

ENHANCEMENT AND BANDWIDTH COMPRESSION OF NOISY SPEECH  
BY ESTIMATION OF SPEECH AND ITS MODEL PARAMETERS

by

Jae Soo Lim

S.B., Massachusetts Institute of Technology  
(1974)

S.M., Massachusetts Institute of Technology  
(1975)

E.E., Massachusetts Institute of Technology  
(1978)

SUBMITTED IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF SCIENCE

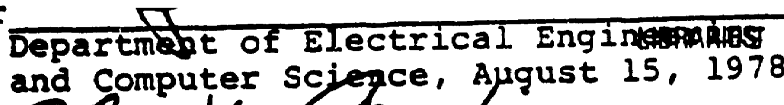
at the

Massachusetts Institute of Technology  
August, 1978

ARCHIVES  
MASSACHUSETTS INSTITUTE  
OF TECHNOLOGY

JAN 31 1979

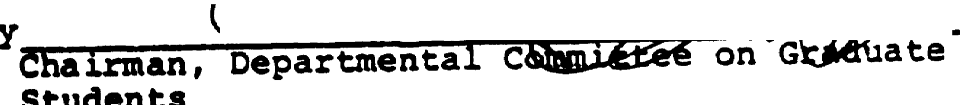
Signature of Author

  
Department of Electrical Engineering  
and Computer Science, August 15, 1978

Certified by

  
Thesis Supervisor

Accepted by

  
Chairman, Departmental Committee on Graduate  
Students

TO MY PARENTS

AND WIFE

ENHANCEMENT AND BANDWIDTH COMPRESSION OF NOISY SPEECH  
BY ESTIMATION OF SPEECH AND ITS MODEL PARAMETERS

by

Jae S. Lim

Submitted to the Department of Electrical Engineering and Computer Science on August 15th, 1978 in partial fulfillment of the requirements for the Degree of Doctor of Science.

ABSTRACT

The problem of enhancement and bandwidth compression of noisy speech is formulated as a parameter estimation problem, in which speech and its model parameters are estimated from the noisy speech based on the MAP estimation procedure. Such an approach leads to two algorithms which require solving sets of linear equations in an iterative manner. Some approximations of the two algorithms lead to two systems which are computationally simpler than the two algorithms by taking advantage of a high speed FFT algorithm. As a preliminary investigation into the performance of the class of systems developed, two systems are implemented and applied to both real and synthetic speech data. An objective and informal subjective evaluation indicate that the systems implemented perform well as enhancement and potential bandwidth compression systems of noisy speech.

THESIS SUPERVISOR: Alan V. Oppenheim

TITLE: Professor of Electrical Engineering

### ACKNOWLEDGEMENT

I am greatly indebted to Professor Alan V. Oppenheim for his supervision and contribution to this thesis. My close association with him through this doctoral research has been an invaluable experience which will have a profound influence on my future career as a researcher. I am also indebted to Professor Louis Braida, Professor Ken Stevens and Professor Alan Willsky for providing valuable comments and advice throughout this thesis work.

A number of discussions with Bruce Musicus and other members of the Digital Signal Processing Group were helpful in learning new material and clarifying various issues. They also contributed to a pleasant environment to work in.

I am deeply indebted to my parents and my aunt who made the dream to study in the United States come true. Without their support, my education at M.I.T. would not have been possible. My wife, KyuHo, contributed to this thesis in several different ways. Her encouragements and understanding at difficult times were invaluable. She also provided all the figures in this thesis.

Finally, I would like to thank Ms. Monica Edelman for her excellent typing of this thesis.

TABLE OF CONTENTS

	PAGE
Abstract	3
Acknowledgement	4
Table of Contents	5
List of Figures and Tables	11
Chapter I Introduction	20
I.1 Introduction	20
I.2 Scope of Thesis	26
I.3 Summary of Chapters	27
Chapter II Survey of Speech Enhancement Techniques	30
II.1 Introduction	30
II.2 Speech Enhancement Techniques	30
II.2.1 Adaptive Comb Filtering Method	30
II.2.2 Correlation Subtraction Method	32
II.2.3 Speech Enhancement by a Voice Excited Vocoder	37
II.2.4 INTEL System	39
II.2.5 SABER Method	41
II.2.6 Other Generalizations of Correlation Subtraction Method	44
II.2.7 SPAC and SPOC	46
II.2.8 Wiener Filtering Method	49
II.2.9 Summary	51

II.3	Summary of Performance Evaluation	52
II.3.1	Adaptive Comb Filtering Method	52
II.3.2	Correlation Subtraction Method and INTEL System	55
II.3.3	SABER Method	57
II.3.4	Other Generalizations of Correlation Subtraction Method	59
II.3.5	Wiener Filtering Method	59
II.4	Bandwidth Compression Systems of Noisy Speech	60
II.5	Motivation for a New Approach	62
Chapter III	Model of Speech and Its Parameter Estimation	65
III.1	Introduction	65
III.2	Model of Speech	66
III.3	Representations of the Model of Speech	68
III.4	Model of Noisy Speech and its Represen- tations	74
III.5	Review of Parameter Estimation Theory	75
III.6	Estimation of Speech Model Parameters	77
Chapter IV	Statistical Parameter Estimation From Noise-Free Speech	81
IV.1	Introduction	81
IV.2	Problem Formulation	81
IV.3	Direct Approach: Maximization of $p(\underline{a} \underline{s}_0)$	82
IV.3.1	Case 1	84
IV.3.2	Case 2	87
IV.3.3	Case 3	90

IV.3.4	Case 4	91
IV.4	State Space Approach: All Pole Coefficients as State Vectors	94
IV.4.1	Kalman Filter: Review	94
IV.4.2	Maximization of $p(\underline{a} \underline{s}_0)$ by a Kalman Filter	95
Chapter V	Statistical Parameter Estimation From Noisy Speech	98
V.1	Introduction	98
V.2	MAP Estimation Procedure: A Non-Linear Problem	99
V.3	Maximization of $p(\underline{a}, \underline{s}_0   \underline{y}_0)$ : Linearized MAP (LMAP) Estimation Procedure	101
V.3.1	An Algorithm to Maximize $p(\underline{a}, \underline{s}_0   \underline{y}_0)$	102
V.3.2	Maximization of $p(\underline{s}_0   \underline{a}, \underline{y}_0)$	104
V.3.3	Linearized MAP Estimation Procedure	108
V.4	Revised Linearized MAP (RLMAP) Estimation Procedure	110
V.4.1	Motivation for the Revision	110
V.4.2	Estimation of $s(i) \cdot s(j)$ by $E[s(i) \cdot s(j)   \underline{a}, \underline{y}_0]$	111
V.4.3	RLMAP Estimation Procedure	114
V.5	Extension to Colored Background Noise Case	115
V.6	Relationship Among Maximization of $p(\underline{a}   \underline{y}_0)$ , LMAP and RLMAP Algorithms	118
Chapter VI	Implementation: Three Noise Reduction Systems	119
VI.1	Introduction	119
VI.2	System A	120

VI.3	System B	128
VI.4	System C	135
Chapter VII Examples and Illustrations		137
VII.1	Introduction	137
VII.2	Application to Synthetic Data	137
VII.2.1	Application of System A to Synthetic Data	141
VII.2.2	Application of System B to Synthetic Data	149
VII.2.3	Application of System C to Synthetic Data	156
VII.3	Application to Real Speech Data	170
Chapter VIII Evaluation		177
VIII.1	Introduction	177
VIII.2	Objective Evaluation	178
VIII.2.1	Systems Evaluated	178
VIII.2.2	Objective Criterion	179
VIII.2.3	Generation of All Pole Coefficients	181
VIII.2.4	Data Acquisition, Analysis and Results	183
VIII.2.5	Discussions	186
VIII.3	Subjective Evaluation: Potential Bandwidth Compression Systems	195
VIII.3.1	Test Sentences	197
VIII.3.2	Speech Analysis/Synthesis System	197
VIII.3.3	Preliminary Comparison	198
VIII.3.3.1	Comparison of Systems A-1, A-2 and A-3	199



VIII.3.3.2	Comparison of Systems B-2, B-5 and B-10	199
VIII.3.3.3	Comparison of Systems C-1, C-2 and C-3	199
VIII.3.3.4	Comparison of Systems A-2, B-10 and C-2	200
VIII.3.4	Evaluation of System A-2 Relative to Conventional LPC Method	200
VIII.3.4.1	Test Material and Procedures	201
VIII.3.4.2	Data Analysis and Results	203
VIII.3.4.3	Discussions	207
VIII.4	Subjective Evaluation: Speech Enhance- ment Systems	207
VIII.4.1	Speech Enhancement Systems	208
VIII.4.2	Preliminary Comparison	208
VIII.4.3	Evaluation of System A-2 as a Speech Enhancement System	210
VIII.5	Additional Studies	214
VIII.5.1	Speech Enhancement by a Complete Analysis/Synthesis System	214
VIII.5.2	System A-2 as a Pre-processor for Other Bandwidth Compression Systems	215
VIII.6	Summary	216
Chapter IX	Future Research	218
IX.1	Introduction	218
IX.2	Improvements	218
IX.2.1	Incorporation of A Priori Information	222
IX.2.2	Smoothing Formant Frequencies	223

IX.2.3	Masking with Random Noise	224
IX.3	Adaptation to Practical Problems	226
IX.3.1	Estimation of $P_d(\omega)$	226
IX.3.2	Estimation of Source Information	227
IX.3.3	Evaluation of Systems	227
IX.4	Further Theoretical Study and Related Work	228
IX.4.1	Implementation of Other Systems	228
IX.4.2	Different Initial Estimates of $\underline{a}$	229
IX.4.3	Incorporation of A Priori Information	230
IX.4.4	Excitation by a Train of Pulses	230
IX.4.5	Pole-zero Modelling	231
Chapter X	Conclusion	232
	References	233
	Appendix	237
	Appendix I	238
	Appendix II	243
	Biographical Note	244

LIST OF FIGURES AND TABLES

FIGURES

- Figure 2.1 A typical speech enhancement system by the correlation subtraction method
- Figure 2.2 INTEL system proposed by Weiss, et al.
- Figure 2.3 A simplified representation of the INTEL system in Figure 2.2
- Figure 2.4 Results of the intelligibility test performed to evaluate the adaptive comb filtering method for enhancement of noisy speech
- Figure 2.5 Results of the intelligibility test performed to evaluate the INTEL system for enhancement of noisy speech
- Figure 2.6 The analysis part of a bandwidth compression system of noisy speech when a speech enhancement system is used as a pre-processor
- Figure 2.7 A possible simplification of the system in Figure 2.6 for some cases; see the text for details
- Figure 3.1 A digital model of sampled speech
- Figure 5.1 One iteration of LMAP and RLMAP algorithms;  $\underline{m}$  and  $\underline{v}$  are given by equation (5-21) in the text

Figure 7.1 (a) A synthetic data segment with random noise excitation

(b) Log magnitude spectrum of the synthetic data in (a) and an all pole fit to the spectrum by the correlation method of the linear prediction analysis

(c) Same as (a) with a pulse train excitation

(d) Same as (b) with a pulse train excitation

Figure 7.2 (a) Log magnitude spectrum of the synthetic data in Figure 7.1(a) and the transfer function that corresponds to the known all pole coefficients

(b) Comparison of the transfer functions that correspond to the known all pole coefficients in (a) and the estimated all pole coefficients in Figure 7.1(b)

(c) Same as (a) with a pulse train excitation

(d) Same as (b) with a pulse train excitation

Figure 7.3 Comparison of System A

(a) Log magnitude spectrum of the synthetic data in Figure 7.1(a) and an all pole fit to the noisy data spectrum after the zeroth iteration of System A at  $S/N=20$  dB

(b) Same as (a) after the first iteration of System A

(c) Same as (a) after the second iteration of System A

(d) Same as (a) after the third iteration of System A

Figure 7.4 Same as Figure 7.3 with the synthetic data of  
Figure 7.1(c)

Figure 7.5 Same as Figure 7.3 with  $S/N=10$  dB

Figure 7.6 Same as Figure 7.4 with  $S/N=10$  dB

Figure 7.7 Same as Figure 7.3 with  $S/N=0$  dB

Figure 7.8 Same as Figure 7.4 with  $S/N=0$  dB

Figure 7.9 Comparison of System B

(a) Log magnitude spectrum of the synthetic data in Figure 7.1(a) (random noise excitation) and an all pole fit to the noisy data spectrum after the zeroth iteration of System B at  $S/N=20$  dB

(b) Same as (a) after the second iteration of System B

(c) Same as (a) after the fifth iteration of System B

(d) Same as (a) after the tenth iteration of System B

Figure 7.10 Same as Figure 7.9 with the synthetic data of Figure 7.1(c) (pulse train excitation)

Figure 7.11 Same as Figure 7.9 with  $S/N=10$  dB

Figure 7.12 Same as Figure 7.10 with  $S/N=10$  dB

Figure 7.13 Same as Figure 7.9 with  $S/N=0$  dB

Figure 7.14 Same as Figure 7.10 with  $S/N=0$  dB

Figure 7.15 Comparison of System C

(a) Log magnitude spectrum of the synthetic data in Figure 7.1(a) (random noise excitation) and

an all pole fit to the noisy synthetic data with  
k=0 of System C at S/N=20 dB

(b) Same as (a) with k=1 of System C

(c) Same as (a) with k=2 of System C

(d) Same as (a) with k=3 of System C

Figure 7.16 Same as Figure 7.15 with the synthetic data  
of Figure 7.1(c) (pulse train excitation)

Figure 7.17 Same as Figure 7.15 with S/N=10 dB

Figure 7.18 Same as Figure 7.16 with S/N=10 dB

Figure 7.19 Same as Figure 7.15 with S/N=0 dB

Figure 7.20 Same as Figure 7.16 with S/N=0 dB

Figure 7.21 (a) A synthetic data segment

(b) Log magnitude spectrum of the synthetic  
data in (a)

(c) Noisy synthetic data of (a) at S/N=-20 dB

(d) Log magnitude spectrum of the noisy  
synthetic data in (c)

Figure 7.22 (a) Log magnitude spectrum of the synthetic  
data in Figure 7.21(a) and an all pole fit to the  
noisy synthetic data with k=0 of System C

(b) Same as (a) with two iterations of  
System A

(c) Same as (a) with ten iterations of  
System B

(d) Same as (a) with k=2 of System C

Figure 7.23 (a) A synthetic data segment  
(b) Log magnitude spectrum of the synthetic data in (a)  
(c) Noisy synthetic data of (a) at  $S/N=10$  dB  
(d) Log magnitude spectrum of the noisy synthetic data in (c)

Figure 7.24 (a) Log magnitude spectrum of the synthetic data in Figure 7.23(a) and an all pole fit to the noisy synthetic data with  $k=0$  of System C  
(b) Same as (a) with two iterations of System A  
(c) Same as (a) with ten iterations of System B  
(d) Same as (a) with  $k=2$  of System C

Figure 7.25 (a) A real data segment of unvoiced speech  
(b) Log magnitude spectrum of the real speech data in (a)  
(c) Noisy speech data of (a) at  $S/N=10$  dB  
(d) Log magnitude spectrum of the noisy speech data in (c).

Figure 7.26 (a) Log magnitude spectrum of the real speech data in Figure 7.25(a) and an all pole fit to the noisy speech data with  $k=0$  of System C  
(b) Same as (a) with two iterations of System A  
(c) Same as (a) with ten iterations of

System B

(d) Same as (a) with  $k=2$  of System C

Figure 7.27 (a) A real data segment of voiced speech

(b) Log magnitude spectrum of the real speech data in (a)

(c) Noisy speech data of (a) at  $S/N=10$  dB

(d) Log magnitude spectrum of the noisy speech data in (c)

Figure 7.28 (a) Log magnitude spectrum of the real speech data in Figure 7.27(a) and an all pole fit to the noisy speech data with  $k=0$  of System C

(b) Same as (a) with two iterations of

System A

(c) Same as (a) with ten iterations of

System B

(d) Same as (a) with  $k=2$  of System C

Figure 8.1 Performance comparison of System A based on

LMSE

(a) random noise excitation case

(b) pulse train excitation case

Figure 8.2 Performance comparison of System B based on

LMSE

(a) random noise excitation case

(b) pulse train excitation case



Figure 8.3 Performance comparison of System C based on  $\overline{\text{LMSE}}$

- (a) random noise excitation case
- (b) pulse train excitation case

Figure 8.4 Performance comparison of Systems A-2, B-10 and C-2 based on  $\overline{\text{LMSE}}$

- (a) random noise excitation case
- (b) pulse train excitation case

Figure 8.5 Performance comparison of Systems A-2, B-10 and C-2 based on Normalized  $\overline{\text{LMSE}}$

- (a) random noise excitation case
- (b) pulse train excitation case

Figure 8.6 Best performance that can be achieved by any combination of Systems A-2, B-10 and C-2 based on  $\overline{\text{LMSE}}$

- (a) random noise excitation case
- (b) pulse train excitation case

Figure 8.7 Results of the speech preference test in which System A-2 is used as a potential bandwidth compression system. The solid line represents  $P_M(R_k)$ , and the distance between the solid line and the dotted line represents  $P_{SD}(R_k)$

Figure 8.8 Data segmentation for the analysis and construction of speech in a speech enhancement system based on System A-2

Figure 8.9 Results of the speech preference test in which System A-2 is used as a speech enhancement system. The solid line represents  $Q_M(R_k)$ , and the distance between the solid line and the dotted line represents  $Q_{SD}(R_k)$

Figure 9.1 Spectrogram of an English sentence "Line up at the screen door." spoken by a male speaker. The spacing between two consecutive lines on the vertical axis corresponds to 1 kHz.

Figure 9.2 Spectrogram of speech synthesized by the conventional LPC analysis/synthesis system at the  $S/N=0$  dB for the same English sentence shown in Figure 9.1. The spacing between two consecutive lines on the vertical axis corresponds to 1 kHz.

Figure 9.3 Spectrogram of speech synthesized by System A as a potential bandwidth compression system at the  $S/N=0$  dB for the same English sentence shown in Figure 9.1. The spacing between two consecutive lines on the vertical axis corresponds to 1 kHz.

TABLES

Table 2.1 Comparison of speech enhancement systems  
related to the correlation subtraction method

Table 8.1 Systems evaluated under an objective criterion

Table 8.2  $\overline{\text{LMSE}}$  and Normalized  $\overline{\text{LMSE}}$

Table 8.3 Results of the speech preference test in which  
System A-2 is used as a potential bandwidth compression  
system

Table 8.4 Results of the speech preference test in which  
System A-2 is used as a speech enhancement system

## CHAPTER I INTRODUCTION

### I.1 Introduction

Degradation of speech by additive noise occurs in a number of practical situations. For example, the speech of a pilot in a plane communicating with the ground control is degraded by the airplane noise. Another example is the speech of a lecturer recorded in a noisy lecture hall. The corrupting noise generally reduces [1] both the intelligibility and the quality of speech. Furthermore, the performance of many narrow-band communication systems degrades quickly [2,3] as the speech to noise ratio decreases. Thus, techniques for enhancement and bandwidth compression of noisy speech have a variety of applications.

In developing systems for speech enhancement, an important task is defining the goal of speech enhancement. A clear definition of this goal can potentially provide an objective criterion on the basis of which speech enhancement systems can be developed. Such a goal also provides a criterion for evaluating the performance of a system for the particular application under consideration. In general, speech enhancement implies a subjective improvement of the speech such as increased intelligibility and quality, reduced listener fatigue, etc. It is important to note that the subjective improvement, even though related, is not necessarily the same as the speech to noise ratio increase. For example, a speech processing

system which eliminates unvoiced segments and low-pass filters voiced segments of speech degraded by wide band additive noise may increase the overall S/N ratio but probably is not a speech enhancement system in most practical applications.

Another important aspect of developing a speech enhancement system is to accurately assess what information can be assumed about the speech and the background noise. Given a noisy speech signal with no assumptions of the speech or noise, there is little that can be done to enhance the speech signal. A general rule for any problem requiring the separation of individual signal components (combined by addition, convolution, etc.) is that the more we know about each component, the better we can solve the problem. Depending on the nature of the corrupting noise, some information of the noise may be obtained from the knowledge of the source, or from actual measurements. About speech, a great deal is known from the vast research efforts in the general area of the speech communications. We know a great deal about the human speech production mechanism and also have some understanding of the human perception of speech. In principle, we can attempt to incorporate everything we know about speech in developing a speech enhancement system. However, some of our knowledge is qualitative or complicated and its incorporation into such a system may be very

difficult. For example, human speech has linguistic constraints imposed by the rules of the language. But to incorporate such knowledge in a system for speech enhancement is probably a difficult task. Thus, the extent of our knowledge of speech that can be incorporated is limited by our capability to develop and implement systems that can exploit such available knowledge.

In developing a speech enhancement system, two different approaches can be taken. One is the "noise removal" approach in which a system is developed to eliminate as much background noise as possible with as little speech degradation as possible. The other approach is the "reconstruction" approach in which the speech parameters sufficient for reconstruction are estimated and then speech is reconstructed based on the estimated parameters. Which approach is better for speech enhancement depends on many factors such as how much we know about speech. However, for relatively high S/N ratios, it is expected that the noise reduction approach is better than the reconstruction approach since the latter generally changes the input speech.

Independent of which approach is taken, the essence of a speech enhancement system is an algorithm that incorporates, in some optimum manner, as much as possible of what we know about speech and the background noise. The optimality condition, ideally, should be based on the

specific goal of speech enhancement. In general, such a condition is unknown or quite complicated since a subjective quantity such as speech intelligibility can not easily be related to a measurable physical quantity that may be used as a criterion for optimality. In the absence of such a criterion or if the resulting system becomes highly complex even in the presence of such a criterion, we may consider a suboptimal procedure or define the optimal condition to be optimum in a different sense such as the maximum likelihood sense.

Suppose we have formulated an algorithm that incorporates our knowledge about speech and the background noise in some optimum manner, then the task remains to evaluate the performance of the system and estimate the implementation cost. In general, the performance improvement of a speech enhancement system can only be shown by an adequate evaluation. Many systems that have been proposed for speech enhancement provide apparent improvement in the S/N ratio, but on careful evaluation [4,5,6] in fact reduce intelligibility. If the system proposed is sufficiently complex such that the implementation cost is too high relative to the system performance, then an alternative procedure has to be considered. Under such a circumstance, we may have to go back to the beginning and redefine the goal of speech enhancement or reconsider the types of knowledge of speech and the background noise

to be incorporated into a speech enhancement system. Thus, developing a speech enhancement system under a specific objective and cost constraints requires a repetitive procedure that begins from a clear definition of the goal of speech enhancement and ends with a decision based on the evaluation of the system performance and estimation of the implementation cost, but probably after some iterations.

The problem of bandwidth compression of noisy speech is closely related to the speech enhancement problem. For example, a successful speech enhancement system with the reconstruction approach has the potential to be used as a bandwidth compression system for noisy speech. Alternatively, the noise reduction approach can be used as a pre-processor for a bandwidth compression system. Consequently, the approach to developing a bandwidth compression system for noisy speech is essentially the same as that for a speech enhancement system except for some additional considerations such as coding the speech parameters, the degree of bandwidth compression desired, etc. In fact, assuming the same knowledge of speech and the background noise, and using the same optimal criterion for both a speech enhancement system and a bandwidth compression system, we would expect that the speech enhancement system would look very similar to the bandwidth compression system. The only major difference would be



that for the speech enhancement system, the speech could be generated either by the noise removal or reconstruction approach whereas for the bandwidth compression system, speech must generally be reconstructed.

The problem of speech enhancement has received a great deal of attention in recent years and numerous systems have been proposed to enhance degraded speech. Nevertheless, significant improvements in speech intelligibility or quality in practical situations have not yet been demonstrated by any of the existing systems. Part of the problem appears to be that the approaches taken in developing various speech enhancement systems capitalize very little on our knowledge of speech. The proposed systems differ primarily in how the small amount of knowledge about the speech incorporated into the system is exploited and how the resulting speech is generated. It will become clear in our discussions in Chapter II that if we follow the same approach that has led to the various existing systems, we can easily generate systems at a faster rate than we can evaluate their performance or even implement them. Regardless of their performances, if we develop a speech enhancement system capitalizing more fully on our knowledge of speech in an "optimal" manner we would expect, in general, a better performance. In this dissertation, we develop systems for enhancement and bandwidth compression of noisy speech by attempting

to "optimally" incorporate a specific underlying speech model. The objective of this dissertation is, of course, to develop speech enhancement and bandwidth compression systems that are potentially applicable to practical situations. An equally important objective of this dissertation is to suggest the direction of other future research efforts by illustrating an example of a structured and theoretical approach for incorporating more of what we know about speech to develop enhancement and bandwidth compression systems of noisy speech.

## I.2 Scope of Thesis

In this dissertation, various speech enhancement systems proposed in the literature are summarized and related to each other in a more common framework. Some of the speech enhancement systems which appeared to be promising were studied more carefully and were evaluated in terms of their performance in improving speech intelligibility. As an attempt to optimally incorporate more of what we know about speech in developing systems for enhancement and bandwidth compression of noisy speech, a parameter estimation problem is formulated. The parameter estimation problem is then considered for both noise-free and noisy speech. For noise-free speech, different points of view such as Maximum Likelihood approach [7,8], Maximum A Posteriori estimation approach,

and Kalman filtering approach [9] are reviewed carefully and related to each other and to the conventional linear prediction analysis. For noisy speech, the parameter estimation problem is shown to be generally non-linear. Therefore, two "suboptimal" procedures which have linear implementations are developed. In addition, two systems for bandwidth compression and enhancement of noisy speech which are computationally simpler than the linear implementations are developed by approximating the linear implementations. As a preliminary investigation into the performance of systems developed in this dissertation, a small subset of the systems are implemented and applied to both synthetic and real speech data. An objective and informal subjective evaluation indicate that the implemented systems perform well as bandwidth compression and speech enhancement systems at various S/N ratios. Finally, a number of potential areas of study which are not performed as a part of the thesis but are within the scope of the theoretical results obtained in the thesis are summarized and a possible direction of future research in this area is suggested.

### I.3 Summary of Chapters

In Chapter II, various existing speech enhancement systems are summarized and related to each other in a common framework. In Chapter III, we discuss a specific model of

speech and the Maximum A Posteriori (MAP) estimation approach taken in this thesis to estimate the speech model parameters. In Chapter IV, the MAP estimation procedure for noise-free speech is discussed. The MAP estimation procedure under different assumptions leads to different sets of equations to solve, two of which are equivalent to the covariance and correlation method of the linear prediction analysis. In Chapter V, we discuss the MAP estimation problem for speech degraded by additive random noise. The theoretical results in this chapter will lead to two algorithms that require solving sets of linear equations in an iterative manner to estimate the speech model parameters from the noisy speech. In Chapter VI, we develop two systems based on the algorithms developed in Chapter V. The two systems developed are approximations of the two algorithms in Chapter V and are computationally simpler than the two algorithms. In addition to the two systems, we develop an "ad-hoc" system primarily for the comparison of the two systems developed in this thesis with other speech enhancement systems previously proposed. In Chapter VII, the performance of the three systems developed in Chapter VI in estimating the speech model parameters is qualitatively demonstrated by various examples based on both synthetic and real speech data. In Chapter VIII, the performance of the three systems is discussed in greater detail and quantitatively based on the

results of both the objective and subjective tests. The objective tests are based on the synthetic data and an objective criterion which reflects the perceptually important aspects of the speech parameters. The subjective tests are divided into two parts, one part corresponding to the bandwidth compression of noisy speech and the second part corresponding to speech enhancement. The comparison of various systems in terms of bandwidth compression are based on the synthesized sentences from the speech model parameters obtained by the developed systems. In the case of speech enhancement, two cases are considered. In the first case, speech is generated by the noise reduction approach. In the second case, speech is generated by a complete analysis/synthesis systems. In all cases of the subjective tests, the evaluation is informal and based on a few sentences spoken by both male and female speakers judged by listeners with no or some previous experience in the subjective tests. In Chapter IX, we suggest a direction and some potential areas of future research. In Chapter X, we conclude the thesis by summarizing the main results of this dissertation.

## CHAPTER II SURVEY OF SPEECH ENHANCEMENT TECHNIQUES

### II.1 Introduction

A number of techniques have been previously proposed for the enhancement of noisy speech. The purpose of this chapter is to summarize various speech enhancement techniques in a common framework and relate them to the bandwidth compression systems of noisy speech. In Section II.2, various speech enhancement systems are summarized and related to each other. In Section II.3, we summarize the performance of some of the systems discussed in Section II.2. Some of the results are based on an informal listening or a formal speech intelligibility test conducted in this research and some others are based on the studies by other researchers. In Section II.4, we discuss various bandwidth compression systems which are based on the speech enhancement systems summarized in Section II.2. In Section II.5, we discuss the motivation for a new approach to the problem of speech enhancement and bandwidth compression of noisy speech.

### II.2 Speech Enhancement Techniques

#### II.2.1 Adaptive Comb Filtering Method

Comb filtering for speech enhancement is based on the notion that voiced sounds are periodic with a period that corresponds to the fundamental frequency. Since the interfering signals in general have energy in the frequency

regions between the speech harmonics, a comb filtering operation in principle can reduce noise while preserving speech signals to the extent that information of the fundamental frequency is available and periodicity of speech is strictly preserved. Capitalizing on this knowledge, a comb filtering operation that passes only the harmonics of speech was first applied by Shields [10] to enhance degraded speech. Frazier [11] later observed that even with accurate fundamental frequency information Shields' adaptive comb filtering method distorts speech signals significantly due to the time varying nature of speech sounds. To reduce some of this distortion, Frazier suggested an adaptive comb filter [11] which adjusts itself to variations in the fundamental frequency. A further improvement on Frazier's algorithm on treating the transition regions between voicing and unvoicing was made by Lim [5]. In Frazier's algorithm, when voiced sounds near the transitions are processed, the adaptive comb filter extends over the unvoiced sounds due to the filter length which causes some distortion. By setting the filter coefficients that extend over unvoiced sounds to zero, Lim [5] found that a better performance can be obtained.

Comb filtering generally requires accurate pitch information. Parsons [12] developed a system which is similar to comb filtering but the pitch information is not

obtained separately but built into the system. More specifically, in an application to a competing speaker environment, each of the local spectral peaks in a high resolution short time Fourier transform of voiced sounds is distinguished between the main speaker and a competing speaker. Then speech is generated based on the spectral contents that correspond to the peaks of the main speaker.

Systems based on comb filtering have been evaluated in this research and by other researchers and the results are summarized in Section II.3.1.

#### II.2.2 Correlation Subtraction Method

The correlation subtraction method for speech enhancement is based on the notion that if additive noise is uncorrelated with the signal, then the correlation of the signal equals the noise correlation subtracted from the correlation of the observed signal. More specifically, when a signal is degraded by additive background noise, a noisy signal  $y(n)$  can be represented by

$$y(n) = s(n) + d(n) \quad (2-1)$$

in which  $s(n)$  and  $d(n)$  represent the signal and the background noise (or disturbance) respectively. Multiplying both sides of equation (2-1) by  $y(n-k)$  and taking the



expected value,

$$\begin{aligned} E[y(n) \cdot y(n-k)] &= E[s(n) \cdot s(n-k)] + E[d(n) \cdot d(n-k)] \\ &+ E[d(n) \cdot s(n-k)] + E[d(n-k) \cdot s(n)] \end{aligned} \quad (2-2)$$

If  $s(n)$  is assumed to be uncorrelated with  $d(n)$ , the last two terms in equation (2-2) disappear and thus

$$E[y(n) \cdot y(n-k)] = E[s(n) \cdot s(n-k)] + E[d(n) \cdot d(n-k)] \quad (2-3)$$

If  $s(n)$  and  $d(n)$  are assumed to be stationary so that the expectation of the two functions depends only on their time differences, equation (2-3) with a change of variables can be written as

$$R_y(n) = R_s(n) + R_d(n) \quad (2-4)$$

in which  $R_x(n)$  represents  $E[x(l) \cdot x(l-n)]$ , the correlation of  $x(n)$ . Fourier transforming equation (2-4) leads to

$$P_y(\omega) = P_s(\omega) + P_d(\omega) \quad (2-5)$$

in which  $P_x(\omega)$  represents  $F[R_x(n)] = \sum_{n=-\infty}^{\infty} R_x(n) \cdot e^{-j\omega n}$ , the power spectrum of  $x(n)$ . It is clear from equation (2-4) that the subtraction of  $R_d(n)$  from  $R_y(n)$  leads to  $R_s(n)$  and

expected value,

$$\begin{aligned} E[y(n) \cdot y(n-k)] &= E[s(n) \cdot s(n-k)] + E[d(n) \cdot d(n-k)] \\ &+ E[d(n) \cdot s(n-k)] + E[d(n-k) \cdot s(n)] \end{aligned} \quad (2-2)$$

If  $s(n)$  is assumed to be uncorrelated with  $d(n)$ , the last two terms in equation (2-2) disappear and thus

$$E[y(n) \cdot y(n-k)] = E[s(n) \cdot s(n-k)] + E[d(n) \cdot d(n-k)] \quad (2-3)$$

If  $s(n)$  and  $d(n)$  are assumed to be stationary so that the expectation of the two functions depends only on their time differences, equation (2-3) with a change of variables can be written as

$$R_y(n) = R_s(n) + R_d(n) \quad (2-4)$$

in which  $R_x(n)$  represents  $E[x(\ell) \cdot x(\ell-n)]$ , the correlation of  $x(n)$ . Fourier transforming equation (2-4) leads to

$$P_y(\omega) = P_s(\omega) + P_d(\omega) \quad (2-5)$$

in which  $P_x(\omega)$  represents  $F[R_x(n)] = \sum_{n=-\infty}^{\infty} R_x(n) \cdot e^{-j\omega n}$ , the power spectrum of  $x(n)$ . It is clear from equation (2-4) that the subtraction of  $R_d(n)$  from  $R_y(n)$  leads to  $R_s(n)$  and

thus the name "correlation subtraction" method.

In the case of speech, the correlation function can not be expressed as  $R_s(n)$  since speech can not be considered stationary. Thus we define the short time correlation of speech  $\phi_s(n)$  as

$$\phi_s(n) = \sum_{\ell=-\infty}^{\infty} s_w(\ell) \cdot s_w(\ell-n) \quad (2-6)$$

in which  $s_w(\ell)$  represents the windowed speech waveform. One important difference between  $\phi_s(n)$  and  $R_s(n)$  is  $\phi_s(n)$  can be defined for non-stationary signals as well as for stationary signals. Since  $y_w(n) = s_w(n) + d_w(n)$ , multiplying both sides with  $y_w(n-k)$  and summing over all  $n$  leads to

$$\phi_s(n) = \phi_y(n) - \phi_d(n) - 2 \cdot \phi_{sd}(n) \quad (2-7)$$

where

$$\phi_y(n) = \sum_{\ell=-\infty}^{\infty} y_w(\ell) \cdot y_w(\ell-n),$$
$$\phi_d(n) = \sum_{\ell=-\infty}^{\infty} d_w(\ell) \cdot d_w(\ell-n),$$

and

$$\phi_{sd}(n) = \sum_{\ell=-\infty}^{\infty} s_w(\ell) \cdot d_w(\ell-n)$$

Equation (2-7) is exact without any approximations. We will find that a number of speech enhancement systems

summarized in this chapter differ primarily in how  $\phi_s(n)$  is specifically estimated and how speech is generated once  $\phi_s(n)$  is estimated. We will also find that in various speech enhancement systems, equation (2-7) is a starting point for estimating  $\phi_s(n)$  from  $y(n)$ . Before we discuss how  $\phi_s(n)$  is specifically estimated in the correlation subtraction method, it is worthwhile to note why it is important to attempt to estimate  $\phi_s(n)$  accurately. From equation (2-6)  $\phi_s(n)$  is related to  $|S_w(\omega)|$ , the magnitude of the discrete time Fourier transform of  $s_w(n)$ , by

$$|S_w(\omega)|^2 = F[\phi_s(n)] = \sum_{n=-\infty}^{\infty} \phi_s(n) \cdot e^{-j\omega n} \quad (2-8)$$

Thus the attempt to estimate  $\phi_s(n)$  more accurately is equivalent to attempting to preserve the short time spectral information of speech  $|S_w(\omega)|$  which is known [13] to be important for both the intelligibility and quality of speech.

In the correlation subtraction method,  $\phi_s(n)$  is estimated based on equation (2-7). From the windowed noisy speech  $y_w(n)$ ,  $\phi_y(n)$  can be directly computed.  $\phi_d(n)$  and  $\phi_{sd}(n)$  can not be obtained exactly from  $y(n)$  unless  $d(n)$  is exactly known and in the correlation subtraction method,  $\phi_d(n)$  and  $\phi_{sd}(n)$  are approximated by  $E[\phi_d(n)]$  and  $E[\phi_{sd}(n)]$ . For a zero mean  $d(n)$  uncorrelated with  $s(n)$ ,  $E[\phi_{sd}(n)]$  equals zero and therefore equation (2-7) can be

approximately written as

$$\phi_s(n) \approx \phi_y(n) - E[\phi_d(n)] \quad (2-9)$$

$E[\phi_d(n)]$  can be obtained either from the assumed known statistics of  $d(n)$  or by an actual measurement from the background noise in the intervals when speech is not present. Fourier transforming equation (2-9),

$$|S_w(\omega)|^2 \approx |Y_w(\omega)|^2 - E[|D_w(\omega)|^2] \quad (2-10)$$

Based on equations (2-9) and (2-10),  $\phi_s(n)$  and  $|S_w(\omega)|^2$  are estimated as

$$\hat{\phi}_s(n) = \phi_y(n) - E[\phi_d(n)] \quad (2-11a)$$

and

$$|\hat{S}_w(\omega)|^2 = |Y_w(\omega)|^2 - E[|D_w(\omega)|^2] \quad (2-11b)$$

From equation (2-11b),  $|\hat{S}_w(\omega)|^2$  is not guaranteed to be non-negative. This is because there is no built-in mechanism in the above estimation procedure to force  $\hat{\phi}_s(n)$  to correspond to the short time correlation of some real sequence. When such a situation does occur, a number of different arbitrary steps may be taken. In some studies, the negative values are made positive by changing the sign.

In some other studies  $|S_w(\omega)|^2$  is set to zero if  $|Y_w(\omega)|^2$  is less than  $E[|D_w(\omega)|^2]$ .

Given an estimate of  $\phi_s(n)$  or  $|S_w(\omega)|$ , there are a number of different ways to generate speech. One method which is popular in the class of systems related to some form of spectral subtraction is to approximate  $\angle S_w(\omega)$ , the phase of  $S_w(\omega)$ , by  $\angle Y_w(\omega)$  and then generate  $s_w(n)$  or  $S_w(\omega)$  by

$$\hat{S}_w(\omega) = |\hat{S}_w(\omega)| \cdot e^{j\angle \hat{S}_w(\omega)} \quad (2-12a)$$

and

$$\hat{s}_w(n) = F^{-1}[\hat{S}_w(\omega)] \quad (2-12b)$$

A typical algorithm for speech enhancement by the correlation subtraction method is shown in Figure 2.1. The system in Figure 2.1 has been evaluated in this research and the results are summarized in Section II.3.2.

Generating  $\hat{s}_w(n)$  by equation (2-12) corresponds to taking the noise reduction approach for speech enhancement. As we discussed in Chapter I, it is possible to take the reconstruction approach as we'll see shortly.

### II.2.3 Speech Enhancement by a Voice Excited Vocoder

Magill and Un [14] developed a speech enhancement system by a voice excited LPC vocoder when the background

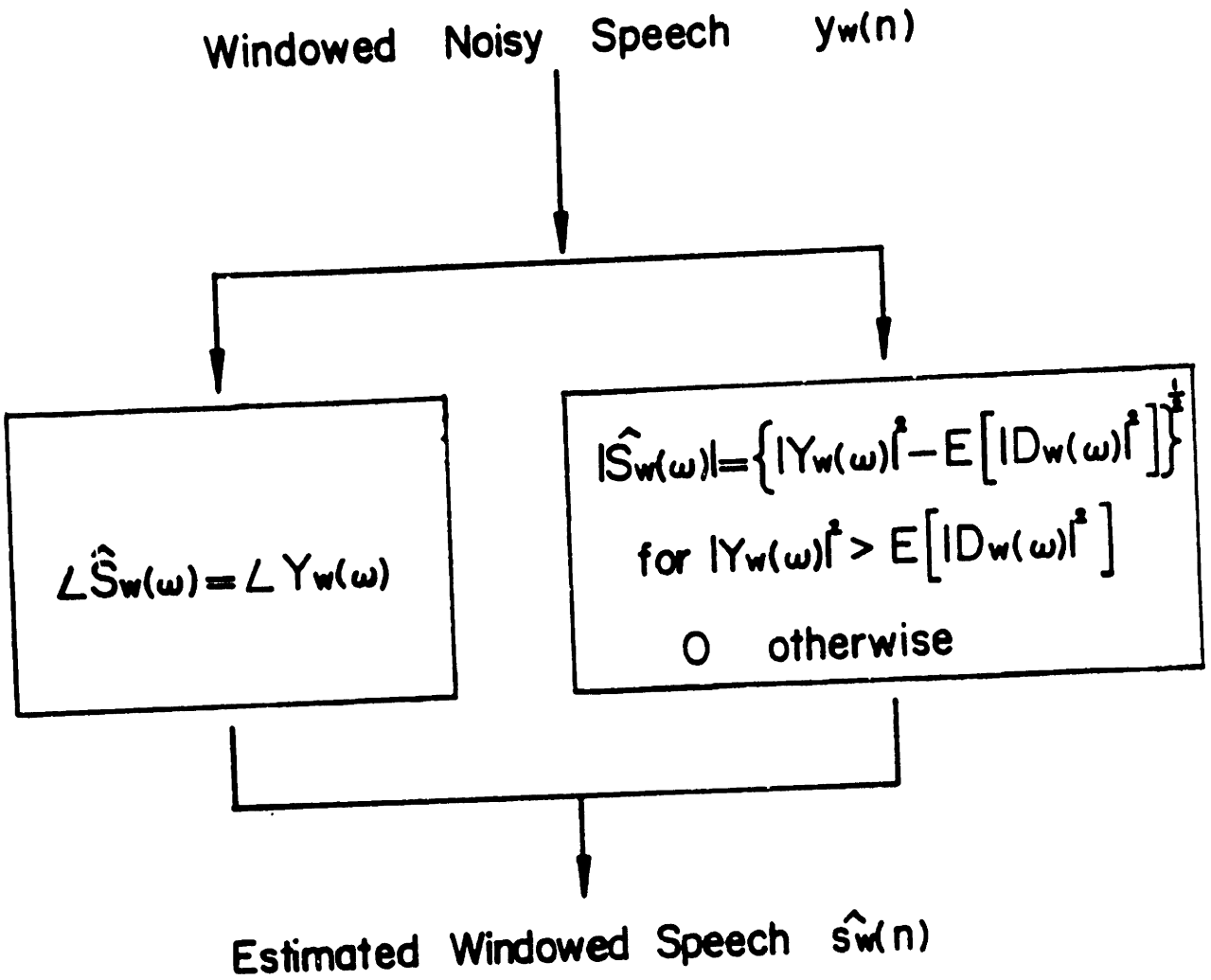


Figure 2.1 A typical speech enhancement system by the correlation subtraction method

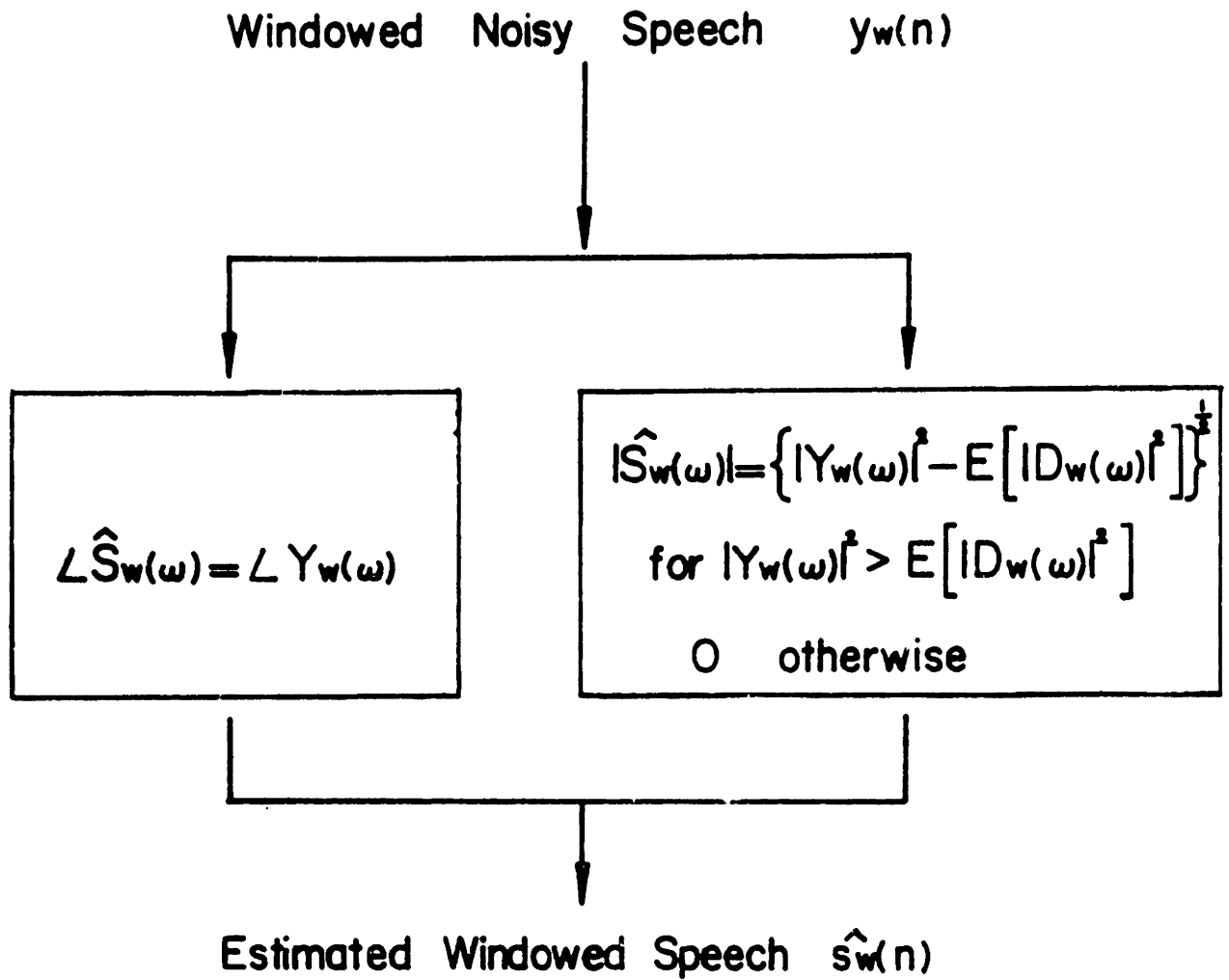


Figure 2.1 A typical speech enhancement system by the correlation subtraction method



noise is white. The system information, namely the LPC coefficients, is obtained by the correlation method of the linear prediction analysis in which the short time correlation of speech is estimated by the correlation subtraction method discussed in Section II.2.2. For the source information, the noisy speech is low pass filtered at 600 Hz and then non-linearly distorted to broaden its bandwidth. This is based on the notion that voiced speech generally decays at 6 db/octave rate and therefore the low frequency components are least degraded by additive white noise. Speech is then generated based on the estimated source and system information.

