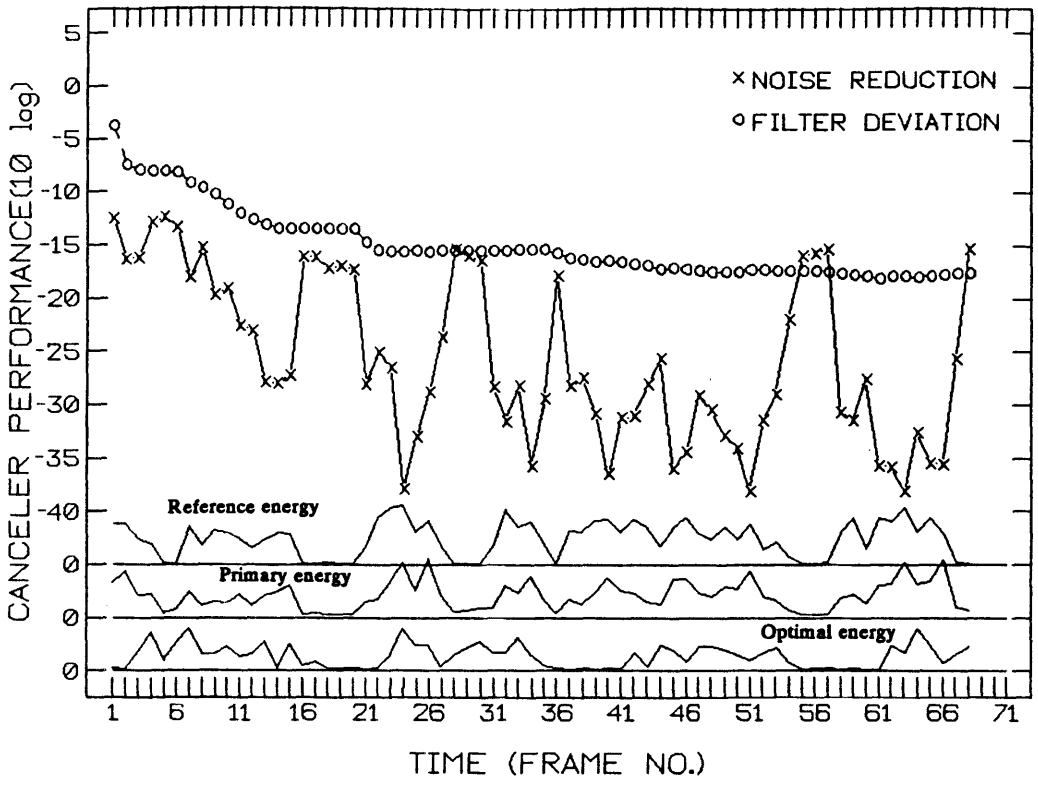


Fig 5.17 The filter deviation γ and the noise reduction ρ in the 1024 point room response with -12 DB SNR newscaster speech