The system by Magill and Un is identical to the correlation subtraction method in estimating  $\phi_s(n)$  from  $y(n)$ . The difference lies in how speech is generated based on the estimated  $\hat{\phi}_s(n)$ . The reconstruction approach taken in this system has a disadvantage in that the source information has to be obtained in some manner. However, it has the advantage that the speech enhancement system can be used not only as a pre-processor for various bandwidth compression systems of noisy-free speech, but also as a bandwidth compression system itself. The performance of the system by Magill and Un is not known.

#### II.2.4 INTEL System

Weiss, et al. [15] developed a speech enhancement

system called INTEL or "Intelligibility Enhancement by Liftering". The INTEL system has several versions. One early version is based on the notion that in the short time correlation domain speech is in general more spread from the origin than the background noise such as white noise. Therefore some form of gating out (liftering) the low time region of the short time correlation of noisy speech may eliminate more noise components than speech and thus may lead to some speech enhancement. When a system based on this method was implemented by Weiss, et al. [15] and also in this research, the performance of the system was found to be rather poor.

Another version of the INTEL system which in a sense is a generalization of the correlation subtraction method has been studied in some detail in this research. The INTEL system referred from this point on corresponds to this version of the INTEL system. In Section II.2.2, it was shown that the correlation subtraction method corresponds to estimating the short time correlation of speech  $\phi_s(n)$  by  $F^{-1}[|Y_w(\omega)|^2] - E[F^{-1}[|D_w(\omega)|^2]]$ . Weiss, et al. simply replaced the squaring operation with an arbitrary positive real constant "a". In this method, then, by defining  $\phi'_s(n)$  to be  $F^{-1}[|S_w(\omega)|^a]$ ,  $\phi'_s(n)$  is estimated by  $F^{-1}[|Y_w(\omega)|^a] - E[F^{-1}[|D_w(\omega)|^a]]$ . Based on this estimate of  $\phi'_s(n)$  and the assumption that  $\hat{S}_w(\omega)$  equals  $\hat{Y}_w(\omega)$ , speech is generated. The speech enhancement system proposed

by Weiss, et al. is shown in Figure 2.2.

The algorithm in Figure 2.2 can be simplified both computationally and conceptually by recognizing that the expectation and Fourier transform operations are linear and hence can be inter-changed. Such a simplified system is shown in Figure 2.3. The figure clearly shows that the INTEL system is one way of estimating the short time spectral magnitude of speech. In Figure 2.3 when  $|S_w(\omega)|$  obtained is not positive, it is set to zero for the similar reason discussed in Section II.2.2. The performance of the INTEL system is summarized in Section II.3.2.

#### II.2.5 SABER Method

Boll [16] developed a speech enhancement system called SABER or "Spectral Averaging for Bias Estimation and Removal". In this method,  $|S_w(\omega)|$  is estimated by subtracting  $E[|D_w(\omega)|]$  from a local average of  $|Y_w(\omega)|$ . More specifically, it is assumed that

$$|S_w(\omega)| \approx \frac{1}{M} \sum_i |Y_w(\omega)|_i - E[|D_w(\omega)|] \quad (2-13a)$$

where  $|Y_w(\omega)|_i$  represents  $|Y_w(\omega)|$  obtained from the  $i$ th segment of the noisy speech and  $M$  is the number of consecutive windows used for local averaging.

To relate the SABER method to the INTEL system, we rewrite equation (2-13a) as follows:

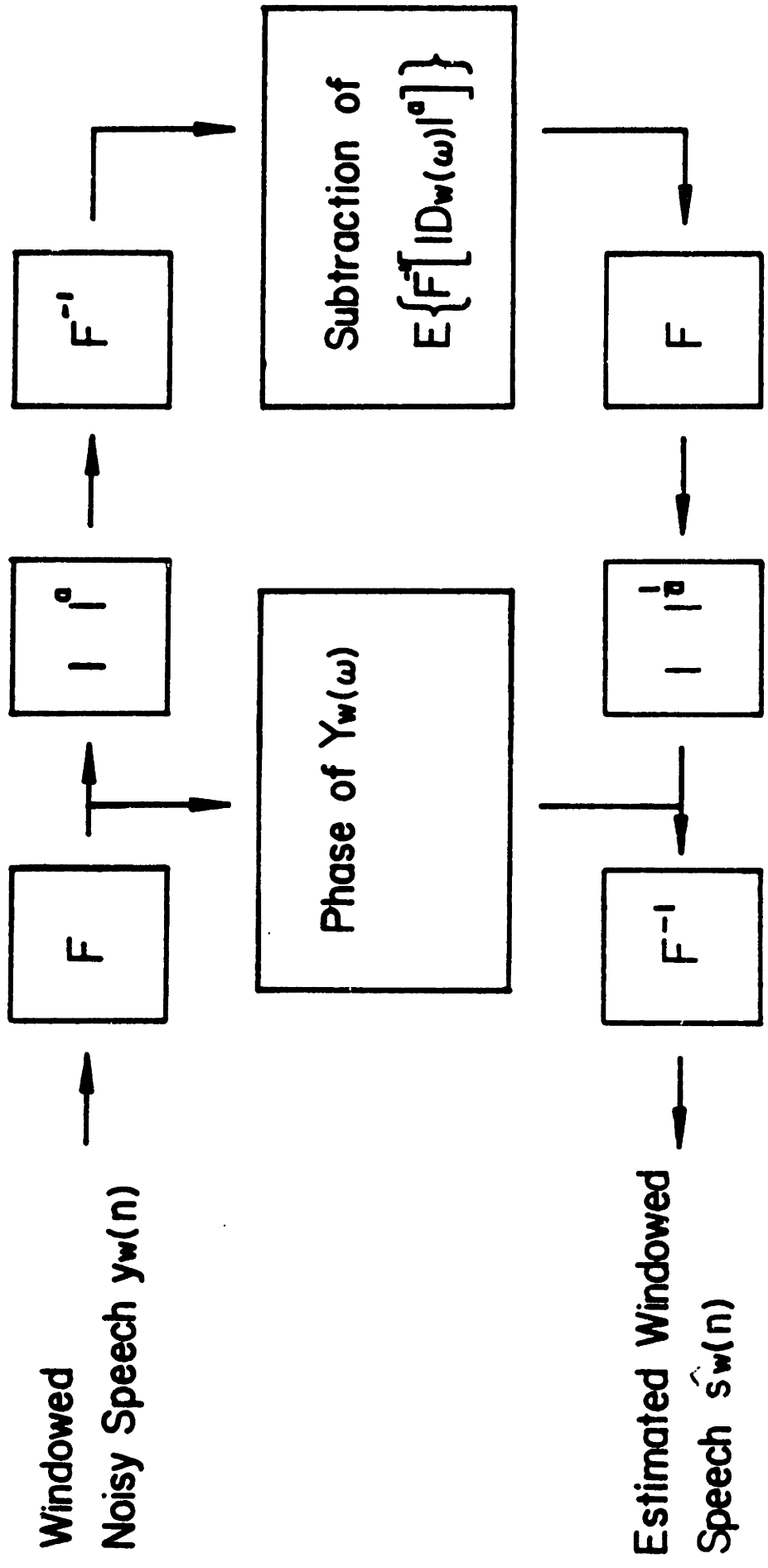


Figure 2.2 INTEL system proposed by Weiss et al.

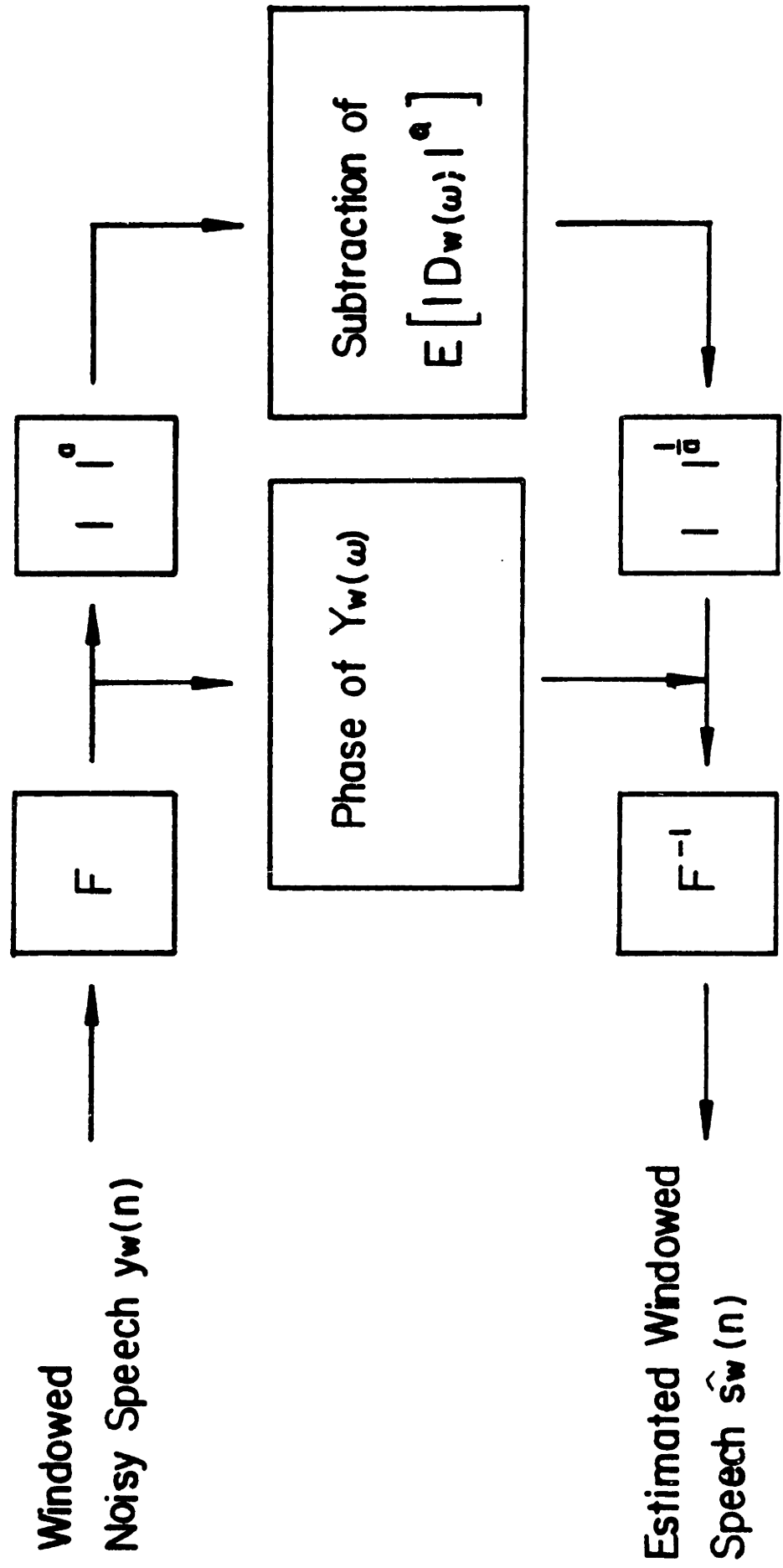


Figure 2.3 A simplified representation of the INTEL system in Figure 2.2

$$|S_w(\omega)| \approx \frac{1}{M} \sum_i (|Y_w(\omega)|_i - E[|D_w(\omega)|]) \quad (2-13b)$$

The term  $|Y_w(\omega)|_i - E[|D_w(\omega)|]$  in equation (2-13) is how  $|S_w(\omega)|_i$  is estimated by the INTEL system with  $a=1$ . Therefore the SABER method is equivalent to estimating  $|S_w(\omega)|$  by a local average of the sets of  $|S_w^\wedge(\omega)|$  obtained by the INTEL system with  $a=1$  if the same windows are used in both cases. In fact, in the implementation of the INTEL system, some form of local averaging is done by applying the windows that are overlapped with each other to the input noisy speech data. In this context, then, the SABER method can be viewed as a variation of a special case of the INTEL system shown in Figure 2.3. The evaluation results of the SABER method reported by Boll are summarized in Section II.3.3.

In a more recent study [17], Boll reported that the local averaging discussed above is not important in his system.

## II.2.6 Other Generalizations of Correlation

### Subtraction Method

The INTEL system discussed in Section II.2.4 is in a sense an arbitrary generalization of the correlation subtraction method. An alternative arbitrary generalization is to estimate  $|S_w(\omega)|^2$  by  $|Y_w(\omega)|^2 - k \cdot E[|D_w(\omega)|^2]$  for some arbitrary constant  $k$  and based on this estimate of

$|S_w(\omega)|$ , speech can be generated in the same manner as in the correlation subtraction method. This system was proposed [18] for possible speech enhancement and studied in this research. The performance of this system is summarized in Section II.3.4.

In a more recent study [19], Schwartz et al. considered for speech enhancement the same system discussed above. In their study, an additional feature is included in that after the subtraction  $|\hat{S}_w(\omega)|^2$  obtained is compared to a threshold level  $\beta \cdot E[|D_w(\omega)|^2]$  for a small arbitrary constant  $\beta$  and if  $|\hat{S}_w(\omega)|^2$  is smaller it is set to  $\beta \cdot E[|D_w(\omega)|^2]$ . Thus in their system,

$$|\hat{S}_w(\omega)|^2 = |Y_w(\omega)|^2 - k \cdot E[|D_w(\omega)|^2]$$

$$\text{for } |Y_w(\omega)|^2 > (k+\beta) \cdot E[|D_w(\omega)|^2],$$

$$\beta \cdot E[|D_w(\omega)|^2] \quad \text{otherwise}$$

Clearly, there exist a number of other arbitrary generalizations. For example, we could estimate  $|S_w(\omega)|^a$  by  $|Y_w(\omega)|^a - k \cdot E[|D_w(\omega)|^a]$  for some arbitrary constants  $a$  and  $k$ . Such a system includes both the INTEL system (by setting  $k=1$ ) and the system discussed in this section (by setting  $a=2$ ) as special cases.

## II.2.7 SPAC and SPOC

Suzuki developed [20] a speech enhancement system called SPAC or "Splicing of Autocorrelation Function". SPOC or "Splicing of Cross-correlation Function" is a revised version [21] of SPAC. The two systems have been used for compression or expansion of the spectrum, or lengthening or shortening the duration of speech, or reducing the noise level in the speech signal. In the discussions in this section only the noise reduction aspect is considered.

SPAC is based on the notion that the short time correlation of speech has common frequency components with the short time speech. Therefore, for voiced sounds that are periodic with the fundamental frequency, the short time correlation properly defined is also periodic with the fundamental frequency. Furthermore, if one replaces each pitch period of speech with the corresponding pitch period of the short time correlation, then the frequency components of speech would be unchanged except that the spectral magnitude at each frequency would be approximately squared. Since the effect of the background noise such as white noise generally degrades more the points near the origin in the short time correlation domain, speech may be enhanced by replacing each pitch period of speech with one pitch period of the corresponding short time correlation beginning some points away from the



origin. Suzuki observed that SPAC causes some distortions due to the squaring operation of the spectral magnitude of speech caused by replacing speech with its short time correlation. SPOC is a revision of SPAC to reduce such distortions.

To appreciate how this method compares to other methods in terms of its performance, we consider a very simple example. Suppose the background noise is zero mean and white Gaussian with the variance of  $\sigma_d^2$  and further assume that  $s(n)$  is periodic with the period of  $T$  such that  $s(n+T) = s(n)$  for all  $n$ . We define the short time correlation of speech  $\phi_s^*(n)$  at  $n_0$  by

$$\phi_s^*(n) \triangleq \sum_{\lambda=n_0}^{n_0+M-1} s(\lambda) \cdot s(\lambda-n)$$

for some fixed  $M$  and  $\phi_y^*(n)$  and  $\phi_d^*(n)$  are similarly defined. Note that  $\phi_s^*(n)$  is slightly different from  $\phi_s(n)$  in that the summation is over  $M$  number of points independent of  $n$ . Three cases are considered. In the first case,  $\phi_s^*(n)$  is simply estimated as  $\phi_y^*(n)$  and thus

$$\hat{\phi}_s^*(n) = \phi_y^*(n) \quad \text{for } 0 \leq n \leq T-1 \quad (2-14)$$

In the second case,  $\phi_s^*(n)$  is estimated by  $\phi_y^*(n) - E[\phi_d^*(n)]$  and therefore

$$\hat{\phi}_s^*(n) = \phi_y^*(n) - M\sigma_d^2 \cdot \delta(n) \quad \text{for } 0 \leq n \leq T-1 \quad (2-15)$$

This case corresponds to the correlation subtraction method. The third case corresponds to estimating  $\phi_s^*(n)$  by SPAC and therefore

$$\begin{aligned} \hat{\phi}_s^*(n) &= \phi_y^*(n+T) & \text{for } n = 0 \\ &\phi_y^*(n) & \text{for } 1 \leq n \leq T-1 \end{aligned} \quad (2-16)$$

Comparing equations (2-14), (2-15) and (2-16),  $\phi_s^*(n)$  estimated is the same for  $1 \leq n \leq T-1$  in all three cases. Defining  $e(0) = \phi_s^*(0) - \hat{\phi}_s^*(0)$ , it can be easily shown for case 1,

$$\begin{aligned} E[e(0)] &= M \cdot \sigma_d^2 \\ \text{Var}[e(0)] &= 4 \cdot \sum_{\ell=n_0}^{n_0+M-1} s^2(\ell) \cdot \sigma_d^2 + 2M \cdot \sigma_d^4 \end{aligned} \quad (2-17a)$$

for case 2,

$$\begin{aligned} E[e(0)] &= 0 \\ \text{Var}[e(0)] &= 4 \cdot \sum_{\ell=n_0}^{n_0+M+1} s^2(\ell) \cdot \sigma_d^2 + 2M \cdot \sigma_d^4 \end{aligned} \quad (2-17b)$$

and for case 3,

$$E[e(0)] = 0$$

$$\text{Var}[e(0)] = 2 \cdot \sum_{\ell=n_0}^{n_0+M-1} s^2(\ell) \cdot \sigma_d^2 + M \cdot \sigma_d^4 + k$$

in which  $k \ll 2 \sum_{\ell=n_0}^{n_0+M-1} s^2(\ell) \cdot \sigma_d^2$

and therefore  $\text{Var}[e(0)] \approx 2 \sum_{\ell=n_0}^{n_0+M-1} s^2(\ell) \cdot \sigma_d^2 + M \cdot \sigma_d^4$  (2-17c)

The above comparison shows that the correlation subtraction method eliminates the bias but does not reduce the error variance. SPAC eliminates the bias and reduces the error variance by about 50%.

On the other hand, SPAC requires an estimation of the fundamental frequency and speech is not strictly periodic even for voiced sounds. Furthermore, SPAC can not be applied to unvoiced sounds and even with the revision made by SPOC, there are some spectral degradations due to replacing speech with the short time correlation type of function. The performance of SPAC or SPOC is not known.

### II.2.8 Wiener Filtering Method

If  $y(n) = s(n) + d(n)$  in which  $s(n)$  and  $d(n)$  are samples obtained from uncorrelated stationary random processes and if  $y(n)$  is available for all time, it is

well known [22] that the optimum linear estimator that minimizes  $E[(s(n) - \hat{s}(n))^2]$  in which  $\hat{s}(n)$  represents the estimate of  $s(n)$  is given by the non-causal Wiener filter whose frequency response is given by

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} \quad (2-18)$$

where  $P_x(\omega)$  represents the power spectrum of  $x(n)$ .

Callahan [23] approximates the non-causal Wiener filter in terms of the average short time energy spectrum and thus

$$H(\omega) = \frac{E[\phi_s(\omega)]}{E[\phi_s(\omega)] + E[\phi_d(\omega)]} \quad (2-19)$$

in which  $\phi_s(\omega)$  and  $\phi_d(\omega)$  are given by  $F[\phi_s(n)]$  and  $F[\phi_d(n)]$ .  $E[\phi_d(\omega)]$  can be obtained either from the assumed known statistics of  $d(n)$  or by averaging many frames of  $\phi_d(\omega)$  during which noise can be assumed to be stationary.  $E[\phi_s(\omega)]$  is estimated by subtracting  $E[\phi_d(\omega)]$  from locally averaged  $\phi_y(\omega)$  over many consecutive windows. Callahan notes that to estimate  $E[\phi_y(\omega)]$  within an acceptable variance,  $\phi_y(\omega)$  should be averaged over at least 100 msec which is a relatively long interval during which speech may not be assumed to be stationary. If  $E[\phi_s(\omega)]$  estimated is negative, it is set to zero. The short time Fourier transform  $S_w(\omega)$  is then estimated by multiplying

$Y_w(\omega)$  with  $H(\omega)$  given in equation (2-19). Thus in this system,  $|S_w(\omega)|$  is estimated by  $|Y_w(\omega)| \cdot H(\omega)$  where  $H(\omega)$  is obtained from equation (2-19) and  $\hat{S}_w(\omega)$  is assumed to be  $\hat{Y}_w(\omega)$ . In the specific algorithm by Callahan, only one point of  $\hat{s}_w(n)$  is obtained from the estimated  $\hat{S}_w(\omega)$  and the window slides through  $y(n)$  by one point at a time. The performance of this system reported by Callahan is summarized in Section II.3.5.

It appears that there are a number of other ways to obtain  $E[\phi_y(\omega)]$  used in estimating  $H(\omega)$  in equation (2-19). Instead of averaging  $\phi_y(\omega)$  over 100 msec, an equally reasonable way appears to be to perform some kind of smoothing on  $\phi_y(\omega)$  and assume the smoothed  $\phi_y(\omega)$  to be  $E[\phi_y(\omega)]$ . Also, if we want to generalize the Wiener filtering method arbitrarily as was done in the case of the correlation subtraction method, there are, of course, numerous possibilities.

### II.2.9 Summary

In this section, various speech enhancement systems discussed in Section II.2 are briefly summarized. The comb filtering method is an attempt to increase the S/N ratio based on the periodicity of voiced sounds. SPAC or SPOC is based on the notion that in the correlation domain the effect of the background noise is typically more pronounced near the origin while speech repeats itself

in each pitch period. In generating speech in SPAC or SPOC, the notion that voiced sounds are periodic and the spectral contents of one period of speech is closely related to one period of its correlation is exploited.

All other methods discussed in Section II.2 differ primarily in how  $\phi_s(n)$  or  $|S_w(\omega)|$  is estimated and how speech is generated based on  $\hat{\phi}_s(n)$  or  $|\hat{S}_w(\omega)|$ . Their differences are summarized in Table 2.1.

## II.3 Summary of Performance Evaluation

### II.3.1 Adaptive Comb Filtering Method

Speech enhancement techniques related to comb filtering have been evaluated more extensively relative to other techniques. Using Frazier's system [11], Perlmutter [4] processed some speech material that consist of nonsense sentences and performed intelligibility tests with interference consisting of the speech of a competing talker. Her results indicate that even with accurate fundamental frequency information, the adaptive comb filtering method decreases intelligibility at the S/N ratios where the intelligibility of unprocessed nonsense sentences range between 20 to 70%.

As a part of this research, Frazier's adaptive comb filtering method with the improvement made by Lim [5] has been evaluated by using nonsense sentences as test materials when the interference is wide band random noise. In Figure

TABLE 2.1 Comparison of Speech Enhancement Systems Related to the Correlation Subtraction Method

System	Estimation of $ S_w(\omega) $ and $\langle S_w(\omega) \rangle$	Speech Generation
Correlation Subtraction Method	$ \hat{S}_w(\omega)  = ( Y_w(\omega) ^2 - E[ D_w(\omega) ^2])^{1/2}$ for $ Y_w(\omega) ^2 \geq E[ D_w(\omega) ^2]$ 0 otherwise. $\hat{S}_w(\omega) = \hat{Y}_w(\omega)$	$s_w(n) = F^{-1} [   \hat{S}_w(\omega)   \cdot e^{j\angle \hat{S}_w(\omega)} ]$ noise reduction approach
System by Magill and Un	Same as Correlation Subtraction Method	use of voice excited vocoder; reconstruction approach
INTEL System	$ \hat{S}_w(\omega)  = ( Y_w(\omega) ^a - E[ D_w(\omega) ^a])^{1/a}$ for $ Y_w(\omega) ^a > E[ D_w(\omega) ^a]$ 0 otherwise. $\hat{S}_w(\omega) = \hat{Y}_w(\omega)$	same as correlation subtraction method
SABER method	Same as the INTEL system with $a=1$	same as correlation subtraction method
Another Generalization of Correlation Subtraction Method	$ \hat{S}_w(\omega)  = ( Y_w(\omega) ^2 - k \cdot E[ D_w(\omega) ^2])^{1/2}$ for $ Y_w(\omega) ^2 > k \cdot E[ D_w(\omega) ^2]$ 0 otherwise. $\hat{S}_w(\omega) = \hat{Y}_w(\omega)$	same as correlation subtraction method

(Table 2.1 continued)

Table 2.1 Continued

System	Estimation of $ S_w(\omega) $ and $\langle S_w(\omega) \rangle$	Speech Generation
Wiener Filtering Method by Callahan	$ S_w^{\wedge}(\omega)  =  Y_w(\omega)  \cdot H(\omega), \text{ where } H(\omega) = \frac{E[\phi_s(\omega)]}{E[\phi_s(\omega)] + E[\phi_d(\omega)]}$ $\text{where } E[\phi_s(\omega)] = \frac{1}{M} \cdot \sum_{i=1}^M  Y_w(\omega) ^2 - E[ D_w(\omega) ^2]$ $\text{for } \frac{1}{M} \cdot \sum_{i=1}^M  Y_w(\omega) ^2 \geq E[ D_w(\omega) ^2]$ $0 \text{ otherwise. } \quad \langle S_w^{\wedge}(\omega) \rangle = \langle Y_w(\omega) \rangle$	Same as Correlation Subtraction Method



2.4 is shown the results of the intelligibility test as a function of the S/N ratio and the length of the adaptive comb filter. The results of the test show that even with carefully hand edited pitch information, an adaptive comb filtering method tends to decrease the speech intelligibility at the S/N ratios where the intelligibility scores of unprocessed nonsense sentences range between 20 and 70%. Since in practice accurate pitch information is not available and can not be expected to be obtained from degraded speech, the intelligibility scores will be even lower than shown in Figure 2.4.

The evaluation results of the systems by Parsons is not available. However, an informal listening indicates that the performance is similar to Frazier's system when applied to a competing speaker environment.

### II.3.2 Correlation Subtraction Method and INTEL System

As we discussed in Section II.2, the INTEL system is in a sense an arbitrary generalization of the correlation subtraction method. More specifically, the case when  $a=2$  in the INTEL system corresponds to the correlation subtraction method. In this research, the performance of the INTEL system in Figure 2.3 has been evaluated [6] by using nonsense sentences as test materials when the interference is wide band random noise. This study was

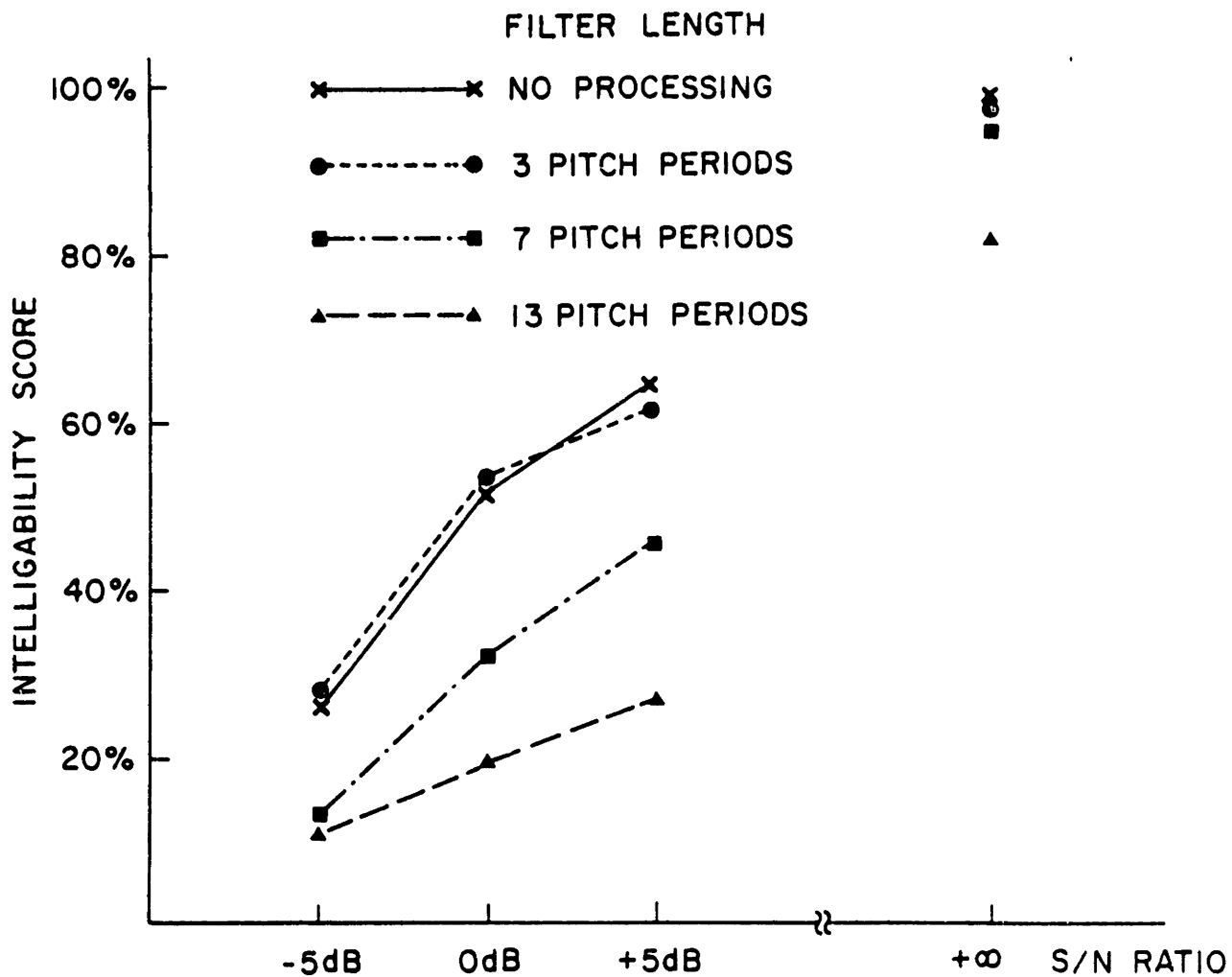


Figure 2.4 Results of the intelligibility test performed to evaluate the adaptive comb filtering method for enhancement of noisy speech

motivated primarily by the subjective impression that substantial noise reduction was achieved by the INTEL system. In Figure 2.5 is shown the results of the intelligibility test as a function of the S/N ratio and the constant "a". The results of the test show that the system does not increase the speech intelligibility at the S/N ratios where the intelligibility scores of unprocessed nonsense sentences range between 20 and 70%. Based on our informal subjective judgement, however, the processed speech by the INTEL system sounds "less noisy" and of higher quality at relatively high S/N ratios. Thus if the system is evaluated at higher S/N ratios, in terms of speech quality or as a pre-processor for a bandwidth compression system, then the system may be found to be useful. There is some indication that the above may be true, as will be discussed in the next section.

### II.3.3 SABER Method

Boll reported [17] the results of a very preliminary evaluation of the SABER method, which corresponds to  $a=1$  of the INTEL system. His results by the Diagnostic Rhyme test indicate that at the S/N ratio at which the intelligibility score of the unprocessed speech material is about 84% the SABER method does not increase speech intelligibility which is consistent with our results of the INTEL system with  $a=1$ . However, when speech quality is

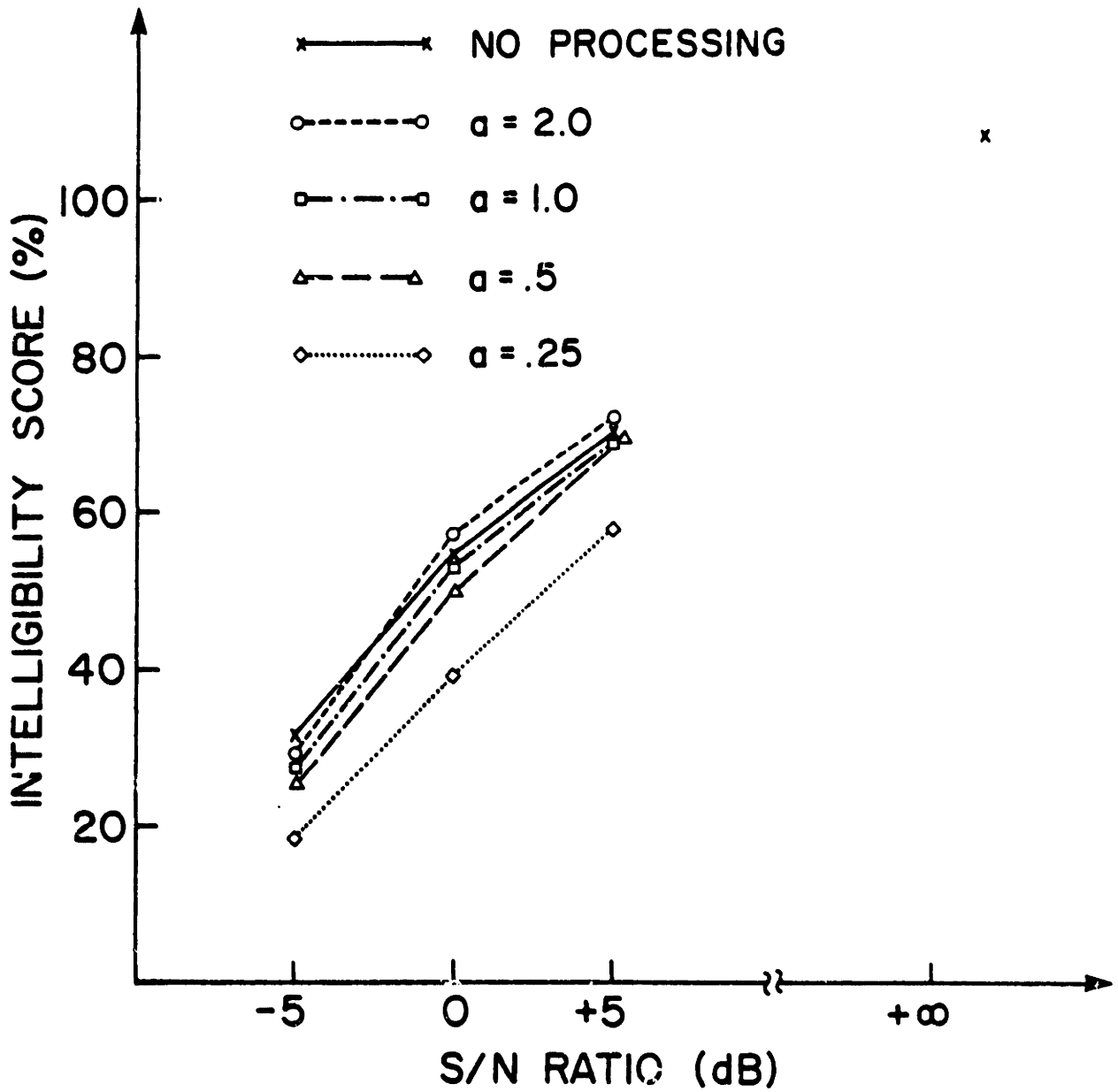


Figure 2.5 Results of the intelligibility test performed to evaluate the INTEL system for enhancement of noisy speech

tested [17] or the SABER method is used as a pre-processor of a bandwidth compression system, some improvement is noted at the above S/N ratio.

#### II.3.4 Other Generalizations of Correlation

##### Subtraction Method

Even though an extensive intelligibility test has not been performed to evaluate the system discussed in Section II.2.6 ( $|S_w^{\wedge}(\omega)|^2 = |Y_w(\omega)|^2 - k \cdot E[|D_w(\omega)|^2]$ ), based on an informal listening test it appears that the performance of this system is similar to the INTEL system, with a higher value of  $k$  generally corresponding to a smaller value of  $a$ . For a wide ranging S/N ratios (below approximately 5 db), a value of  $k$  less than 2 appears to be better. A large value of  $k$  at low S/N ratios has the effect of essentially eliminating the unvoiced sounds and higher formants of voiced sounds. Further details on the performance of this system will be discussed later in this thesis.

The system by Schwartz et al. which has an additional parameter  $\beta$  is reported [19] to eliminate some perceptually unpleasant speech degradation in the processing by a proper choice of  $\beta$ .

#### II.3.5 Wiener Filtering Method

Callahan applied the Wiener filtering method discussed

in Section II.2.8 to reduce surface noise of a 1907 recording by Enrico Caruso and reported [23] that the technique "greatly reduces" the surface noise. The performance of the system when applied to enhance noisy speech is not known.

#### II.4 Bandwidth Compression Systems of Noisy Speech

Our discussions in Sections II.2 and II.3 have been primarily concerned with speech enhancement systems. However, most of the discussions apply equally well to the bandwidth compression systems of noisy speech, since the two are closely related to each other, as we discussed in Chapter I. A successful speech enhancement system can in general be used as a part of a bandwidth compression system of noisy speech. This point is obvious for a class of speech enhancement systems based on an analysis/synthesis system. Alternatively, a successful speech enhancement system can potentially be used as a pre-processor for a bandwidth compression system of noise-free speech, in which case we can represent an overall bandwidth compression system of noisy speech as shown in Figure 2.6.

In some cases, the system in Figure 2.6 can be simplified. For example, a speech enhancement system such as the correlation subtraction method is directed towards estimating  $|S_w(\omega)|$  more accurately. In a bandwidth compression system such as an LPC vocoder [24,25], a homomorphic

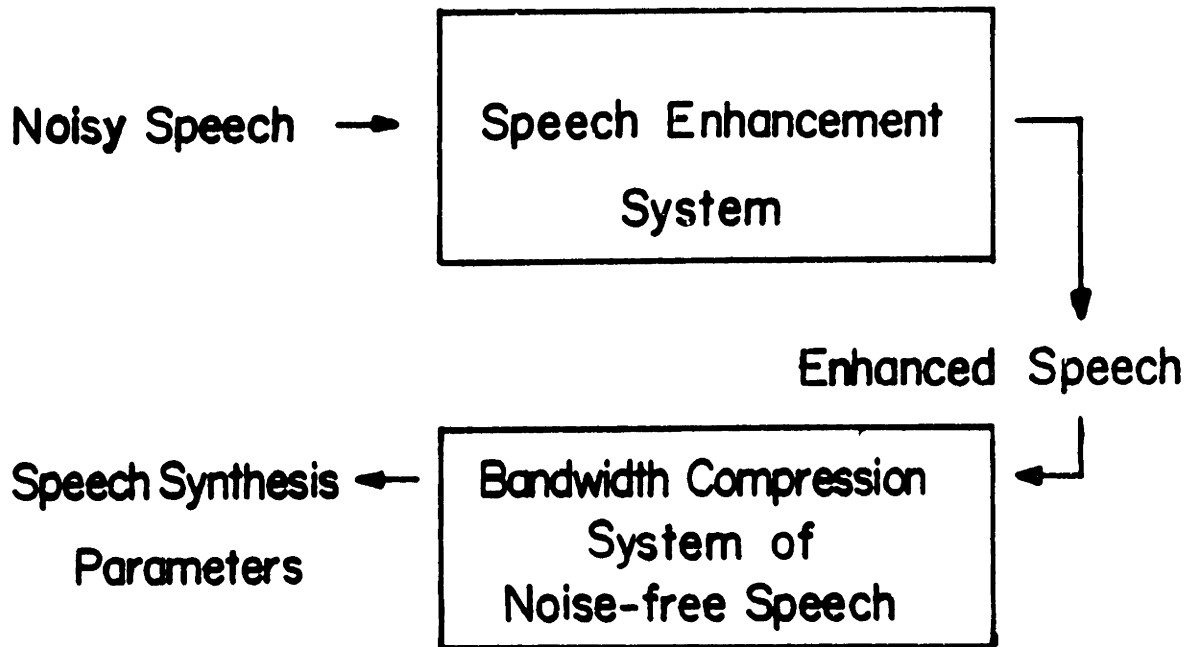


Figure 2.6 The analysis part of a bandwidth compression system of noisy speech when a speech enhancement system is used as a pre-processor

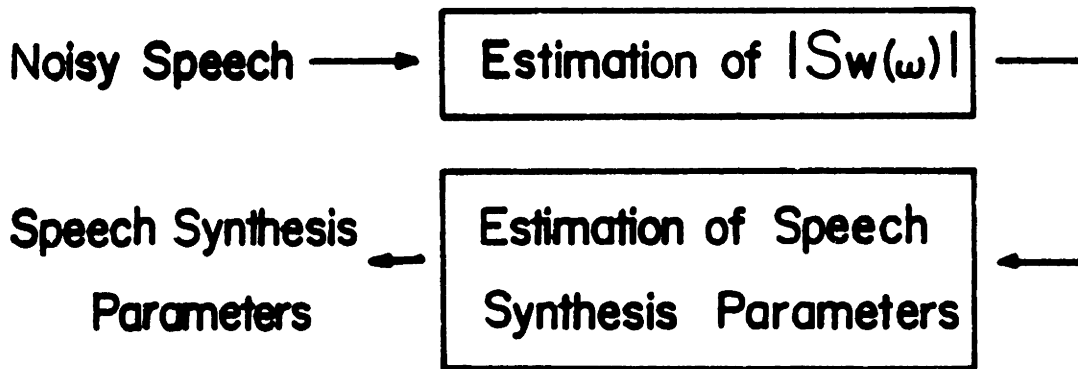


Figure 2.7 A possible simplification of the system in Figure 2.6 for some cases. See the text for the details

vocoder [26] and a spectral root vocoder [27],  $|\hat{S}_w(\omega)|$  can be directly used to obtain the speech synthesis parameters. Then the system in Figure 2.6 can be simplified to Figure 2.7. The main advantage of the system in Figure 2.7 relative to the system in Figure 2.6 is the computational simplicity in that the speech generation process from  $|\hat{S}_w(\omega)|$  in the speech enhancement system can be avoided. A disadvantage is that an existing bandwidth compression system of noise-free speech has to be modified.

From the above discussions, any speech enhancement system discussed in Section II.2 may be used in one form or another for the bandwidth compression of noisy speech. Little data exist in the literature on the performance evaluation of such a bandwidth compression system.

## II.5 Motivation for a New Approach

In this chapter, we have summarized various speech enhancement systems previously proposed. Even though the list of the speech enhancement systems summarized in Section II.2 is not complete, they illustrate the basic philosophy behind currently available speech enhancement systems and raise a number of important questions. One question is in the incorporation of more knowledge of speech. As we have seen in Section II.2, the speech enhancement systems previously proposed are typically based on the periodicity of voiced sounds, uncorrelation of speech with



the background noise or the importance of the short time spectral information for the human speech perception. A natural question is if other knowledge of speech can be incorporated in developing speech enhancement systems. Another question is in how we incorporate what we know about speech. As we discussed in Chapter I, it is desirable to incorporate our knowledge of speech in a manner consistent with the goal of speech enhancement. In the speech enhancement systems previously proposed, a serious attempt has not been made to "optimally" incorporate what we know about speech. A third question is on developing a bandwidth compression system. In our discussions of the bandwidth compression systems of noisy speech in Section II.4, we have considered using the speech enhancement systems as pre-processors. Such a system typically requires generating enhanced speech and then using the enhanced speech as input to a bandwidth compression system of noise-free speech. A natural question that arises is if we can estimate the speech synthesis parameters directly from the noisy speech.

In this dissertation, we develop systems for enhancement and bandwidth compression of noisy speech by attempting to estimate the speech synthesis parameters directly from the noisy speech based on a well known estimation procedure. Such an approach leads to the incorporation of more knowledge of speech in an "optimum" manner. In the next chapter, we

discuss the basic approach taken in this thesis for enhancement and bandwidth compression of noisy speech.

CHAPTER III MODEL OF SPEECH AND ITS PARAMETER  
ESTIMATION

III.1 Introduction

Many successful speech processing systems rely at least to some extent on a model of the speech as the response of a quasi-stationary linear system to a pulse-like excitation for voiced sounds or a noise-like excitation for unvoiced sounds. To develop systems for enhancement and bandwidth compression of noisy speech, it is reasonable to capitalize on the underlying speech model. Thus in this chapter, we formulate the problem of speech enhancement and bandwidth compression of noisy speech as a parameter estimation problem of the speech model parameters. In Section III.2, we present the model of speech which has been studied in great detail [7,13] and has been used extensively [7,13] in many practical applications. In Section III.3, we represent the speech model discussed in Section III.2 in several different forms which we'll find useful in the later chapters. In Section III.4, we discuss the model of noisy speech and its several different representations. In Section III.5, we review briefly the theory of the general parameter estimation problem and three standard estimation rules that have been studied extensively in the literature. In Section III.6, we discuss the estimation of the speech model parameters and its relation to the problem of enhancement and bandwidth compression of

noisy speech.

### III.2 Model of Speech

A digital model of sampled speech that has been used in a number of practical applications and has a basis [7,13] in the human speech production system is shown in Figure 3.1. In the model, the excitation source is either a quasi-periodic train of pulses for voiced sounds or random noise for unvoiced sounds. The digital filter represents the effects of the vocal tract, lip radiation, and in addition the glottal source in the case of voiced sounds. Since the vocal tract changes in shape as a function of time, the digital filter in Figure 3.1 is in general time varying. However, over a short interval of time, we may approximate the digital filter as a linear time invariant system that can be represented as

$$\begin{aligned} H(z) &= G(z) \cdot V(z) \cdot R(z) && \text{for voiced sounds} \\ &V(z) \cdot R(z) && \text{for unvoiced sounds} \end{aligned}$$

where  $G(z)$ ,  $V(z)$  and  $R(z)$  represent the effects of the glottal source, the vocal tract and the lip radiation, respectively.

In general  $H(z)$  consists of both poles and zeroes. However, for non-nasal voiced sounds,  $H(z)$  can be shown [7] to be reasonably well modelled by an all pole system.

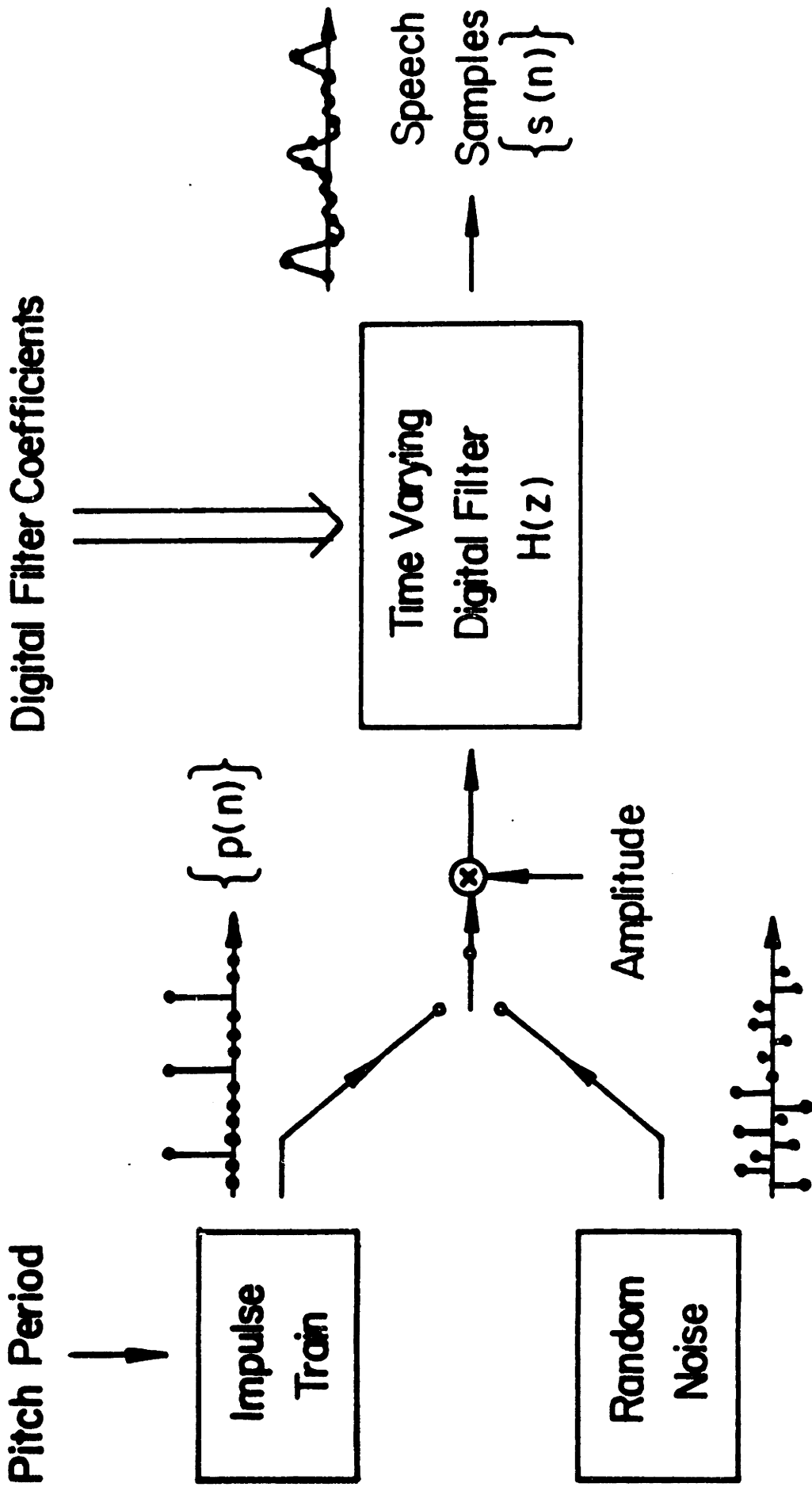


Figure 3.1 A digital model of sampled speech

Furthermore, even for those cases such as nasal sounds or unvoiced sounds in which  $H(z)$  can not adequately be modelled [7] as an all pole system, experience [7,13] has shown that speech analysis based on an all pole system  $H(z)$  leads to many useful results and speech synthesized based on the all pole model is highly intelligible and of high quality. Since the analysis in general is much simpler for an all pole system than a more general system that includes zeroes as well as poles,  $H(z)$  will be modelled as an all pole system. Thus in this thesis, speech is modelled on the short time basis as the response of a stationary all pole system to a pulse-like excitation for voiced sounds or a noise-like excitation for unvoiced sounds.

### III.3 Representations of the Model of Speech

The model of speech discussed in Section III.2 can be represented in many different forms. In this section, we discuss four different representations of the speech model.

In the speech model discussed in Section III.2, the transfer function  $H(z)$  is modelled to be all-pole of the form

$$H(z) = \frac{1}{1 - \sum_{k=1}^p a_k \cdot z^{-k}} \quad (3-1)$$

Thus, on a short time basis the speech waveform  $s(n)$  is assumed to satisfy a difference equation of the form

$$s(n) = \sum_{k=1}^p a_k \cdot s(n-k) + u(n) \quad (3-2)$$

where  $u(n)$  is a pulse train or random noise. Notationally, it is convenient to represent equation (3-2) in a matrix form as

$$s(n) = \underline{a}^T \cdot \underline{s}(n-1, n-p) + u(n) \quad (3-3)$$

and  $\underline{a}$  is the parameter vector<sup>1</sup>

$$\underline{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} \quad (3-4)$$

and  $\underline{s}(n_1, n_2)$  denotes the vector of speech samples

$$\underline{s}(n_1, n_2) = \begin{pmatrix} s(n_1) \\ \vdots \\ s(n_2) \end{pmatrix} \quad (3-5)$$

---

<sup>1</sup>A summary of various notations used throughout the thesis is in Appendix 1.

The vector of observations is assumed to consist of N values  $s(N-1), s(N-2), \dots, s(0)$ , i.e.,  $\underline{s}(N-1,0)$ , which will be denoted by  $\underline{s}_0$ . Equation (3-3) for  $0 \leq n \leq N-1$  is one representation of the speech model.

Equation (3-3) can be represented in various different forms. One form comes from rewriting equation (3-3) as

$$\underline{s}(N-1,0) = A \cdot \underline{s}(N-1,0) + A_I \cdot \underline{s}_I + \underline{u}(N-1,0) \quad (3-6a)$$

where A is an NxN matrix given by

$$A = \begin{bmatrix} 0 & a_1 & a_2 & \dots & a_p & 0 & 0 & \dots & 0 \\ 0 & 0 & a_1 & a_2 & & & & & 0 \\ & & 0 & & & & & & a_p \\ & & & & & & & & a_2 \\ & & & & & & & & a_1 \\ & & & & & & & & 0 \\ & & \underline{0} & & & & & & 0 \end{bmatrix} \quad (3-6b)$$

and  $A_I$  is an Nx $p$  matrix given by



$$A_I = \begin{bmatrix} 0 & & & & & \\ \vdots & & & & & \\ \vdots & & & & & \\ \vdots & & & & & \\ 0 & & \underline{0} & & & \\ a_p & & & & & \\ a_{p-1} & & & & & \\ \vdots & & & & & \\ \vdots & & & & & \\ \vdots & & & & & \\ a_1, \dots, a_{p-1}, a_p \end{bmatrix} \tag{3-6c}$$

and  $\underline{s}_I$  is a  $p \times 1$  matrix given by

$$\underline{s}_I = \underline{s}(-1, -p) = \begin{pmatrix} s(-1) \\ s(-2) \\ \vdots \\ \vdots \\ \vdots \\ s(-p) \end{pmatrix} \tag{3-6d}$$

Therefore,

$$\underline{s}(N-1, 0) = (I-A)^{-1} A_I \cdot \underline{s}_I + (I-A)^{-1} \underline{u}(N-1, 0) \tag{3-6e}$$

Equation (3-6) is another representation of the speech model.

Two other forms can be derived by representing equation (3-3) in a state space form as shown in the following

equation:

$$\begin{aligned}\underline{x}(n) &= F(n) \cdot \underline{x}(n-1) + G(n) \cdot \underline{u}(n) \\ \underline{z}(n) &= H(n) \cdot \underline{x}(n) + \underline{v}(n) \text{ for } 0 \leq n \leq N-1\end{aligned}\quad (3-7)$$

where  $\underline{x}(n)$  is a state vector,  
 $\underline{z}(n)$  is an observation vector,  
 $\underline{u}(n)$  is an excitation vector,  
 $\underline{v}(n)$  is an observation noise vector,  
and  $\underline{x}(-1)$  is an initial condition vector.

Equation (3-3) can be represented in the form of equation (3-7) by using  $\underline{a}$  as a state vector and thus

$$\begin{aligned}\underline{a}(n) &= \underline{a}(n-1) \\ \underline{s}(n) &= \underline{s}^T(n-1, n-p) \cdot \underline{a}(n) + u(n) \text{ for } 0 \leq n \leq N-1\end{aligned}\quad (3-8)$$

Alternatively,  $\underline{s}(n, n-p+1)$  can be used as a state vector  $\underline{x}(n)$  and thus

$$\begin{aligned}\underline{x}(n) &= F \cdot \underline{x}(n-1) + G \cdot u(n) \\ \underline{s}(n) &= H \cdot \underline{x}(n) \quad \text{for } 0 \leq n \leq N-1\end{aligned}\quad (3-9a)$$

$$\text{where } \underline{x}(n) = \begin{pmatrix} x_1(n) \\ x_2(n) \\ \vdots \\ \vdots \\ x_p(n) \end{pmatrix} = \begin{pmatrix} s(n) \\ s(n-1) \\ \vdots \\ \vdots \\ s(n-p+1) \end{pmatrix} \quad (3-9b)$$

$$F = \begin{bmatrix} a_1, a_2, \dots, a_p \\ 1, 0, 0, 0, \dots, 0 \\ 0, 1, 0, 0, \dots, 0 \\ 0, 0, 1, 0, 0, \dots, 0 \\ 0, 0, 0, 1, 0, 0, \dots, 0 \\ \quad \quad \quad \underline{0} \quad 0, 1, 0, \quad \cdot \\ \quad \quad \quad \quad \quad \quad \quad \cdot \\ \quad \quad \quad \quad \quad \quad \quad 0 \\ \quad \quad \quad \quad \quad \quad \quad 1 \end{bmatrix} \quad (3-9c)$$

$$G = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{pmatrix} \quad (3-9d)$$

and  $H = [1, 0, 0, \dots, 0] \quad (3-9e)$

In the above, we have seen that the speech model can be represented in at least four different forms, namely equations (3-3), (3-6), (3-8) and (3-9). These different representations will be found to be useful at various points

in our later discussions.

### III.4 Model of Noisy Speech and its Representations

When the background noise is added to speech, the noisy speech can be represented as

$$y(n) = s(n) + d(n) \quad (3-10a)$$

where  $y(n)$  represents noisy speech and  $d(n)$  represents the background noise or disturbance. The observation vector  $\underline{y}(N-1,0)$  which will alternatively be denoted as  $\underline{y}_0$ , then, consists of the sum of speech and background noise, i.e.,

$$\underline{y}(N-1,0) = \underline{s}(N-1,0) + \underline{d}(N-1,0) \quad (3-10b)$$

Combining equations (3-3) and (3-10),

$$y(n) = \underline{a}^T \cdot \underline{y}(n-1, n-p) - \underline{a}^T \cdot \underline{d}(n-1, n-p) + u(n) \\ \text{for } 0 \leq n \leq N-1 \quad (3-11)$$

Like equation (3-3), equation (3-11) can alternatively be represented in various different forms. Two convenient representations which parallel equations (3-6e) and (3-9a) are

$$\underline{y}(N-1,0) = (\underline{I}-\underline{A})^{-1} \cdot \underline{A}_I \cdot \underline{s}_I + (\underline{I}-\underline{A})^{-1} \cdot \underline{u}(N-1,0) + \underline{d}(N-1,0) \quad (3-12)$$

where  $A$ ,  $A_I$  and  $\underline{s}_I$  are given by equation (3-6) and

$$\begin{aligned}\underline{x}(n) &= F \cdot \underline{x}(n-1) + G \cdot u(n) \\ y(n) &= H \cdot \underline{x}(n) + d(n) \quad \text{for } 0 \leq n \leq N-1\end{aligned}\tag{3-13}$$

where  $\underline{x}(n)$ ,  $F$ ,  $G$  and  $H$  are given by equation (3-9).

Equation (3-11), (3-12) or (3-13) represents the model of noisy speech and will be found to be useful in the later discussions.

### III.5 Review of Parameter Estimation Theory

In this section, we review very briefly the general parameter estimation theory. Let  $A$  and  $R$  denote the parameter space and the observation space, and suppose that there is a probabilistic mapping between the parameter space and the observation space. Assume that a point  $\alpha$  in the parameter space was mapped to a point  $r$  in the observation space. The parameter estimation problem is to estimate the value of  $\alpha$  after observing  $r$  by some estimation rule.

Three different estimation rules known as Maximum Likelihood (ML), Maximum A Posteriori (MAP) and Minimum Mean Square Error (MMSE) estimation have many desirable properties and thus have been studied [22,28] extensively in the literature. For non-random parameters, the ML estimation rule is often used. In the ML estimation, the

parameter value is chosen such that the chosen value most likely resulted in the observation  $r$ . Thus, the value of  $\alpha$  is chosen such that  $p_{R|A}(r|\alpha)$ , the probability density function of  $R$  conditioned on  $A$ , is maximized at the observed  $r$  and the chosen value of  $\alpha$ . The MAP and MMSE estimation rules are commonly used for the parameters that can be considered as random variables whose a priori density function is known. In the MAP estimation rule, the parameter value is chosen such that the a posteriori density  $p_{A|R}(\alpha|r)$  is maximized at the observed  $r$  and the chosen value of  $\alpha$ . Even though the MAP estimation rule is based on a random parameter assumption and the ML estimation rule is based on a non-random parameter assumption, the two estimation rules lead to identical estimates of the parameter value when the a priori density of the parameter in the MAP estimation rule is assumed to be flat over the parameter space. For this reason, the ML estimation rule is often viewed as a special case of the MAP estimation rule. In the MMSE estimation rule  $\hat{\alpha}(R)$ , the estimate of  $\alpha$ , is obtained by minimizing the mean square error  $E[(\hat{\alpha}(R)-\alpha)^2]$ . The MMSE estimate of  $\alpha$  is given by  $E[\alpha|r]$ , the a posteriori mean of  $\alpha$  given  $r$ . Therefore, when the maximum of the a posteriori density function  $p_{A|R}(\alpha|r)$  coincides with its mean, the MAP estimation and MMSE estimation rules lead to identical estimates.

The three estimation procedures briefly discussed above have been applied [22,28] to a number of practical

parameter estimation problems. Detailed discussions on their properties, relations and application areas can be found in [22,28].

### III.6 Estimation of Speech Model Parameters

The model of speech discussed in Section III.2 is completely specified if we determine the parameters related to the excitation  $u(n)$  and the system parameters  $\underline{a}$  in  $H(z)$  of equation (3-1). The basic problem that has been considered in this dissertation is the estimation of the all pole coefficients  $a_k$ .

Ideally, the all pole coefficients should be estimated based on a rule consistent with the subjective aspects of speech. Since a function of  $\underline{a}$  that relates the degree of speech degradation in the subjective domain is not well understood, developing such an estimation rule is difficult. However, we may attempt to use other well known estimation rules discussed in Section III.5 which are optimum in a different sense but which have been successfully applied to a number of other practical problems. In this dissertation, we take the approach to use the MAP estimation procedure. The parameter to be estimated is  $\underline{a}$  and the observation is the noisy speech.

The MAP estimation procedure is based on the philosophy to maximize  $p(\underline{a}|\underline{y}_0)$  where  $\underline{a}$  and  $\underline{y}_0$  represent the all pole coefficient vector and the noisy speech vector.

The approach to use the MAP estimation procedure to estimate the all pole coefficients has a number of advantages.

First, the procedure and properties of the MAP estimation are well established [22] and can be applied to speech processing. Second, the Maximum Likelihood (ML)

estimation procedure can be viewed as a special case of the MAP estimation procedure since the two estimates are the same when the a priori density of  $\underline{a}$  is assumed to be flat.

One property of the ML estimation which is useful for speech processing is that if  $f(\underline{a})$  has a one to one correspondence with  $\underline{a}$ , then  $f_{\text{ML}}^{\hat{}}(\underline{a}) = f(\hat{\underline{a}}_{\text{ML}})$  where  $\hat{\underline{a}}_{\text{ML}}$

represents the ML estimate of  $\underline{a}$ . Therefore, if the perceptually important parameters have a one to one correspondence with  $\underline{a}$ , then the ML estimates for such perceptually important parameters are automatically obtained by obtaining

$\hat{\underline{a}}_{\text{ML}}$ . Further, as will be discussed in greater detail in Chapter IV, for noise-free speech  $\hat{\underline{a}}_{\text{ML}}$  under appropriate assumptions are equivalent to the  $\hat{\underline{a}}$  obtained by the covariance [7,29] or correlation [7,29] method both of which have been successfully applied to the Linear Predictive Coding of speech.

Third, the MAP estimation procedure provides a theoretical framework in which some a priori information about  $\underline{a}$  can be incorporated. Due to the temporal and spectral characteristics of speech, some a priori information of the all pole coefficients  $\underline{a}$  when properly incorporated may in fact aid in estimating  $\underline{a}$ .



In estimating  $\underline{a}$  by the MAP estimation procedure, the excitation  $u(n)$  is assumed to be zero-mean white Gaussian noise. In the context of the speech model discussed in Section III.2, this assumption is valid only for unvoiced speech since the excitation is assumed to be random noise. There are several reasons behind this particular choice of the excitation. First, the analysis of the MAP estimation procedure is relatively simple in the case of the random noise excitation if the excitation is assumed to be generated by a white Gaussian process. The case when the excitation is a pulse train is considerably more difficult. Second, as will be discussed in Chapter IV, in the absence of background noise with the excitation treated random one set of the MAP estimation procedures corresponds exactly to the linear prediction analysis which is well known to be successful for both voiced and unvoiced speech. Further, as will be discussed in Chapters VII and VIII, the theoretical results developed in the thesis for the system parameter estimation in the presence of background noise when the excitation is random noise can be applied with similar performance to the case of the pulse train excitation.

If the all pole coefficients can be "better" estimated through the MAP estimation procedure by accounting for the presence of noise, then we in fact have a better bandwidth compression system of noisy speech in the context of an LPC

vocoder. Even though a complete vocoding system requires the estimation of the source parameters as well as the system parameters, the problem of estimating the source parameters accounting for the presence of background noise will not be treated in this thesis. For the enhancement of noisy speech, there are two ways that the estimation of a can lead to speech enhancement. If we in fact have a successful bandwidth compression system, then the bandwidth compression system itself can be used as a speech enhancer. Alternatively, in the systems that we develop for the estimation of the all pole coefficients, the speech  $s(n)$  is estimated in the process of estimating the all pole coefficients. Thus if speech enhancement is desired, then the estimated  $\hat{s}(n)$  can be used as the enhanced speech. The fact that  $s(n)$  is also estimated is important not only in the context of speech enhancement, but in the context of bandwidth compression of noisy speech. If we estimate only the all pole coefficients, then we are limited to a class of vocoding systems known as LPC vocoders. Since speech is estimated as well as the all pole coefficients, the systems developed can also be used as pre-processors for any vocoding system. Therefore, the systems developed in this thesis are potentially applicable for both bandwidth compression through a variety of vocoding systems and speech enhancement of noisy speech.

CHAPTER IV STATISTICAL PARAMETER ESTIMATION FROM  
NOISE-FREE SPEECH

IV.1 Introduction

In this chapter, we review and relate various ways of estimating the speech model parameters from the noise-free speech. In Section IV.2, the problem of parameter estimation from the noise-free speech is formulated. In Sections IV.3 and IV.4 are discussed two different approaches for the same parameter estimation problem formulated in Section IV.2.

IV.2 Problem Formulation

Speech is modelled as the response of a linear quasi-stationary system to a noise-like excitation. From equation (3-3) with  $u(n)$  corresponding to white Gaussian noise,

$$s(n) = \underline{a}^T \cdot \underline{s}(n-1, n-p) + g \cdot w(n) \quad (4-1)$$

where  $w(n)$  is white Gaussian noise with zero mean and unit variance (i.e.,  $E[w(n)] = 0$  and  $E[w(n) \cdot w(m)] = \delta(n-m)$ ).

Equation (4-1) implies that  $s(n)$  depends on a total of  $2p+1$  parameters, specifically the  $p$  values in the coefficient vector  $\underline{a}$ , the initial conditions  $\underline{s}_I = \underline{s}(-1, -p)$ , and the gain factor  $g$ . We assume that these unknown parameters are random with associated a priori Gaussian probabil-

ity densities. The basic problem treated in this thesis is to estimate the system parameters  $a_k$  from the observation vector  $\underline{s}_0$  by the MAP estimation procedure. Thus the system parameters  $\underline{a}$  are chosen to maximize  $p(\underline{a}|\underline{s}_0)$ , the probability density function<sup>2</sup> of  $\underline{a}$  conditioned on  $\underline{s}_0$ . There are several approaches that can be taken in maximizing  $p(\underline{a}|\underline{s}_0)$ . In Sections IV.3 and IV.4, we consider two different approaches.

#### IV.3 Direct Approach: Maximization of $p(\underline{a}|\underline{s}_0)$

$p(\underline{a}|\underline{s}_0)$  can be written as

$$p(\underline{a}|\underline{s}_0) = \int \int p(\underline{a}, g, \underline{s}_I | \underline{s}_0) dg d\underline{s}_I \quad (4-2)$$

over  $g$   
and  $\underline{s}_I$

From Bayes' rule,  $p(\underline{a}, g, \underline{s}_I | \underline{s}_0)$  is given by:

$$p(\underline{a}, g, \underline{s}_I | \underline{s}_0) = \frac{p(\underline{s}_0 | \underline{a}, g, \underline{s}_I) p(\underline{a}, g, \underline{s}_I)}{p(\underline{s}_0)} \quad (4-3)$$

The conditional density function  $p(\underline{s}_0 | \underline{a}, g, \underline{s}_I)$  can be evaluated by noting that

---

<sup>2</sup>For a more accurate representation, a probability density function  $p_x(\cdot)$  and the density function evaluated at  $x=x_0$  should be distinguished. For the notational convenience,  $p(x_0)$  will be used in both cases and the distinction will be left to the context in which it is used.

$$\begin{aligned}
 p(\underline{s}_0 | \underline{a}, g, \underline{s}_I) &= p(\underline{s}(N-1, 0) | \underline{a}, g, \underline{s}(-1, -p)) \\
 &= \prod_{n=0}^{N-1} p(\underline{s}(n) | \underline{a}, g, \underline{s}(n-1, -p)) \\
 &\approx \prod_{n=0}^{N-1} p(\underline{s}(n) | \underline{a}, g, \underline{s}(n-1, n-p)) \quad (4-4)
 \end{aligned}$$

From the model of equation (4-1) and the assumption that  $w(n)$  is white Gaussian noise with unit variance,

$$\begin{aligned}
 p(\underline{s}(n) | \underline{a}, g, \underline{s}(n-1, n-p)) \\
 = \frac{1}{(2\pi g^2)^{1/2}} \exp\left[-\frac{1}{2g^2} \cdot (\underline{s}(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2\right] \quad (4-5)
 \end{aligned}$$

From equations (4-4) and (4-5),

$$p(\underline{s}_0 | \underline{a}, g, \underline{s}_I) = \frac{1}{(2\pi g^2)^{N/2}} \exp\left[-\frac{1}{2g^2} \cdot \sum_{n=0}^{N-1} (\underline{s}(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2\right] \quad (4-6)$$

$p(\underline{a}, g, \underline{s}_I)$  in equation (4-3) represents the a priori knowledge of the three unknown parameters. For a general Gaussian density of  $p(\underline{a}, g, \underline{s}_I)$ , it can be shown<sup>3</sup> that maximizing

<sup>3</sup>Consider a special case in which  $g$  is known,  $\underline{s}_0 = [s(0)]$  and  $p=1$ . For a Gaussian density of  $p(\underline{a}, \underline{s}_I)$ ,  $p(\underline{a} | \underline{s}_0)$  is in the form of

$$k_1 \frac{1}{(a_1^2 + k_4)^{1/2}} e^{-k_2 (a_1 - k_3)^2} \cdot f(a_1)$$

where  $k_1$ ,  $k_2$ ,  $k_3$  and  $k_4$  are constants. Maximizing  $p(a_1 | s(0))$  in the above highly simplified case involves solving a non-linear equation.

$p(\underline{a}|\underline{s}_0)$  given by equations (4-2), (4-3), (4-4) and (4-6) in general requires solving a set of non-linear equations.

The problem can be made linear, however, by making some specific assumptions of  $p(\underline{a}, g, \underline{s}_I)$  and/or including as the parameters for estimation the auxiliary parameters such as  $g$  and  $\underline{s}_I$  which are unwanted in the sense that our primary interest is in estimating  $\underline{a}$ . In the remainder of this section, four such cases<sup>4</sup> are examined. In case 1, all of the parameters  $\underline{a}$ ,  $g$  and  $\underline{s}_I$  are jointly estimated assuming no a priori information of the parameters. The estimate for  $\underline{a}$  that results corresponds exactly to the covariance method of the linear prediction analysis. In case 2,  $\underline{s}_I$  is assumed to be known and  $\underline{a}$  and  $g$  are estimated jointly assuming no a priori information of  $\underline{a}$  and  $g$ . Depending on specifically how  $\underline{s}_I$  is assumed known, this corresponds to estimating  $\underline{a}$  using either the covariance method or correlation method of the linear prediction analysis. In case 3,  $g$  is assumed to be known and  $\underline{a}$  and  $\underline{s}_I$  are jointly estimated assuming no a priori information of  $\underline{s}_I$ . In case 4, only  $\underline{a}$  is estimated assuming  $g$  and  $\underline{s}_I$  are known.

#### IV.3.1 Case 1

In this case,  $p(\underline{a}, g, \underline{s}_I | \underline{s}_0)$  is maximized with respect

---

<sup>4</sup>These are the only four cases in which the solution can be obtained by solving a set of linear equations.

to  $\underline{a}$ ,  $g$  and  $\underline{s}_I$  with the assumption that no a priori information of  $\underline{a}$ ,  $g$  or  $\underline{s}_I$  is available. This corresponds to the case when  $p(\underline{a}, g, \underline{s}_I)$  is constant<sup>5</sup>. From equation (4-3), since  $p(\underline{s}_0)$  is not a function of  $\underline{a}$ ,  $g$ , or  $\underline{s}_I$  and  $p(\underline{a}, g, \underline{s}_I)$  is assumed to be constant, maximizing  $p(\underline{a}, g, \underline{s}_I | \underline{s}_0)$  is equivalent to maximizing  $p(\underline{s}_0 | \underline{a}, g, \underline{s}_I)$ . Thus, the MAP estimation of  $\underline{a}$ ,  $g$  and  $\underline{s}_I$  in the absence of a priori information reduces to the ML estimation of those parameters.

From equation (4-6), maximizing  $p(\underline{s}_0 | \underline{a}, g, \underline{s}_I)$  with respect to  $g$  leads to

$$g^2 = \frac{1}{N} \cdot \sum_{n=0}^{N-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2 \quad (4-7)$$

Maximization of  $p(\underline{s}_0 | \underline{a}, g, \underline{s}_I)$  with respect to  $\underline{a}$  and  $\underline{s}_I$  is equivalent to minimizing  $\epsilon_p$  given by

$$\epsilon_p = \frac{1}{g^2} \cdot \sum_{n=0}^{N-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2 \quad (4-8)$$

Thus we choose the parameters  $\underline{a}$  and  $\underline{s}_I$  to satisfy the set

<sup>5</sup>As the variance becomes larger, the density function becomes wider and flatter approaching a constant. More formally, however, it should be assumed that  $p(\underline{a}, g, \underline{s}_I)$  is Gaussian whose covariance approaches an arbitrarily large value. In all the cases in this thesis where we assume that no a priori information of some parameters can be modelled by a uniform density of the parameters, it can be shown that the same theoretical results are obtained by first solving the case of finite variance and then letting the variance approach  $\infty$ .

of equations

$$\frac{\partial \epsilon_p}{\partial a_i} = 0 \quad \text{for } i = 1, 2, \dots, p \quad (4-9a)$$

$$\frac{\partial \epsilon_p}{\partial s(-j)} = 0 \quad \text{for } j = 1, 2, \dots, p \quad (4-9b)$$

Rewriting equation (4-8) as

$$\begin{aligned} \epsilon_p = & \frac{1}{g^2} \cdot \sum_{n=0}^{p-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2 \\ & + \frac{1}{g^2} \cdot \sum_{n=p}^{N-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2, \end{aligned} \quad (4-10)$$

only the first of these summations involves the initial condition vector  $\underline{s}_I$ . It is straightforward to show algebraically that for any non-zero solution of the parameter vector  $\underline{a}$ ,  $\underline{s}_I$  can be chosen so that the first summation in equation (4-10) is zero. Since these are the values which minimize  $\epsilon_p$  with respect to  $\underline{s}_I$ , they would then correspond to the estimate of these parameters. Since we are only interested in explicitly estimating the coefficient vector  $\underline{a}$ , it is not necessary to solve for  $\underline{s}_I$ . Since the first term in equation (4-10) will always be zero when  $\epsilon_p$  is minimized, the minimization of equation (4-10) corresponds to minimizing with respect to  $\underline{a}$ , the function

$$\frac{1}{g^2} \cdot \sum_{n=p}^{N-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2 \quad (4-11)$$



Setting the partial derivatives of equation (4-11) with respect to each of the coefficients  $a_i$  to zero results in a set of linear equations given by

$$\sum_{n=p}^{N-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p)) \cdot s(n-i) = 0, \quad i=1, \dots, p \quad (4-12)$$

Equation (4-12) corresponds exactly to the equations obtained by the covariance method of the linear prediction analysis [7,24].

#### IV.3.2 Case 2

In this case, we assume that the initial condition vector  $\underline{s}_I$  is known and no a priori knowledge of  $\underline{a}$  and  $g$  is available. Then  $p(\underline{a}, g | \underline{s}_0)$  is maximized with respect to  $\underline{a}$  and  $g$ . From Bayes' rule,

$$p(\underline{a}, g | \underline{s}_0) = \frac{p(\underline{s}_0 | \underline{a}, g) \cdot p(\underline{a}, g)}{p(\underline{s}_0)} \quad (4-13)$$

and since  $\underline{s}_I$  is assumed to be known  $p(\underline{s}_0 | \underline{a}, g)$  represents  $p(\underline{s}_0 | \underline{a}, g, \underline{s}_I)$  evaluated at  $\underline{s}_I$  equal to its assumed known value. Assuming  $p(\underline{a}, g)$  is constant, maximizing  $p(\underline{a}, g | \underline{s}_0)$  is equivalent to maximizing  $p(\underline{s}_0 | \underline{a}, g)$  corresponding again to the ML estimation of  $\underline{a}$  and  $g$ . From equation (4-6) with known  $\underline{s}_I$ , maximization of  $p(\underline{s}_0 | \underline{a}, g)$  with respect to  $g$  leads to equation (4-7) for  $g^2$ . Maximization with respect to  $\underline{a}$  is identical to minimizing  $\epsilon_p$  given by equation (4-8).

However, the minimization is now carried out with respect to a alone. Comparing equations (4-10) and (4-11), we see that the function to be minimized with respect to a is similar in both cases, differing only in the lower limit of the summation. The linear set of equations for a is now given by

$$\sum_{n=0}^{N-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p)) \cdot s(n-i) = 0$$

$$i=1, 2, \dots, p \quad (4-14)$$

If the initial conditions are indeed known, then we in fact have available  $N+p$  observations of  $s(n)$ . From the  $N+p$  observations, we use the first  $p$  observations to form the initial condition vector  $\underline{s}_I$  and the remaining  $N$  observations to form the observation vector  $\underline{s}_O$ . If we consider the relationship between case 1 and case 2 on the basis of the same total number of observations, then in fact they lead to identical functions to be minimized and consequently identical estimates.

In the above case, we have assumed that  $p(\underline{a}, g)$  is constant and  $\underline{s}_I$  is exactly known. Therefore, maximization of  $p(\underline{a}, g | \underline{s}_O)$  was identical to maximizing  $p(\underline{s}_O | \underline{a}, g)$ . Because maximization of  $p(\underline{s}_O | \underline{a}, g)$  with respect to a and  $g$  in this case corresponds to the ML estimation for a and  $g$  given (conditioned on) the initial condition vector  $\underline{s}_I = \underline{s}(-1, -p)$ , it is sometimes referred to as the

Conditional Maximum Likelihood (CML) estimate of  $\underline{a}$ .

As an alternative to using the first  $p$  observations in each analysis frame to form the initial condition vector, we can assume that the response was zero prior to the observation interval. In this case, assuming that we have a total of  $N$  actual observations, we augment these with  $p$  additional zero values. Now, if we further extend the data by  $p$  points and augment  $\underline{s}(N+p-1, N)$  with zeroes, then maximization of  $p(\underline{a}, g | \underline{s}(N+p-1, 0))$  with respect to  $\underline{a}$  and  $g$  leads to

$$\sum_{n=0}^{N+p-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p)) \cdot s(n-i) = 0$$

for  $i=1, 2, \dots, p$  (4-15)

and  $\underline{s}(N+p-1, N)$  and  $\underline{s}(-1, -p)$  are all  $\underline{0}$ . This is exactly the same equations given by the correlation method of the linear prediction analysis. In the context of the linear prediction analysis, the principal advantage of the correlation method over the covariance method has been that in that case, the solution of the set of equations involves the inversion of a Toeplitz matrix for which there are particularly efficient methods [30]. In addition, the resulting all-pole model is guaranteed to be stable. From equations (4-12) and (4-15) the resulting linear equations to be solved in both methods are given by

$$\sum_n (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p)) \cdot s(n-i) = 0, \quad i=1, 2, \dots, p \quad (4-16)$$

and the summation extends from  $p$  to  $N-1$  for the covariance method and from  $0$  to  $N+p-1$  for the correlation method.

#### IV.3.3 Case 3

Now we consider the case when  $g$  is known so that  $p(\underline{a}, \underline{s}_I | \underline{s}_O)$  is maximized with respect to  $\underline{a}$  and  $\underline{s}_I$  and no a priori information of  $\underline{s}_I$  is available so that  $p(\underline{s}_I)$  is constant. Assuming  $p(\underline{a}, \underline{s}_I) = p(\underline{a}) \cdot p(\underline{s}_I)$ , from Bayes' rule

$$p(\underline{a}, \underline{s}_I | \underline{s}_O) = \frac{p(\underline{s}_O | \underline{a}, \underline{s}_I) \cdot p(\underline{a}) \cdot p(\underline{s}_I)}{p(\underline{s}_O)} \quad (4-17)$$

where  $p(\underline{s}_O | \underline{a}, \underline{s}_I)$  represents  $p(\underline{s}_O | \underline{a}, g, \underline{s}_I)$  evaluated at  $g$  equal to its assumed known value. Since  $p(\underline{s}_I)$  is assumed constant, maximizing  $p(\underline{a}, \underline{s}_I | \underline{s}_O)$  is equivalent to maximizing  $p(\underline{s}_O | \underline{a}, \underline{s}_I) \cdot p(\underline{a})$ . Assuming that  $\underline{a}$  has a Gaussian density with mean  $\bar{\underline{a}}$  and covariance function  $P_0$ ,  $p(\underline{a})$  is of the form

$$p(\underline{a}) = \frac{1}{(2\pi)^{P/2} \cdot |P_0|^{1/2}} \cdot \exp\left[-\frac{1}{2}(\underline{a}-\bar{\underline{a}})^T \cdot P_0^{-1} \cdot (\underline{a}-\bar{\underline{a}})\right] \quad (4-18)$$

Combining equations (4-6), (4-17) and (4-18), it can be seen that maximizing equation (4-17) is equivalent to

minimizing  $\epsilon_p$  given by

$$\epsilon_p = \frac{1}{g^2} \cdot \sum_{n=0}^{N-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2 + (\underline{a} - \bar{\underline{a}})^T \cdot P_0^{-1} \cdot (\underline{a} - \bar{\underline{a}}) \quad (4-19)$$

$\epsilon_p$  in equation (4-19) is similar to  $\epsilon_p$  in equation (4-8) or (4-10) but with the additional term  $(\underline{a} - \bar{\underline{a}})^T \cdot P_0^{-1} \cdot (\underline{a} - \bar{\underline{a}})$ . Since this extra term is not a function of  $\underline{s}_I$ , minimization of  $\epsilon_p$  in equation (4-19) with respect to  $\underline{s}_I$  requires that  $\underline{s}_I$  be such that

$$\sum_{n=0}^{p-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2 = 0$$

Therefore minimization of  $\epsilon_p$  in equation (4-19) with respect to  $\underline{a}$  reduces to minimization of  $\epsilon_p$  given by

$$\frac{1}{g^2} \sum_{n=p}^{N-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2 + (\underline{a} - \bar{\underline{a}})^T \cdot P_0^{-1} \cdot (\underline{a} - \bar{\underline{a}}) \quad (4-20)$$

Partial differentiation with respect to  $a_i$  for  $i=1, 2, \dots, p$  results in a set of linear equations.

If no a priori information on  $\underline{a}$  is assumed so that  $P_0 = \sigma_d^2 \cdot I$  with  $\sigma_d^2$  arbitrarily large, the  $\hat{\underline{a}}$  obtained in this case would be identical to  $\hat{\underline{a}}$  in case 1.

#### IV.3.4 Case 4

Now we maximize  $p(\underline{a} | \underline{s}_0)$  with respect to  $\underline{a}$  assuming

that  $g$  and  $\underline{s}_I$  are known. From Bayes' rule,

$$p(\underline{a}|\underline{s}_0) = \frac{p(\underline{s}_0|\underline{a}) p(\underline{a})}{p(\underline{s}_0)} \quad (4-21)$$

and since  $g$  and  $\underline{s}_I$  are assumed known  $p(\underline{s}_0|\underline{a})$  represents  $p(\underline{s}_0|\underline{a}, g, \underline{s}_I)$  evaluated at  $g$  and  $\underline{s}_I$  equal to their respective assumed known values. Therefore maximizing  $p(\underline{a}|\underline{s}_0)$  is equivalent to maximizing  $p(\underline{s}_0|\underline{a}) \cdot p(\underline{a})$ . Assuming  $p(\underline{a})$  is of the form given by equation (4-18), maximizing  $p(\underline{a}|\underline{s}_0)$  in equation (4-21) is the same as minimizing the same  $\epsilon_p$  in equation (4-19), which can be easily seen by comparing equations (4-17) and (4-21). Here, however, we minimize  $\epsilon_p$  with respect to  $\underline{a}$  alone, which again corresponds to solving a set of linear equations. The difference between equations (4-19) and (4-20) is in the limit of the summation, analogous to the difference between equations (4-10) and (4-11). If we assume no a priori information of  $\underline{a}$ , then the second term in equation (4-20) would be eliminated and the estimate for  $\underline{a}$  obtained in this case would be identical to that obtained in case 2.

If we assume that  $\underline{s}_I = \underline{0}$  and further extend the data by  $p$  points with  $\underline{0}$  (i.e.,  $\underline{s}(N+p-1, N) = \underline{0}$ ) as we did in case 2, then the equation to be minimized is given by

$$\epsilon_p = \frac{1}{g^2} \cdot \sum_{n=0}^{N+p-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2 + (\underline{a} - \bar{\underline{a}})^T \cdot P_0^{-1} \cdot (\underline{a} - \bar{\underline{a}}) \quad (4-22)$$

with  $\underline{s}_I$  and  $\underline{s}(N+p-1,0)$  both equal to  $\underline{0}$ . In the limiting case, as  $P_0$  approaches  $\infty \cdot I$ , corresponding to no a priori information of  $\underline{a}$ , the minimization of  $\epsilon_p$  in equation (4-22) reduces to equation (4-15) which corresponds to the correlation method of the linear prediction analysis.

In the above discussion, we saw that maximizing  $p(\underline{a}|\underline{s}_0)$  leads to a set of linear equations only when  $g$  and  $\underline{s}_I$  are known. In practice these parameters may not be known exactly. However we might expect to make some reasonable guess of  $g$  and  $\underline{s}_I$ . Alternatively, we can solve the linear equations in case 1, assume that these estimates of  $g$  and  $\underline{s}_I$  are exact and maximize equation (4-21) with respect to  $\underline{a}$ . A third possibility for obtaining  $\underline{s}_I$  is to use the first  $p$  data points as  $\underline{s}_I$  and use the remaining  $N-p$  points as  $\underline{s}_0$ , which leads to the same estimate of  $\underline{a}$  as in case 3.

In this section, we have seen that maximizing  $p(\underline{a}|\underline{s}_0)$  in general is a non-linear problem. However the problem can be linearized if we make some specific assumptions about the a priori density of the parameters and/or include as parameters for estimation some auxiliary parameters such as  $g$  and  $\underline{s}_I$ . As will be discussed in Chapter V, the notion of including as parameters for estimation some auxiliary parameters and making some specific assumptions of the a priori information on the parameters will again lead to two linear implementations when we deal with the statistical

parameter estimation from noisy speech. In Section IV.4, we investigate an alternative way to solve the same parameter estimation problem discussed in Section IV.3.

#### IV.4 State Space Approach: All Pole Coefficients as State Vectors

In Section IV.3,  $g$  and  $\underline{s}_1$  were assumed to be known and estimating  $\underline{a}$  by maximizing  $p(\underline{a}|\underline{s}_0)$  led to solving a set of linear equations. By representing the model of speech in  $r$  state space form, the same solution can be obtained in a recursive manner by a Kalman filter. In Section IV.4.1, the properties of a Kalman filter relevant to our discussions in this thesis are briefly summarized. In Section IV.4.2, based on the properties of a Kalman filter discussed in Section IV.4.1, it is discussed that a Kalman filter applied to the proper model of speech maximizes  $p(\underline{a}|\underline{s}_0)$ .

##### IV.4.1 Kalman Filter: Review

Suppose a system can be represented by a state equation of the following:

$$\begin{aligned}\underline{x}(n) &= F(n) \cdot \underline{x}(n-1) + G(n) \cdot \underline{u}(n) \\ \underline{z}(n) &= H(n) \cdot \underline{x}(n) + \underline{v}(n) \quad \text{for } 0 \leq n \leq N-1\end{aligned} \quad (4-23)$$

where  $\underline{x}(n)$  is a state vector,



$\underline{z}(n)$  is an observation vector,

$\underline{u}(n)$  is a vector of zero mean white Gaussian noise

with a given covariance function,

$\underline{v}(n)$  is a vector of zero mean white Gaussian noise

with a given covariance function uncorrelated with

$\underline{u}(n)$ ,

and  $\underline{x}(-1)$  is the initial condition vector which is

Gaussian with a given mean and covariance.

If  $F(n)$ ,  $G(n)$  and  $H(n)$  are known, then  $E[\underline{x}(n) | \underline{z}(n_1, 0)]$

which is the optimum under the MMSE criterion can be obtained

by a linear solution known as the "Kalman filter".

Depending on whether  $n$  is greater than, equal to, or

less than  $n_1$ , the solution is known as a predictor,

filter or smoother, respectively. For a Gaussian  $\underline{x}(n)$

which is the case in equation (4-23), the MMSE estimator

is equivalent to the MAP estimator since  $p(\underline{x}(n) | \underline{z}(n_1, 0))$

is symmetric about the conditional expectation  $E[\underline{x}(n) | \underline{z}(n_1, 0)]$ .

The detailed linear solutions of a Kalman filter and its

properties can be found in [22,31,32,33,34].

#### IV.4.2 Maximization of $p(\underline{a} | \underline{s}_0)$ by a Kalman Filter

Equation (3-8) of the speech model with  $u(n) = g \cdot w(n)$

is given by

$$\underline{a}(n) = \underline{a}(n-1)$$

$$\underline{s}(n) = \underline{s}^T(n-1, n-p) \cdot \underline{a}(n) + g \cdot w(n) \quad \text{for } 0 \leq n \leq N-1 \quad (4-24)$$

Equation (4-24) is a special case of equation (4-23). If  $g$  and  $\underline{s}_1$  are assumed known, then  $F(n)$ ,  $G(n)$  and  $H(n)$  are completely specified. Therefore,  $E[\underline{a}|\underline{s}(N-1,0)]$  which corresponds to both the MMSE and MAP estimates of  $\underline{a}$  can be obtained by a Kalman filter. The filtering form [31, 32] of a Kalman filter applied to equation(4-24) is given by an iterative solution;

$$\hat{\underline{a}}(n+1) = \hat{\underline{a}}(n) + \underline{k}(n+1) \cdot (s(n+1) - \underline{s}^T(n, n+1-p) \cdot \hat{\underline{a}}(n)) \quad (4-25)$$

where  $\hat{\underline{a}}(n)$  represents  $E[\underline{a}(n)|\underline{s}(n,0)]$  and  $\underline{k}(n+1)$  is the Kalman filter gain which is a function of the covariance matrix of  $\underline{a}(n)$ . The covariance matrix of  $\underline{a}(n)$  can also be updated and the initial starting values  $\hat{\underline{a}}(-1)$  and the covariance of  $\underline{a}(-1)$  are, of course, the a priori mean and covariance of  $\underline{a}$ . For each  $n$ ,  $\underline{a}(n)$  obtained in this manner is identical to  $\underline{a}$  estimated by minimizing the function

$$\frac{1}{g^2} \cdot \sum_{m=0}^n (s(m) - \underline{a}^T \cdot \underline{s}(m-1, m-p))^2 + (\underline{a} - \bar{\underline{a}})^T \cdot P_0^{-1} (\underline{a} - \bar{\underline{a}}) \quad (4-26)$$

In particular,  $\hat{\underline{a}}(N-1)$  is the estimate of  $\underline{a}$  obtained by minimizing equation (4-19) with respect to  $\underline{a}$ . The filtering form of the Kalman filter solution discussed above is also known as a recursive least squares procedure and the primary

advantage of a recursive solution is that the data can be sequentially processed as they appear.

CHAPTER V STATISTICAL PARAMETER ESTIMATION FROM  
NOISY SPEECH

V.1 Introduction

In chapter IV, a framework was established for the MAP parameter estimation of the noise-free speech. In two of its forms, leading to equations (4-12) and (4-15), there has been extensive experience in the context of the linear prediction speech analysis with considerable success and are currently the basis for many speech processing systems [7,8,12,14,24,25,29]. It is well known, however, that these procedures degrade quickly in the presence of additive background noise [2,3]. Consequently, it is of interest to consider whether the same basic approach and philosophy can be applied when the observations are recognized to be corrupted by the background noise. Thus, in this chapter, we consider the statistical parameter estimation from the noisy speech based on the MAP estimation procedure.

In our discussions in this chapter, we first consider the case of the white Gaussian background noise and then extend the theoretical results obtained to a more general case when the background noise is colored. In Section V.2, the MAP estimation procedure that maximizes the probability density function of the parameters to be estimated conditioned on the noisy speech vector will be shown to be a non-linear problem. In Section V.3, we

develop a linear iterative algorithm which approximates the MAP estimation procedure. In Section V.4, we develop another linear iterative algorithm by revising the method discussed in Section V.3. In Section V.5, we extend the theoretical results discussed in Sections V.2, V.3, and V.4 to a more general case when the background noise is colored. In Section V.6, we relate the two linear iterative algorithms to the MAP estimation procedure.

## V.2 MAP Estimation Procedure: A Non-linear Problem

Speech is again assumed to be generated by the model of equation (4-1) and the coefficient vector  $\underline{a}$  are the basic parameters to be estimated. The observation vector  $\underline{y}(N-1,0)$  which will alternatively be denoted as  $\underline{y}_0$  consists of the sum of the speech and background noise, i.e.,

$$\underline{y}(N-1,0) = \underline{s}(N-1,0) + \underline{d}(N-1,0) \quad (5-1)$$

where  $d(n)$  is zero mean white Gaussian background noise with variance of  $\sigma_d^2$  and is assumed to be uncorrelated with  $s(n)$ .

Following a procedure similar to that of case 4 (Section IV.2.4), we can consider choosing the parameters  $\underline{a}$  to maximize  $p(\underline{a}|\underline{y}_0)$ . In Chapter IV when we assumed that  $\underline{g}$  and  $\underline{s}_1$  were known and  $p(\underline{a})$  was Gaussian, the resulting

equations were linear. For the current situation, this will no longer be the case. Specifically, from equations (4-1) and (5-1),

$$y(n) = \underline{a}^T \cdot \underline{s}(n-1, n-p) + g \cdot w(n) + d(n) \quad (5-2)$$

$$\text{or } y(n) = \underline{a}^T \cdot \underline{y}(n-1, n-p) + g \cdot w(n) + d(n) - \underline{a}^T \cdot \underline{d}(n-1, n-p) \quad (5-3)$$

Expressing  $p(\underline{y}_0 | \underline{a}, g, \underline{s}_I)$  in a manner similar to equation (4-4),

$$p(\underline{y}_0 | \underline{a}, g, \underline{s}_I) = \prod_{n=p}^{N-1} p(y(n) | \underline{a}, g, \underline{s}_I, \underline{y}(n-1, 0)) \cdot \prod_{n=1}^{p-1} p(y(n) | \underline{a}, g, \underline{s}_I, \underline{y}(n-1, 0)) \cdot p(y(0) | \underline{a}, g, \underline{s}_I) \quad (5-4)$$

From equation (5-2), for  $n \geq p$ ,  $p(y(n) | \underline{a}, g, \underline{s}_I, \underline{y}(n-1, 0))$  is Gaussian with mean of  $\underline{a}^T \cdot E[\underline{s}(n-1, n-p) | \underline{a}, g, \underline{s}_I, \underline{y}(n-1, 0)]$  and variance of  $g^2 + \sigma_d^2 + \underline{a}^T \cdot \text{Var}[\underline{s}(n-1, n-p) | \underline{a}, g, \underline{s}_I, \underline{y}(n-1, 0)] \cdot \underline{a}$  where  $E[\underline{s}(n-1, n-p) | \underline{a}, g, \underline{s}_I, \underline{y}(n-1, 0)]$  and  $\text{Var}[\underline{s}(n-1, n-p) | \underline{a}, g, \underline{s}_I, \underline{y}(n-1, 0)]$  denote the mean and covariance of  $\underline{s}(n-1, n-p)$  conditioned on  $\underline{a}, g, \underline{s}_I$  and  $\underline{y}(n-1, 0)$ . Since the variance is a function of  $\underline{a}$ , and will likewise be so for the remaining terms, the resulting equations for maximizing

$p(\underline{a}|\underline{y}_0)$  will by necessity be non-linear.