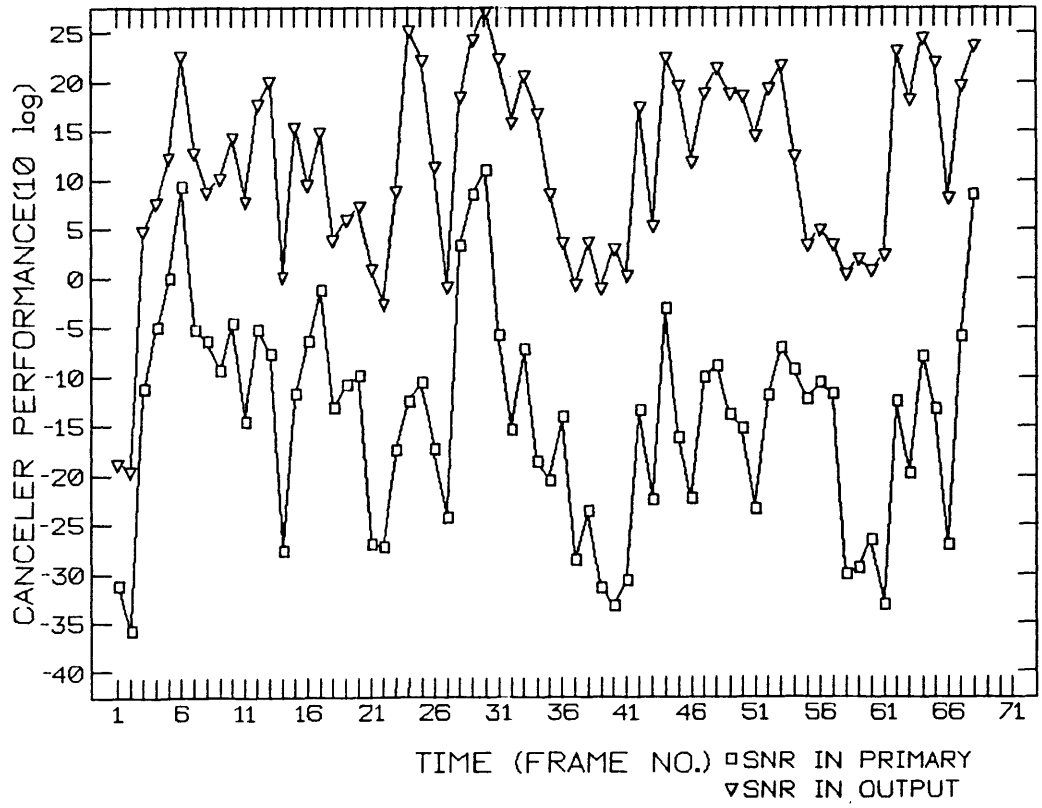


Fig 5.18 The SNR in the primary and the SNR in the output in the 1024 point room response with -12 DB SNR newscaster speech