Even though we have only shown that maximizing  $p(\underline{a}|\underline{y}_0)$  which corresponds to case 4 in Chapter IV is a non-linear problem, it is easy to see that maximizing  $p(\underline{a}, \underline{g}, \underline{s}_I|\underline{y}_0)$ ,  $p(\underline{a}, \underline{g}|\underline{y}_0)$  or  $p(\underline{a}, \underline{s}_I|\underline{y}_0)$  corresponding to cases 1, 2 and 3 in the previous chapter is also a non-linear problem. This is partly because each of the three density functions  $p(\underline{a}, \underline{g}, \underline{s}_I|\underline{y}_0)$ ,  $p(\underline{a}, \underline{g}|\underline{y}_0)$ , or  $p(\underline{a}, \underline{s}_I|\underline{y}_0)$  is a product of several terms, one of which is

$$\prod_{n=p}^{N-1} p(y(n) | \underline{a}, \underline{g}, \underline{s}_I, \underline{y}(n-1, 0)).$$

It was shown above that  $p(y(n) | \underline{a}, \underline{g}, \underline{s}_I, \underline{y}(n-1, 0))$  for  $p \leq n \leq N-1$  has the variance which is a function of  $\underline{a}$ .

### V.3 Maximization of $p(\underline{a}, \underline{s}_0|\underline{y}_0)$ : Linearized MAP (LMAP) Estimation Procedure

To maximize  $p(\underline{a}|\underline{y}_0)$  which was shown to be a non-linear problem in Section V.2, one approach is to determine  $p(\underline{a}|\underline{y}_0)$  for any set of specific  $\underline{a}$  and then use some form of hill searching algorithm [35,36,37]. In general, solving such a non-linear problem is computationally undesirable. Thus, we are led to consider another method which has a linear implementation, but which may not be optimum in the sense that  $p(\underline{a}|\underline{y}_0)$  is not maximized. In Chapter IV, we have seen that maximizing  $p(\underline{a}|\underline{s}_0)$  is in

general a non-linear problem. However, by incorporating some auxiliary parameters as parameters for estimation and/or making some specific assumptions on the a priori knowledge of the parameters, the resulting equations can be made linear. When the resulting equations (4-12, 4-15, 4-20, 4-22) are used to estimate  $\underline{a}$  and speech is synthesized based on the estimated  $\underline{a}$ , experience [7,8,24,25] has shown that intelligible speech with high quality can be generated. Motivated by the apparent success in the case of noise-free speech, we take a similar approach in the case of noisy speech. More specifically, we assume that  $g$  and  $\underline{s}_I$  are known, and include the speech vector  $\underline{s}_O$  as an additional parameter to be estimated. Thus we maximize  $p(\underline{a}, \underline{s}_O | Y_O)$ <sup>6</sup> jointly with respect to  $\underline{a}$  and  $\underline{s}_O$ . In this section, we show that maximizing  $p(\underline{a}, \underline{s}_O | Y_O)$  is still a non-linear problem but can be implemented by a linear iterative procedure.

### V.3.1 An Algorithm to Maximize $p(\underline{a}, \underline{s}_O | Y_O)$

Suppose we begin with an assumed set of initial

---

<sup>6</sup>A linear implementation for  $\underline{a}$  can also be obtained essentially in a parallel manner by maximizing  $p(\underline{a}, \underline{s}_O, g, \underline{s}_I | Y_O)$ ,  $p(\underline{a}, \underline{s}_O, g | Y_O)$  or  $p(\underline{a}, \underline{s}_O, \underline{s}_I | Y_O)$  with the appropriate a priori density assumptions of the unknown parameters. This situation is analogous to the four cases considered in Chapter IV and allow us to estimate the other parameters ( $g, \underline{s}_I$ ) in the same manner as  $\underline{a}$  if such an approach is desired. In the discussions in this chapter, we concentrate primarily on maximizing  $p(\underline{a}, \underline{s}_O | Y_O)$ .



values  $\hat{\underline{a}}_0$  for the coefficient vector  $\underline{a}$  and based on this, estimate  $\underline{s}_0$  by maximizing  $p(\underline{s}_0 | \hat{\underline{a}}_0, \underline{y}_0)$ . Denoting this first estimate of  $\underline{s}_0$  by  $\hat{\underline{s}}_{01}$ , we then form a first estimate  $\hat{\underline{a}}_1$  of  $\underline{a}$ . This procedure can then be continued iteratively to obtain the final estimate  $\hat{\underline{a}}_\infty$  of the coefficients. We now show that this procedure for estimating  $\underline{a}$  (and  $\underline{s}_0$ ) always increases  $p(\underline{a}, \underline{s}_0 | \underline{y}_0)$  at each iteration unless a converging solution is obtained. Specifically, since  $\hat{\underline{a}}_i$  is obtained by maximizing  $p(\underline{a} | \hat{\underline{s}}_{0i})$ ,

$$p(\hat{\underline{a}}_i | \hat{\underline{s}}_{0i}, \underline{y}_0) \cdot p(\hat{\underline{s}}_{0i} | \underline{y}_0) \geq p(\hat{\underline{a}}_{i-1} | \hat{\underline{s}}_{0i}, \underline{y}_0) \cdot p(\hat{\underline{s}}_{0i} | \underline{y}_0) \quad (5-5a)$$

and therefore

$$p(\hat{\underline{a}}_i, \hat{\underline{s}}_{0i} | \underline{y}_0) \geq p(\hat{\underline{a}}_{i-1}, \hat{\underline{s}}_{0i} | \underline{y}_0) \quad (5-5b)$$

The equality sign in equation (5-5b) holds only if  $\hat{\underline{a}}_i = \hat{\underline{a}}_{i-1}$  since  $p(\underline{a} | \underline{s}_0, \underline{y}_0)$  is Gaussian in  $\underline{a}$ . Since  $\hat{\underline{s}}_{0i}$  is obtained by maximizing  $p(\underline{s}_0 | \hat{\underline{a}}_{i-1}, \underline{y}_0)$ ,

$$p(\hat{\underline{s}}_{0i} | \hat{\underline{a}}_{i-1}, \underline{y}_0) \cdot p(\hat{\underline{a}}_{i-1} | \underline{y}_0) \geq p(\hat{\underline{s}}_{0i-1} | \hat{\underline{a}}_{i-1}, \underline{y}_0) \cdot p(\hat{\underline{a}}_{i-1} | \underline{y}_0) \quad (5-6a)$$

and therefore

$$p(\hat{\underline{a}}_{i-1}, \hat{\underline{s}}_{0i} | \underline{y}_0) \geq p(\hat{\underline{a}}_{i-1}, \hat{\underline{s}}_{0i-1} | \underline{y}_0) \quad (5-6b)$$

The equality sign in equation (5-6b) holds only if  $\hat{\underline{s}}_{0i} = \hat{\underline{s}}_{0i-1}$  since  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$  is Gaussian in  $\underline{s}_0$ . From equations (5-5b) and (5-6b),

$$p(\hat{\underline{a}}_i, \hat{\underline{s}}_{0i} | \underline{y}_0) \geq p(\hat{\underline{a}}_{i-1}, \hat{\underline{s}}_{0i-1} | \underline{y}_0) \quad (5-7)$$

in which the equality sign holds if  $\hat{\underline{a}}_i = \hat{\underline{a}}_{i-1}$  and  $\hat{\underline{s}}_{0i} = \hat{\underline{s}}_{0i-1}$ . Equation (5-7) shows that the iterative procedure discussed above always increases  $p(\underline{a}, \underline{s}_0 | \underline{y}_0)$  at each iteration unless a converging solution is reached. If the initial guess for  $\underline{a}$  and the shape of  $p(\underline{a}, \underline{s}_0 | \underline{y}_0)$  is such that this procedure converges to the global maximum, then this procedure will in fact correspond to that joint MAP estimate of the parameters  $\underline{a}$  and  $\underline{s}_0$ . Thus, in essence, this attempt to simplify the problem computationally corresponds to augmenting the desired set of parameters  $\underline{a}$  with the additional parameters  $\underline{s}_0$ .

### V.3.2 Maximization of $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$

From the discussions in Chapter IV, maximizing  $p(\underline{a} | \underline{s}_0, \underline{y}_0)$  which is equivalent to maximizing  $p(\underline{a} | \underline{s}_0)$  requires the solution of a set of  $p$  linear equations for  $\underline{a}$ . To show that the algorithm requires solving only linear equations, we now show that maximizing  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$

is also a linear problem.

From Bayes' rule,  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$  can be denoted as

$$p(\underline{s}_0 | \underline{a}, \underline{y}_0) = p(\underline{y}_0 | \underline{a}, \underline{s}_0) \cdot \frac{p(\underline{s}_0 | \underline{a})}{p(\underline{y}_0 | \underline{a})} \quad (5-8)$$

Denoting  $p(\underline{y}_0 | \underline{a}, \underline{s}_0)$  by

$$p(\underline{y}_0 | \underline{a}, \underline{s}_0) = \prod_{n=1}^{N-1} p(y(n) | \underline{a}, \underline{s}_0, \underline{y}(n-1, 0)) \cdot p(y(0) | \underline{a}, \underline{s}_0) \quad (5-9)$$

and noting that  $p(y(n) | \underline{a}, \underline{s}_0, \underline{y}(n-1, 0))$  is Gaussian with mean of  $s(n)$  and variance of  $\sigma_d^2$  for  $1 \leq n \leq N-1$  and  $p(y(0) | \underline{a}, \underline{s}_0)$  is Gaussian with mean of  $s(0)$  and variance of  $\sigma_d^2$ ,  $p(\underline{y}_0 | \underline{a}, \underline{s}_0)$  can be denoted as

$$p(\underline{y}_0 | \underline{a}, \underline{s}_0) = \frac{1}{(2\pi\sigma_d^2)^{N/2}} \cdot \exp\left(-\frac{1}{2\sigma_d^2} \sum_{n=0}^{N-1} (y(n) - s(n))^2\right) \quad (5-10)$$

Combining equations (4-6) and (5-10) with equation (5-8) with the assumption that  $g$  and  $\underline{s}_I$  are given and noting that  $p(\underline{y}_0 | \underline{a})$  is not a function of  $\underline{s}_0$ ,

$$p(\underline{s}_0 | \underline{a}, \underline{y}_0) = \text{constant} \cdot \frac{1}{(4\pi^2 \cdot g^2 \cdot \sigma_d^2)^{N/2}} \cdot \exp\left(-\frac{1}{2} \epsilon_p\right) \quad (5-11a)$$

and

$$\epsilon_p = \frac{1}{g^2} \cdot \sum_{n=0}^{N-1} (s(n) - \underline{a}^T \cdot \underline{s}(n-1, n-p))^2 + \frac{1}{\sigma_d^2} \cdot \sum_{n=0}^{N-1} (y(n) - s(n))^2 \quad (5-11b)$$

Maximizing  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$  is equivalent to minimizing  $\epsilon_p$  in equation (5-11b) and thus we choose  $\underline{s}_0$  that satisfy the set of linear equations,

$$\frac{\partial \epsilon_p}{\partial s(i)} = 0 \quad \text{for } i = 0, 1, 2, \dots, N-1 \quad (5-12)$$

A closed form expression for the solution of equation (5-12) can be obtained by representing the speech model with equation (3-6). From equation (5-1),

$$\begin{aligned} p(\underline{y}(N-1, 0) | \underline{a}, \underline{s}(N-1, 0)) &= p(\underline{y}(N-1, 0) | \underline{s}(N-1, 0)) \\ &= N(\underline{s}(N-1, 0), \sigma_d^2 \cdot I) \end{aligned} \quad (5-13a)$$

From equation (3-6e) with  $u(n) = g \cdot w(n)$ ,

$$p(\underline{s}(N-1, 0) | \underline{a}) = N((I-A)^{-1} \cdot A_I \cdot \underline{s}_I, g^2 \cdot (I-A)^{-1} \cdot ((I-A)^{-1})^T) \quad (5-13b)$$

We now combine equations (5-13) with equation (5-8)

assuming that  $g$  and  $\underline{s}_I$  are given and noting that  $p(\underline{y}_0 | \underline{a})$  is not a function of  $\underline{s}_0$ . The result is that

$$\begin{aligned} p(\underline{s}(N-1, 0) | \underline{a}, \underline{y}_0) &= N\left(\left(R_s^{-1} + \frac{1}{\sigma_d^2} \cdot I\right)^{-1} \cdot \left(\frac{1}{\sigma_d^2} \underline{y}_0 + R_s^{-1} \cdot \underline{m}_s\right), \right. \\ &\quad \left. \left(R_s^{-1} + \frac{1}{\sigma_d^2} \cdot I\right)^{-1}\right) \end{aligned} \quad (5-14a)$$

where

$$\underline{m}_s = (I-A)^{-1} \cdot A_I \cdot \underline{s}_I \quad \text{and} \quad R_s = g^2 \cdot (I-A)^{-1} \cdot ((I-A)^{-1})^T \quad (5-14b)$$

Therefore, maximizing  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$  is equivalent to estimating  $\underline{s}_0$  by

$$\underline{s}_0 = E[\underline{s}_0 | \underline{a}, \underline{y}_0] = (R_s^{-1} + \frac{1}{\sigma_d^2} I)^{-1} \cdot (\frac{1}{\sigma_d^2} \underline{y}_0 + R_s^{-1} \cdot \underline{m}_s) \quad (5-15)$$

An alternative way to maximize  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$  is from the smoothing form [33,34] of a Kalman filter. As we discussed in Section III.3, equation (3-13) of the noisy speech model can be represented in the form of equation (3-7) with  $\underline{x}(n)$ ,  $F(n)$ ,  $G(n)$ ,  $\underline{u}(n)$ ,  $\underline{z}(n)$ ,  $H(n)$  and  $\underline{v}(n)$  given by equation (3-9). As we discussed in Section IV.4.1, it is well known that for equation (3-7) with zero mean white Gaussian  $\underline{u}(n)$  and  $\underline{v}(n)$  uncorrelated with each other (this corresponds to equation (4-23)), the smoothing form of a Kalman filter leads to  $E[\underline{x}(n) | \underline{z}(N-1,0), F(n)]$  for  $n=0,1,2,\dots,N-1$  which corresponds to  $E[\underline{s}_0 | \underline{a}, \underline{y}_0]$ . Since  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$  is jointly Gaussian,  $E[\underline{s}_0 | \underline{a}, \underline{y}_0]$  is also the MAP estimate of  $\underline{s}_0$  that maximizes  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$ . The Kalman filtering approach has an advantage in that only  $p \times p$  matrix (the state  $\underline{x}(n)$  has  $p$  elements) operations are required while equation (5-15) requires  $N \times N$  matrix operations.

### V.3.3 Linearized MAP Estimation Procedure

Summarizing the steps involved in the linear implementation method, we have

- Step 1: Begin with  $\hat{\underline{a}}_i$ , the  $i$ th estimate of  $\underline{a}$ .
- Step 2: Obtain  $\hat{\underline{s}}_{0i+1}$ , the  $i+1$ st estimate of  $\underline{s}_0$ , by solving equation (5-12), from equation (5-15), or from the smoothing form of a Kalman filter.
- Step 3: Obtain  $\hat{\underline{a}}_{i+1}$ , the  $i+1$ st estimate of  $\underline{a}$ , by minimizing equation (4-19) with  $\hat{\underline{s}}_0$  obtained in Step 2.

The above steps complete one iteration and the procedure can be continued for as many desirable number of iterations. The initial estimate  $\hat{\underline{a}}_0$  may be obtained by simply applying the correlation method of the linear prediction analysis to  $\underline{y}_0$ . We'll refer to this algorithm as the "Linearized MAP" (LMAP) estimation procedure.

In our discussions so far, we have assumed that  $g$  and  $\underline{s}_I$  are known. Even though these parameters are not known exactly, we might expect to make some reasonable guess of  $g$  and  $\underline{s}_I$ . For example, in the LMAP estimation procedure, for each iteration when Step 2 is completed, we have an estimate of  $\underline{s}_0$ . Before going to Step 3, we could maximize  $p(\underline{a}, g, \underline{s}_I | \underline{s}_0)$  that leads to equations (4-7) and (4-9) from which  $g$  and  $\underline{s}_I$  can be estimated. Then we

can assume that these estimates are exact and use them in step 3 in the current iteration and step 2 of the next iteration. Another possibility for estimation  $g$  and  $\underline{s}_I$  is to jointly estimate  $\underline{a}, g$  and  $\underline{s}_I$  in step 3 from  $\underline{s}_0$  estimated in step 2 with the assumption of no a priori information of  $g$  and a general Gaussian density assumption of  $\underline{a}$  and  $\underline{s}_I$ . An example of  $p(\underline{s}_I)$  could be  $N(\underline{y}(-1, -p), \sigma_d^2 I)$ . In Section IV.3.1, it was shown that  $p(\underline{a}, g, \underline{s}_I | \underline{s}_0)$  could be maximized by solving a set of linear equations if no a priori information of  $\underline{a}, g$  and  $\underline{s}_I$  is available. When a priori information of  $\underline{a}$  and  $\underline{s}_I$  is available, jointly maximizing  $p(\underline{a}, g, \underline{s}_I | \underline{s}_0)$  is a non-linear problem. However we can again solve iteratively by maximizing  $p(\underline{a}, \underline{s}_I | g, \underline{s}_0)$  with respect to  $\underline{a}$  and  $\underline{s}_I$  and then maximizing  $p(g | \underline{a}, g, \underline{s}_I, \underline{s}_0)$  with respect to  $g$  for each iteration. Maximizing  $p(\underline{a}, \underline{s}_I | g, \underline{s}_0)$  again involves an iterative procedure in which  $p(\underline{a} | \underline{s}_I, g, \underline{s}_0)$  is maximized with respect to  $\underline{a}$  and then  $p(\underline{s}_I | \underline{a}, g, \underline{s}_0)$  is maximized with respect to  $\underline{s}_I$  for each iteration. It can be shown<sup>7</sup> that the above procedure never decreases  $p(\underline{a}, g, \underline{s}_I | \underline{s}_0)$  at each iteration. Maximizing  $p(\underline{a} | \underline{s}_I, g, \underline{s}_0)$ ,  $p(\underline{s}_I | \underline{a}, g, \underline{s}_0)$ , or  $p(g | \underline{a}, \underline{s}_I, \underline{s}_0)$  involves<sup>8</sup> solving a set of linear equations.

---

<sup>7</sup>This statement can be proved in an analogous manner as in equations (5-5), (5-6) and (5-7).

<sup>8</sup>The derivations are similar to the derivations in the four cases (Sections IV.3.1, IV.3.2, IV.3.3, and IV.3.4) and they begin from equation (4-6).

A third possibility for  $\underline{s}_I$  or  $g$  is simply to assume that  $\underline{s}_I = \underline{0}$  as we did in case 2 (Section IV.3.2) which led to the correlation method of the linear prediction analysis, and estimate  $g$  from the energy considerations which will be discussed further in Chapter VI.

The discussions so far were based on the assumption that the primary interest is in the estimation of  $\underline{a}$ . It is important to note, however, that the LMAP estimation procedure estimates  $\underline{s}_0$  in the process of estimating  $\underline{a}$  by  $\hat{\underline{s}}_0 = E[\underline{s}_0 | \underline{a}, \underline{y}_0]$ .  $\hat{\underline{s}}_0$  estimated in this manner can be directly used as enhanced speech. Therefore the LMAP algorithm discussed in this section can be used not only for the bandwidth compression but also for the enhancement of noisy speech.

#### V.4 Revised Linearized MAP (RLMAP) Estimation Procedure

##### V.4.1 Motivation for the Revision

A careful observation of the LMAP estimation procedure discussed in Section V.3 leads to another estimation procedure that again requires solving a set of linear equations in an iterative manner. In step 2 of the LMAP estimation procedure, we estimate  $\underline{s}_0$  by  $E[\underline{s}_0 | \underline{a}, \underline{y}_0]$ . In step 3, we note that the MAP estimate of  $\underline{a}$  corresponding to maximizing  $p(\underline{a} | \underline{s}_0)$  uses the values  $\underline{s}_0$  to form products of the form  $s(i) \cdot s(j)$ . Thus estimating  $\underline{s}_0$  in step 2 by  $E[\underline{s}_0 | \underline{a}, \underline{y}_0]$  corresponds to estimating  $s(i) \cdot s(j)$  as



$$s(i) \hat{s}(j) = E[s(i) | \underline{a}, \underline{y}_0] \cdot E[s(j) | \underline{a}, \underline{y}_0] \quad (5-16)$$

As an alternative, we can consider generating directly the MMSE estimate of the product  $s(i) \cdot s(j)$ . Thus the estimate of  $s(i) \cdot s(j)$  is given by

$$s(i) \hat{s}(j) = E[s(i) \cdot s(j) | \underline{a}, \underline{y}_0] \quad (5-17)$$

In this method, then, we follow the same procedure as we did in the LMAP method with the difference in that  $s(i) \cdot s(j)$  is estimated by equation (5-17) rather than equation (5-16).

#### V.4.2 Estimation of $s(i) \cdot s(j)$ by $E[s(i) \cdot s(j) | \underline{a}, \underline{y}_0]$

In this section, we show that  $E[s(i) \cdot s(j) | \underline{a}, \underline{y}_0]$  can be obtained by solving sets of linear equations.

From the expression of  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$  in equation (5-11),  $\epsilon_p$  in equation (5-11b) can be written as

$$\epsilon_p = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \beta_{ij} (s(i) - m_i) \cdot (s(j) - m_j) + \text{constant} \quad (5-18)$$

Since  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$  is jointly Gaussian in  $\underline{s}_0$ ,  $[\beta_{ij}]^{-1}$  is a covariance matrix for  $\underline{s}_0$  conditioned on  $\underline{a}$  and  $\underline{y}_0$  where  $[\beta_{ij}]^{-1}$  represents the inverse of a matrix whose  $ij$ th element is  $\beta_{ij}$ . Denoting this covariance matrix by  $[\gamma_{ij}]$

(i.e.,  $[\gamma_{ij}] = [\beta_{ij}]^{-1}$ ),

$$\begin{aligned} \gamma_{ij} &= E[(s(i) - E[s(i) | \underline{a}, \underline{y}_0]) \cdot (s(j) - E[s(j) | \underline{a}, \underline{y}_0]) | \underline{a}, \underline{y}_0] \\ &= E[s(i) \cdot s(j) | \underline{a}, \underline{y}_0] - E[s(i) | \underline{a}, \underline{y}_0] \cdot E[s(j) | \underline{a}, \underline{y}_0] \end{aligned} \quad (5-19)$$

Therefore,

$$E[s(i) \cdot s(j) | \underline{a}, \underline{y}_0] = \gamma_{ij} + E[s(i) | \underline{a}, \underline{y}_0] \cdot E[s(j) | \underline{a}, \underline{y}_0] \quad (5-20)$$

in which  $[\gamma_{ij}]$  is given by  $[\beta_{ij}]^{-1}$ .

A closed form expression for  $\gamma_{ij}$  and therefore for  $E[s(i) \cdot s(j) | \underline{a}, \underline{y}_0]$  can be obtained by representing the speech model with equation (3-6). From equation (5-14),  $p(\underline{s}_0 | \underline{a}, \underline{y}_0)$  is given by

$$p(\underline{s}_0 | \underline{a}, \underline{y}_0) = N(\underline{m}, \underline{V}) \quad (5-21a)$$

in which

$$\underline{m} = (R_s^{-1} + \frac{1}{\sigma_d^2} \cdot I)^{-1} \cdot (\frac{1}{\sigma_d^2} \cdot \underline{y}_0 + R_s^{-1} \cdot \underline{m}_s) \quad (5-21b)$$

and

$$\underline{V} = [\gamma_{ij}] = (R_s^{-1} + \frac{1}{\sigma_d^2} \cdot I)^{-1} \quad (5-21c)$$

where  $\underline{m}_s$  and  $R_s$  are given by equation (5-14b)

Since

$$\underline{V} = E[(\underline{s}_0 - E[\underline{s}_0 | \underline{a}, \underline{y}_0]) \cdot (\underline{s}_0 - E[\underline{s}_0 | \underline{a}, \underline{y}_0])^T | \underline{a}, \underline{y}_0]$$

$$= E[\underline{s}_0 \cdot \underline{s}_0^T | \underline{a}, \underline{y}_0] - E[\underline{s}_0 | \underline{a}, \underline{y}_0] \cdot E[\underline{s}_0^T | \underline{a}, \underline{y}_0], \quad (5-22)$$

$$E[\underline{s}_0 \cdot \underline{s}_0^T | \underline{a}, \underline{y}_0] = \underline{V} + E[\underline{s}_0 | \underline{a}, \underline{y}_0] \cdot E[\underline{s}_0^T | \underline{a}, \underline{y}_0] = \underline{V} + \underline{m} \cdot \underline{m}^T \quad (5-23)$$

in which  $\underline{m}$  and  $\underline{V}$  are given by equation (5-21). Equation (5-23) is a closed form expression for  $E[s(i) \cdot s(j) | \underline{a}, \underline{y}_0]$  for  $0 \leq i, j \leq N-1$ .

An alternative way to obtain  $\gamma_{ij}$  in equation (5-20) is by representing the noisy speech model with equation (3-13). When  $u(n) = g \cdot w(n)$ , equation (3-13) is a special case of equation (4-23). Then from the smoothing form of a Kalman filter, we can obtain the covariance function of the states conditioned on all the observations and known matrices such as  $F(n)$ , which in our case directly leads to  $\gamma_{ij}$ . The Kalman filtering approach has an advantage in that only  $p \times p$  matrix operations are required while equation (5-23) requires  $N \times N$  matrix operations.

### V.4.3 RLMAP Estimation Procedure

Summarizing the steps involved in the linear implementation,

- Step 1: Begin with  $\hat{\underline{a}}_i$ , the  $i$ th estimate of  $\underline{a}$ .
- Step 2: A. Obtain  $\hat{\underline{s}}_{0i+1}$ , the  $i+1$ st estimate of  $\underline{s}_0$ , by solving equation (5-12), from equation (5-15) or from the smoothing form of a Kalman filter.
- B. Obtain  $\beta_{ij}$  from equation (5-18) and  $\gamma_{ij}$  from  $[\gamma_{ij}] = [\beta_{ij}]^{-1}$ , or obtain  $\gamma_{ij}$  from equation (5-21c), or from the smoothing form of a Kalman filter.
- C. Estimate  $s(i) \cdot s(j)$  from equation (5-20) with the results obtained in the steps A. and B. above, or estimate  $\underline{s}_0 \cdot \underline{s}_0^T$  from Equation (5-23).
- Step 3: Obtain  $\hat{\underline{a}}_{i+1}$ , the  $i+1$ st estimate of  $\underline{a}$ , by minimizing equation (4-19) with  $s(i) \cdot s(j)$  or  $\underline{s}_0 \cdot \underline{s}_0^T$  obtained in Step 2.

The above steps can be continued for as many desirable iterations. The initial estimate  $\hat{\underline{a}}_0$  can be obtained by simply applying the correlation method of the linear prediction analysis to  $\underline{y}_0$ . Like the LMAP case,

there are a number of ways of obtaining  $g$  and  $\underline{s}_I$  which are the assumed known variables in the algorithm. The possible methods discussed in Section V.3 are equally applicable to the algorithm discussed in this section. We'll refer to this algorithm as the Revised Linearized MAP" (RLMAP) estimation procedure.

To emphasize the difference between the LMAP and RLMAP algorithms, a block diagram that represents one iteration of the two algorithms is shown in Figure 5.1. The only difference between the two algorithms is an additional term  $\underline{v}$  in estimating  $\underline{s}_0 \cdot \underline{s}_0^T$  in the RLMAP algorithm. Compared with the LMAP algorithm discussed in Section V.3, the RLMAP algorithm is computationally less tractable. As will be discussed in Chapter VI, however, when  $N$  is assumed to approach  $\infty$ , the RLMAP algorithm is slightly more complex in its computation than the LMAP algorithm. In the RLMAP algorithm, there are at least two ways  $\underline{s}_0$  can be estimated. One way is to use  $\hat{\underline{s}}_0$  obtained in Step 2A. This is equivalent to estimating  $\underline{s}_0$  by  $E[s_0 | \underline{a}, \underline{y}_0]$ . Alternatively,  $\underline{s}_0$  can be estimated by forming  $\phi_s(n)$  from  $s(i) \cdot \hat{s}(j)$  and assuming some phase of  $\underline{s}_0$ . The estimated  $\hat{\underline{s}}_0$  can be used as enhanced speech if speech enhancement is desired.

#### V.5 Extension to Colored Background Noise Case

Our discussions in Sections V.2, V.3 and V.4 are

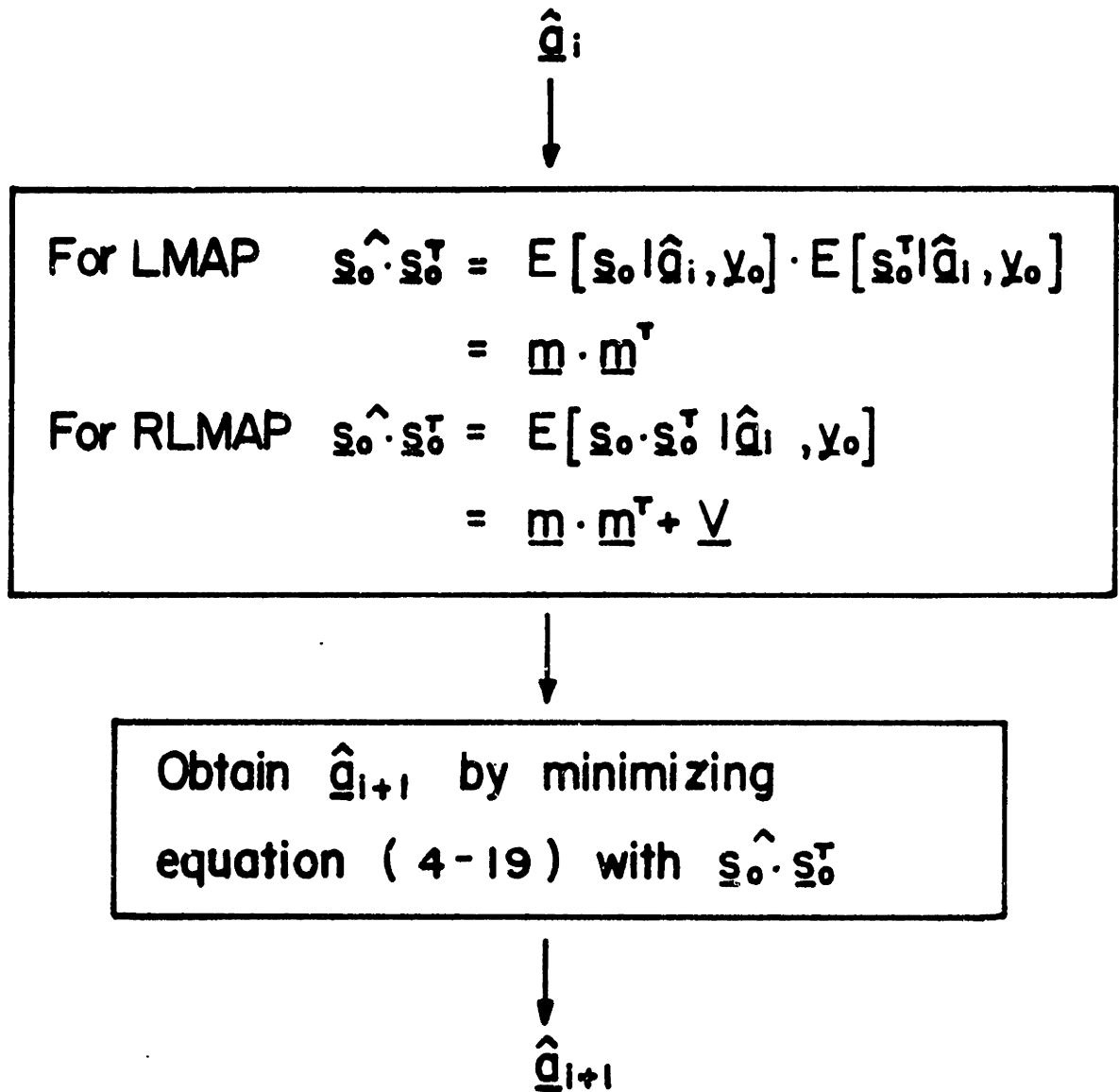


Figure 5.1 One iteration of LMAP and RLMAP algorithms.  $\underline{m}$  and  $\underline{V}$  are given by equation (5-21) in the text.

based on white Gaussian noise as the additive background noise. In this section, the theoretical results are extended to the case when the background noise is Gaussian but colored. When the background noise is colored, all the discussions in the previous three sections remain unchanged except that estimating  $\underline{s}_0 \cdot \underline{s}_0^T$  by  $E[\underline{s}_0 | \underline{a}, \underline{y}_0]$   $\cdot E[\underline{s}_0^T | \underline{a}, \underline{y}_0]$  or  $E[\underline{s}_0 \cdot \underline{s}_0^T | \underline{a}, \underline{y}_0]$  should again be shown to be a linear problem.

From equation (3-6e) with  $u(n) = g \cdot w(n)$ , equation (5-14) can be easily generalized as

$$p(\underline{s}(N-1,0) | \underline{a}, \underline{y}_0) = N((R_s^{-1} + R_d^{-1})^{-1} \cdot (R_d^{-1} \cdot \underline{y}_0 + R_s^{-1} \cdot \underline{m}_s), (R_s^{-1} + R_d^{-1})^{-1}) \quad (5-24)$$

in which

$$\begin{aligned} \underline{m}_s &= (I - A)^{-1} \cdot A_I \cdot \underline{s}_I \\ R_s &= g^2 \cdot (I - A)^{-1} \cdot ((I - A)^{-1})^T \\ R_d &= E[\underline{d}(N-1,0) \cdot \underline{d}^T(N-1,0)] \end{aligned}$$

which is obtained from the assumed known statistics of  $d(n)$ , and  $A$  and  $A_I$  are defined in equation (3-6).

Equation (5-24) can be used to show that estimating  $\underline{s}_0$  by  $E[\underline{s}_0 | \underline{a}, \underline{y}_0]$  and  $\underline{s}_0 \cdot \underline{s}_0^T$  by  $E[\underline{s}_0 \underline{s}_0^T | \underline{a}, \underline{y}_0]$  are still linear problems since

$$E[\underline{s}_0 | \underline{a}, \underline{y}_0] = (R_s^{-1} + R_d^{-1})^{-1} \cdot (R_d^{-1} \cdot \underline{y}_0 + R_s^{-1} \cdot \underline{m}_s) \quad (5-25)$$

$$\text{and } E[\underline{s}_0 \cdot \underline{s}_0^T | \underline{a}, \underline{y}_0] = (R_s^{-1} + R_d^{-1})^{-1} \\ + E[\underline{s}_0 | \underline{a}, \underline{y}_0] \cdot E[\underline{s}_0^T | \underline{a}, \underline{y}_0] \quad (5-26)$$

#### V.6 Relationship Among Maximization of $p(\underline{a} | \underline{y}_0)$ , LMAP, and RLMAP Algorithms

The LMAP and RLMAP algorithms have been developed in this chapter by attempting to suboptimally maximize  $p(\underline{a} | \underline{y}_0)$ . Some recent theoretical work by Musicus [38] carried out in parallel with this dissertation shows that a close relationship exists among the LMAP and RLMAP algorithms and the problem of maximizing  $p(\underline{a} | \underline{y}_0)$ . More specifically, suppose that  $g$  is known and  $\underline{s}_I = \underline{0}$ . Representing  $p(\underline{a} | \underline{y}_0)$  by  $f(\underline{a}) \cdot \exp(g(\underline{a}))$ , the LMAP and RLMAP algorithms increase  $g(\underline{a})$  and  $p(\underline{a} | \underline{y}_0)$  respectively at each iteration unless a converging solution is reached. Therefore if  $g$  is assumed known and  $\underline{s}_I$  is assumed to be  $\underline{0}$ , then the RLMAP algorithm is one way to maximize  $p(\underline{a} | \underline{y}_0)$ . Further theoretical work related to the above discussions is currently under way and will be reported by Musicus [38].



CHAPTER VI IMPLEMENTATION: THREE NOISE  
REDUCTION SYSTEMS

VI.1 Introduction

In this chapter, three noise reduction systems that are implemented and evaluated are discussed. Two systems discussed in Sections VI.2 and VI.3 are derived by approximating the LMAP and RLMAP algorithms discussed in Chapter V. Even though the LMAP and RLMAP algorithms require solving only sets of linear equations or implementing a Kalman filter, some approximations lead to computationally simpler systems by making use of an FFT algorithm. In Section VI.4, a speech enhancement system discussed in Section II.2.6 is summarized. The primary purpose of implementing this system is to compare it with the other two systems discussed in Sections VI.2 and VI.3. Since the system summarized in Section VI.4 is probably as good in its performance as any other speech enhancement system summarized in Chapter II, such a comparison can provide an indication of the performance of the two systems derived from the theoretical framework of this dissertation relative to other speech enhancement systems previously proposed. The results of the evaluation of the three systems will be presented in Chapters VII and VIII.

## VI.2 System A

For each iteration of the LMAP algorithm discussed in Chapter V, it is in general necessary to solve a set of  $p$  linear equations to estimate  $\underline{a}$  from  $\hat{\underline{s}}_0$  and  $N$  linear equations to estimate  $\underline{s}_0$  from  $\hat{\underline{a}}$  and  $\underline{y}_0$ . Since  $N$  in general is in the order of several hundred for a typical application in speech, solving a set of  $N$  linear equations simultaneously can be computationally tedious. Thus we develop a procedure that approximates solving the set of  $N$  linear equations.

From equations (5-11b) and (5-12),

$$\begin{aligned} \frac{g^2}{\sigma_d^2} \cdot y(i) = & s(i) - \sum_{k=1}^p a_k \cdot s(i-k) - \sum_{k=1}^p a_k \cdot s(i+k) \\ & + \sum_{k=1}^p \sum_{\ell=1}^p a_k \cdot a_\ell \cdot s(i+k-\ell) + \frac{g^2}{\sigma_d^2} \cdot s(i) \end{aligned}$$

for  $0 \leq i \leq N-p-1$  (6-1a)

$$\begin{aligned} \frac{g^2}{\sigma_d^2} \cdot y(i) = & s(i) - \sum_{k=1}^p a_k \cdot s(i-k) - \sum_{k=1}^{N-1-i} a_k \cdot s(i+k) \\ & + \sum_{k=1}^p \sum_{\ell=1}^p a_k \cdot a_\ell \cdot s(i+k-\ell) + \frac{g^2}{\sigma_d^2} \cdot s(i) \end{aligned}$$

with  $\underline{s}(N+p-3, N) = \underline{0}$  for  $N-p < i < N-2$  (6.1b)

and

$$\frac{g^2}{\sigma_d^2} \cdot y(i) = s(i) - \sum_{k=1}^p a_k \cdot s(i-k) + \frac{g^2}{\sigma_d^2} \cdot s(i) \text{ for } i=N-1 \quad (6-1c)$$

Solving equation (6-1) for  $\underline{s}_0$  in general requires solving N simultaneous linear equations. However, if we assume that  $\underline{s}(p-1,0)$  is also given as well as  $\underline{s}_I$ , then the N equations do not have to be solved simultaneously. More specifically, rearranging equation (6-1a),

$$a_p \cdot s(i+p) = s(i) - \sum_{k=1}^p a_k \cdot s(i-k) - \sum_{k=1}^{p-1} a_k \cdot s(i+k) + \sum_{k=1}^p \sum_{\ell=1}^p a_k \cdot a_\ell \cdot s(i+k-\ell) + \frac{g^2}{\sigma_d^2} \cdot s(i) - \frac{g^2}{\sigma_d^2} \cdot y(i)$$

for  $0 \leq i \leq N-p-1$  (6-2)

$s(i+p)$  in equation (6-2) for  $0 \leq i \leq N-p-1$  can be solved individually if  $\underline{s}(p-1,0)$  is given since the right hand side of equation (6-2) involves terms of  $s(n)$  for  $n < i+p$ .  $\underline{s}(p-1,0)$ , of course, is not given, but we could assume  $\underline{s}(p-1,0) = \underline{y}(p-1,0)$ . For N sufficiently large relative to p, we would in general expect that the effect of a specific assumption of  $\underline{s}(p-1,0)$  is rather small.

In the above, we have developed a procedure which does not require solving a set of N linear equations simultaneously. However, solving for  $\underline{s}_0$  from equation (6-2) still requires in the order of  $N \cdot p^2$  multiplications. Furthermore, once  $\underline{s}_0$  is estimated, the correlation function has to be formed from  $\underline{\hat{s}}_0$ . An alternative approach which is computationally simpler and leads to a system with

a simple interpretation is to consider the problem in the frequency domain. More specifically, z transforming equation (6-1a) with the assumption that the difference equation holds for all i (i.e.,  $N=\infty$ ),

$$\begin{aligned} \frac{g^2}{\sigma_d^2} \cdot Y(z) = S(z) - \sum_{k=1}^p a_k \cdot S(z) \cdot z^{-k} - \sum_{k=1}^p a_k \cdot S(z) \cdot z^k \\ + \sum_{k=1}^p \sum_{\ell=1}^p a_k \cdot a_\ell \cdot S(z) \cdot z^{k-\ell} + \frac{g^2}{\sigma_d^2} \cdot S(z) \end{aligned} \quad (6-3)$$

and therefore

$$S(\omega) = Y(\omega) \cdot \frac{P_s(\omega)}{P_s(\omega) + \sigma_d^2} \quad (6-4a)$$

where

$$\begin{aligned} P_s(\omega) &= \frac{g^2}{1 - 2 \cdot \sum_{k=1}^p a_k \cdot \cos k\omega + \sum_{k=1}^p \sum_{\ell=1}^p a_k \cdot a_\ell \cdot \cos(k-\ell)\omega} \\ &= \frac{g^2}{|1 - \sum_{k=1}^p a_k \cdot e^{-j\omega k}|^2} \end{aligned} \quad (6-4b)$$

Equation (6-4) is a non-causal Wiener filter. This result is quite reasonable since it is well known that when  $y(n) = s(n) + \hat{d}(n)$  where  $s(n)$  is uncorrelated with  $d(n)$  and the power spectral densities of  $s(n)$  and  $d(n)$  are known, the MMSE estimate of  $s(n)$  from  $y(n)$  can be obtained by a Wiener filter. For this reason, then, for a more general

case when the background noise is colored, the procedure for obtaining  $s(n)$  by equation (5-20) is equivalent to estimating  $s(n)$  by filtering  $y(n)$  with a linear, time invariant filter with the frequency response given by

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} \quad (6-5)$$

and  $P_d(\omega)$  represents the power spectral density of the background noise and  $P_s(\omega)$  represents the power spectral density of speech given by equation (6-4b).

Theoretically, the non-causal Wiener filter requires an infinite amount of data. In practice, we have only  $N$  points of data that can be modelled as  $y_w(n) = y(n) \cdot w_s(n)$  where  $w_s(n)$  represents a sufficiently smooth analysis window over the effective length of  $h(n)$ . For a sufficiently large  $N$  and small effective length of  $h(n)$  relative to  $N$ ,

$$\begin{aligned} (y(n) \cdot w_s(n)) * h(n) &\approx (y(n) * h(n)) \cdot w_s(n) \\ &= s(n) \cdot w_s(n) \end{aligned} \quad (6-6a)$$

$$\text{and therefore } y_w(n) * h(n) \approx s_w(n) \quad (6-6b)$$

Based on equation (6-6b),  $s_w(n)$  is estimated by

$$\hat{s}_w(n) = y_w(n) * h(n) \quad (6-6c)$$

We expect that the approximation given by equation (6-6b) is not good for  $n$  close to 0 or  $N-1$  but is adequate for  $0 \ll n \ll N-1$ . For a sufficiently large  $N$ , it is expected that the poor approximation at the edges of the window do not have a large effect. From equation (6-6c),

$$\hat{S}_w(\omega) = Y_w(\omega) \cdot H(\omega) \quad (6-7a)$$

and 
$$|\hat{S}_w(\omega)|^2 = |Y_w(\omega)|^2 \cdot |H(\omega)|^2 \quad (6-7b)$$

Now if equation (4-22) is used rather than equation (4-19) in step 2 of the LMAP algorithm, the function that is directly used in minimizing  $\epsilon_p$  in equation (4-22) can be expressed as

$$\hat{\phi}_s(n) = \sum_{i=-\infty}^{\infty} \hat{S}_w(i) \cdot \hat{S}_w(i-n).$$

Then  $\hat{s}_w(n)$  and  $\hat{\phi}_s(n)$  can be obtained by inverse Fourier transforming  $\hat{S}_w(\omega)$  and  $|\hat{S}_w(\omega)|^2$ , i.e.,

$$\hat{s}_w(n) = F^{-1}[\hat{S}_w(\omega)] = F^{-1}[Y_w(\omega) \cdot H(\omega)] \quad (6-8a)$$

$$\hat{\phi}_s(n) = F^{-1}[|\hat{S}_w(\omega)|^2] = F^{-1}[|Y_w(\omega)|^2 \cdot |H(\omega)|^2] \quad (6-8b)$$

Denoting the M point Discrete Fourier Transform (DFT) of a sequence  $x(n)$  by  $X(k)$ ,

$$X(k) = \sum_{n=0}^{M-1} x(n) \cdot e^{-j\frac{2\pi}{M} \cdot k \cdot n} = X(\omega) \Big|_{\omega = \frac{2\pi}{M}k}$$

Since  $x(n)$  is related [39] to the Inverse Discrete Fourier Transform (IDFT) of  $X(k)$  by

$$\sum_{k=-\infty}^{\infty} x(n+k \cdot M) = \text{IDFT}[X(k)],$$

equation (6-8) leads to

$$\sum_{k=-\infty}^{\infty} \hat{s}_w(n+k \cdot M) = \text{IDFT}[\hat{S}_w(k)] = \text{IDFT}[Y_w(k) \cdot H(k)] \quad (6-9a)$$

$$\sum_{k=-\infty}^{\infty} \hat{\phi}_s(n+k \cdot M) = \text{IDFT}[\hat{\phi}_s(k)] = \text{IDFT}[|Y_w(k)|^2 \cdot |H(k)|^2] \quad (6-9b)$$

For a finite effective length of  $\hat{s}_w(n)$  and  $\hat{\phi}_s(n)$  and for a sufficiently large  $M$ ,

$$\sum_{k=-\infty}^{\infty} \hat{s}_w(n+k \cdot M) \approx \hat{s}_w(n)$$

and

$$\sum_{k=-\infty}^{\infty} \hat{\phi}_s(n+k \cdot M) \approx \hat{\phi}_s(n)$$

With this assumption, we estimate  $s_w(n)$  and  $\phi_s(n)$  by

$$\hat{s}_w(n) = \text{IDFT} (Y_w(k) \cdot H(k)) \quad (6-10a)$$

$$\hat{\phi}_s(n) = \text{IDFT} (|Y_w(k)|^2 \cdot |H(k)|^2) \quad (6-10b)$$

$\hat{s}_w(n)$  in equation (6-10a) can be used as enhanced speech.

$\hat{\phi}_s(n)$  in equation (6-10b) can be used to estimate the

a by minimizing  $\epsilon_p$  in equation (4-22).



Now, we summarize the specific algorithm that has been implemented and evaluated.

Step 0: Obtain  $\hat{\underline{a}}_0$ , the initial estimate, by the correlation method of the linear prediction analysis assuming  $s_w(n) = y_w(n)$

Step 1: Begin from  $\underline{a}_i$ , the  $i$ th estimate of  $\underline{a}$ .

Step 2: A. Estimate  $g$  by an energy measurement;

$$\frac{\sum_n w_s^2(n)}{2\pi} \int_{-\pi}^{\pi} \frac{g^2}{|1 - \sum_{k=1}^p a_k \cdot e^{-jk\omega}|^2} \cdot d\omega = \sum_n y_w^2(n) - \sum_n w_s^2(n) \cdot \sigma_d^2$$

where  $\underline{a}$  corresponds to  $\hat{\underline{a}}_i$ .

B. Estimate  $\phi_s(n)$  by IDFT  $[|Y_w(k)|^2 \cdot |H(k)|^2]$

where

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} \quad \text{with} \quad P_s(\omega) = \frac{g^2}{|1 - \sum_{k=1}^p a_k \cdot e^{-jk\omega}|^2}$$

and  $\underline{a}$  corresponds to  $\hat{\underline{a}}_i$ . If  $\hat{s}_w(n)$  is desired for speech enhancement and if this is the last iteration desired,  $\hat{s}_w(n) = \text{IDFT}(Y(k) \cdot H(k))$ .

Step 3: With the first  $p+1$  points of  $\hat{\phi}_s(n)$  and  $\bar{\underline{a}}$  and  $P_0$  given by the available a priori knowledge of  $\underline{a}$ , estimate the LPC coefficient vector  $\hat{\underline{a}}_{i+1}$  by minimizing equation (4-22). In the case when no a priori information is available, we let  $P_0$

approach  $\infty$  thus reducing the minimization of equation (4-22) to the correlation method of the linear prediction analysis.

The above steps complete one iteration and we'll refer to the above system as System A. It is noted that System A can be used to estimate  $s_w(n)$  as well as  $\underline{a}$ , and that System A does not require an estimate of  $\underline{s}_I$ . Further, it is noted that the phase of  $S(\omega)$  estimated in System A is the same as the phase of  $Y(\omega)$ . This is because the frequency response of a non-causal Wiener filter  $H(\omega)$  is real and positive and thus zero phase. In various speech enhancement systems discussed in Chapter II, we have seen that the phase of  $S(\omega)$  used is the same as the phase of  $Y(\omega)$ .

### VI.3 System B

In Section VI.2, System A was developed based on the LMAP estimation procedure discussed in Section V.3. In this section we develop a system that is based on the RLMAP estimation procedure discussed in Section V.4. From equation (5-20), the difference between the LMAP and RLMAP estimation procedure is the additional term  $\gamma_{ij}$  in estimating  $s(i) \cdot s(j)$ . The system developed in this section is a modification of System A that incorporates the term  $\gamma_{ij}$ .

In general, to obtain  $\gamma_{ij}$  it is necessary to perform at least an inversion of an  $N \times N$  matrix since  $[\gamma_{ij}] = [\beta_{ij}]^{-1}$ . However, as we let  $N$  approach  $\infty$ , a computationally simpler procedure can be developed. More specifically for a very large  $N$ , from equations (5-11b) and (5-18),  $\beta_{ij}$  can be expressed as

$$\beta_{ij} = \alpha_{ij} + \theta_{ij} \tag{6-11a}$$

where

$$\begin{aligned} \alpha_{ij} &= \frac{1}{g^2} \cdot (1 + \sum_{k=1}^p a_k^2) && \text{for } |i-j|=0 \\ &= -\frac{1}{g^2} \cdot a_{|i-j|} + \frac{1}{g^2} \cdot \sum_{k=1}^{p-|i-j|} a_k \cdot a_{k+|i-j|} && \text{for } 0 < |i-j| < p \\ &= -\frac{1}{g^2} \cdot a_{|i-j|} && \text{for } |i-j|=p \\ &= 0 && \text{for } |i-j| > p \end{aligned} \tag{6-11b}$$

and

$$\begin{aligned} \theta_{ij} &= \frac{1}{\sigma_d^2} && \text{for } |i-j|=0 \\ &= 0 && \text{otherwise} \end{aligned} \tag{6-11c}$$

From equation (6-11), since  $\alpha_{ij}$  and  $\theta_{ij}$  depend only on the time difference  $i-j$ , representing  $\alpha_{ij}$ ,  $\theta_{ij}$  and  $\beta_{ij}$  by

$$\alpha(n) = \alpha(i-j) = \alpha_{ij}, \quad \theta(n) = \theta(i-j) = \theta_{ij} \quad \text{and} \quad \beta(n) = \beta(i-j) = \beta_{ij},$$

$$\beta(n) = \alpha(n) + \theta(n) \tag{6-12a}$$

$$\begin{aligned} \text{where } \alpha(n) &= \frac{1}{g^2} \cdot \left(1 + \sum_{k=1}^p a_k^2\right) && \text{for } |n|=0 \\ &= -\frac{1}{g^2} \cdot a_{|n|} + \frac{1}{g^2} \cdot \sum_{k=1}^{p-|n|} a_k \cdot a_{k+|n|} && \text{for } 0 < |n| < p \\ &= -\frac{1}{g^2} \cdot a_{|n|} && \text{for } |n|=p \\ &= 0 && \text{for } |n| > p \end{aligned} \tag{6-12b}$$

$$\begin{aligned} \text{and } \theta(n) &= \frac{1}{\sigma_d^2} && \text{for } n=0 \\ &= 0 && \text{otherwise} \end{aligned} \tag{6-12c}$$

From equation (6-12b), taking the Fourier transform of  $\alpha(n)$ <sup>9</sup>,

$$A(\omega) = F[\alpha(n)] = \left( \frac{g^2}{\left|1 - \sum_{k=1}^p a_k \cdot e^{-jk\omega}\right|^2} \right)^{-1} \tag{6-13}$$

<sup>9</sup>This result can alternatively be obtained by noting that  $[\alpha_{ij}]^{-1}$  is the covariance matrix of  $p(\underline{s}_0 | \underline{a})$  and that for  $N$  approaching  $\infty$ ,  $\alpha_{ij}$  depends only on the time difference  $i-j$ .

From equations (6-4b) and (6-13),

$$A(\omega) = \frac{1}{P_s(\omega)} \quad (6-14a)$$

and

$$P_s(\omega) = \frac{g^2}{|1 - \sum_{k=1}^p a_k \cdot e^{-jk\omega}|} \quad (6-14b)$$

From equations (6-12) and (6-14),

$$B(\omega) = A(\omega) + \theta(\omega) \quad \text{and therefore,}$$

$$B(\omega) = \frac{1}{P_s(\omega)} + \frac{1}{\sigma_d^2} \quad (6-15)$$

Since  $\gamma_{ij} = [\beta_{ij}]^{-1}$  and  $\beta_{ij}$  depends on the time difference  $i-j$ , representing  $\gamma_{ij}$  by  $\gamma(n) = \gamma(i-j) = \gamma_{ij}$ ,

$$\gamma(n) * \beta(n) = \delta(n) \quad \text{and therefore,}$$

$$\Gamma(\omega) \cdot B(\omega) = 1 \quad (6-16)$$

From equations (6-15) and (6-16),

$$\Gamma(\omega) = \frac{P_s(\omega) \cdot \sigma_d^2}{P_s(\omega) + \sigma_d^2} \quad (6-17)$$

and  $P_s(\omega)$  is given by equation (6-14b).

Since  $[\theta_{ij}]^{-1}$  represents the covariance matrix of the background noise, for  $N$  approaching  $\infty$ ,  $\theta(\omega) = 1/P_d(\omega)$

where  $P_d(\omega)$  is the power spectrum of the background noise. Therefore, a more general result of equation (6-17) that also applies to the colored background noise is given by

$$\Gamma(\omega) = \frac{P_s(\omega) \cdot P_d(\omega)}{P_s(\omega) + P_d(\omega)} \quad (6-18)$$

and  $P_s(\omega)$  is given by equation (6-14b). From equation (6-18),

$$\gamma(n) = \gamma(i-j) = \gamma_{ij} = F^{-1} \left[ \frac{P_s(\omega) \cdot P_d(\omega)}{P_s(\omega) + P_d(\omega)} \right] \quad (6-19)$$

From equation (5-20),

$$\begin{aligned} s(i) \hat{\cdot} s(j) &= E[s(i) \cdot s(j) | \underline{a}, \underline{y}_0] = E[s(i) | \underline{a}, \underline{y}_0] \\ &\cdot E[s(j) | \underline{a}, \underline{y}_0] + \gamma_{ij}. \end{aligned}$$

Denoting  $s(i) \cdot w_s(i)$  by  $s_w(i)$  and letting  $\hat{\phi}_s(n) = \sum_{i=-\infty}^{\infty} s_w(i) \hat{\cdot} s_w(i-n)$  for  $0 \leq n \leq p$ ,

$$\begin{aligned} \hat{\phi}_s(n) &= \sum_{i=-\infty}^{\infty} E[s(i) | \underline{a}, \underline{y}_0] \cdot E[s(i-n) | \underline{a}, \underline{y}_0] \cdot w_s(i) \cdot w_s(i-n) \\ &+ \sum_{i=-\infty}^{\infty} \gamma_{i \ i-n} \cdot w_s(i) \cdot w_s(i-n) \quad \text{for } 0 \leq n \leq p. \end{aligned} \quad (6-20)$$

Approximating  $\gamma_{i \ i-n}$  by  $\gamma(n)$  for  $0 \leq n \leq p$ ,

$$\sum_{i=-\infty}^{\infty} E[s_w(i) | \underline{a}, \underline{y}_0] \cdot E[s_w(i-n) | \underline{a}, \underline{y}_0]$$

by  $\hat{\phi}_s(n)$  in equation (6-8b), and  $\sum_{i=-\infty}^{\infty} w_s(i) \cdot w_s(i-n)$

by  $\sum_{i=-\infty}^{\infty} w_s^2(i)$  for  $0 \leq n \leq p$ ,

$$\begin{aligned} \hat{\phi}_s(n) = F^{-1} [ & |Y_w(\omega)|^2 \cdot \left( \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} \right)^2 + \sum_{i=-\infty}^{\infty} w_s^2(i) \\ & \cdot \frac{P_s(\omega) \cdot P_d(\omega)}{P_s(\omega) + P_d(\omega)} ] \text{ for } 0 \leq n \leq p \end{aligned} \quad (6-21)$$

In an analogous manner as equation (6-10) was obtained from equation (6-8),  $\hat{\phi}_s(n)$  is estimated from equation (6-21) by

$$\begin{aligned} \hat{\phi}_s(n) = \text{IDFT} [ & |Y_w(k)|^2 \cdot \left| \frac{P_s(k)}{P_s(k) + P_d(k)} \right|^2 + \sum_{i=-\infty}^{\infty} w_s^2(i) \\ & \cdot \frac{P_s(k) \cdot P_d(k)}{P_s(k) + P_d(k)} \text{ for } 0 \leq n \leq p \end{aligned} \quad (6-22)$$

and  $P_s(\omega)$  is given by equation (6-14b). Equation (6-22) can be used in minimizing  $\epsilon_p$  in equation (4-22).

Now we summarize the specific algorithm that has been implemented and evaluated.

Step 0: Obtain  $\hat{\underline{a}}_0$  by the correlation method of the linear prediction analysis assuming  $s_w(n) = y_w(n)$ .

Step 1: Begin from  $\hat{\underline{a}}_i$ , the  $i$ th estimate of  $\underline{a}$ .

Step 2: A. Estimate  $g$  by an energy measurement;

$$\frac{\sum_n w_s^2(n)}{2\pi} \cdot \int_{-\pi}^{\pi} \frac{g^2}{|1 - \sum_{k=1}^p a_k \cdot e^{-jk\omega}|^2} \cdot d\omega$$

$$= \sum_n y_w^2(n) - \sum_n w_s^2(n) \cdot \sigma_d^2$$

where  $\underline{a}$  corresponds to  $\hat{\underline{a}}_i$ .

B. Estimate  $\phi_s(n)$  by

$$\text{IDFT}[|Y(k)|^2 \cdot |H(k)|^2 + \sum_{i=-\infty}^{\infty} w_s^2(i) \cdot \frac{P_s(k) \cdot P_d(k)}{P_s(k) + P_d(k)}]$$

$$\text{where } H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)},$$

$$P_s(\omega) = \frac{g^2}{|1 - \sum_{k=1}^p a_k \cdot e^{-jk\omega}|^2}$$

and  $\underline{a}$  corresponds to  $\hat{\underline{a}}_i$ . If  $s_w^{\hat{}}(n)$  is desired and if it is the last iteration to be performed, estimate  $s_w(n)$  by  $\text{IDFT}[Y_w(k) \cdot H(k)]$ .

Step 3: With the first  $p+1$  points of  $\phi_s^{\hat{}}(n)$ , and  $\bar{\underline{a}}$  and  $P_0$  given by the available a priori knowledge of  $\underline{a}$ , estimate  $\underline{a}_{i+1}$  by minimizing equation (4-22).



The above steps complete one iteration and we'll refer to the above system as System B. It is noted that System B can be used to estimate  $s_w(n)$  as well as  $\underline{a}$ .

#### VI.4 System C

This system is based on a speech enhancement method discussed in Section II.2.6. The specific algorithm implemented and evaluated is given below.

Step 1: Estimate  $|S_w(\omega)|^2$  by

$$|\hat{S}_w(\omega)|^2 = |Y_w(\omega)|^2 - k \cdot E[|D_w(\omega)|^2]$$

$$\text{for } |Y_w(\omega)|^2 \geq k \cdot E[|D_w(\omega)|^2]$$

0 otherwise

for some constant  $k$ .  $S_w(\omega)$ ,  $Y_w(\omega)$  and  $D_w(\omega)$  represent the Fourier Transform of the windowed segment of speech, noisy speech, and noise respectively.

Step 2: Obtain  $\hat{\phi}_s(n)$  by IDFT[ $|S_w(k)|^2$ ]. If  $\hat{s}_w(n)$  is desired, then  $s_w(n)$  is estimated from  $|S_w(\omega)|$  in Step 1 and the phase of  $Y_w(\omega)$ .

Step 3: Estimate  $\underline{a}$  by minimizing  $\epsilon_p$  in equation (4-22) with the first  $p+1$  points of  $\hat{\phi}_s(n)$  obtained in Step 2.

We'll refer to the above system as System C. Compared with System A or System B, System C is computationally simpler. It is also noted that when  $k=0$  in System C and no a priori information is assumed, it corresponds to estimating  $\underline{a}$  by the correlation method of the linear prediction analysis with the assumption of  $s_w(n) = y_w(n)$ .

## CHAPTER VII EXAMPLES AND ILLUSTRATIONS

### VII.1 Introduction

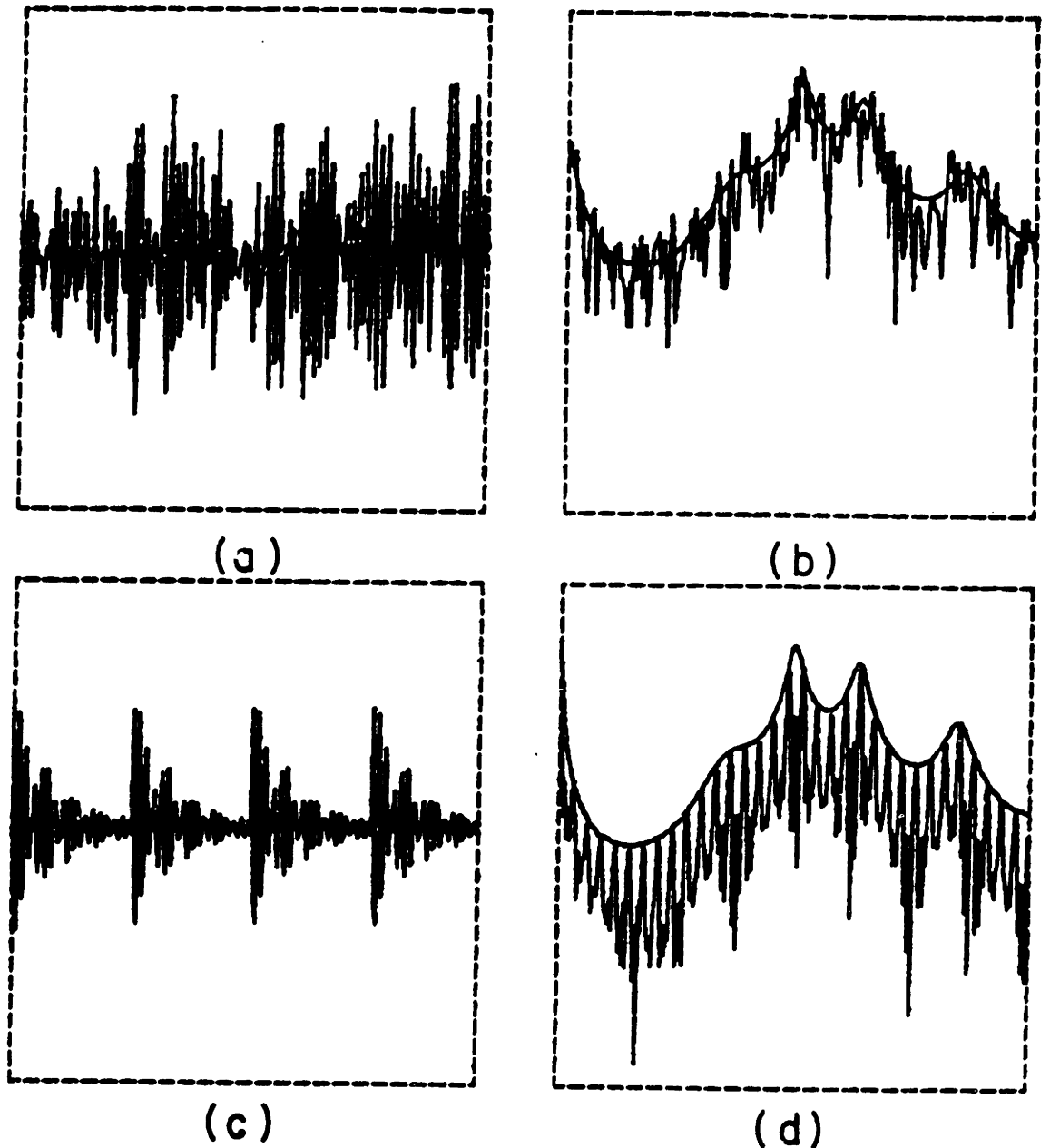
The three systems developed in Chapter VI have been applied to both synthetic and real speech data at various S/N ratios and in this chapter a few examples are illustrated. In Section VII.2, examples in which the systems are applied to synthetic data are illustrated. In Section VII.3, examples in which the systems are applied to real speech data are illustrated. In all the examples considered in this chapter, noisy data are generated by adding zero mean white Gaussian background noise and the S/N in dB is defined as  $10 \cdot \log(\frac{\sum_n s^2(n)}{\sum_n d^2(n)})$  where the summation is over the length of the analysis segment. In all the figures in which a time waveform is displayed, the duration is 25.6 msec. In all the figures in which the log magnitude spectrum is displayed, the range is approximately 50 dB and the angular frequency is between 0 and  $\pi$  that corresponds to the analog frequency between 0 and 5 kHz at 10 kHz sampling rate.

### VII.2 Application to Synthetic Data

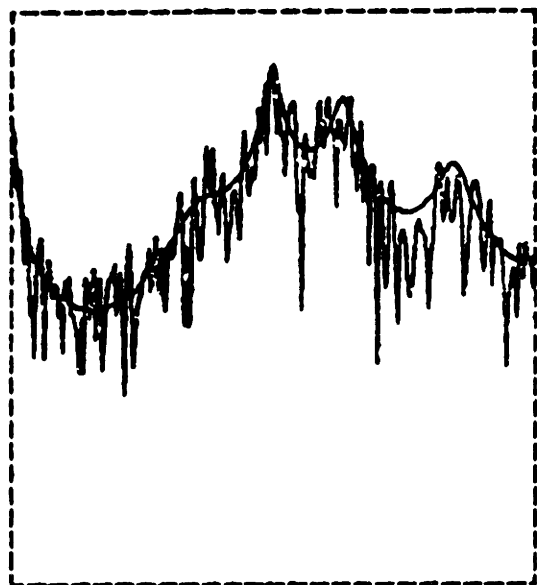
The synthetic data used in the examples are based on a 10 kHz sampling rate and are generated by exciting a tenth order all pole filter whose coefficients were derived from segments of real speech data. The excitation was

chosen in one set of examples to be white Gaussian noise and in the other set of examples to be a periodic impulse train. As we discussed in Chapter III, all the theoretical results in Chapters IV, V and VI were derived assuming a stochastic excitation. For speech without background noise, systems derived from this point of view have empirically been shown to perform well even when the excitation is a periodic impulse train and it will be seen in this chapter and Chapter VIII that this statement generally applies to the three systems under consideration. In the examples considered in this section, the analysis is based on 256 synthetic data points, the order of the all pole system is assumed to be 10, and the S/N ratios considered are 20 dB, 10 dB and 0 dB.

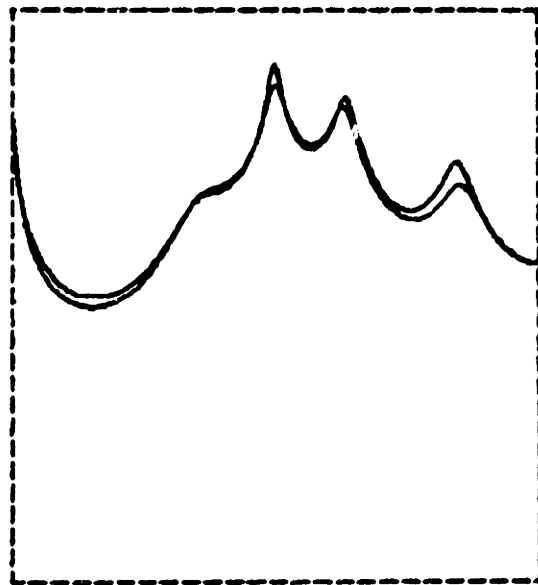
In Sections VII.2.1, VII.2.2 and VII.2.3, the performance of the three systems are discussed and illustrated individually based on one specific synthetic data segment and then later a few more examples are illustrated. The synthetic data used in Sections VII.2.1, VII.2.2 and VII.2.3 are shown in Figures 7.1 and 7.2. In Figure 7.1(a) is shown the synthetic data when the excitation is random noise. In Figure 7.1(b) is shown the log magnitude spectrum of the data in Figure 7.1(a) and a tenth order all pole fit to the spectrum by the correlation method of the linear prediction analysis. In Figure 7.1(c) is shown the synthetic data generated by the same all pole coefficients as in



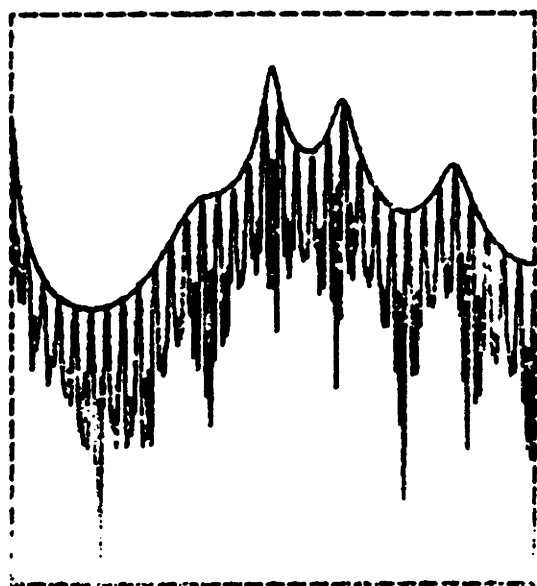
**Figure 7.1** (a) Synthetic data segment with random noise excitation; (b) Log magnitude of the spectrum of the synthetic data in (a) and an all pole fit to the spectrum by the correlation method of the linear prediction analysis; (c) Same as (a) with a pulse train excitation; (d) Same as (b) with a pulse train excitation



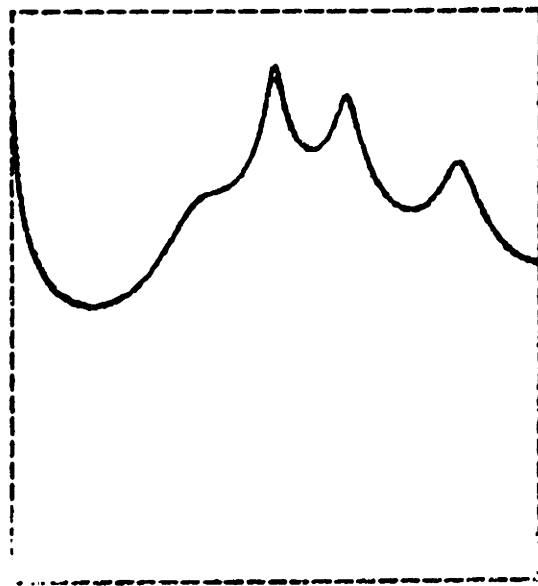
(a)



(b)



(c)



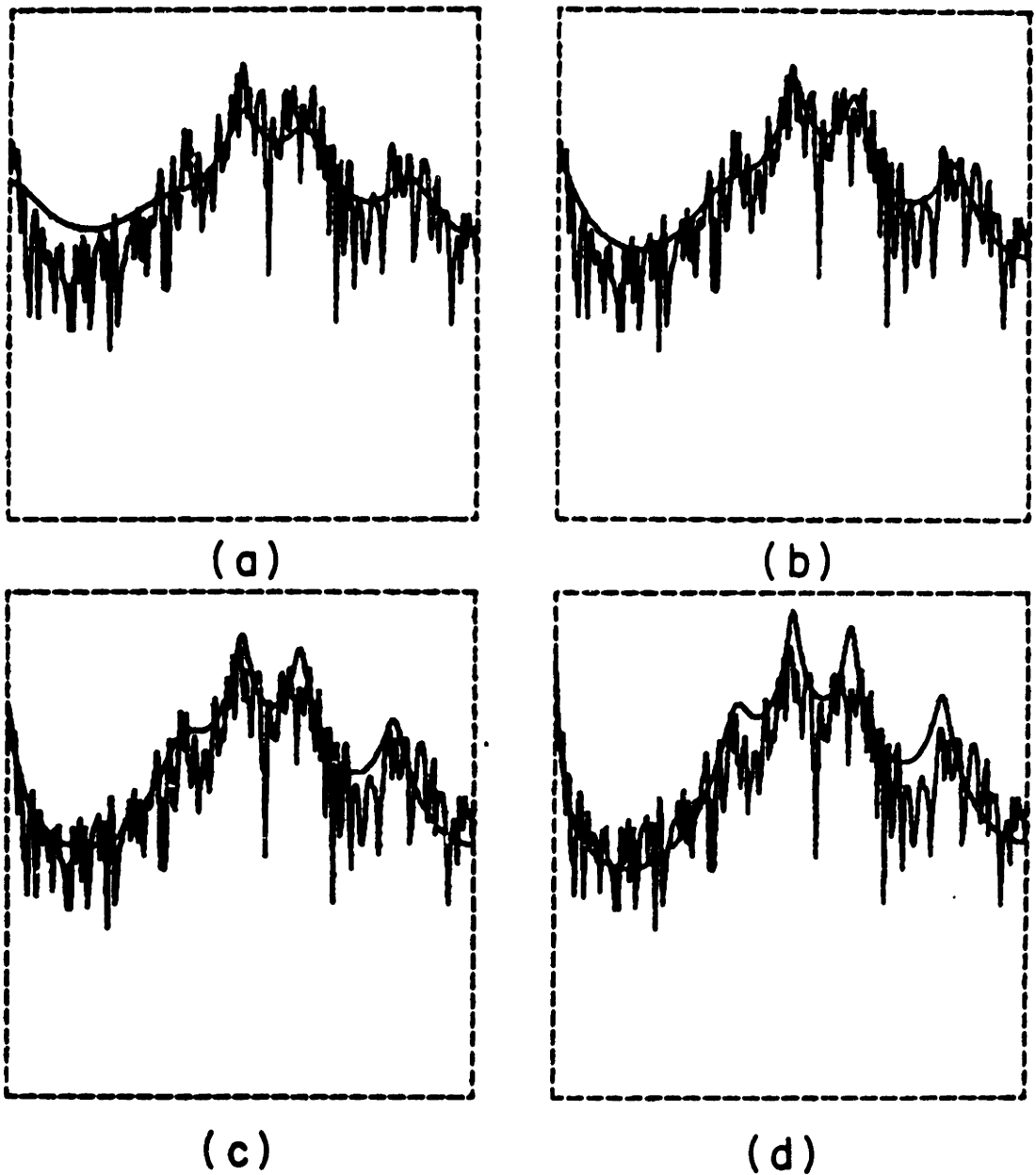
(d)

**Figure 7.2** (a) Log magnitude spectrum of the synthetic data in Figure 7.1(a) and the transfer function that corresponds to the known all pole coefficients; (b) Comparison of the transfer functions that correspond to the known all pole coefficients in (a) and the estimated all pole coefficients in Figure 7.1(b); (c) Same as (a) with a pulse train excitation; (d) Same as (b) with a pulse train excitation

Figure 7.1(a) but the excitation is now a train of pulses whose fundamental frequency is 150 Hz typical of an adult male speech. In Figure 7.1(d) is illustrated a tenth order all pole fit to the spectrum by the correlation method of the linear prediction analysis. Since the data used are synthetic, the all pole coefficients from which the synthetic data were generated are known. In Figure 7.2(a) is shown the log magnitude spectrum of the synthetic data in Figure 7.1(a) and the transfer function that corresponds to the known all pole coefficients. In Figure 7.2(b) is shown the two transfer functions that correspond to the known all pole coefficients and the all pole coefficients estimated from the synthetic data by the correlation method of the linear prediction analysis. Figures 7.2(c) and 7.2(d) are equivalent to Figures 7.2(a) and 7.2(b) with the difference in that the excitation is a train of pulses.

#### VII.2.1 Application of System A to Synthetic Data

In Figure 7.3 is shown the results of the analysis based on System A as a function of the number of iterations when the S/N ratio is 20 dB and the excitation is random noise. More specifically, in Figure 7.3(a) is shown the all pole fit to the noisy synthetic data by the correlation method of the linear prediction analysis with the assumption that  $s_w(n) = y_w(n)$ , i.e. zeroth iteration. Figures 7.3(b),



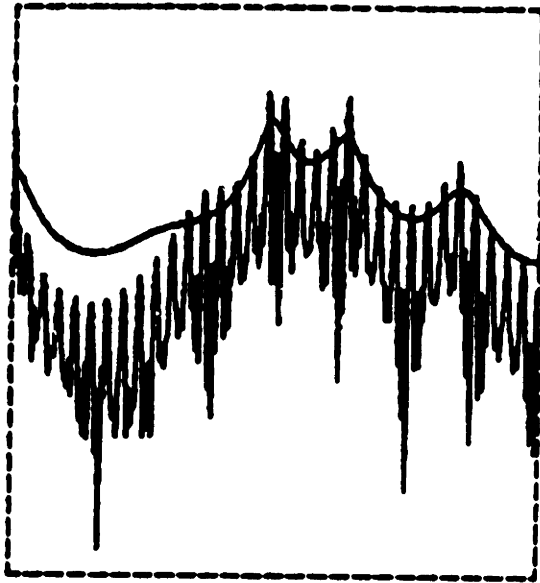
**Figure 7.3** Comparison of System A

- (a) Log magnitude spectrum of the synthetic data in Figure 7.1(a) and an all pole fit to the noisy data spectrum after the zeroth iteration of System A at  $S/N \approx 20$  dB;
- (b) Same as (a) after the first iteration of System A;
- (c) Same as (a) after the second iteration of System A;
- (d) Same as (a) after the third iteration of System A

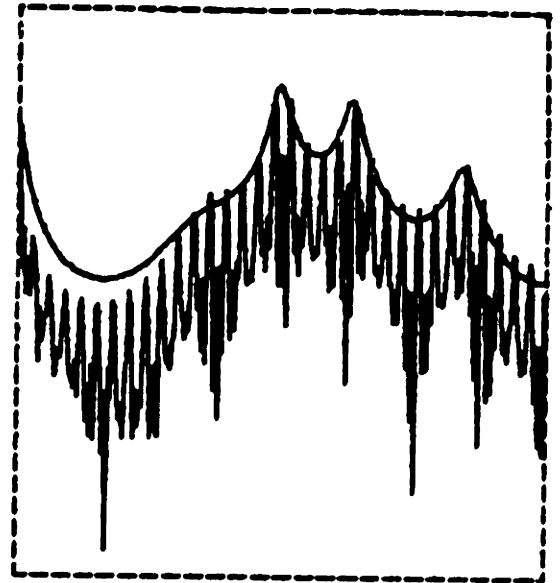


(c) and (d) represent the transfer functions obtained by applying System A to the noisy synthetic data after one, two and three iterations, respectively. In each of the four figures ((a), (b), (c) and (d)), the true log magnitude spectrum corresponding to the excitation of random noise is also shown to facilitate the comparisons. Figure 7.4 is the same as Figure 7.3 with the difference in that the excitation is a train of pulses. Figures 7.5 and 7.6 are the same as Figures 7.3 and 7.4 with the difference in that the S/N ratio is 10 dB. Figures 7.7 and 7.8 are the same as Figures 7.3 and 7.4 with the difference in that the S/N ratio is 0 dB. In all the Figures 7.3 through 7.8, the analysis is based on the assumption that no a priori information of the coefficient vector is available. From the figures, it can be observed that for the three S/N ratios considered a good fit to the true log magnitude spectrum can be obtained after two iterations of System A. It is also observed that the performance of the system when applied to the synthetic data generated by an excitation of a train of pulses is similar to the case of the random noise excitation.

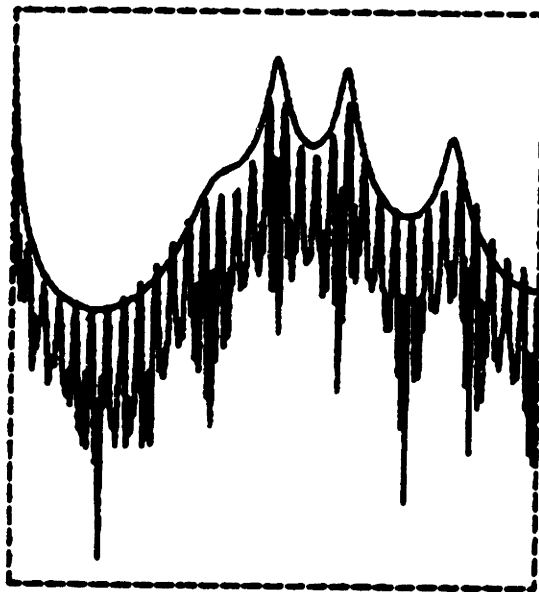
From the theoretical point of view, it is expected that a converging solution after many iterations is more desirable. In general, however, it has been observed that the converging solution of System A generates the transfer function for which the bandwidths of the poles are smaller than those associated with real speech. Such a phenomenon



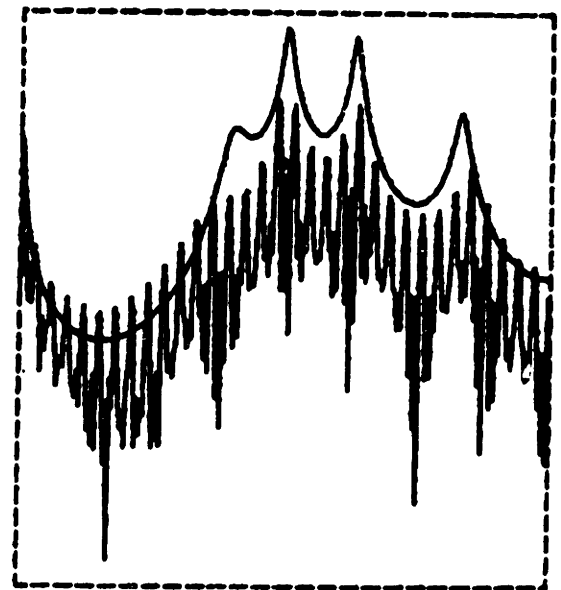
(a)



(b)

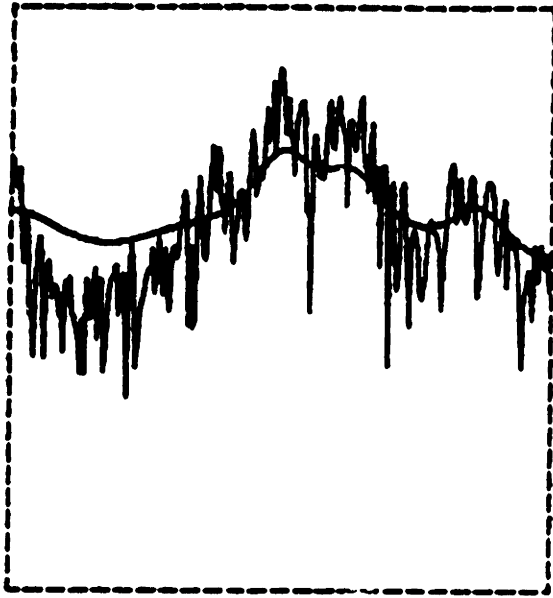


(c)

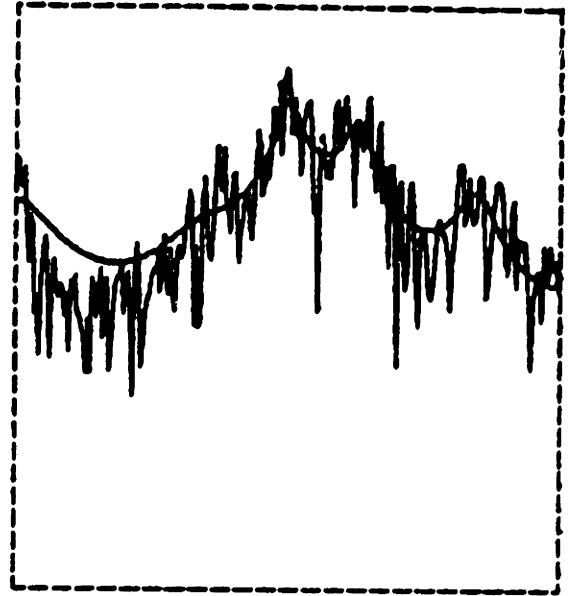


(d)

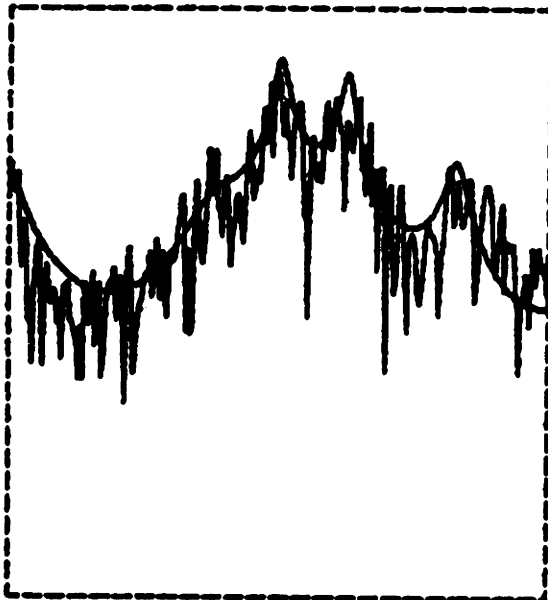
**Figure 7.4** Same as Figure 7.3 with the synthetic data of Figure 7.1(c)



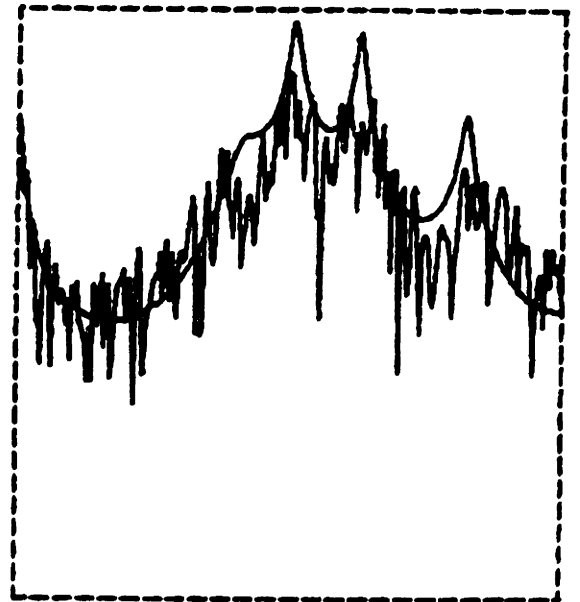
(a)



(b)

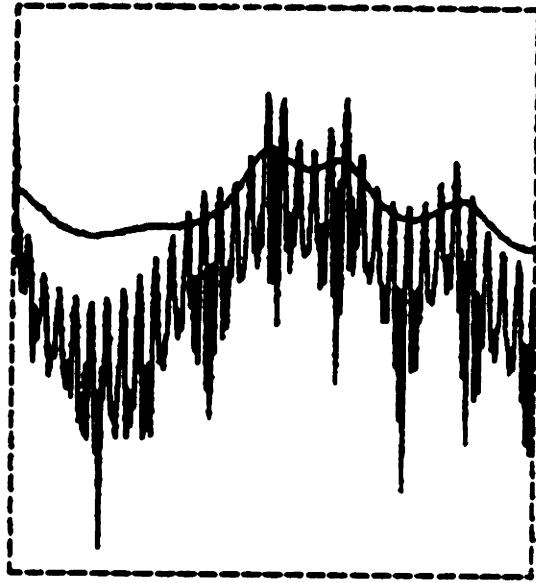


(c)

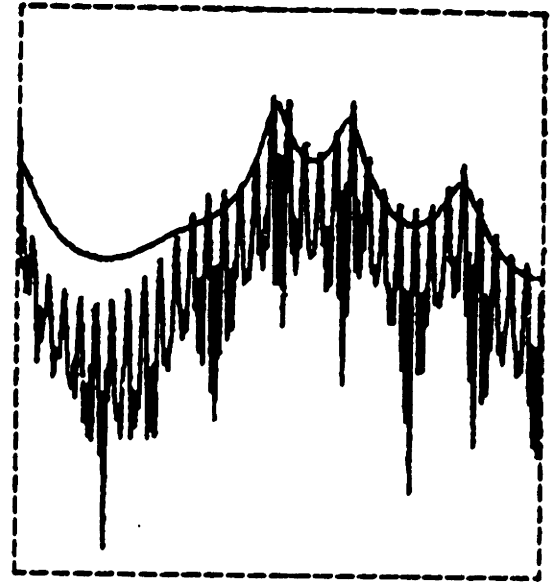


(d)

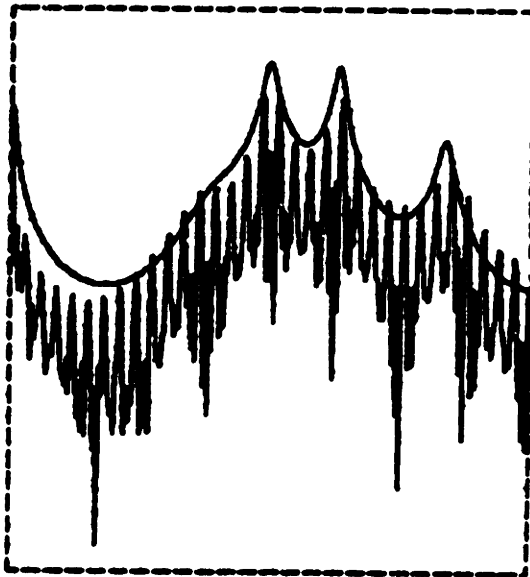
Figure 7.5 Same as Figure 7.3 with  $S/N = 10$  dB



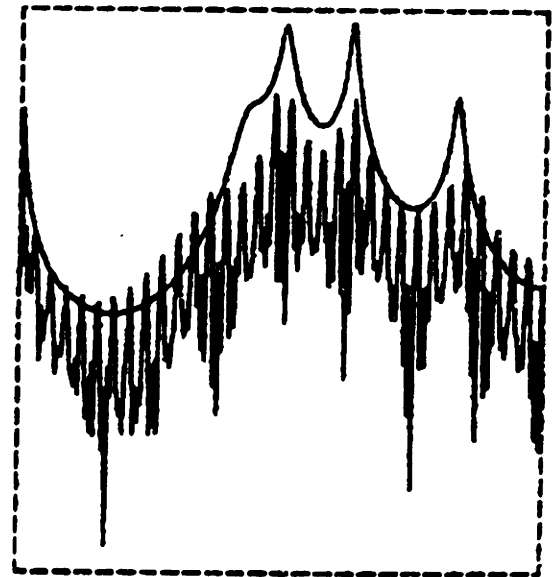
(a)



(b)

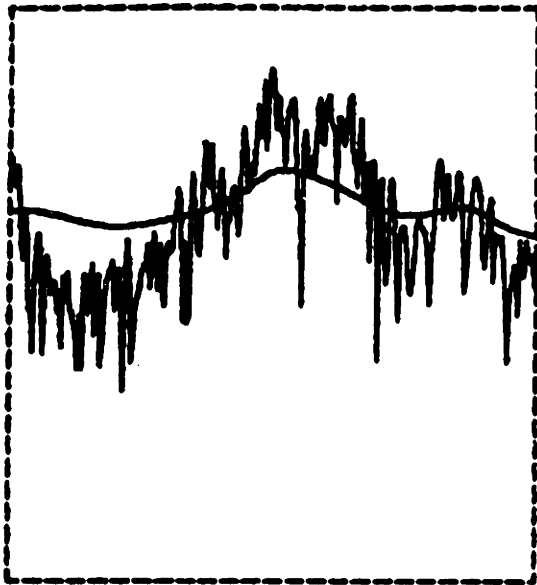


(c)

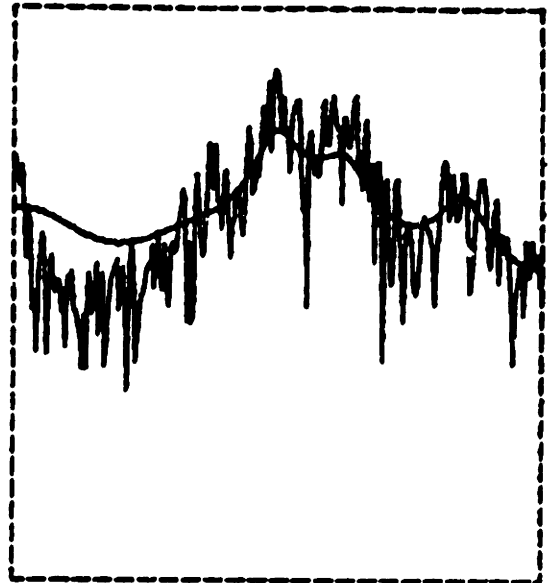


(d)