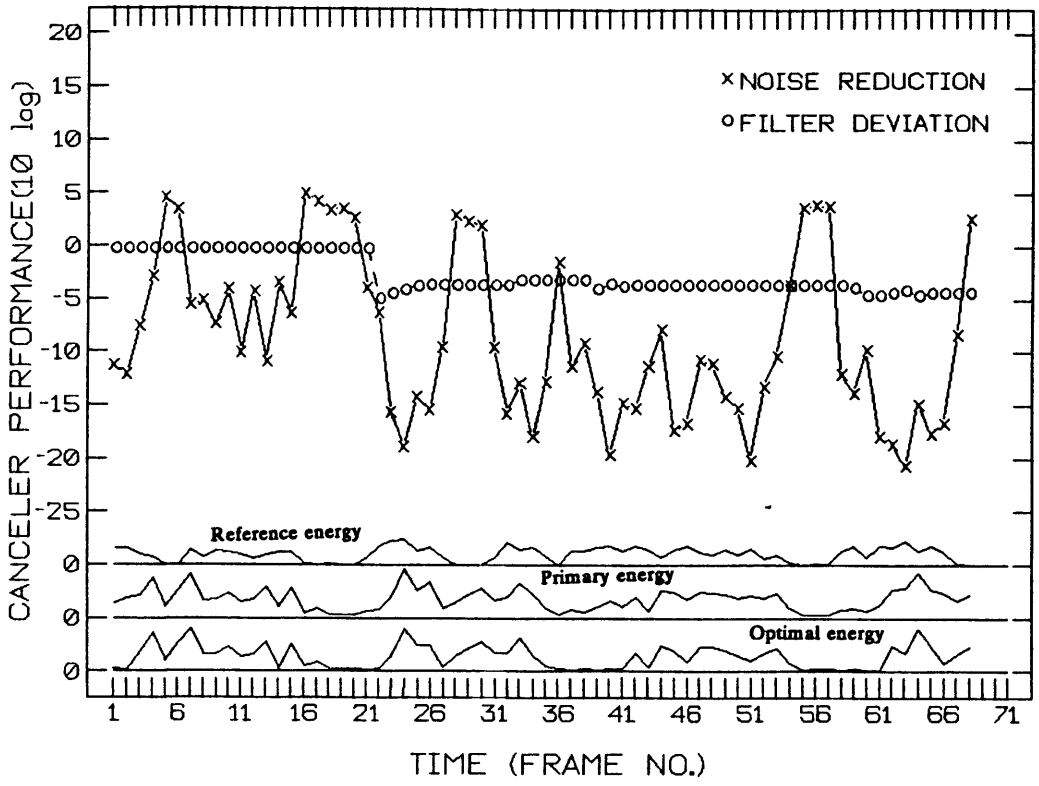


Fig 5.19 The filter deviation γ and the noise reduction ρ in the 1024 point room response with 6 DB SNR newscaster speech

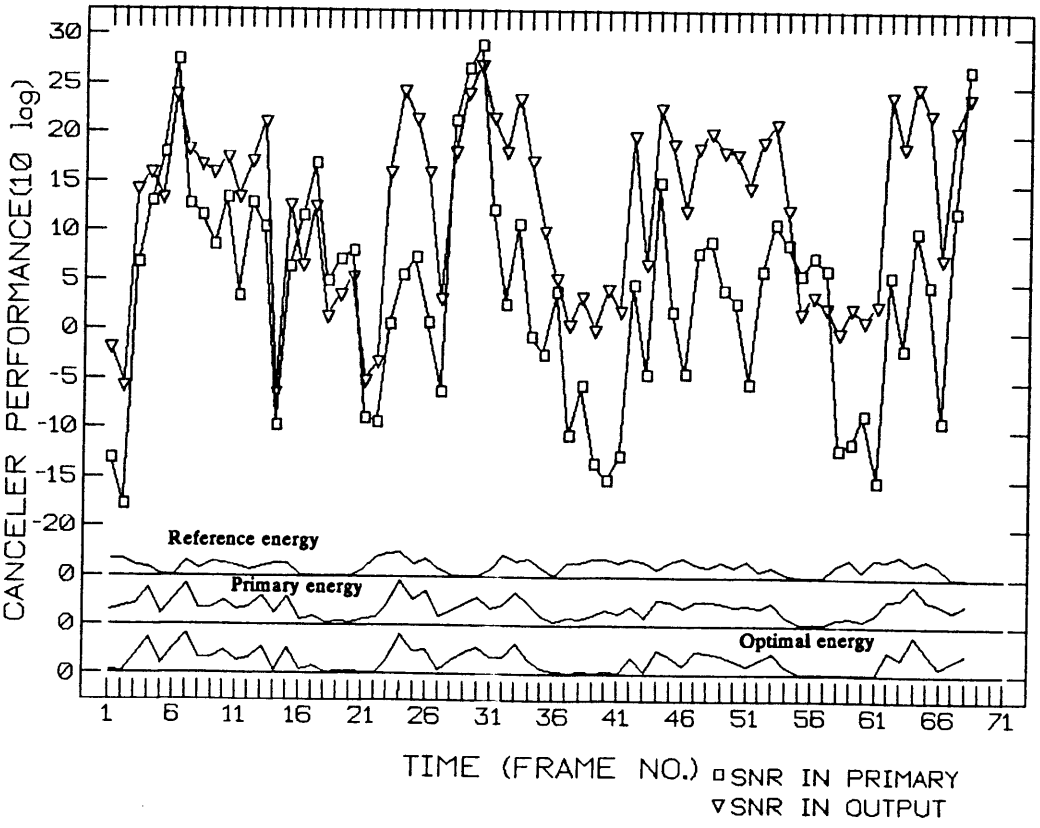


Fig 5.20 The SNR in the primary and the SNR in the output in the 1024 point room response with 6 DB SNR newscaster speech

in the primary it is not necessary to perform the noise reduction as well as at low SNR in the primary. This argument explains the result that regardless of the SNR in the primary, the algorithm achieves around 20 DB SNR in the output.

An informal listening test indicates that artifacts due to processing are not noticeable in the output. It means that the signal element of newsman's speech in the primary is considered to be transmitted without any degradation due to the processing. So the quality of the desired speech is maintained in the output. Thus, the factor which contaminates the output is merely the remaining reference signal which fails to have been removed from the primary. The SNR in the output is almost same as in the simple 32 point delayed delta function case, which suggests that the algorithm is capable of noise cancellation in a wide range of room type. The listening test also suggest that there is no difference between the two room filter cases.

5.5.4. The actual speech data experiment

In the second category, the primary speech is collected in the microphone in a real room experiment, offering two major trials. One is the case where the primary speech is comprised of only the transmitted version of the reference speech, but not the newsman's speech. The other case is that the primary speech is the combination of the transmitted version of the reference speech and the newsman's speech.

The result is, however, too obscure enough to make any conclusive statements. This is not because of the lack of capability of the algorithm, but due to some unidentified problems either in the execution of the program or in recording the speech data. Further analysis of this case is needed.

Chapter 6 Summary and Conclusion

The two-channel noise canceler approach to a broadcast room problem leads to a problem of transfer function estimation with a long impulse response of more than 1000 sample points. Although there are several possible algorithms for solving the transfer function estimation problem, the nature of the long impulse response limits which methods are practical. Constraints arise mainly from both the computation time and the need for robust estimation of such a large number of coefficients. The finite impulse response estimation techniques often fail when identifying more than 1000 taps. Two common FIR methods are the covariance method and the least mean square (LMS) method. The covariance method can deal with up to 50 parameters, but has trouble when there are more than 50 taps, because of the increase of the computation time, and the loss of robustness. The LMS method might cope with the large number of taps[1] (around 1000), but in the circumstance with a heavy disturbance signal (the newsman's speech signal), LMS lacks an efficient method to distinguish the portion of the signal which offers a good estimate with low variance.

The pole-zero model technique is much more questionable, since the fitness of the model to the room characteristics is not guaranteed.

The spectral analysis estimation technique is promising in the sense that it can deal with any long transfer function with various types of filter shapes. Apparently, it can cope with much longer impulse response than the other FIR techniques can handle. Because speech signals are non-stationary, they are difficult to deal with in comparison with stationary signals. The least square minimization in the conventional system identification model is found equivalent to be an approach of Maximum likelihood estimation method

with the assumption that the desired speech is Gaussian white noise, whereas the reference speech is a known deterministic signal. This suggests that the conventional minimization approach does not offer an effective result due to its view of the speech as white noise. Instead, we proposed the method which assumes that the desired speech is modeled as Gaussian colored noise with a power spectral density which changes frame by frame. This offers a better analogy to the non-stationarity of the speech signal.

The implementation is realized by a frame-by-frame process. New information on the filter is obtained from the current frame observation, then the filter estimate is updated based on the new information, and then the updated filter is used to restore the current frame newscaster's speech. The method of updating the filter estimation in the proposed algorithm is based on the ratio in the primary signal of the signal levels in the reverberant reference speech signal and the newscaster's speech signal at each individual frequency point. If at a frequency k, the energy of the pre-estimated newsman's speech seems much weaker than the reference speech, then the current information about the filter at that frequency is judged reliable. Then that information is used to update the filter estimate. On the contrary, if the relation in the energy of two speech is reversed, then the current information is regarded as poor. That information contributes to the updating of the estimate to only a small extent. This procedure is performed at every frequency and the filter is updated frequency sample by frequency sample.

If the frame length is approximately infinite, smoothing windows are not used, and the desired speech is not present at all, then the proposed algorithm offers the exact filter values at the first frame, and cancels the reference speech perfectly, leaving nothing in the output. If the desired speech is present instead, then the algorithm improves the filter estimate frame by frame. It leads to the true filter and restores the desired speech exactly when an infinite number of frames are observed. However, in reality, one can not either

obtain a numerous number of infinitely long observations or , even if one could, computing with such an enormous length of data is not realistic. In practical implementation with finite length frames of data, significant problems are: (i) the essential inaccuracy in the estimation due to the end effect of windowing input data. This inaccuracy is a function of the filter length and the length of input data read into a buffer for a frame. Even when no newsman's speech is present — this is the best situation for the estimation process — the true filter can never be obtained in the first frame due to this end effect. (ii) The bias effect of a finite length window on the filter estimate. It causes the filter estimate to be updated in the slightly wrong direction. The algorithm does not compute the exact values of the true filter.