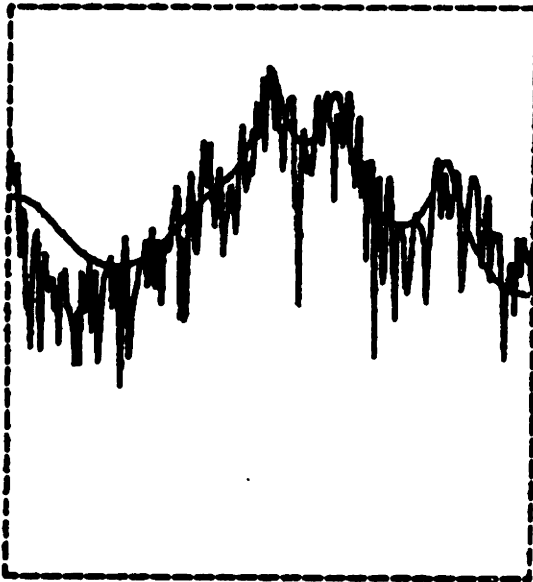
Figure 7.6 Same as Figure 7.4 with  $S/N = 10$  dB



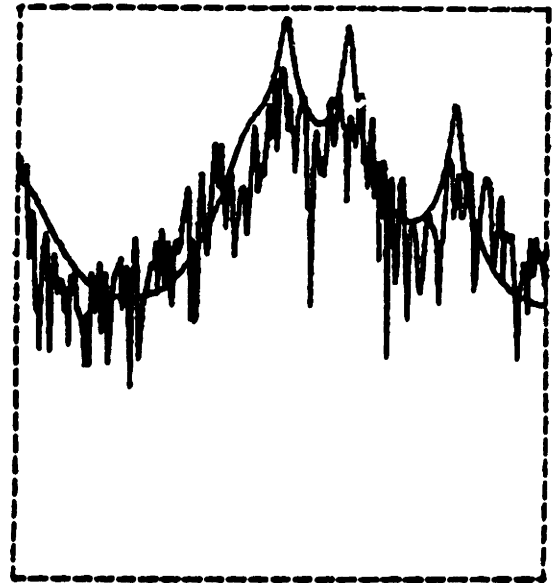
(a)



(b)

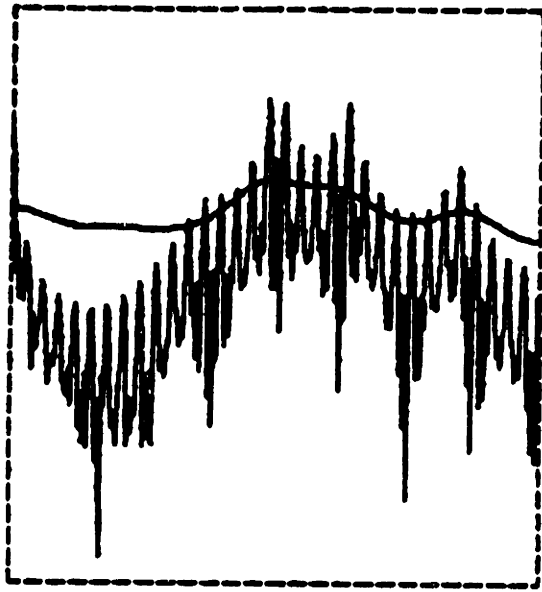


(c)

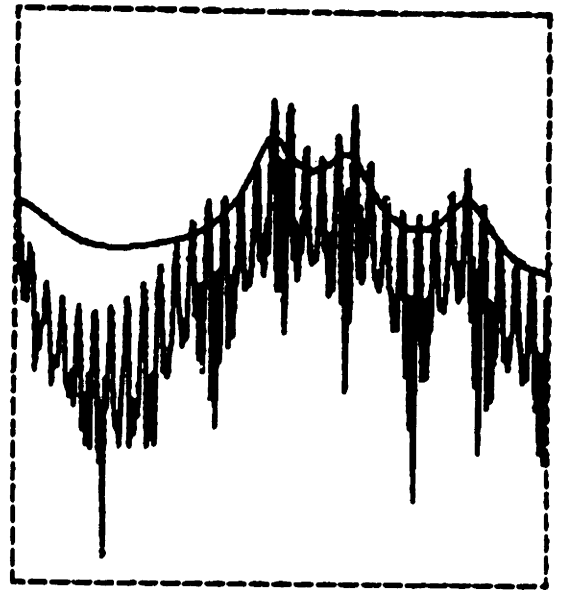


(d)

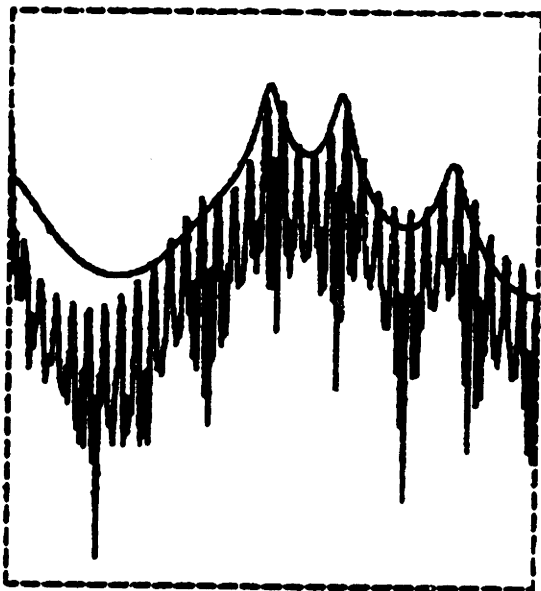
Figure 7.7 Same as Figure 7.3 with  $S/N = 0$  dB



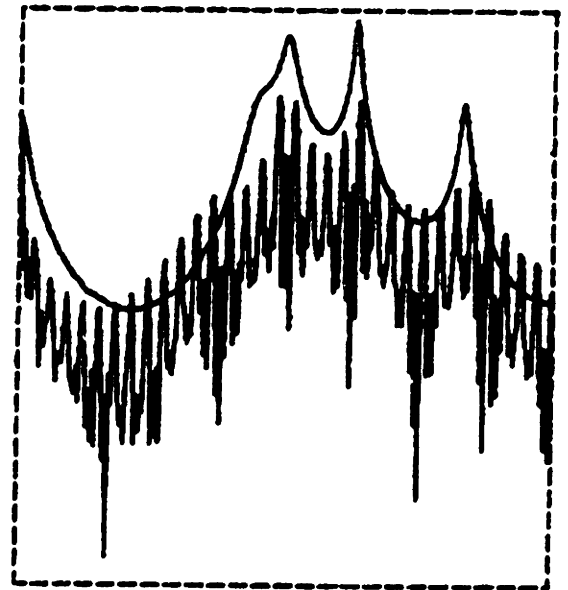
(a)



(b)



(c)



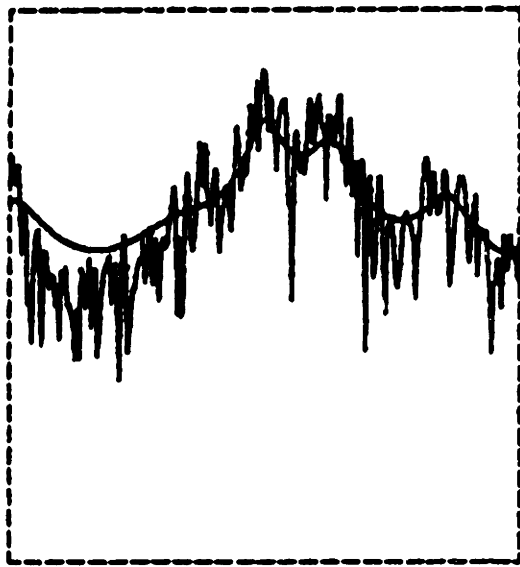
(d)

Figure 7.8 Same as Figure 7.4 with  $S/N = 0$  dB

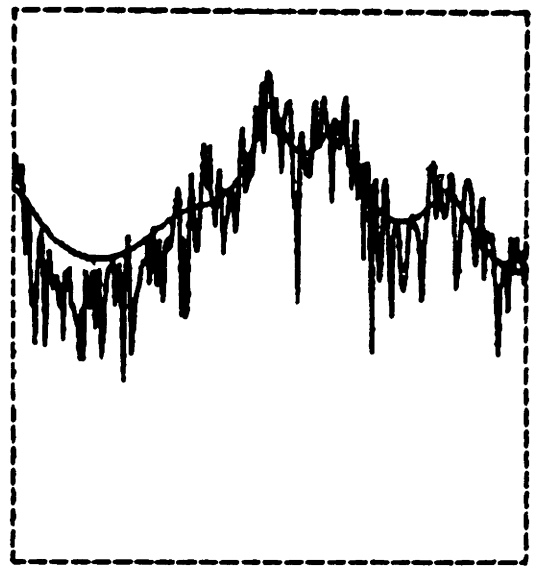
can be observed by the general trend of the estimated transfer functions shown in Figures 7.3 through 7.8 as the number of iterations increases. Thus, in the actual implementation of System A, it seems desirable to limit the number of iterations to two.

### VII.2.2 Application of System B to Synthetic Data

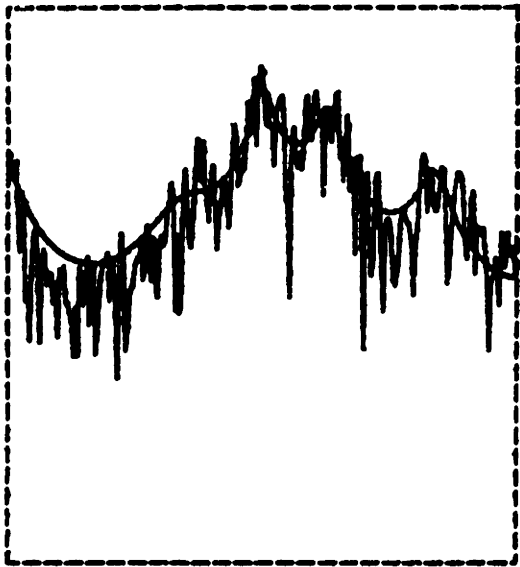
In Figure 7.9 is shown the results of the analysis based on System B as a function of the number of iterations when the S/N ratio is 20 dB and the excitation is random noise. More specifically, in Figure 7.9(a) is shown the all pole fit to the noisy synthetic data by the correlation method of the linear prediction analysis with the assumption that  $s_w(n) = y_w(n)$ , i.e. zeroth iteration. Figures 7.9(b), (c) and (d) represent the estimated transfer functions obtained by applying System B to the noisy synthetic data after two, five and ten iterations, respectively. In each of the four figures ((a), (b), (c) and (d)), the true log magnitude spectrum corresponding to the excitation of random noise is also shown to facilitate the comparisons. Figure 7.10 is the same as Figure 7.9 with the difference in that the excitation is a train of pulses. Figures 7.11 and 7.12 are the same as Figures 7.9 and 7.10 with the difference in that the S/N ratio is 10 dB. Figures 7.13 and 7.14 are the same as Figures 7.9 and 7.10 with the difference in that the S/N ratio is 0 dB. Again the



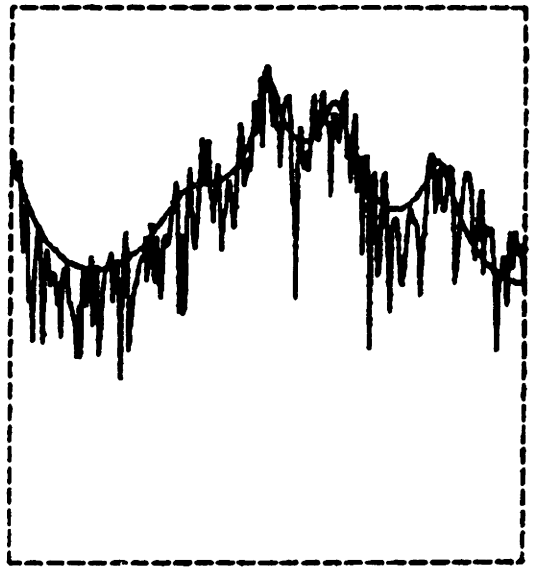
(a)



(b)



(c)



(d)

**Figure 7.9** Comparison of System B

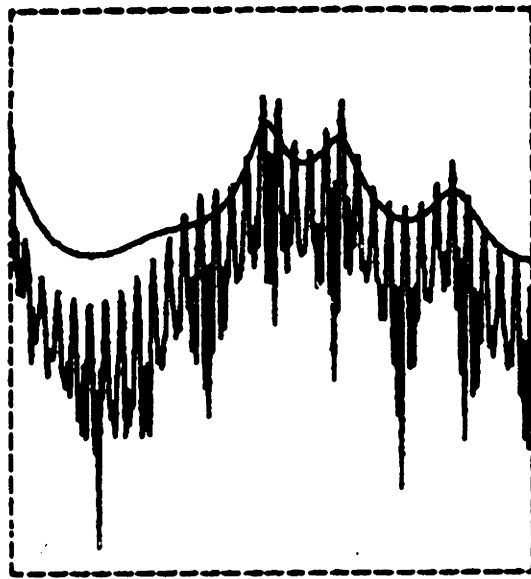
(a) Log magnitude spectrum of the synthetic data in Figure 7.1(a) (random noise excitation) and an all pole fit to the noisy data spectrum after the zeroth iteration of System B at  $S/N = 20$  dB;

(b) Same as (a) after the second iteration of System B;

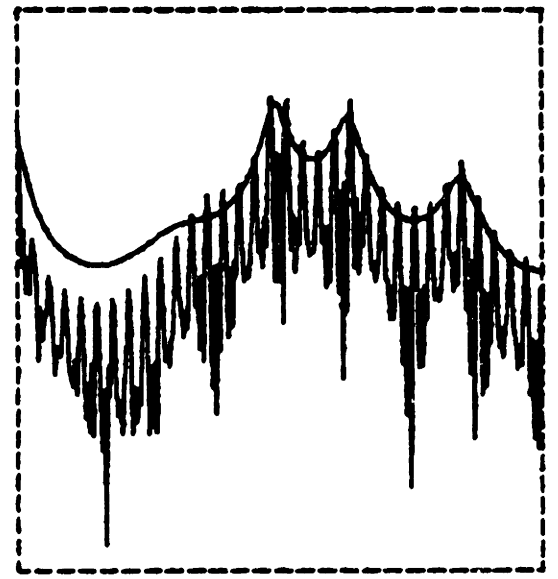
(c) Same as (a) after the fifth iteration of System B;

(d) Same as (a) after the tenth iteration of System B

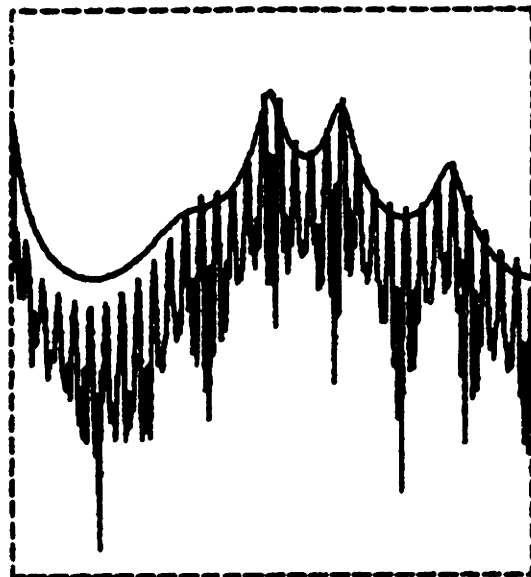




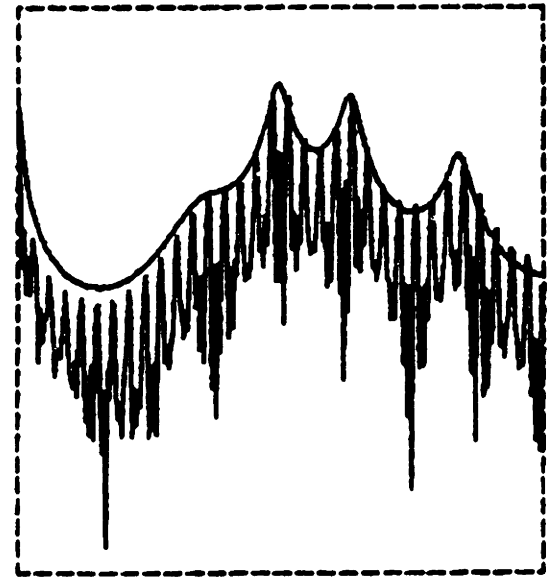
(a)



(b)

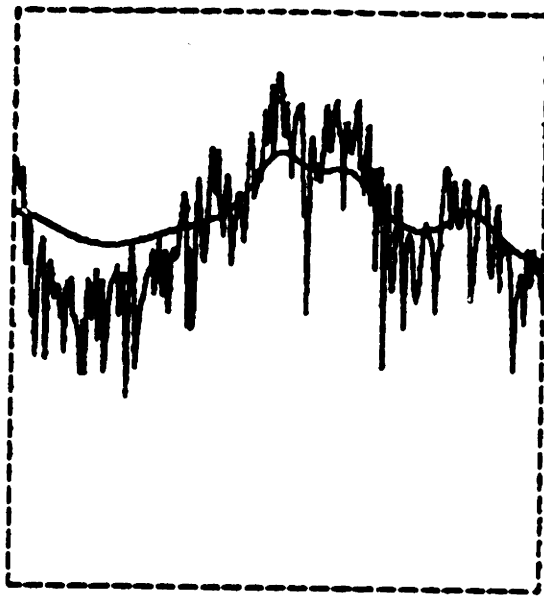


(c)

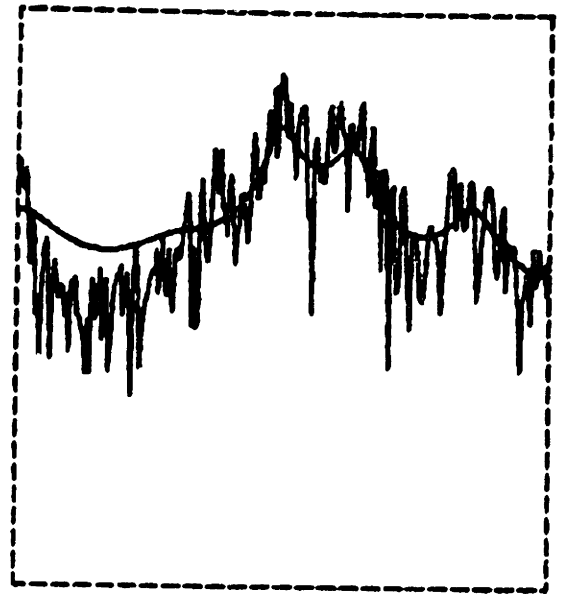


(d)

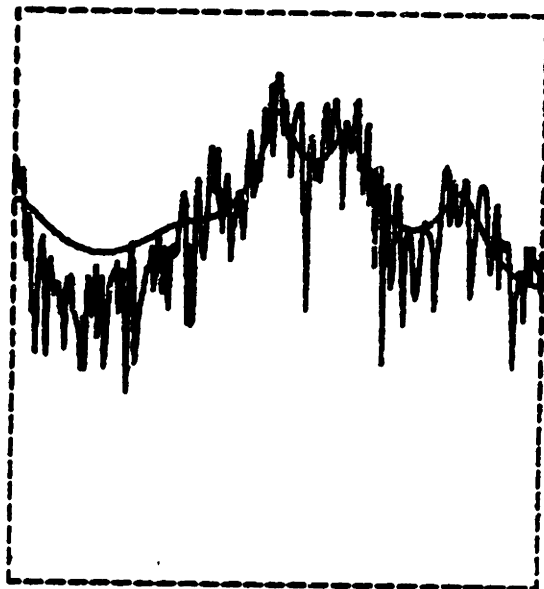
**Figure 7.10** Same as Figure 7.9 with the synthetic data of Figure 7.1(c) (pulse train excitation)



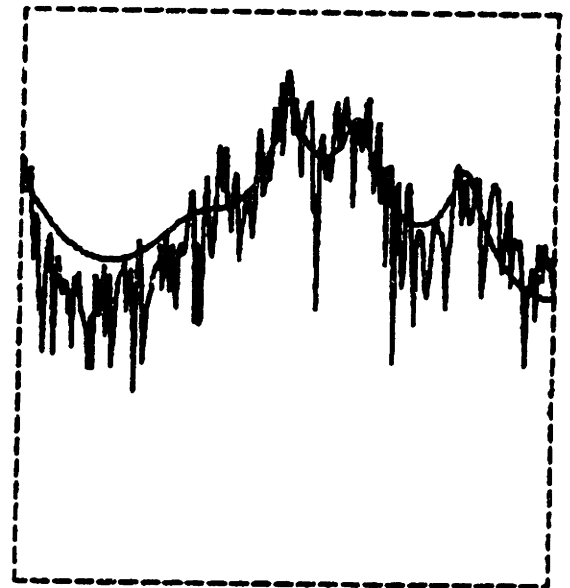
(a)



(b)

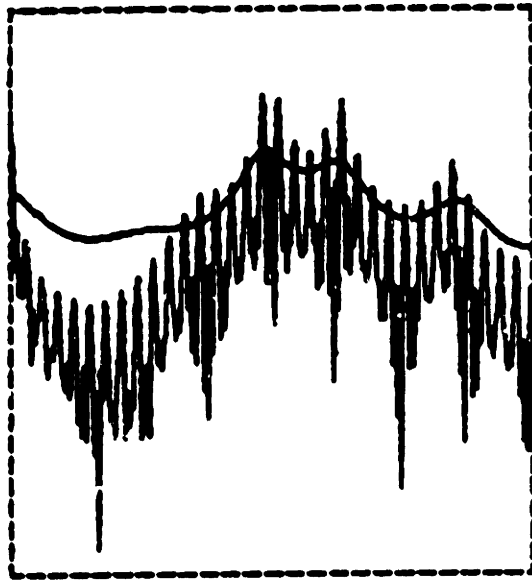


(c)

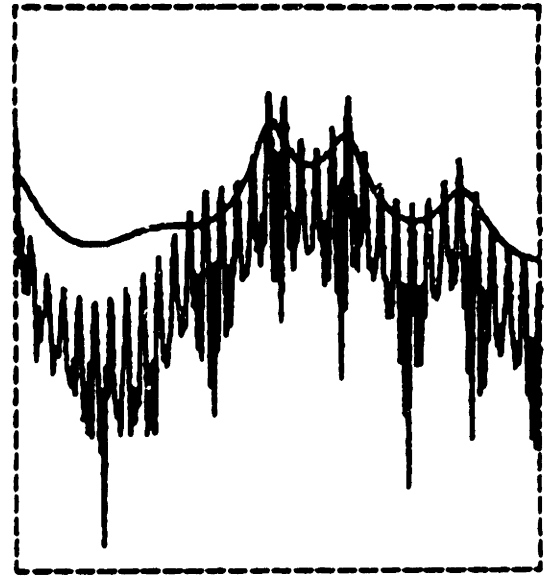


(d)

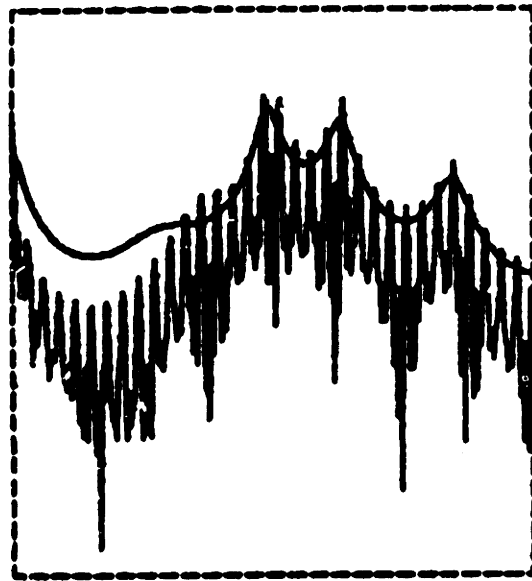
Figure 7.11 Same as Figure 7.9 with  $S/N = 10$  dB



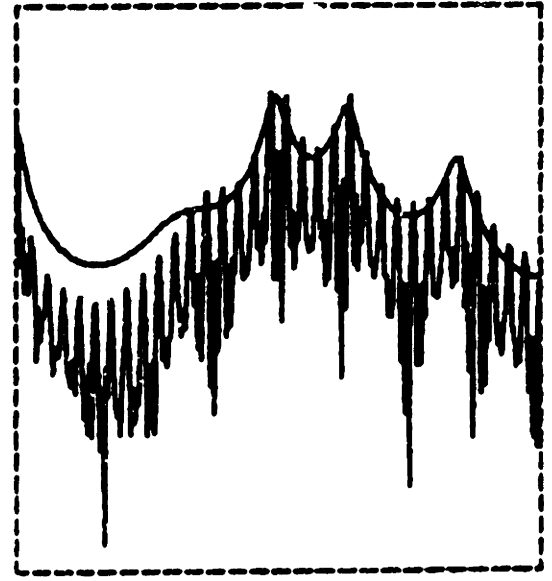
(a)



(b)

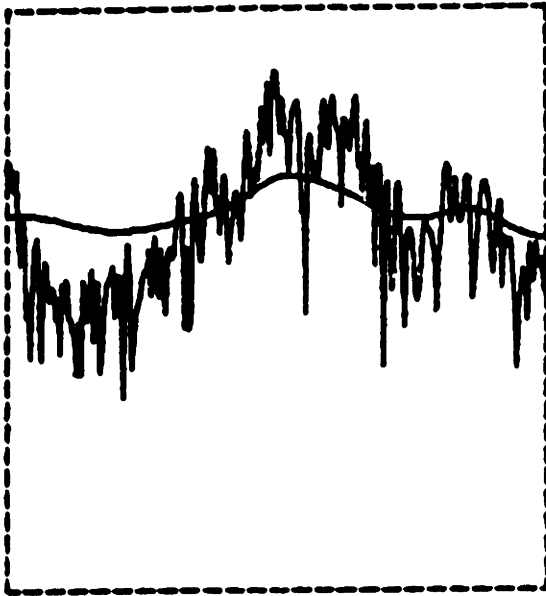


(c)

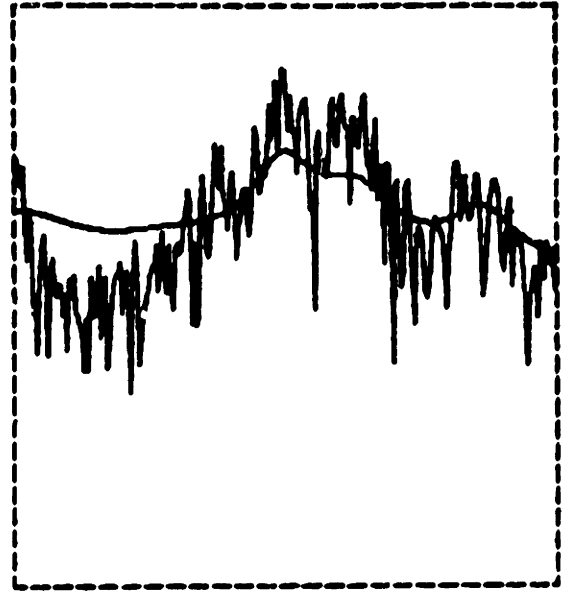


(d)

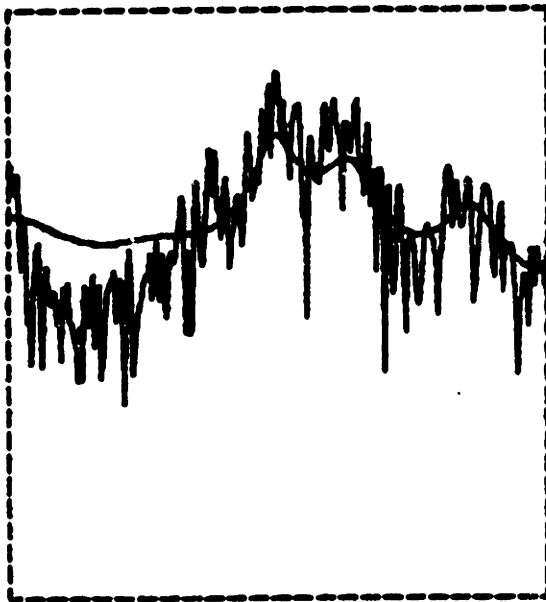
Figure 7.12 Same as Figure 7.10 with  $S/N = 10$  dB



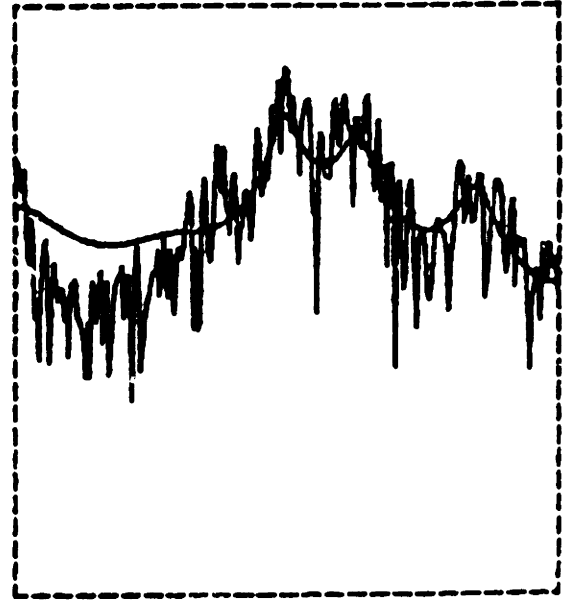
(a)



(b)

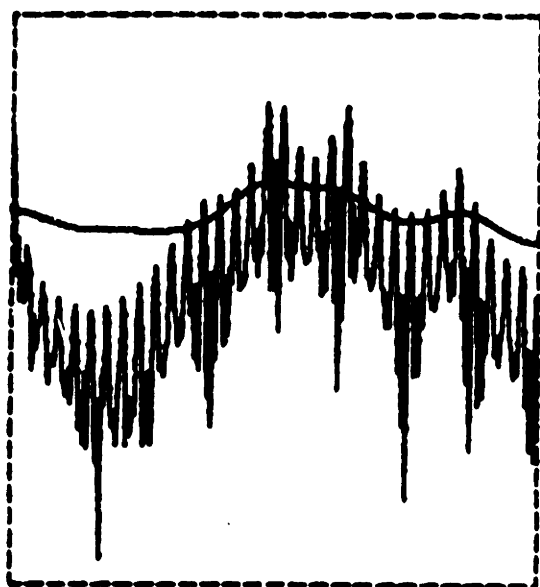


(c)

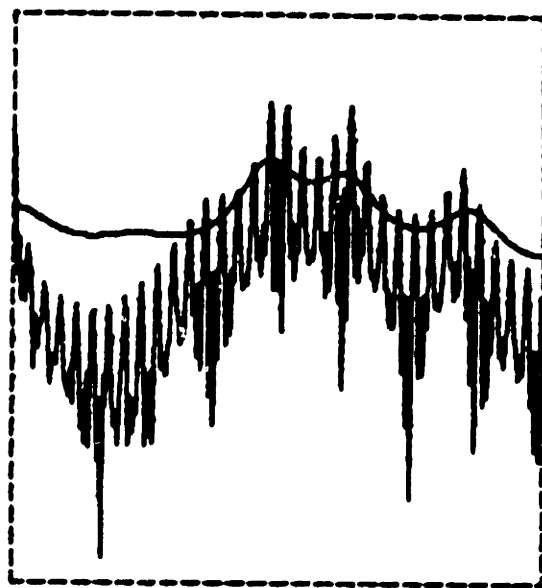


(d)

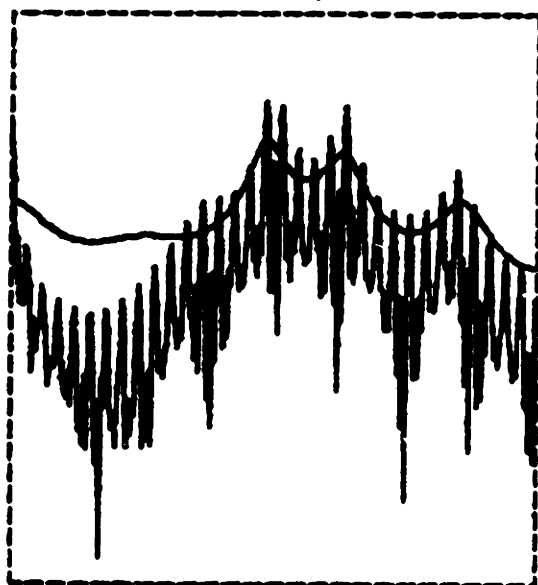
Figure 7.13 Same as Figure 7.9 with  $S/N = 0$  dB



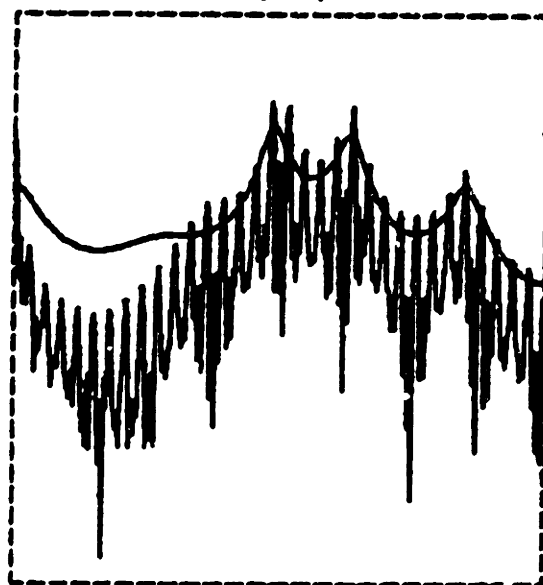
(a)



(b)



(c)



(d)

Figure 7.14 Same as Figure 7.10 with  $S/N = 0$  dB

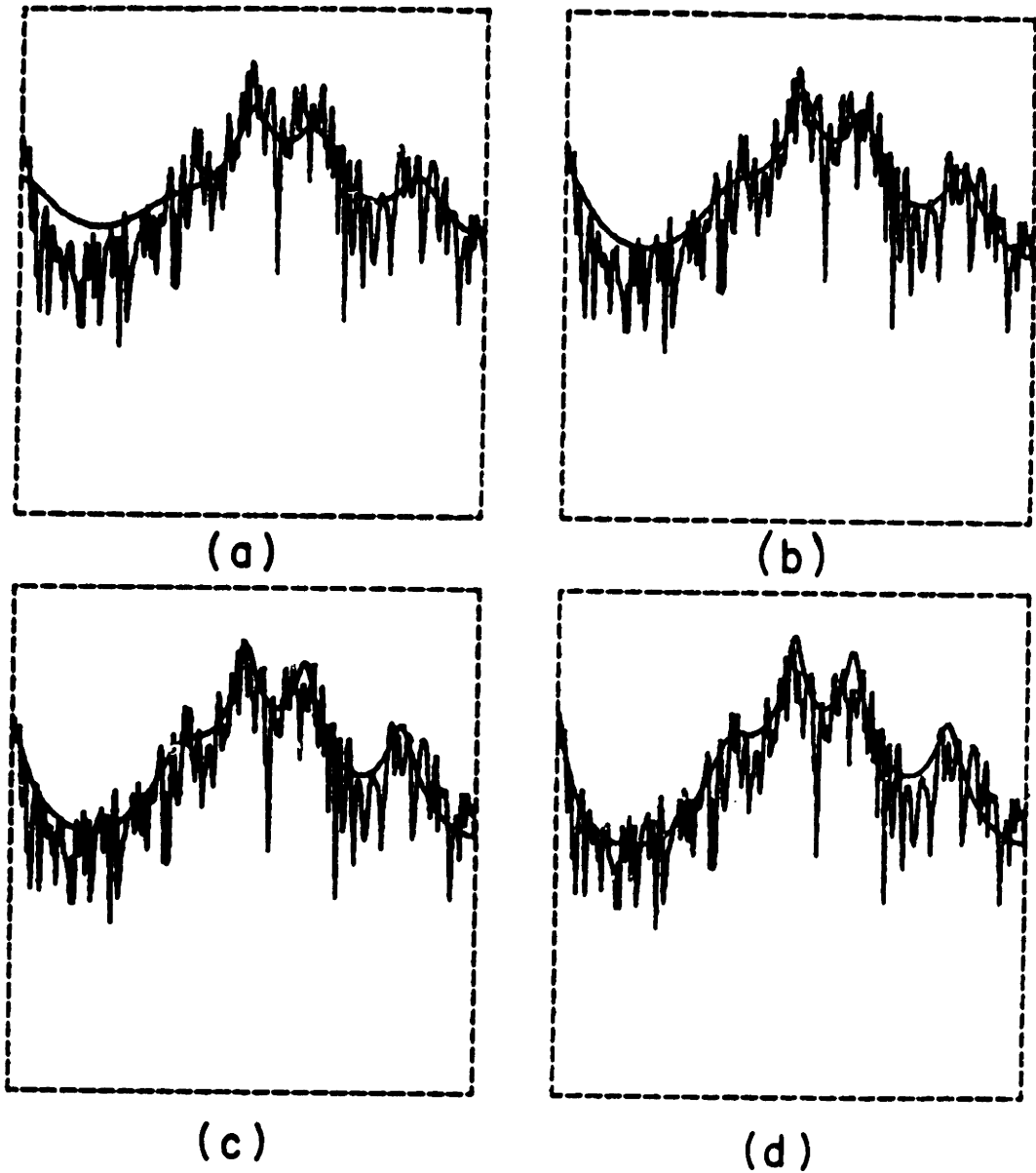
analysis used in Figures 7.9 through 7.14 is based on the assumption that no a priori knowledge of the coefficients is available.

From the figures, it can be observed that for the S/N ratios considered, a good fit to the true spectrum can be obtained after five or more iterations of System B. It can also be observed that the performance of the system is similar to both cases of excitation, i.e. random noise and a train of pulses even though the system was developed based on the assumption of the random noise excitation.

It is not theoretically known if System B converges to a solution. In all the synthetic and real speech data that have been considered, however, it has been observed that System B appears to converge and the estimate after many iterations in general fits better to the true log magnitude spectrum than the estimate obtained after a few iterations. It has also been observed that the results after ten iterations correspond reasonably well to the final estimate.

### VII.2.3 Application of System C to Synthetic Data

In Figure 7.15 is shown the results of the analysis based on System C as a function of the scaling constant "k", a parameter of System C, when the S/N ratio is 20 dB and the excitation is random noise. More specifically,



**Figure 7.15** Comparison of System C

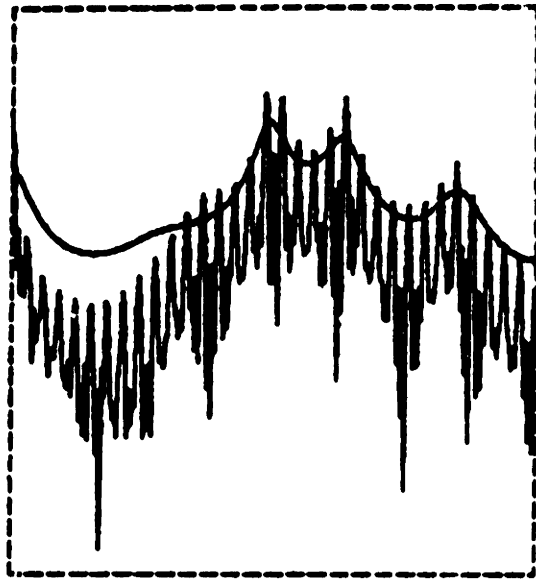
- (a) Log magnitude spectrum of the synthetic data in Figure 7.1(a) (random noise excitation) and an all pole fit to the noisy synthetic data with  $k=0$  of System C at  $S/N = 20$  dB;
- (b) Same as (a) with  $k=1$  of System C;
- (c) Same as (a) with  $k=2$  of System C;
- (d) Same as (a) with  $k=3$  of System C

in Figure 7.15(a) is shown the all pole fit to the log magnitude spectrum of the noisy synthetic data by the correlation method of the linear prediction analysis with the assumption that  $s_w(n)=y_w(n)$ . Figures 7.15(b), (c) and (d) represent the estimated transfer functions obtained by applying System C to the noisy synthetic data at  $k=1, 2$ , and  $3$  respectively. In each of the four figures ((a), (b), (c), (d)), the true log magnitude spectrum corresponding to the excitation of random noise is shown to facilitate the comparisons. Figure 7.16 is the same as Figure 7.15 with the difference in that the excitation is a train of pulses. Figures 7.17 and 7.18 are the same as Figures 7.15 and 7.16 with the difference in that the S/N ratio is 10 dB. Figures 7.19 and 7.20 are the same as Figures 7.15 and 7.16 with the difference in that the S/N ratio is 0 dB. In all the Figures 7.15 through 7.20, the analysis is based on the assumption that no a priori information of the coefficient vector is available.

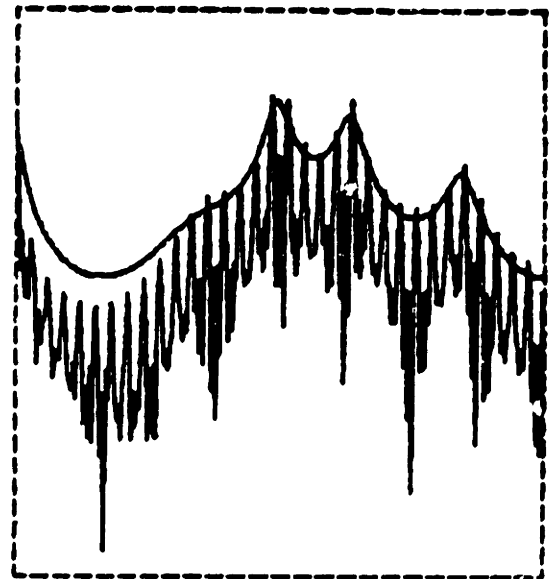
From the figures, it can be observed that for the S/N ratios considered a good fit to the true log magnitude spectrum can be obtained when  $k=2$  in System C. It is also observed that the performance of the system is similar in both cases of excitation, i.e. random noise and a train of pulses.