Owing to these two major problems, the algorithm shows several distinct types of behavior. First of all, it presents gradual adaptation to the true filter over time. It never immediately locks on the true value in the first frame nor converges to the true values at all. Secondly, this end effect causes circular aliasing in the filter estimate and in the desired speech estimate. Even though the true filter is M points long, the estimate $\hat{H}_i(k)$ is calculated using N point DFT's, and does not necessarily correspond to the true M point long sequence. Rather, it gives N non-zero points in the time domain due to the circular aliasing. Therefore non-zero data at sample points where values should be exact zeroes cause echoes of the interfering speech to remain in the output. The truncation of the last N-M in the $\hat{h}_i(n)$ is expected to limit the effect of echoes. However it does not necessarily cure the estimated filter of the circular aliasing problem, since truncation causes more windowing effects. Moreover, since the circular aliasing also causes non-zero values at sample points during the initial delay time. Thus the longer the delay time, the more circular aliasing is noticed as an echo in the output speech.

Aims of the experiments with synthetic data are to examine the performance of the algorithm, and to find out its applicability to the real situation.

According to the empirical results, the end effect of sectioning data into frames produces a slow improvement of the global measure γ. The experimental results show that the proposed method achieves above 20 DB SNR in the output in the case of the 1024 point time invariant filter case, no matter what SNR is in the primary. The speech with 20 DB SNR is tolerable, although it is not perfect. The 32 point delayed delta function calculates a similar SNR in the output. This suggests that the algorithm is potentially capable of estimating the large number of the sample points in the impulse response of the transfer function, at least 1000 points. The ideal will be applicable to any long transfer filter, as long as appropriate values of parameters are set accordingly.

The estimated newsman's voice is still slightly degraded by several factors. The first factor arises from smearing the reference speech in a such way that the remaining reference speech contains highly reverberated reference speech. The second is, although not easy to notice, the newsman's voice itself is also influenced by the filter estimation, because the filter estimation process implicitly makes the filter estimate depend on the primary signal and thus on the newsman's voice. The first factor is easy to recognize in the noise reduction curve with different value of the filter length parameter M. For the 32 point delayed delta function case, the noise reduction with M=32 is much better than that with M=1024. The second factor is difficult to observe. One way is to compare the newsman's voice in the primary speech with that in the output. In the experiments done here, the output speech does not seem to have noticeable degradation of the newsman's voice. Therefore this second factor is assumed to be negligible.

Finally we summarize topics necessary for future work. They are pointed out by stat-

ing the extent to which the proposed spectral analysis method satisfies the requirements for the broadcasting room problem mentioned in section 2.1

We already mention the robustness of the algorithm, saying that it achieves above 20 DB SNR in the output. According to Rabiner2 for the white noise case, the noise reduction ρ is equal to the filter deviation γ, and expressed as the

$$\rho|_{white\ noise\ input} = \gamma|_{white\ noise\ input} \tag{6.0.1}$$

$$= 10\log_{10}\left[\frac{M}{N_{total}}\right] + SNR_{pri} \tag{6.0.2}$$

where M is the length of the filter, N_{total} is the total number of sample points used for calculating the filter estimate. In our synthetic experiments, $M = 1024$.. Since $\alpha = 0.25$, and $F = 2048$, the total number of different sample points integrated into each filter estimate, N_{total} is to $\frac{F}{\alpha} = 8192$. Then ρ is

$$\rho|_{white\ noise\ input} = -9.0 + SNR_{pri}\ (DB) \tag{6.0.3}$$

Therefore if SNR_{pri} = -12, or 0, or 6 DB, the above equation gives $\rho|_{white\ noise\ input}$ = -21, or -9, or -3 DB. On the other hand, the experimental data show that in our algorithm, if SNR_{pri} = -12, or 0, or 6 DB, the noise reduction ρ reaches -30, or -20, or -15 DB respectively. It means that speech case is much better than white noise case. The reason is that for the white noise case a disturbance signal (newscaster's speech) always exists at any frequency k with constant energy, whereas for the speech case that disturbance signal drastically changes its signal level. Thus the instantaneous disturbance SNR at different frequencies may be much poorer than the overall SNR would indicate, which is much suitable to the filter estimation. This comparison of the performance between the synthetic experiments and the theoretical analysis with white noise, suggests that the algorithm can successfully deal with the speech case, selecting only the good information to estimate the filter on the frequency-by-frequency basis. It becomes a fairly robust filter estimate for

speech. One big question to be answered is why the signal-to-noise ratio in the output from the algorithm is limited to 20 DB SNR_{out}, no matter what SNR_{pri} is. This would be a key to improving the performance of the algorithm further.

The second requirement is the efficiency of the computation. Despite using the fast computational tool of the FFT, large FFT buffers N=8192 were used and about 12 FFT or inverse FFT operations are required in each frame. Two filtering operations are needed for $\bar{s}_i(n)$ ans $\hat{s}_i(n)$, two are needed to smooth each of the spectra, $\hat{\sigma}^2{}_i(k)$, $P_{x_i x_i}(k)$, and $P_{x_i \bar{s}_i}(k)$, and one is needed to truncate the filter estimate. Since a real FFT operation is used, each N point FFT computation costs $2 \times N \log_2 N$ real multiplications. The total amount of computation in each frame is thus about $12 \times 2 \times N \log_2 N + 8N$ real multiplications. Since N=8192, it turns out to be $12 \times 2 \times 8192 \times \log_2 8192 + 8 \times 8192 \approx 320 \times 8192 \approx 2.6$ million real multiplications. Each frame produces $F = 2048$ points, then the computational cost is about 13 K multiplications per point. For comparison, let us calculate the computational cost in the auto correlation method, using the Levinson's recursion. The calculation of the correlations using L data points costs $4 \times 2 \times (2 \times L \log_2 2L) + 2(2L)$ real multiplications. The Levinson's recursion for computing the M filter coefficients requires $2M^2$ multiplications. The overlap-save method for filtering operation requires $3 \times 2 \times (L+M) \log_2 (L+M) + (L+M)$. When L=4096 and M=1024, then one frame computation amounts to 3.3 million multiplications. If the frame shift F=4096, it leads to 0.75 K multiplications per point. Consequently, our algorithm is more expensive than the auto correlation method. It suggests that the implementation of the algorithm still encounters computational difficulties. To reduce such expensive computation, one way is to calculate the smoothing of spectra by directly convolving the spectrum with a short smoothing sequence in the frequency domain.

The third point is the adaptability. The adaptive behavior of the filter estimate for

the true time-invariant filter from the initial state of zero is examined in terms of γ in the case with no disturbance signal (newscaster's speech). The γ in the synthetic experiments for the case with no disturbance signal shows the slow but monotonic adaptive behavior of the filter estimate. For the case with newsman's speech, then the γ fluctuates locally, however the overall trend goes downward. We did not examine a time-variant room reverberation case with synthetic speech data. Further work should be done to study the adaptability of the algorithm with a time-variant filter.

Finally experiments using recorded speech data in an actual room fails. This case should be examined further to see what caused the problem.

References

1. S. F. Boll and D. C. Pulsipher, "Suppression of Acoustic Noise in Speech Using Two Microphone Adaptive Noise Cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, Dec. 1980.

2. L. R. Rabiner, R. E. Crochiere, and J. B. Allen, "FIR System Modeling and Identification in the Presence of Noise and with Band-Limited Inputs," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 4, August 1978.