When  $k$  equals zero, System C corresponds to the correlation method of the linear prediction analysis that does

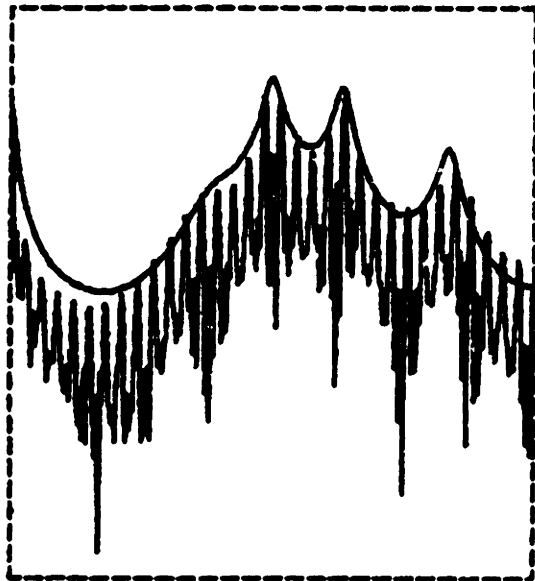




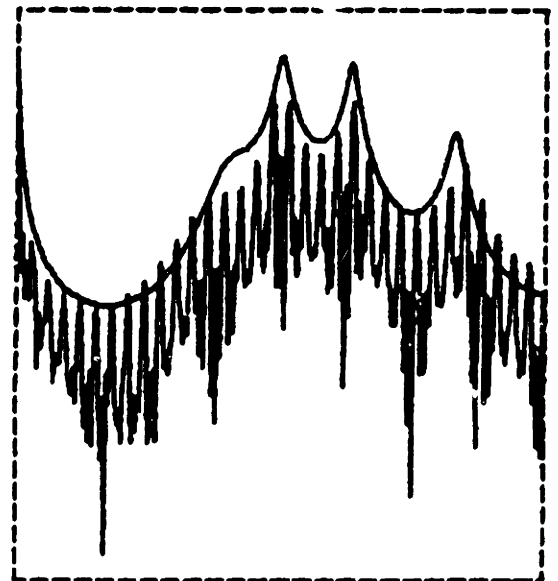
(a)



(b)

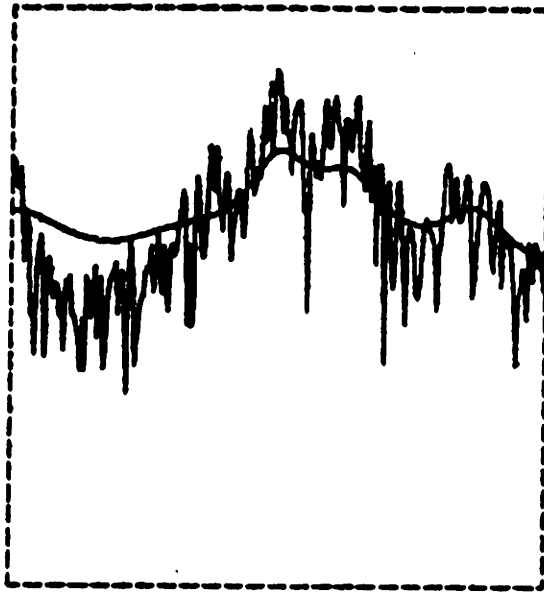


(c)

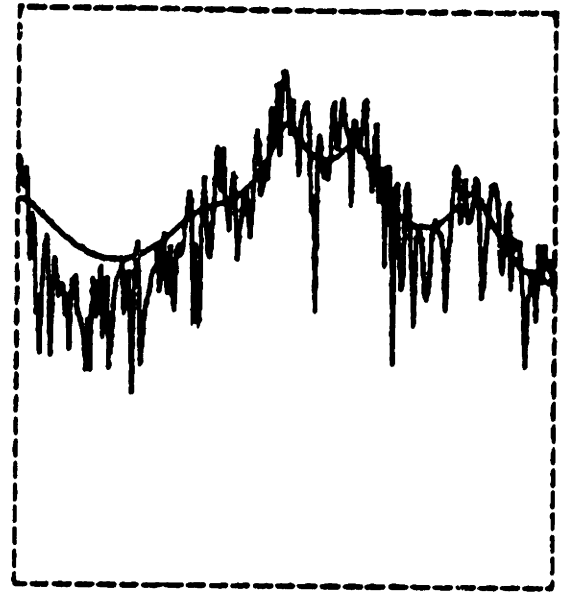


(d)

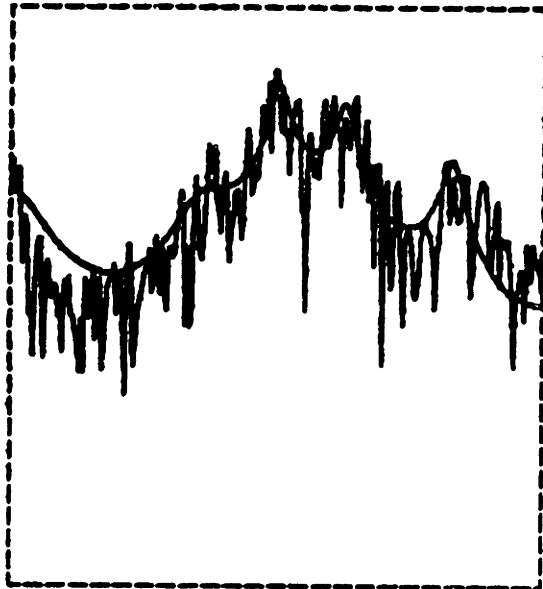
**Figure 7.16** Same as Figure 7.15 with the synthetic data of Figure 7.1(c) (pulse train excitation)



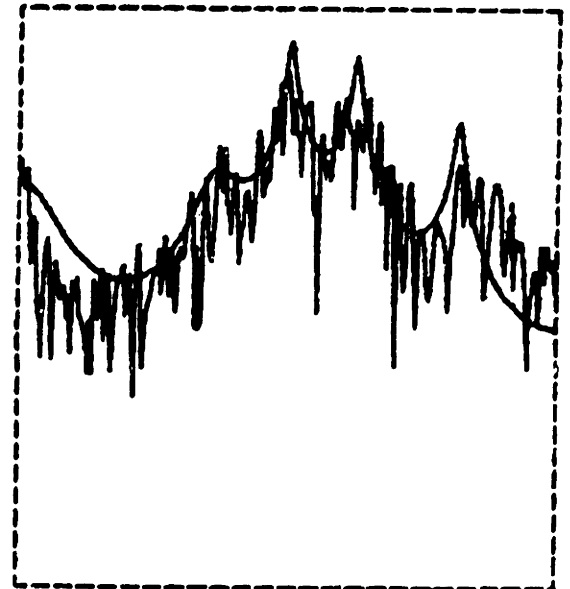
(a)



(b)

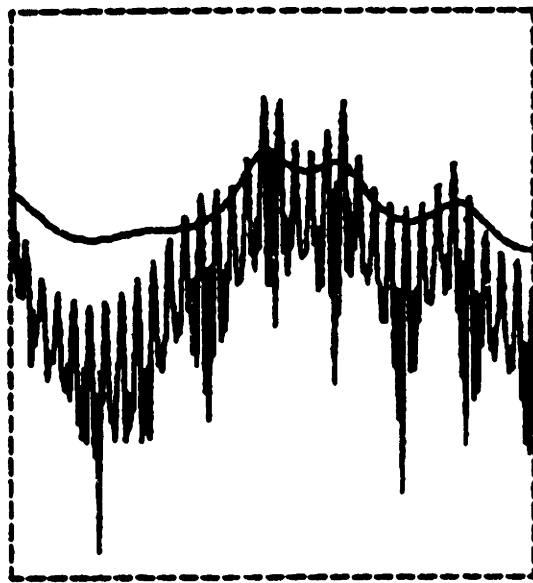


(c)

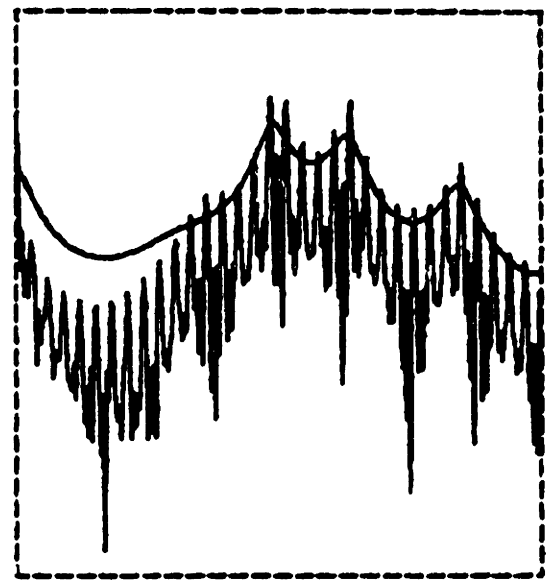


(d)

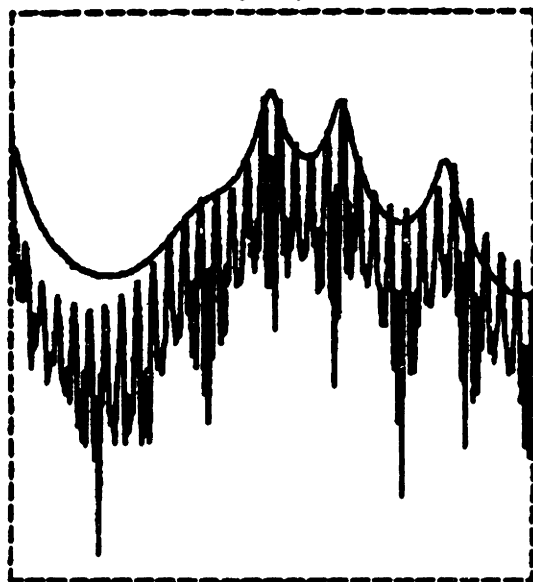
Figure 7.17 Same as Figure 7.15 with  $S/N = 10$  dB



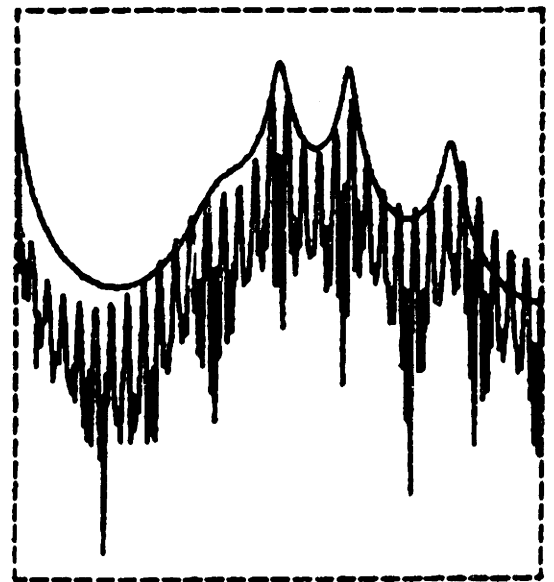
(a)



(b)

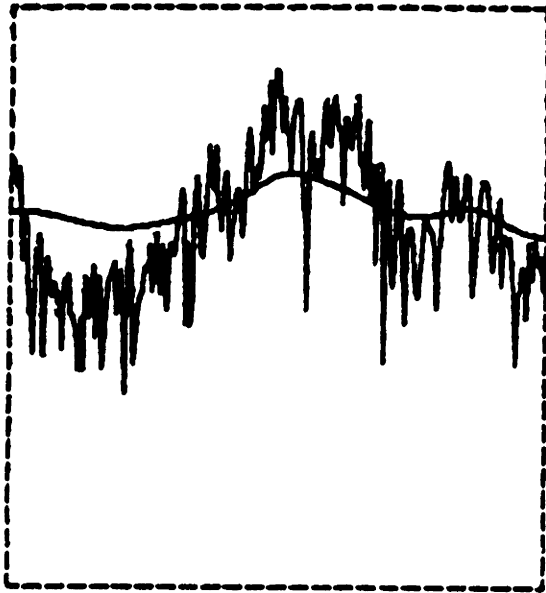


(c)

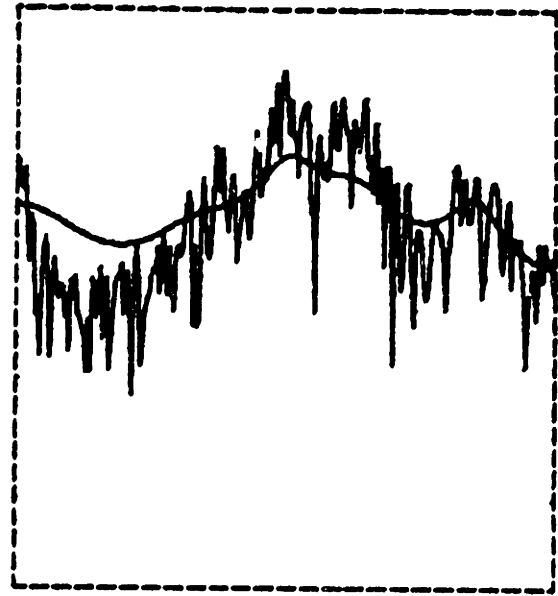


(d)

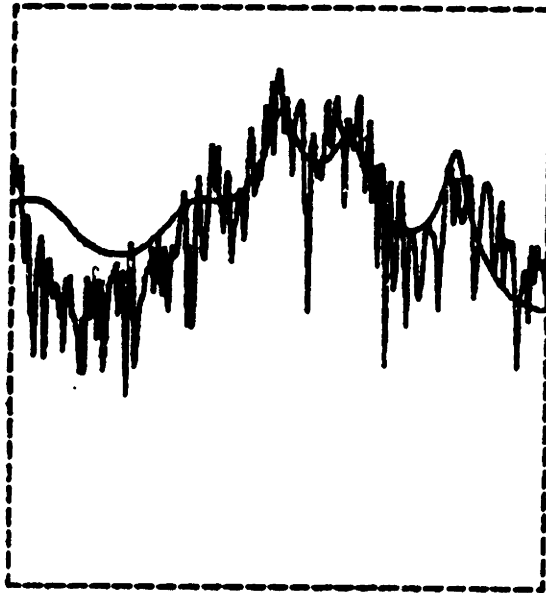
Figure 7.18 Same as Figure 7.16 with  $S/N = 10$  dB



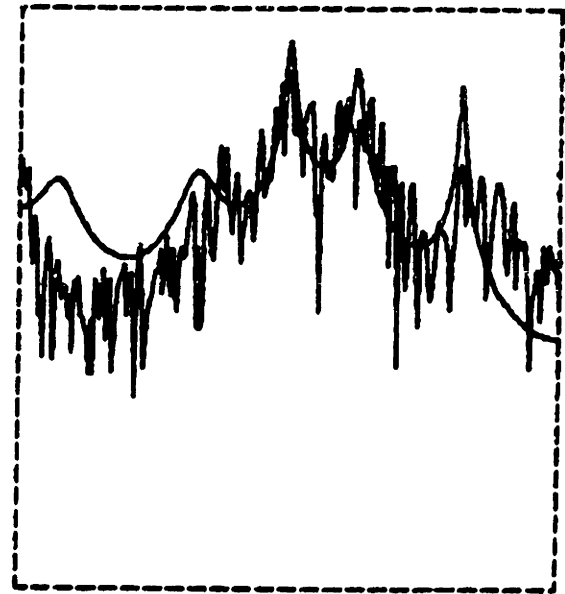
(a)



(b)

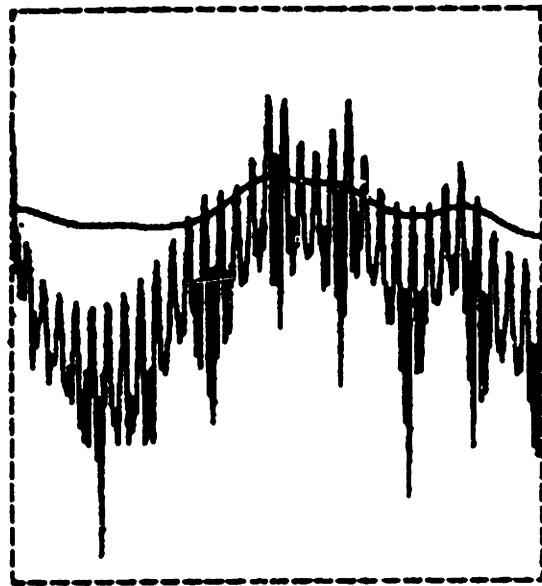


(c)

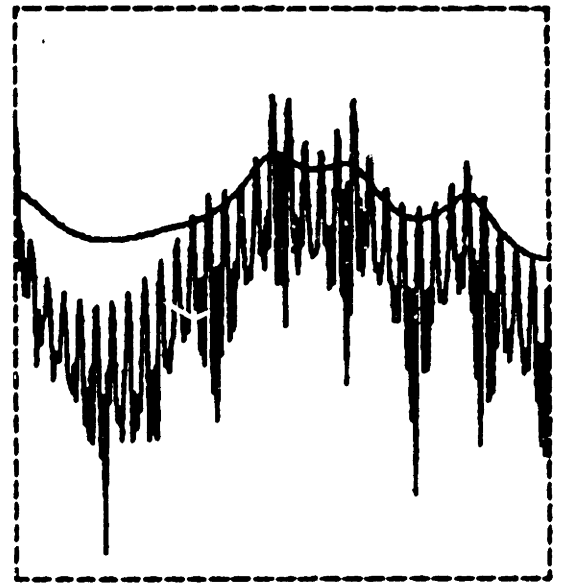


(d)

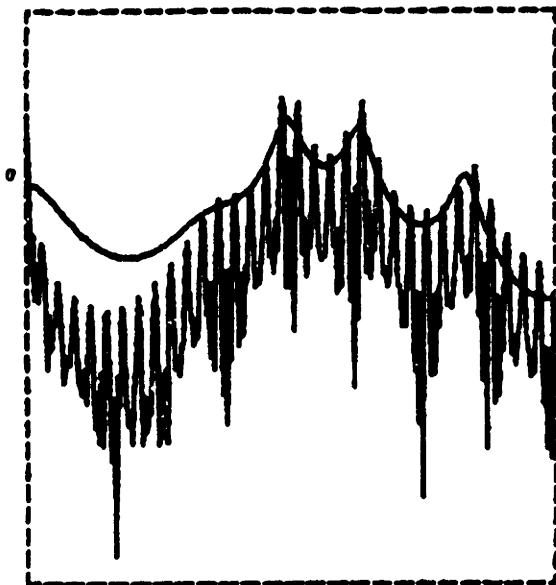
Figure 7.19 Same as Figure 7.15 with  $S/N = 0$  dB



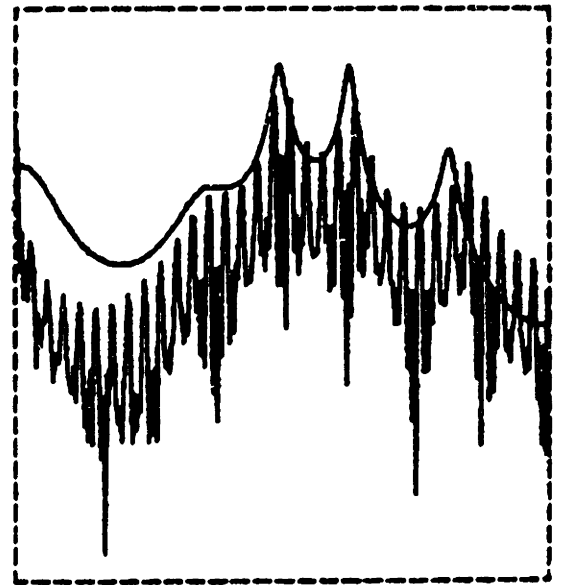
(a)



(b)



(c)



(d)

Figure 7.20 Same as Figure 7.16 with  $S/N = 0$  dB

not account for the presence of background noise. Thus, the estimated transfer functions shown in (a) of Figures 7.15 through 7.20, correspond to the case when  $k$  equals zero. From many examples of synthetic data, it has been observed that the performance of System C in terms of the log magnitude spectrum fit is poor when  $k$  is greater than 3. It has also been observed that the log magnitude spectrum fit at  $k=2$  is generally better than the fit when  $k=1$  which corresponds to the correlation subtraction method.

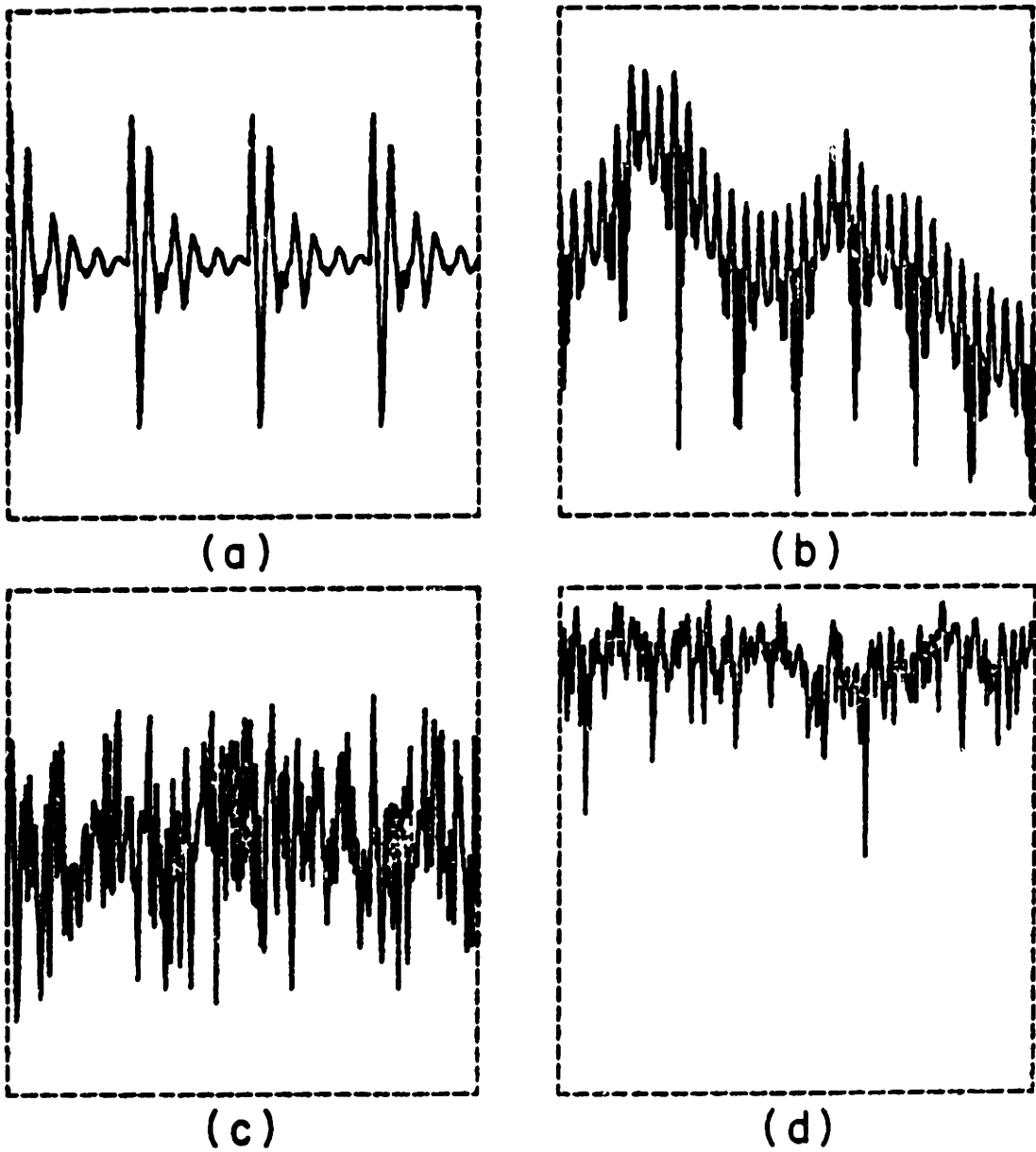
In the specific example of the synthetic data that has been considered in Sections VII.2.1, VII.2.2 and VII.2.3, a reasonably good fit to the log magnitude spectrum can be obtained by any of the three systems with a proper choice of the system parameter (i.e. the number of iterations for Systems A and B, and the value of  $k$  for System C). However, when the noisy data have no spectral peaks or spectral peaks that are different from the pole locations of the original data, then the application of the three systems can result in the estimated transfer functions whose pole frequencies are different from those of the original data. This situation can occur when the overall S/N ratio is sufficiently low in which case all the pole frequencies can be affected, or when

the local<sup>10</sup> S/N ratios near some pole locations are sufficiently low in which case the local poles can be affected. In Figures 7.21 through 7.24 are illustrated two such examples. In Figures 7.21(a) and (b) are shown an example of a segment of the synthetic data and its log magnitude spectrum. In Figures 7.21(c) and (d) are illustrated the noisy synthetic data at the S/N ratio of -20 dB and its log magnitude spectrum. In Figure 7.22(a) is illustrated the transfer function estimated from the noisy synthetic data in Figure 7.21(c) by the correlation method of the linear prediction analysis. In Figures 7.22(b), (c) and (d) are shown the transfer functions estimated by System A after two iterations, System B after ten iterations and System C with k=2. In each of the four figures of Figure 7.22, the true log magnitude spectrum of Figure 7.21(b) is also illustrated to facilitate the comparisons. From Figure 7.22, it is clear that the transfer function generated by any of the three systems does not fit the true spectrum well. Figures 7.23 and 7.24 are equivalent to Figures 7.21 and 7.22 with the

---

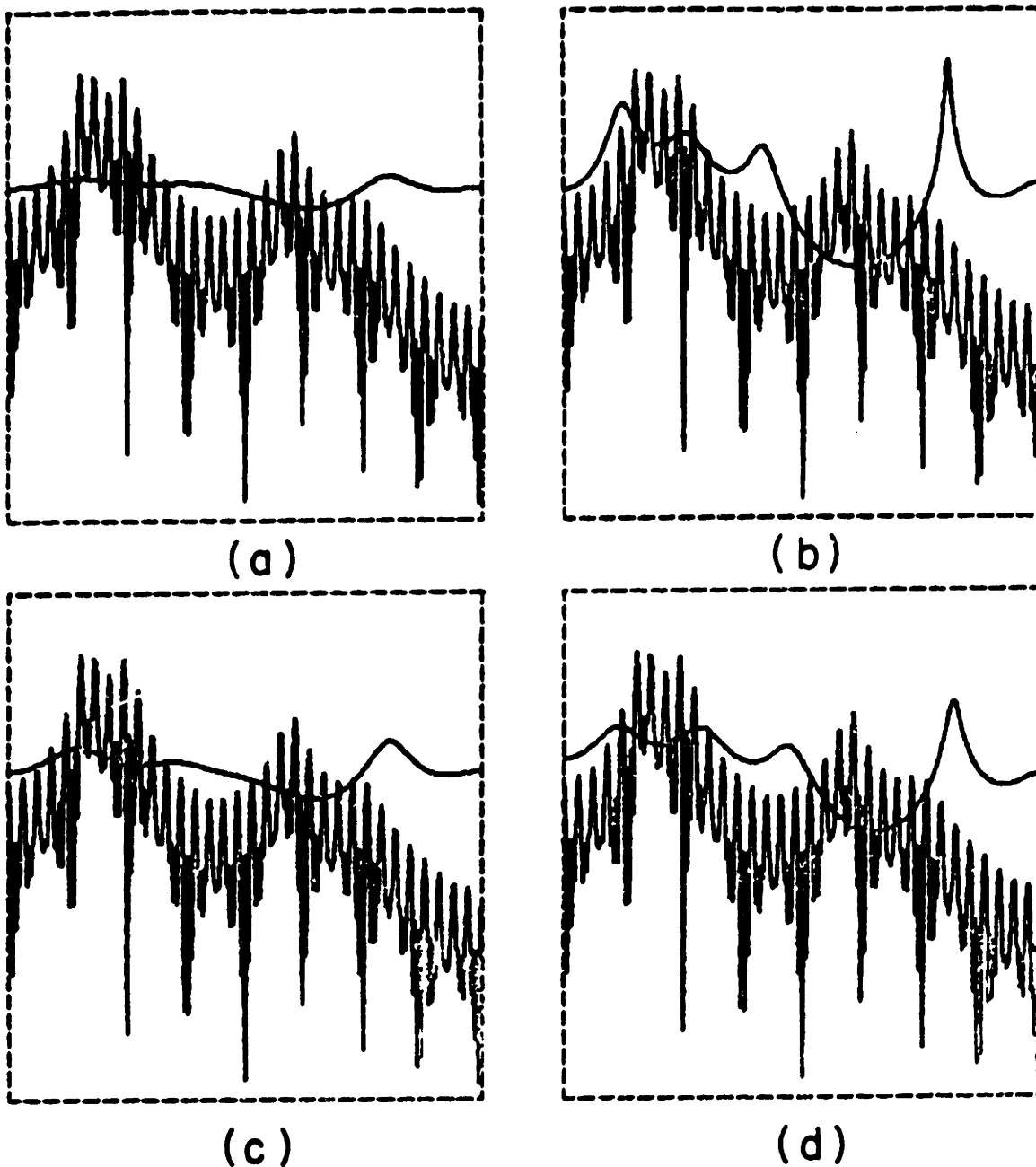
<sup>10</sup>The local S/N ratio between two angular frequencies  $\omega_1$  and  $\omega_2$  is defined by

$$\text{Local S/N ratio in dB} = 10 \cdot \log \frac{\int_{\omega_1}^{\omega_2} |S(\omega)|^2 \cdot d\omega}{\int_{\omega_1}^{\omega_2} |D(\omega)|^2 \cdot d\omega}$$



**Figure 7.21** (a) A synthetic data segment;  
(b) Log magnitude spectrum of the synthetic data in (a);  
(c) Noisy synthetic data of (a) at  $S/N = -20$  dB;  
(d) Log magnitude spectrum of the noisy synthetic data in (c)



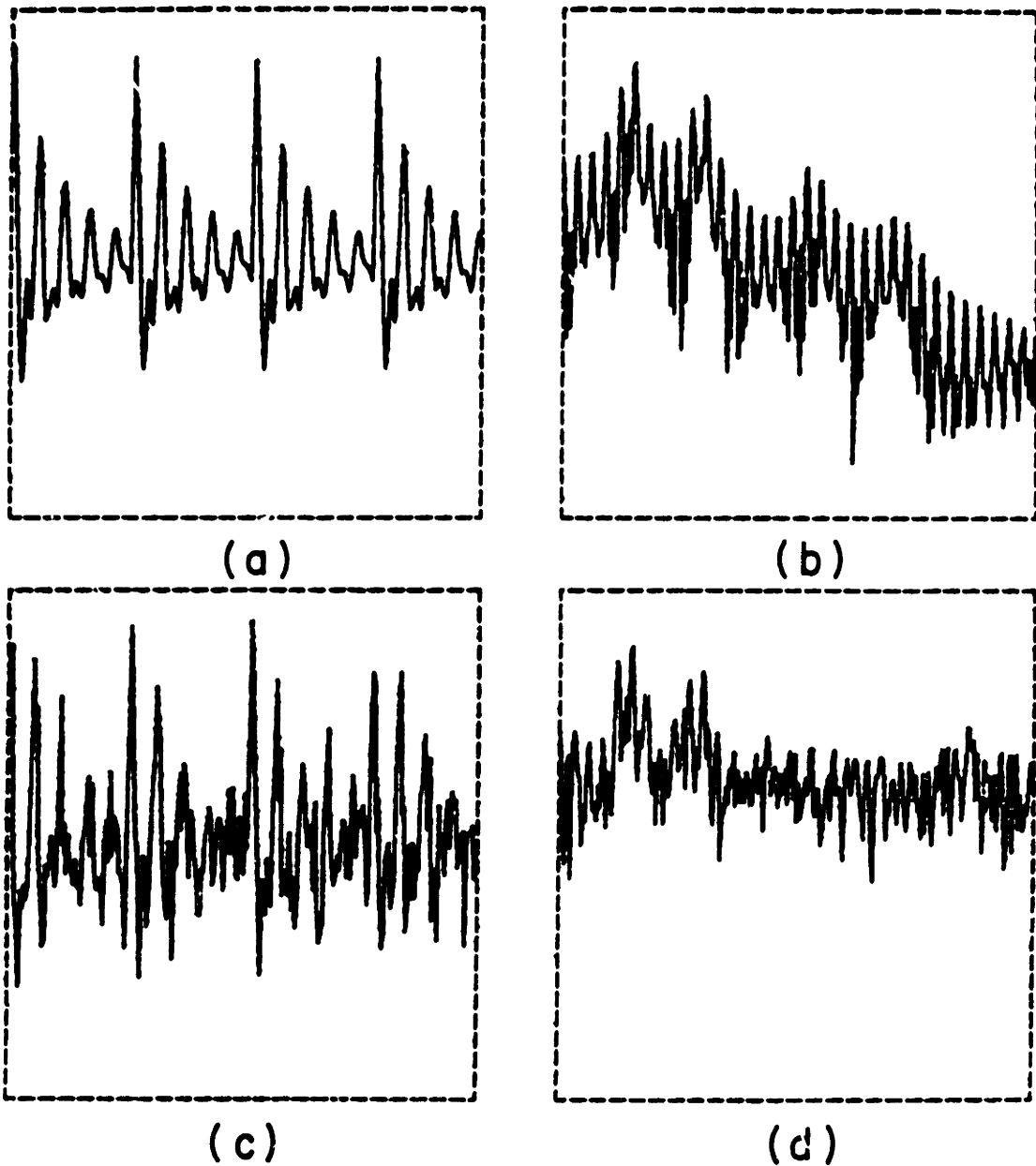


**Figure 7.22** (a) Log magnitude spectrum of the synthetic data in Figure 7.21(a) and an all pole fit to the noisy synthetic data with  $k=0$  of System C;

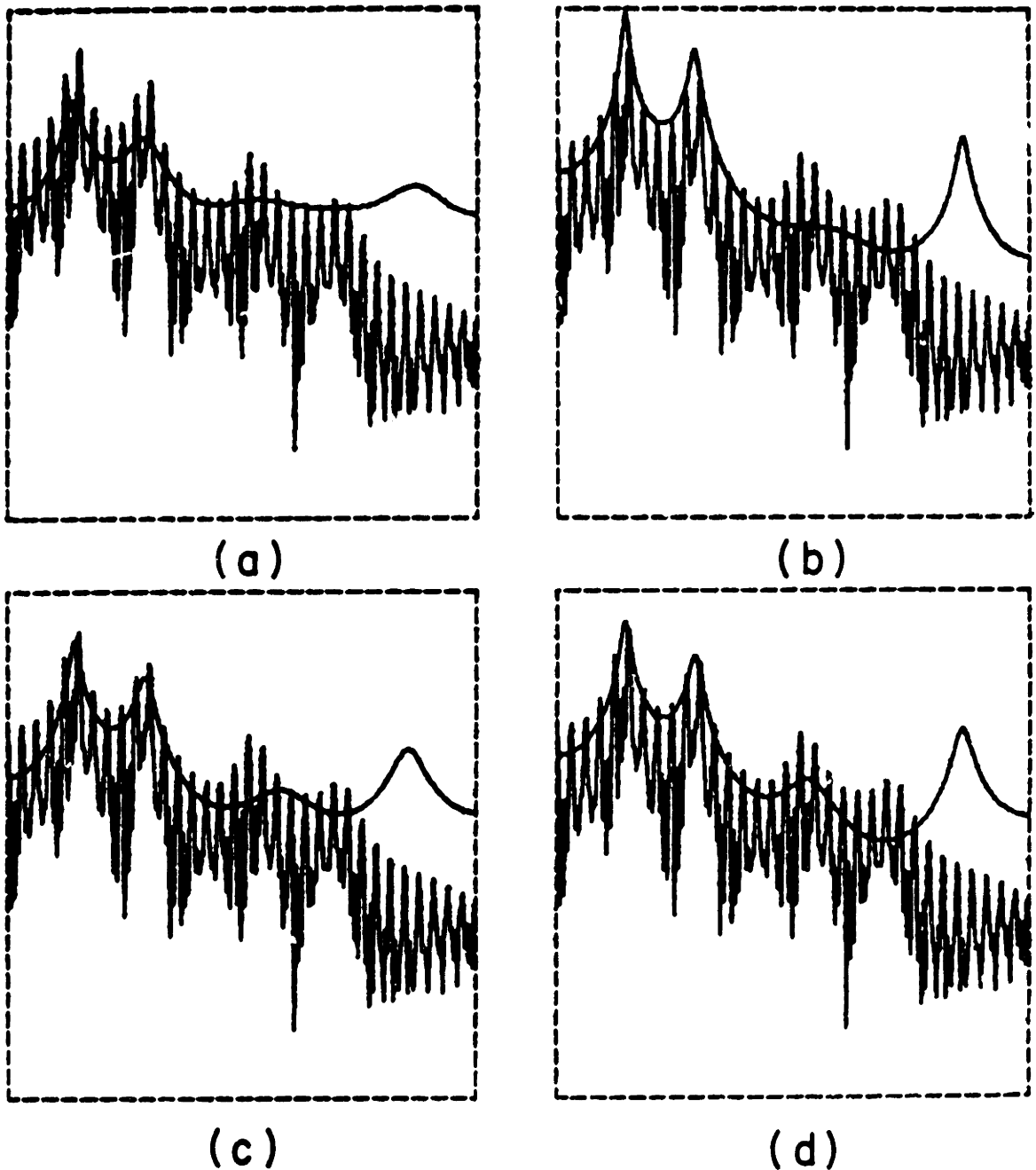
(b) Same as (a) with two iterations of System A;

(c) Same as (a) with ten iterations of System B;

(d) Same as (a) with  $k=2$  of System C



**Figure 7.23** (a) A synthetic data segment;  
(b) Log magnitude spectrum of the synthetic data in (a);  
(c) Noisy synthetic data of (a) at  $S/N = 10$  dB;  
(d) Log magnitude spectrum of the noisy synthetic data in (c)



**Figure 7.24** (a) Log magnitude spectrum of the synthetic data in Figure 7.23(a) and an all pole fit to the noisy synthetic data with  $k=0$  of System C;  
(b) Same as (a) with two iterations of System A;  
(c) Same as (a) with ten iterations of System B;  
(d) Same as (a) with  $k=2$  of System C

difference in that a different synthetic data segment is used at the S/N ratio of 10 dB. From Figure 7.24, it is clear that the lower formants where the local S/N ratio is relatively high are well recovered by the three systems but the performance is poor for the higher formants where the S/N ratio is relatively low.

At a high S/N ratio, the types of errors discussed above do not occur frequently. As the S/N ratio decreases, the errors occur more frequently and eventually a point is reached at which the systems are no longer useful for the analysis of noisy speech. In Chapter VIII, this issue will be discussed in greater detail as the performance of the three systems is evaluated by some objective and subjective tests.

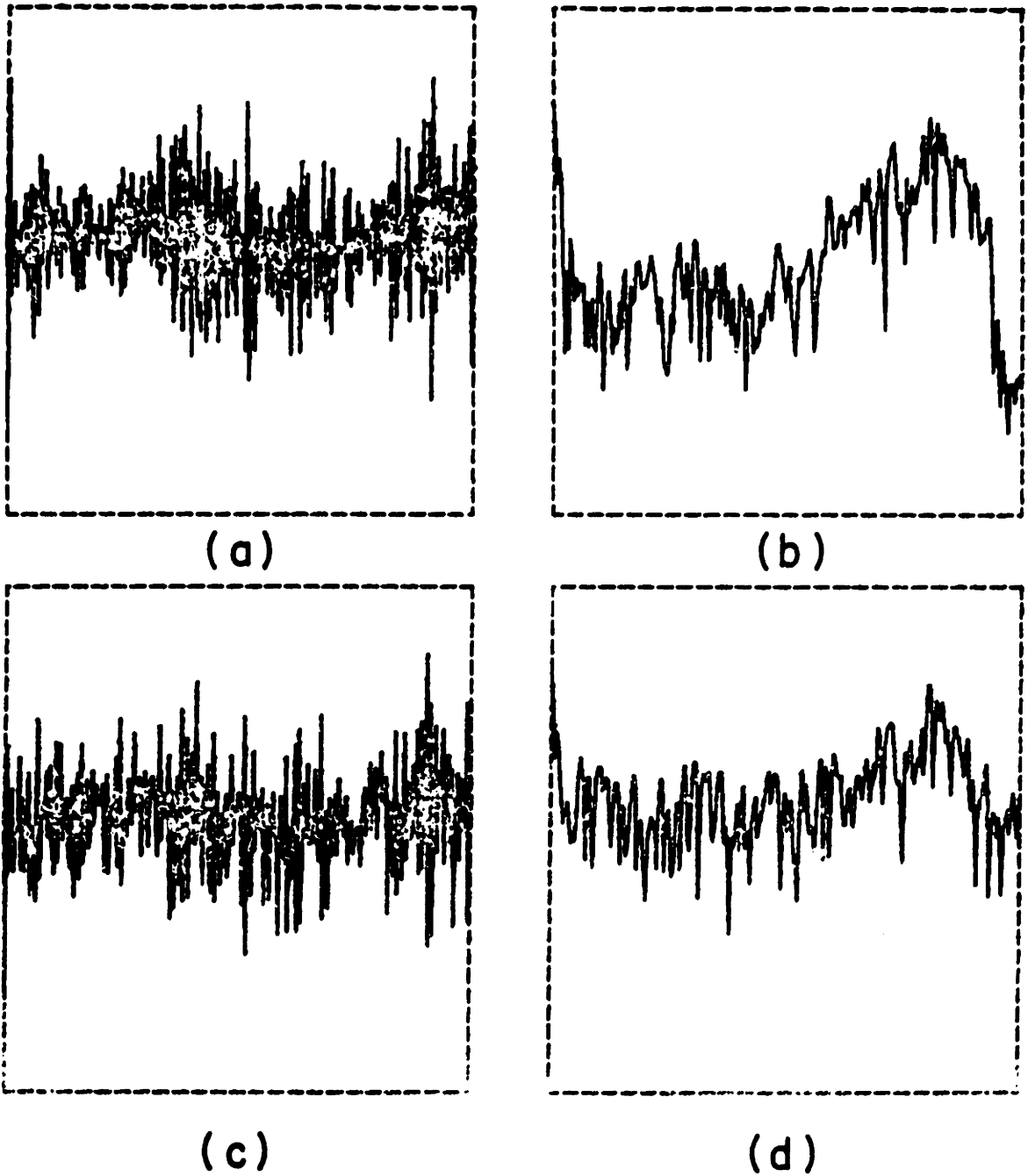
### VII.3 Application to Real Speech Data

A number of discussions in Section VII.2 on the performance of the three systems when applied to the synthetic data in general also apply to the real speech data. Therefore, only two examples of real speech data at the S/N ratio of 10 dB will be illustrated primarily to demonstrate that the performance of the systems when applied to the real speech data is similar to the case of the synthetic data. Again, the real speech data are based on a 10 kHz sampling rate, the order of the all pole model is assumed to be 10, the analysis is based on 256

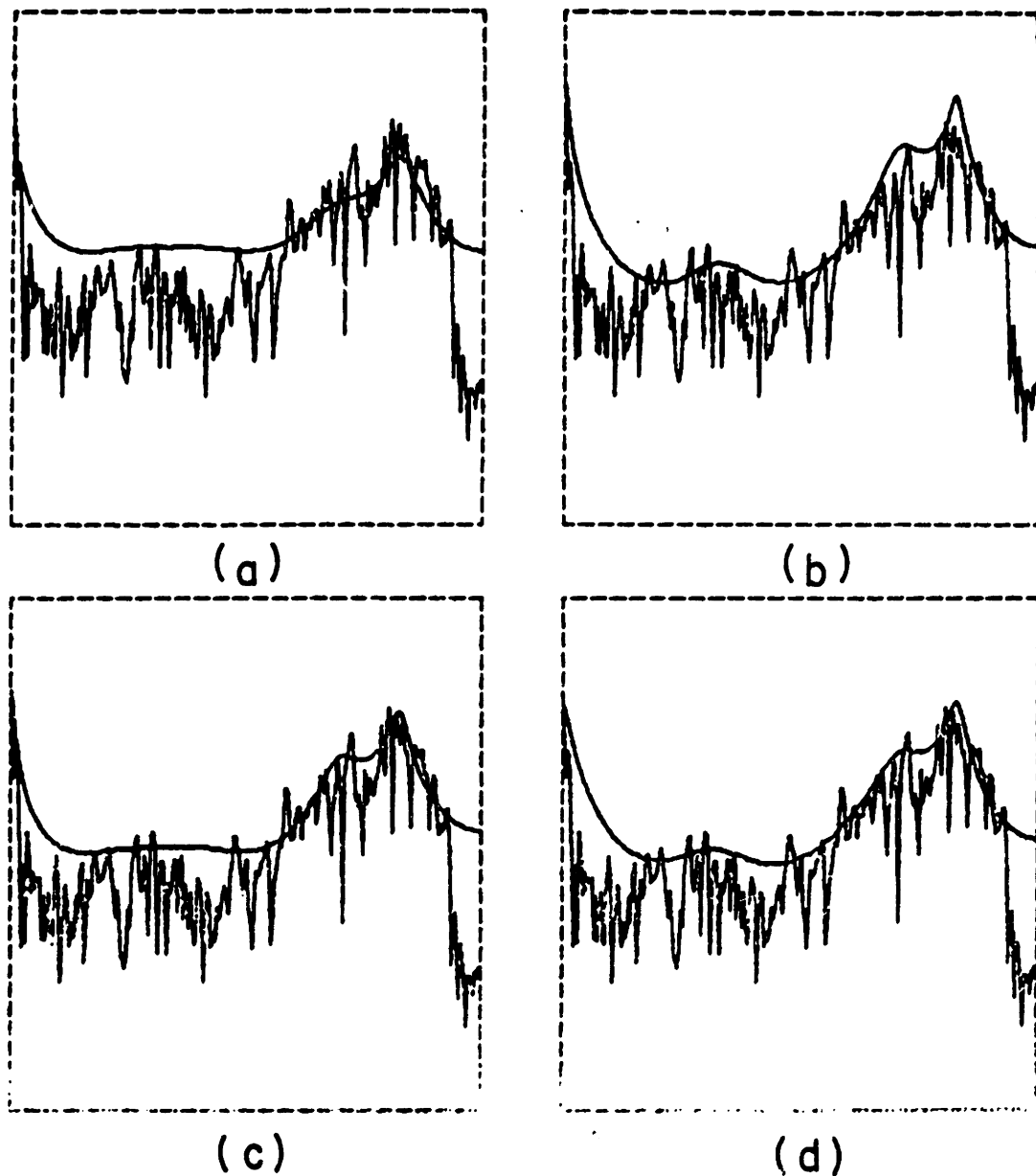
data points, and no a priori information of the all pole coefficient vector is assumed to be available.

In Figures 7.25(a) and (b) are shown an example of a segment of unvoiced speech and its log magnitude spectrum. In Figures 7.25(c) and (d) are illustrated the noisy synthetic data and its log magnitude spectrum. In Figure 7.26(a) is illustrated the transfer function estimated from the noisy speech data in Figure 7.25(c) by the correlation method of the linear prediction analysis. In Figures 7.25(b), (c) and (d) are shown the transfer functions estimated by System A after two iterations, System B after ten iterations and System C with  $k=2$ . In each of the four figures of Figure 7.26, the true log magnitude spectrum of Figure 7.25(b) is also illustrated to facilitate the comparisons. Figures 7.27 and 7.28 are equivalent to Figures 7.25 and 7.26 with the difference in that a different real speech data which is voiced is used. In the two examples considered, a good fit to the spectrum can be obtained by the three systems. Again when a sufficiently large amount of background noise is added to speech, the errors discussed in Section VII.2 also occur. This can be observed to some extent for the higher formants in Figure 7.28.

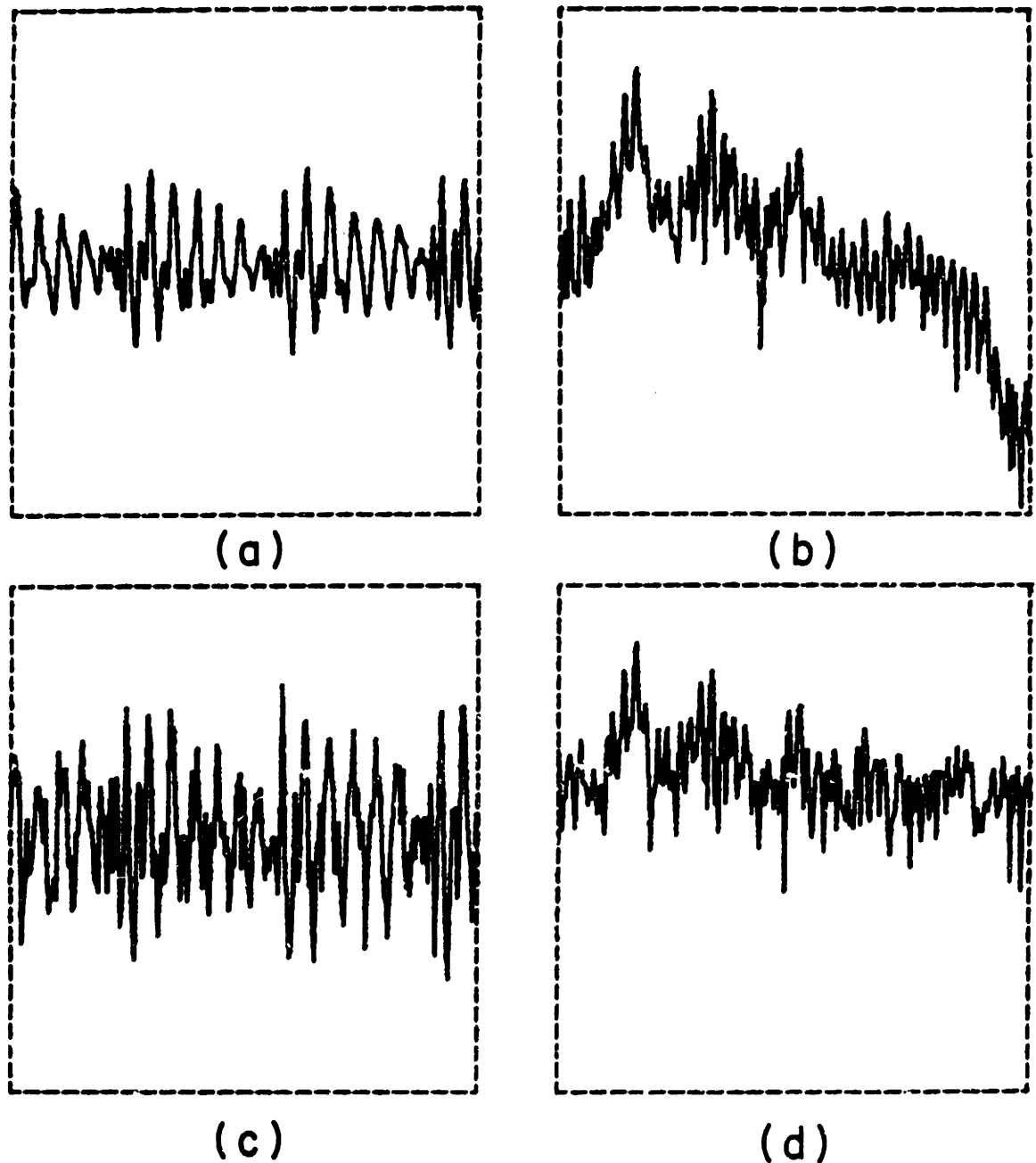
In this chapter, various examples were shown to qualitatively illustrate the performance of the three systems when applied to both synthetic and real speech data.



**Figure 7.25** (a) A real data segment of unvoiced speech;  
(b) Log magnitude spectrum of the real speech data in (a);  
(c) Noisy speech data of (a) at  $S/N = 10$  dB;  
(d) Log magnitude spectrum of the noisy speech data in (c)

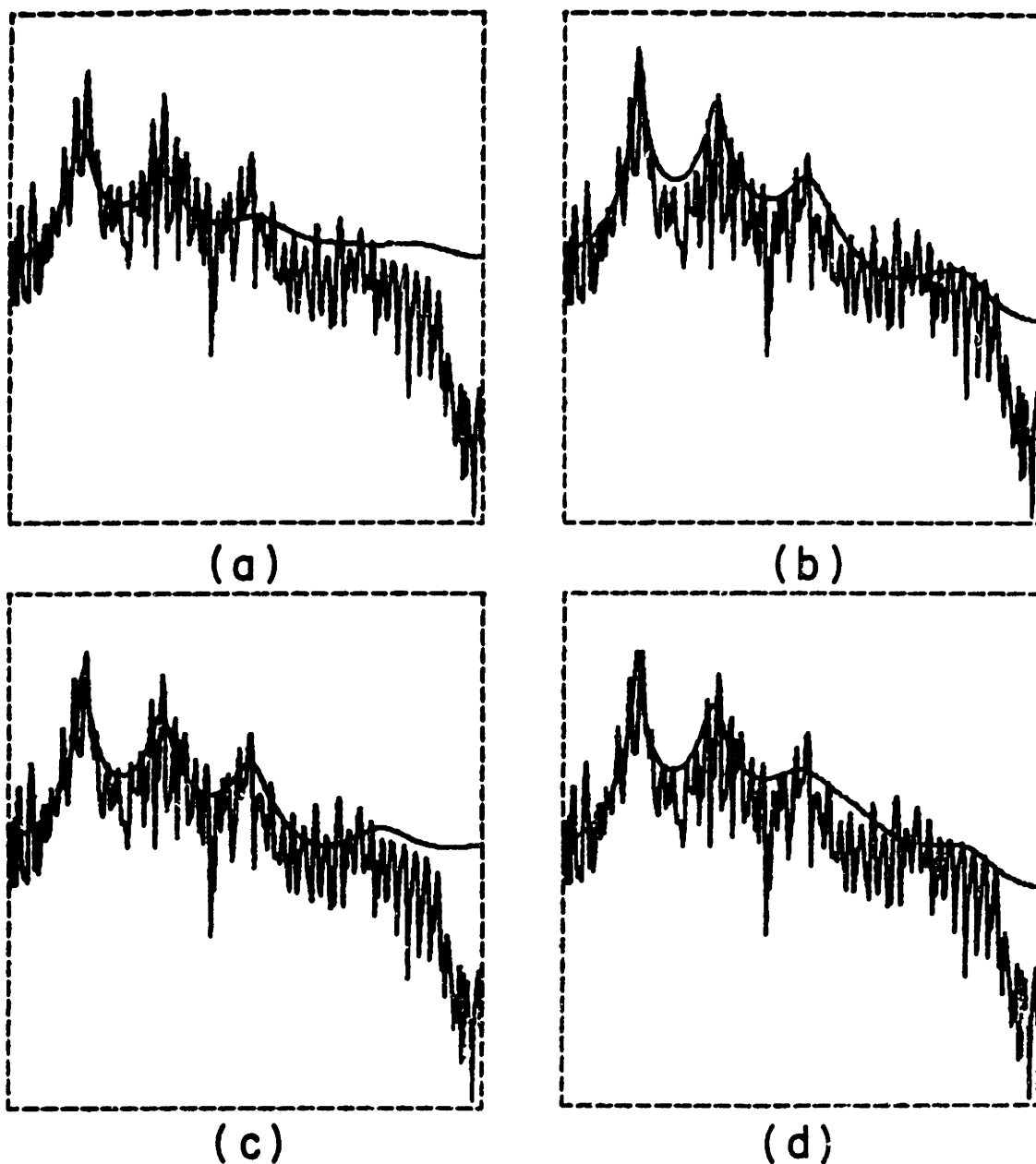


**Figure 7.26** (a) Log magnitude spectrum of the real speech data in Figure 7.25(a) and an all pole fit to the noisy speech data with  $k=0$  of System C;  
(b) Same as (a) with two iterations of System A;  
(c) Same as (a) with ten iterations of System B;  
(d) Same as (a) with  $k=2$  of System C



**Figure 7.27** (a) A real data segment of voiced speech;  
(b) Log magnitude spectrum of the real speech data in (a);  
(c) Noisy speech data of (a) at  $S/N = 10$  dB;  
(d) Log magnitude spectrum of the noisy speech data in (c)





**Figure 7.28** (a) Log magnitude spectrum of the real speech data in Figure 7.27(a) and an all pole fit to the noisy speech data with  $k=0$  of System C;  
(b) Same as (a) with two iterations of System A;  
(c) Same as (a) with ten iterations of System B;  
(d) Same as (a) with  $k=2$  of System C

In Chapter VIII, a more detailed and quantitative discussion on the performance of the three systems will be presented based on some objective and subjective tests.

CHAPTER VIII EVALUATION

VIII.1 Introduction

In this chapter, the performance of the three systems developed in Chapter VI is discussed in greater detail and more quantitatively based on some objective and subjective tests. Even though the theoretical results can be applied to colored noise as well as white noise, the background noise considered here is white Gaussian background noise. In Section VIII.2, the results of an objective test are discussed. In the objective test, the synthetic data are generated from the known all pole coefficients and the estimated all pole coefficients by the three systems are compared with the known all pole coefficients under a reasonable criterion. In Section VIII.3, the results of a subjective test to evaluate the three systems as analysis/synthesis systems (potential bandwidth compression systems) of noisy speech are discussed. If the estimated speech parameters are properly coded, then they would correspond to true bandwidth compression systems. In Section VIII.4, the three systems are evaluated as speech enhancement systems. In Section VIII.5, some additional studies are discussed, in which a complete analysis/synthesis system is used as input to a channel vocoder. In Section VIII.6, the main results obtained in Chapter VIII are summarized.

## VIII.2 Objective Evaluation

In this section, we discuss the performance of the three systems developed in Chapter VI based on an objective criterion. In Section VIII.2.1, the systems and their parameters that are objectively evaluated are listed. In Section VIII.2.2, we describe the objective criterion used for the system evaluation. In Section VIII.2.3, we discuss how the all pole coefficients are obtained to generate synthetic data. In Section VIII.2.4, we describe how the synthetic data generated are used to obtain a measure that leads to the system evaluation under the objective criterion discussed in Section VIII.2.2. In Section VIII.2.5, we discuss the results of the objective evaluation.

### VIII.2.1 Systems Evaluated

All three systems discussed in Chapter VII are evaluated for three cases per system. System A is evaluated based on the results obtained after one, two and three iterations. System B is evaluated based on the results obtained after two, five and ten iterations. System C is evaluated for the cases when  $k=1,2$ , and 3. The above nine cases are compared with each other and with the case of System C when  $k=0$  which corresponds to the conventional linear prediction analysis. In all cases, it is assumed that the a priori information of the

all pole coefficient vector is not available.

The systems and their parameters for which the objective test is performed are summarized in Table 8.1.

### VIII.2.2 Objective Criterion

One measurement is made for the objective evaluation. The measurement made is LMSE which represents the square error of the log magnitude spectrum. More specifically, LMSE is given by

$$\begin{aligned} \text{LMSE} &= \frac{1}{4\pi} \cdot M \cdot \int_{-\pi}^{\pi} (\log |1 - \sum_{i=1}^p a_i \cdot e^{-j\omega i}| - \log |1 - \sum_{i=1}^p \hat{a}_i \cdot e^{-j\omega i}|)^2 \cdot d\omega \\ &= \frac{1}{4\pi} \cdot M \cdot \int_{-\pi}^{\pi} (\log \left| \frac{1 - \sum_{i=1}^p a_i \cdot e^{-j\omega i}}{1 - \sum_{i=1}^p \hat{a}_i \cdot e^{-j\omega i}} \right|)^2 \cdot d\omega \end{aligned} \quad (8-1a)$$

and  $\underline{a}$  and  $\underline{\hat{a}}$  represent the known all pole coefficients from which the synthetic data are generated and the estimated all pole coefficients by one of the ten cases listed in Table 8.1. In evaluating LMSE in equation (8-1a), the integral is replaced by a summation by sampling at  $\omega = \frac{2\pi}{M} k$  where  $M=512$ . The  $M$  used here is the same  $M$  in equation (8-1a). Thus LMSE is evaluated by

$$\text{LMSE} = \frac{1}{2} \sum_{k=0}^{M-1} (\log \left| \frac{1 - \sum_{i=1}^p a_i \cdot e^{-j\frac{2\pi}{M} k \cdot i}}{1 - \sum_{i=1}^p \hat{a}_i \cdot e^{-j\frac{2\pi}{M} k \cdot i}} \right|)^2 \quad (8-1b)$$

Table 8.1  
Systems Evaluated under an Objective Criterion

Cases	System	Parameters	A Priori Information
A-1	A	one iteration	none
A-2	A	two iterations	none
A-3	A	three iterations	none
B-2	B	two iterations	none
B-5	B	five iterations	none
B-10	B	ten iterations	none
C-1	C	k=1	none
C-2	C	k=2	none
C-3	C	k=3	none
C-0	C	k=0	none

The criterion used here for the performance evaluation is based on the studies [7] that indicate that the square error of the log magnitude spectrum reflects reasonably well the degradation of the perceptually important aspects of speech.

In addition to the LMSE measure, another measurement, LCSE, which represents the LPC Coefficient Square Error was also made. LCSE is defined as

$$\text{LCSE} \triangleq \sum_{i=1}^p (a_i - \hat{a}_i)^2 \quad (8-2)$$

and  $\underline{a}$  and  $\underline{\hat{a}}$  represent the known all pole coefficients from which the synthetic data are generated and the estimated all pole coefficients by one of the ten cases listed in Table 8.1. The results based on this measure will not be used for the system performance evaluation in the context of speech analysis. However, LCSE is an interesting quantity in that the all pole coefficients are the parameters that are directly estimated in the systems developed in this thesis. The results based on LCSE are summarized in Appendix 2.

### VIII.2.3 Generation of All Pole Coefficients

The following two steps are used to obtain one hundred sets of the all pole coefficients that are used for generating synthetic data. The first step involves

generating a tenth order all pole function in the form of

$$H(z) = \frac{1}{\prod_{k=1}^5 (1-b_k \cdot z^{-1})(1-b_k^* \cdot z^{-1})} \quad (8-3)$$

where  $b_k$  is chosen randomly from within a circle with the radius of 0.98 in the  $z$  plane with equal a priori probability for each point in the circle. The second step involves generating the synthetic data of 256 points long by exciting  $H(z)$  in equation (8-3) with white Gaussian noise and then estimating the all pole coefficients based on the synthetic data by the correlation method of the linear prediction analysis. In generating the all pole coefficients, the second step was necessary since some all pole coefficients generated by the first step alone were quite large in their magnitudes (sometimes greater than 20) and the error measurement LCSE in equation (8-2) was dominated by the error due to a few such coefficients. It was found that the second step essentially forced the magnitudes of all the all pole coefficients generated to be less than 4 without significantly changing the locations and bandwidths of the poles generated by the first step. One hundred sets of tenth order all pole coefficients were obtained by the above two step procedure and were used in generating the synthetic data for the



objective evaluation.

#### VIII.2.4 Data Acquisition, Analysis and Results

Based on the one hundred all pole transfer functions obtained in the manner discussed in Section VIII.2.3, two hundred sequences were generated, one hundred sequences by exciting with zero mean white Gaussian noise and the remaining one hundred sequences by exciting with a train of pulses with the pulse spacing that corresponds to the fundamental frequency of 150 Hz. Then for each of the two hundred sequences, noisy synthetic data were generated by adding zero mean white Gaussian background noise at the S/N ratios of -20, 0, 10, 20, and 40 dB. For each sequence of the noisy synthetic data (one thousand sequences), the ten systems in Table VIII.1 were used to estimate the all pole coefficients. They were then compared with the known all pole coefficients from which the synthetic data were generated. Thus LMSE in equation (8-1) and LCSE in equation (8-2) were obtained for each of the one hundred sets of known all pole coefficients as a function of the system type (ten cases in Table 8.1), the type of excitation (random noise or a pulse train) and the S/N ratio (-20, 0, 10, 20, and 40 dB). For notational convenience, we denote LMSE ( $\underline{a}_n, S_i, E_j, R_k$ ) and LCSE( $\underline{a}_n, S_i, E_j, R_k$ ) to represent LMSE and LCSE that correspond to  $\underline{a}_n, S_i, E_j$  and  $R_k$ , where

$\underline{a}_n$  represents the nth set of the all pole coefficients  
and thus  $1 \leq n \leq 100$ ,

$S_i$  represents the ith system in Table 8.1 and  
thus  $1 \leq i \leq 10$ ,

$E_j$  represents the excitation type with  $E_1$  and  $E_2$   
corresponding to random noise and a pulse train  
respectively,

and  $R_k$  represents the kth S/N ratio with  $R_1, R_2, R_3, R_4$  and  
 $R_5$  corresponding to -20, 0, 10, 20, and 40 dB  
respectively.

Using this notation, we define  $\overline{\text{LMSE}}$  and  $\overline{\text{LCSE}}$  by

$$\overline{\text{LMSE}}(S_i, E_j, R_k) = \frac{1}{100} \cdot \sum_{n=1}^{100} \text{LMSE}(\underline{a}_n, S_i, E_j, R_k) \quad (8-4)$$

$$\overline{\text{LCSE}}(S_i, E_j, R_k) = \frac{1}{100} \cdot \sum_{n=1}^{100} \text{LCSE}(\underline{a}_n, S_i, E_j, R_k) \quad (8-5)$$

From equations (8-4) and (8-5),  $\overline{\text{LMSE}}$  and  $\overline{\text{LCSE}}$  represent  
the mean LMSE and LCSE averaged over the one hundred sets  
of the all pole coefficients obtained in Section VIII.2.3  
as a function of the system type, excitation type and S/N  
ratio.  $\overline{\text{LMSE}}$  obtained in this manner is tabulated in Table  
8.2 and figures based on Table 8.2 are illustrated in  
Section VIII.2.5 where we discuss the performance of  
different systems under the objective criterion.  $\overline{\text{LCSE}}$   
is tabulated in Appendix 2. In Table 8.2 is also shown  
the normalized LMSE which is defined by



$$\text{Normalized } \overline{\text{LMSE}}(S_i, E_j, R_k) = \frac{\overline{\text{LMSE}}(S_i, E_j, R_k)}{\overline{\text{LMSE}}(S_{10}, E_j, R_k)} \quad (8-6)$$

Since  $S_{10}$  corresponds to the conventional linear prediction analysis that does not account for the presence of noise, Normalized  $\overline{\text{LMSE}}(S_i, E_j, R_k)$  smaller than 1 indicates the improvement of System  $S_i$  over System  $S_{10}$ . Normalized  $\overline{\text{LCSE}}(S_i, E_j, R_k)$  defined in an analogous manner as in equation (8-6) is also tabulated in Appendix 2.

#### VIII.2.5 Discussions

In Figure 8.1 is shown  $\overline{\text{LMSE}}(S_i, E_j, R_k)$  for  $1 \leq i \leq 3$  that corresponds to System A,  $1 \leq j \leq 2$  and  $1 \leq k \leq 5$ . In Figure 8.1(a) is illustrated the case of  $j=1$  that corresponds to random noise excitation and in Figure 8.1(b) is shown the case of  $j=2$  that corresponds to the case of the pulse train excitation. In the figures,  $\overline{\text{LMSE}}(S_{10}, E_j, R_k)$  is also shown by a solid line to facilitate the comparison in terms of improvement over the conventional linear prediction analysis. From Figure 8.1, the following points are noted. First, System A is capable of performing better than the conventional linear prediction analysis for a wide range of S/N ratios. Second, System A shows a better performance after two iterations than after one iteration or three iterations at the S/N ratios above -10 dB. This result is consistent with our observations in

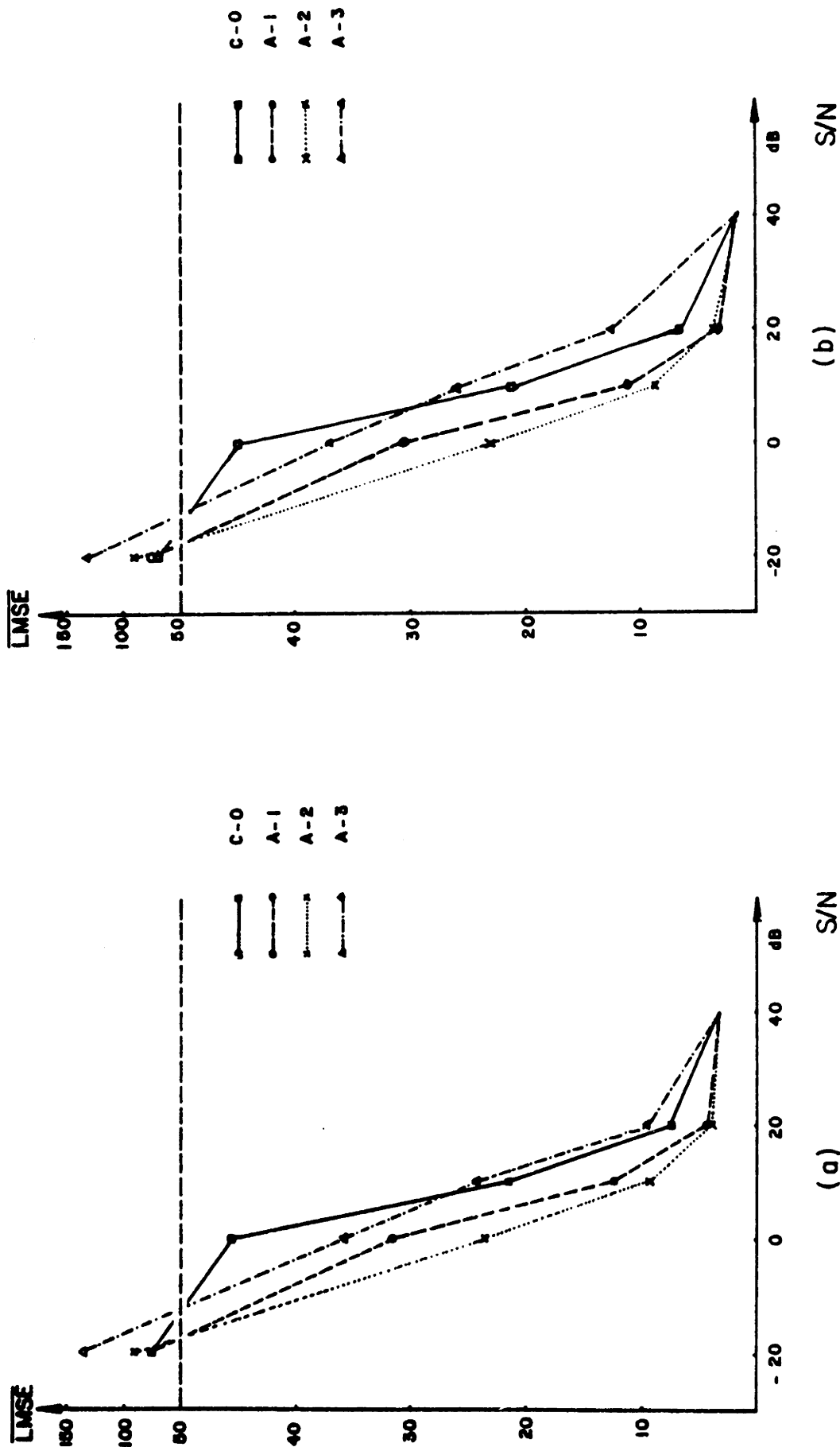


Figure 8.1.1 Performance comparison of System A based on LMSE

(a) random noise excitation case;

(b) pulse train excitation case

Chapter VII. Third, System A degrades quickly below 0 dB of S/N ratio and eventually performs worse than the conventional linear prediction analysis. Therefore, -10 dB seems to be the lowest S/N ratio at which System A shows some improvement over the conventional linear prediction analysis. Fourth, even though there are detailed quantitative differences, qualitatively speaking, the performance of System A is essentially the same for both types of excitations which are consistent with our observations in Chapter VII.

Figure 8.2 is essentially the same as Figure 8.1 with the difference in that  $\overline{\text{LMSE}}$  is plotted to determine the performance of System B. The three systems plotted are B-2, B-5 and B-10 listed in Table 8.1. From Figure 8.2, the following points are noted. First, System B is capable of performing better than the conventional linear prediction analysis for a wide range of S/N ratios. Second, System B performs better after more iterations are carried out. Therefore, it appears that the converging solution is the optimum under the objective criterion. This is consistent with our observations in Chapter VII. It also appears that the results after ten iterations are reasonably close to the converging solution. In Figure 8.2(a) is plotted a point (x) at the S/N ratio of 10 dB after 20 iterations and it is slightly better than the results after 10 iterations. Third, System B degrades

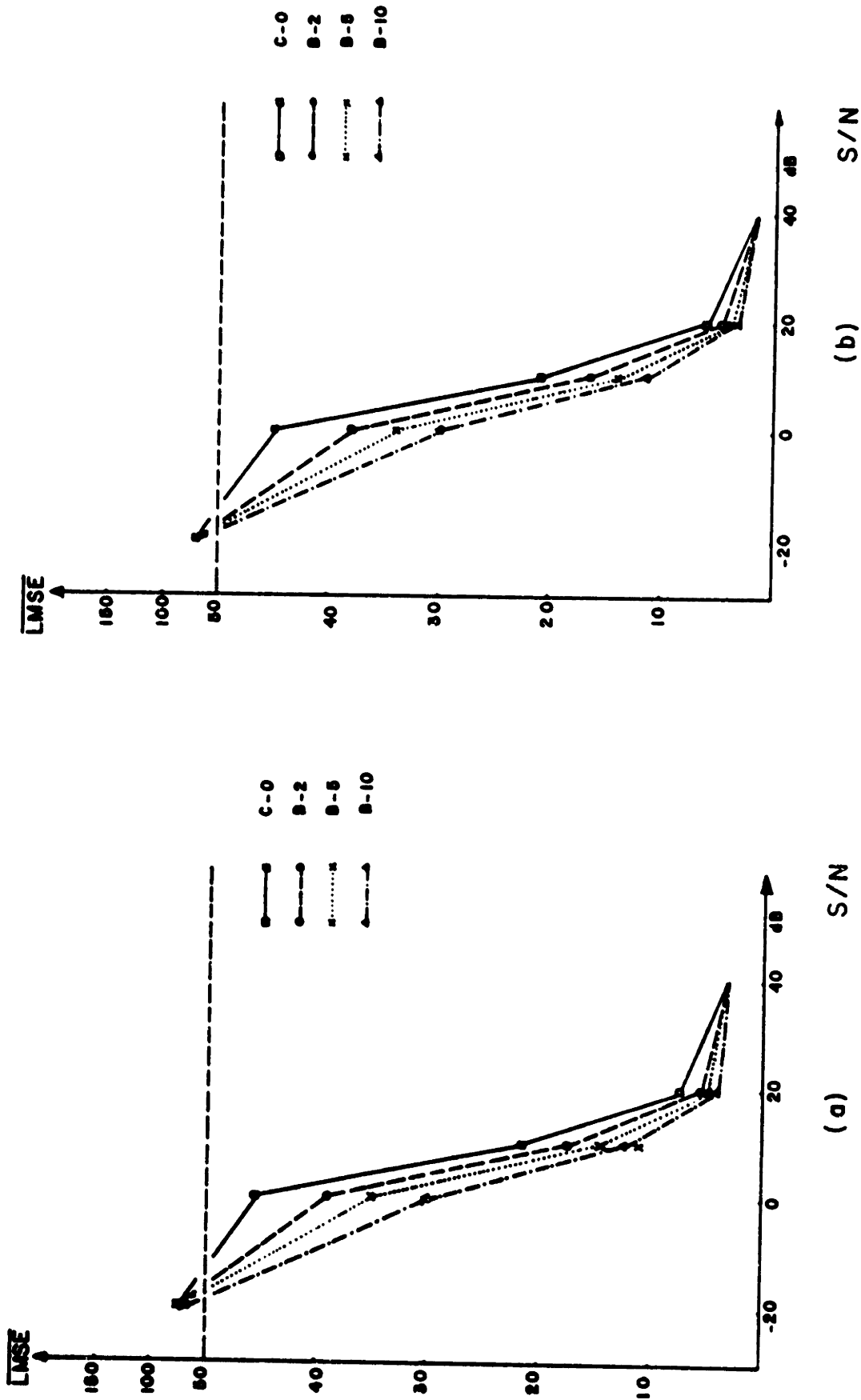


Figure 8.2 Performance Comparison of System B based on LMSE  
(a) random noise excitation case;  
(b) pulse train excitation case

quickly below 0 dB of S/N ratio and eventually performs similarly to the conventional linear prediction analysis. Therefore, -20 dB seems to be the lowest S/N ratio at which System B shows some improvement over the conventional linear prediction analysis. Fourth, the performance of System B is essentially the same for both types of excitations, which is consistent with our observations in Chapter VII.

Figure 8.3 is essentially the same as Figure 8.1 with the difference in that  $\overline{\text{LMSE}}$  is plotted to determine the performance of System C. The three systems plotted are C-1, C-2 and C-3 listed in Table 8.1. From Figure 8.3, the following points are noted. First, System C is capable of performing better than the conventional linear prediction analysis for a wide range of S/N ratios. Second, System C with  $k=2$  shows a better performance than with  $k=1$  or 3 at the S/N ratios above -10 dB. This result is consistent with our observations in Chapter VII. Since  $k$  is a real number, there may be a more optimum  $k$  which is not an integer. To understand how much more improvement can be made by a different choice of  $k$ ,  $\overline{\text{LMSE}}$  ( $S_1, E_1, S/N=10$  dB) was computed for  $k$  between 1.0 and 3.0 sampled at twenty equally spaced points (i.e.,  $k=1.1, 1.2, \dots, 2.8, 2.9$ ). It was found that  $k=2.0$  is the optimum among the 20 different values of  $k$ . Even though  $k$  has not been varied for all its possible values at the S/N



quickly below 0 dB of S/N ratio and eventually performs similarly to the conventional linear prediction analysis. Therefore, -20 dB seems to be the lowest S/N ratio at which System B shows some improvement over the conventional linear prediction analysis. Fourth, the performance of System B is essentially the same for both types of excitations, which is consistent with our observations in Chapter VII.

Figure 8.3 is essentially the same as Figure 8.1 with the difference in that  $\overline{\text{LMSE}}$  is plotted to determine the performance of System C. The three systems plotted are C-1, C-2 and C-3 listed in Table 8.1. From Figure 8.3, the following points are noted. First, System C is capable of performing better than the conventional linear prediction analysis for a wide range of S/N ratios. Second, System C with  $k=2$  shows a better performance than with  $k=1$  or 3 at the S/N ratios above -10 dB. This result is consistent with our observations in Chapter VII. Since  $k$  is a real number, there may be a more optimum  $k$  which is not an integer. To understand how much more improvement can be made by a different choice of  $k$ ,  $\overline{\text{LMSE}}$  ( $S_1, E_1, S/N=10$  dB) was computed for  $k$  between 1.0 and 3.0 sampled at twenty equally spaced points (i.e.,  $k=1.1, 1.2, \dots, 2.8, 2.9$ ). It was found that  $k=2.0$  is the optimum among the 20 different values of  $k$ . Even though  $k$  has not been varied for all its possible values at the S/N

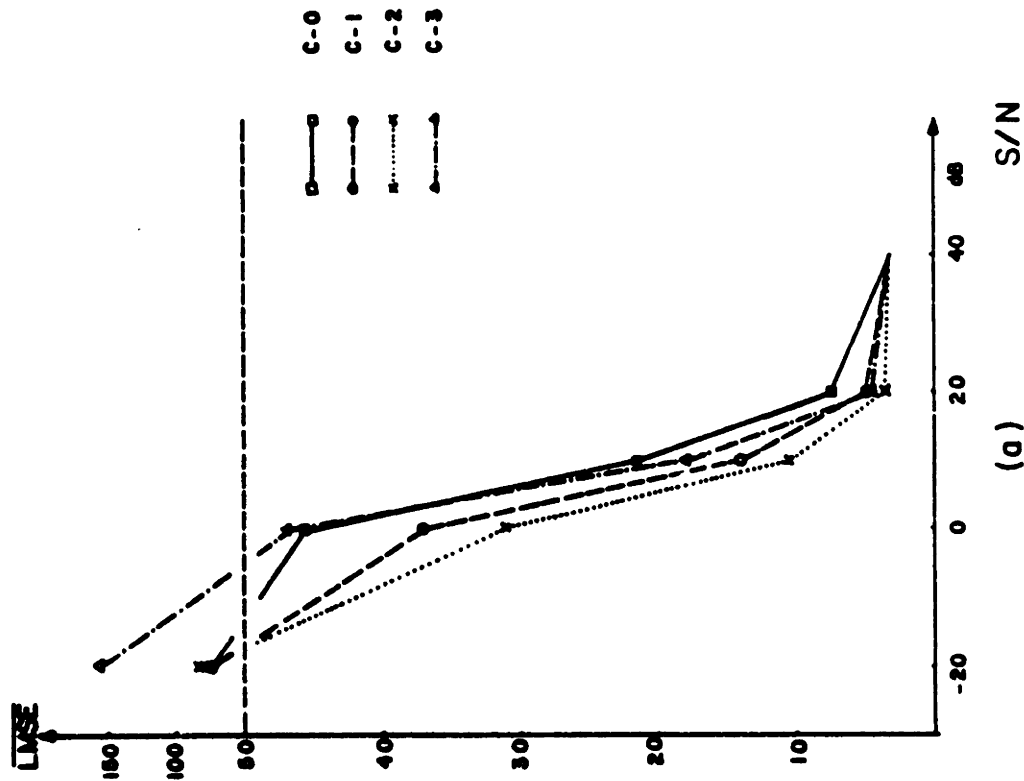
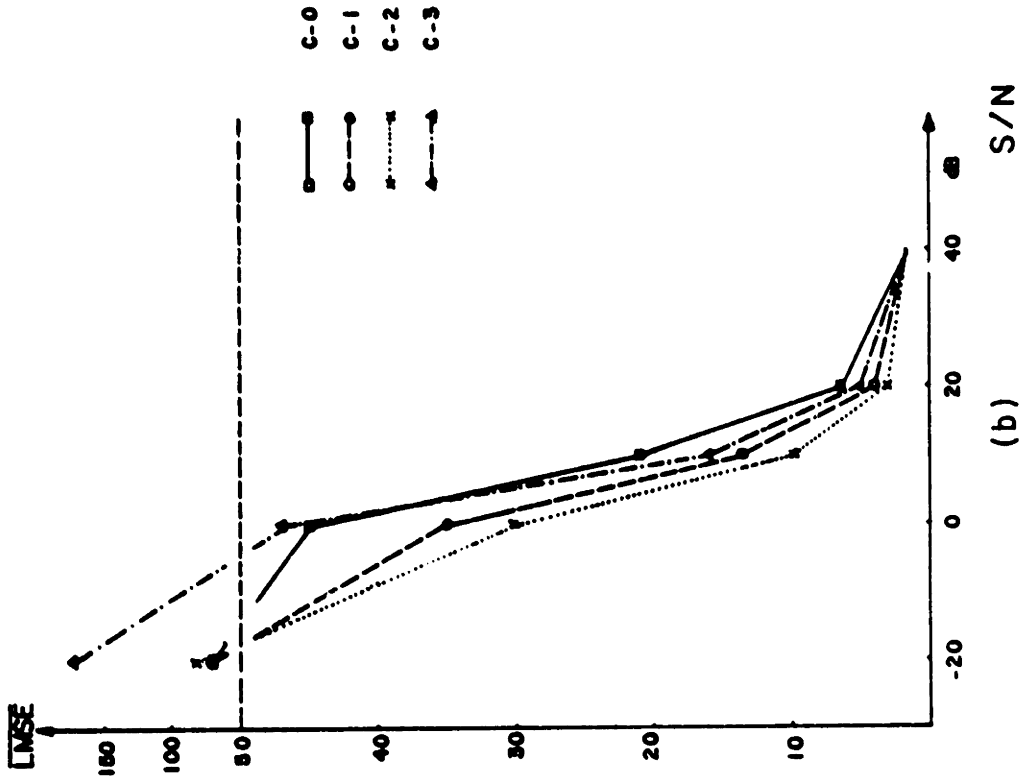


Figure 8.3 Performance comparison of System C based on LMSE

(a) random noise excitation case;

(b) pulse train excitation case

ratios considered, it appears that other choices of  $k$  do not significantly improve the performance of System C. Third, System C degrades quickly below 0 dB of S/N ratio and eventually performs worse than the conventional linear prediction analysis. Therefore, -10 dB seems to be the lowest S/N ratio at which System C with  $k=2$  shows some improvement over the conventional linear prediction analysis. Fourth, the performance of System C is essentially the same for both types of excitation which are consistent with our observations in Chapter VII.

In Figure 8.4 are shown the results of cases A-2, B-10 and C-2 which seem to be approximately the best that can be achieved by the three systems. The case of C-0 is also shown to facilitate the comparison with the conventional linear prediction analysis. Figure 8.5 is equivalent to Figure 8.4 except that Normalized  $\overline{LMSE}$  is plotted instead of  $\overline{LMSE}$ . From Figures 8.4 and 8.5, the following points are noted. First, below S/N ratio of -20 dB, none of the three systems performs better than the conventional linear prediction analysis. Between -20 and -10 dB of S/N ratio, System B after ten iterations performs best. Approximately from -10 dB to 20 dB of S/N ratio, System A after two iterations shows the best performance. Between 20 to 40 dB of S/N ratio, System C with  $k=2$  performs best. However, the improvement of System C over System A or System B is not large. Above the

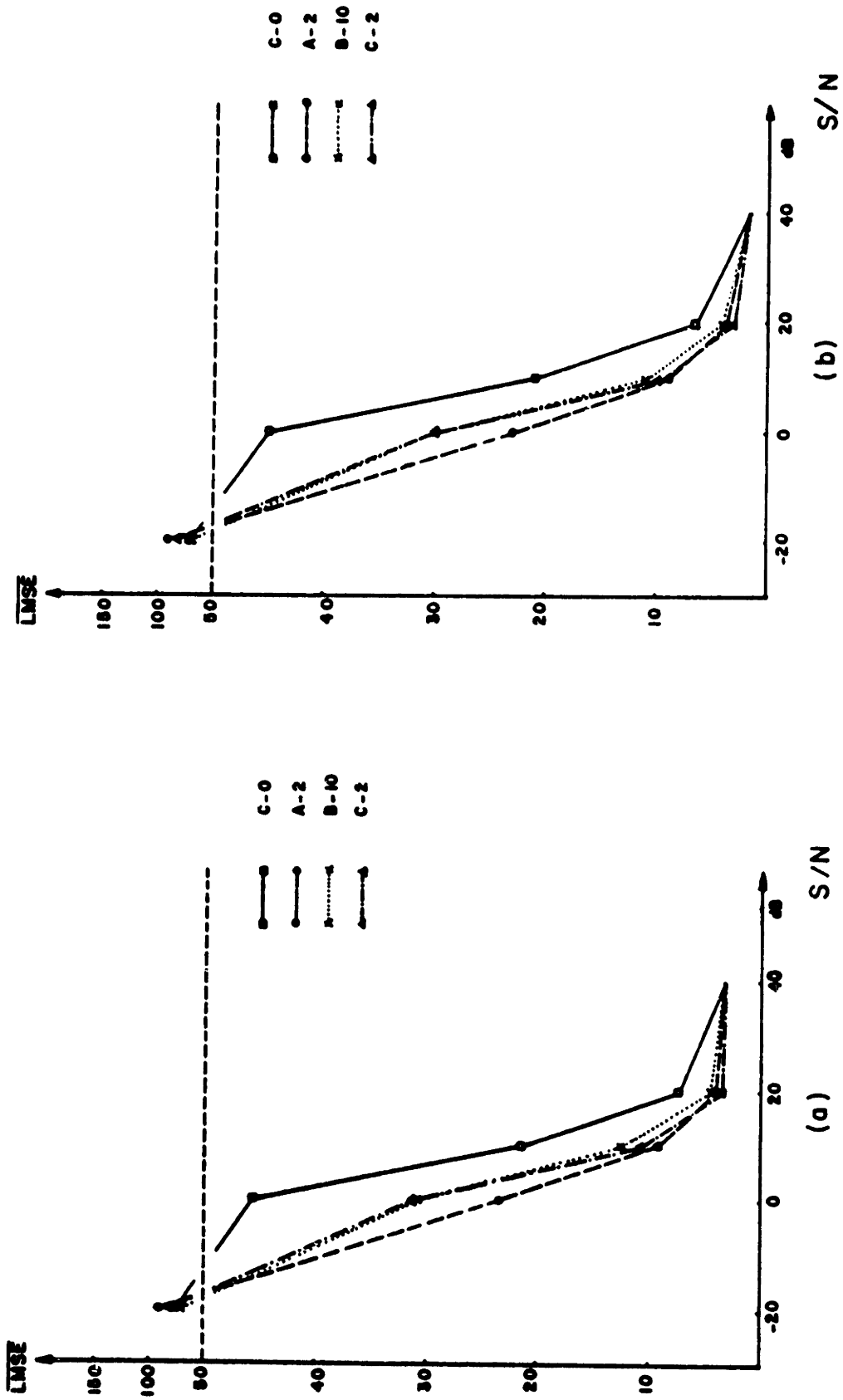


Figure 8.4 Performance comparison of Systems A-2, B-10, and C-2 based on LMSE

- (a) random noise excitation case;
- (b) pulse train excitation case

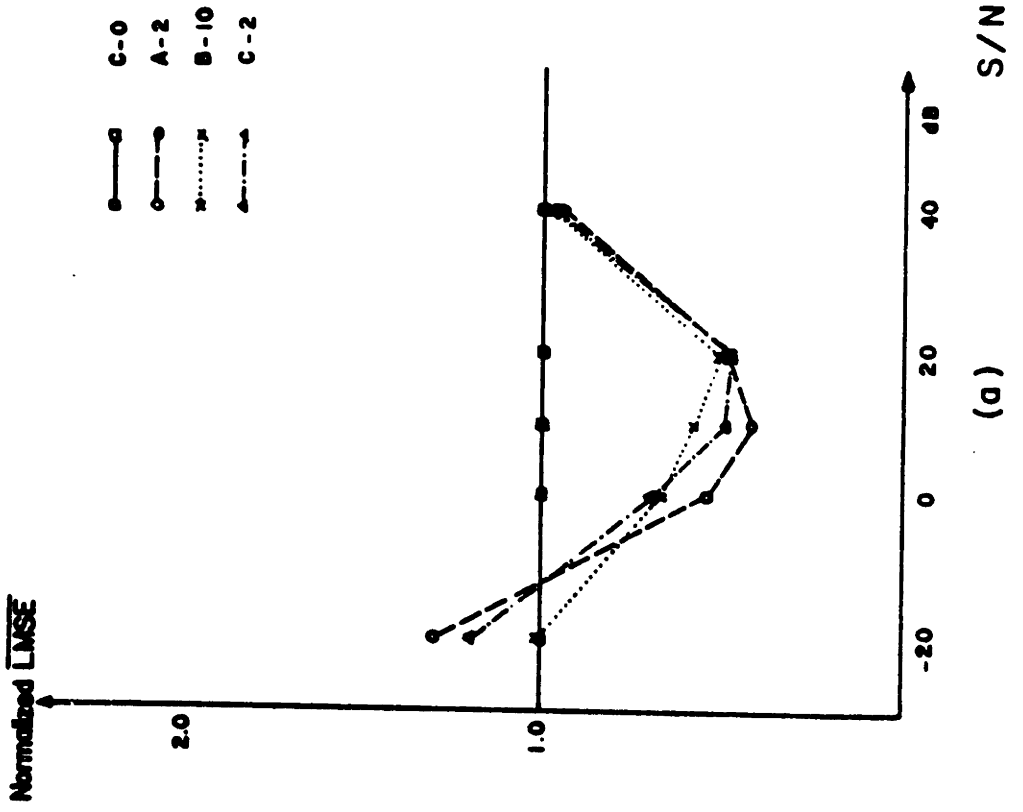
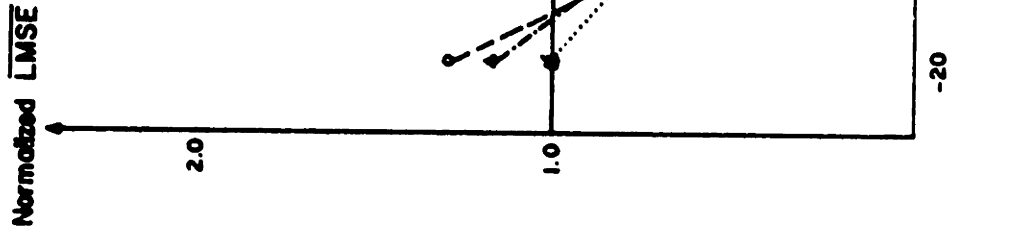


Figure 8.5 Performance comparison of Systems A-2, B-10 and C-2 based on Normalized LMSE

(a) random noise excitation case;

(b) pulse train excitation case

S/N ratio of 40 dB, all the three systems essentially reach the same performance. This result indicates that no one system performs best at all S/N ratios. Since the intelligibility of speech changes between essentially zero to near perfect in the range of S/N ratios between -10 and 20 dB, System A after two iterations would be most useful for various practical applications under the objective criterion.

In Figure 8.6, the dotted line shows the best that can be achieved by any combination of the three systems discussed in Figures 8.4 and 8.5. The solid line corresponds to the conventional linear prediction analysis. Therefore, the difference between the solid line and the dotted line shows the improvement that can be achieved by any combination of the three systems developed in Chapter VI. How this improvement translates to the improvement in the listener's subjective domain is the topic of the next section.

### VIII.3 Subjective Evaluation: Potential Bandwidth Compression Systems

In this section, we discuss the performance of the three systems as potential bandwidth compression systems of noisy speech. When the speech model parameters are properly coded they would correspond to true bandwidth compression systems. In Section VIII.3.1, the test sentences

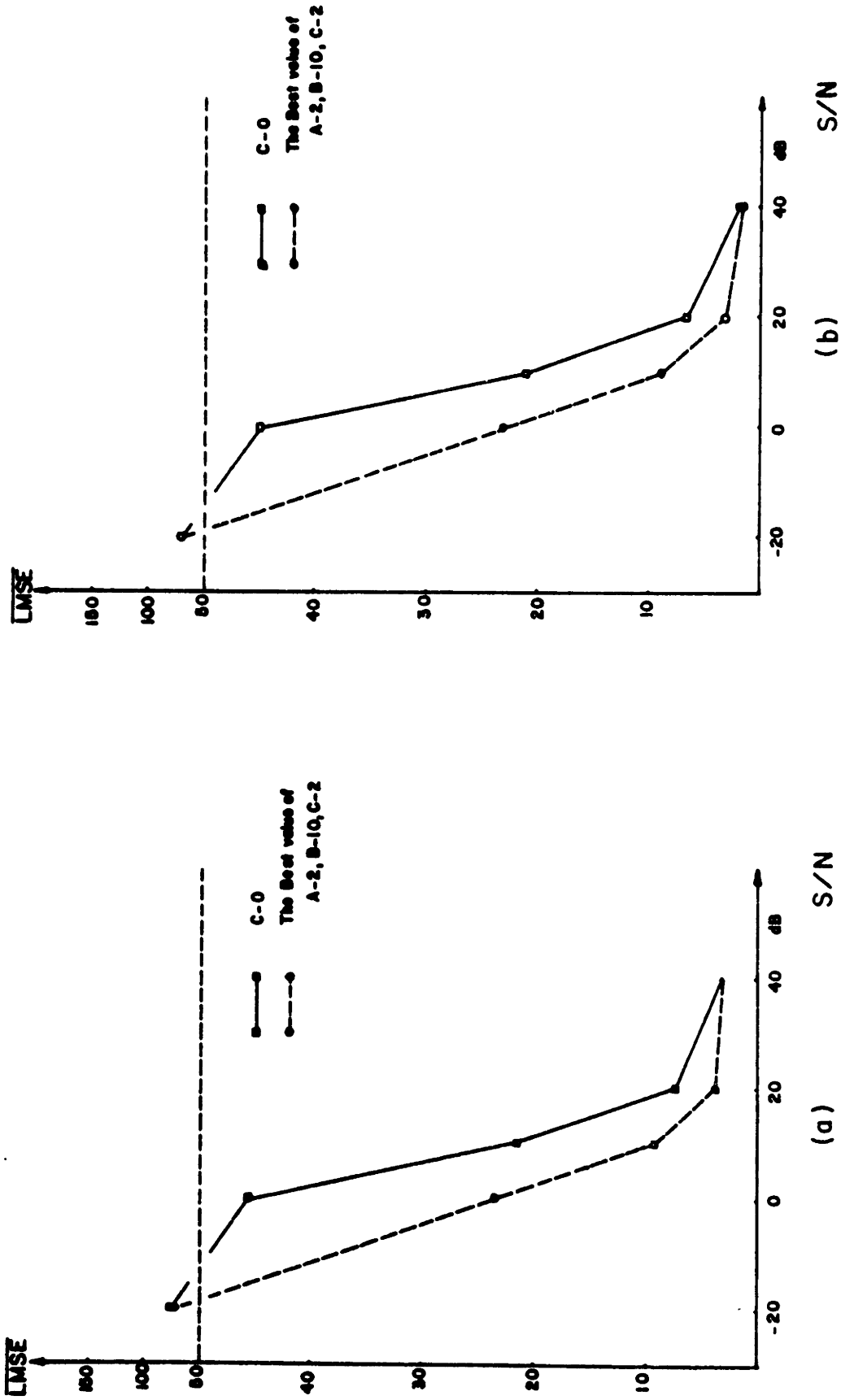


Figure 8.6 Best performance that can be achieved by any combination of Systems A-2, B-10, and C-2 based on LMSE  
(a) random noise excitation case;  
(b) pulse train excitation case

that have been used in all the subjective tests are listed. In Section VIII.3.2, the speech synthesis system used in synthesizing speech based on the estimated all pole coefficients by various potential bandwidth compression systems is discussed. In Section VIII.3.3, various systems are compared with each other and based on a very informal listening, the potential bandwidth compression system that performs best is determined. In Section VIII.3.4, the system chosen in Section VIII.3.3 is compared with the conventional linear prediction analysis by fifteen listeners and the results obtained are discussed.

#### VIII.3.1 Test Sentences

In all the subjective comparisons discussed in this chapter, the following five English sentences are used:

sentence 1: They took the cross town bus.

sentence 2: That shirt seems much too long.

sentence 3: He has the bluest eyes.

sentence 4: The ball dropped from his hands.

sentence 5: Line up at the screen door.

Sentences 1, 3, and 5 are spoken by adult male speakers and sentences 2 and 4 are spoken by adult female speakers.

#### VIII.3.2 Speech Analysis/Synthesis System

In the analysis of speech, the all pole coefficients are estimated by various different systems. The gain factor  $g$  is



estimated by an energy consideration such that the synthesized speech has the energy that is approximately equal to  $\sum_n y^2(n) - \sum_n E[d^2(n)]$ . In the case of the conventional linear prediction analysis, the gain  $g$  is obtained such that the synthesized speech has the energy that is approximately equal to  $\sum_n y^2(n)$ . The source information consists of the voicing/unvoicing decision and the pitch period in the case of voicing. The source information is obtained from the noise-free speech and the same source information is used in all cases.

In the analysis, the number of all pole coefficients  $p$  is assumed to be 10, the analysis window used is a rectangular window of 256 points long and after each analysis, the window is moved by 128 points and therefore the current analysis window overlaps with the previous analysis window by 128 data points. Other choices of windows such as Hamming window were also considered. The subjective improvements by other choices of windows were minor in all cases.

In the speech synthesis, the system in Figure 3.1 is used to generate speech.

### VIII.3.3 Preliminary Comparison

The synthesized speech at three S/N ratios (i.e. 20 dB, 10 dB, 0 dB) by various systems listed in Table 8.1 has been compared informally with each other by a few listeners and the following subjective judgements were made.

#### VIII.3.3.1 Comparison of Systems A-1, A-2 and A-3

As the number of iterations of System A increases from one iteration to three iterations, it has been observed that speech sounds clearer and noise is reduced more. However, some "musical tone" like background noise becomes more apparent and intenser as the number of iterations increases. It appears that such speech degradation is primarily due to the possible incorrect estimation of the formant frequencies, particularly at higher formants. As a reasonable compromise, System A-2 appears to be better than either System A-1 or System A-3.

#### VIII.3.3.2 Comparison of Systems B-2, B-5 and B-10

As the number of iterations of System B increases from one iteration to ten iterations, it has been observed that speech appears clearer and noise seems to be reduced more. For System B, it appears that the performance of System B-10 is better than System B-2 or System B-5.

#### VIII.3.3.3 Comparison of Systems C-1, C-2 and C-3

For the S/N ratios of 10 dB and 20 dB, it appears that the performance of System C-2 is better than System C-1. At the S/N ratio of 0 dB, System C-2 appears to generate clearer voiced sounds. However, many segments of unvoiced sounds and the higher formants of voiced sounds essentially disappear due to the subtraction of twice as much average short time

energy spectrum of noise from the short time energy spectrum of noisy speech when the noise level is high relative to the signal level.

A comparison between System C-2 and C-3 indicates that the performance of System C-2 is better than System C-3 at all three S/N ratios considered.

#### VIII.3.3.4 Comparison of Systems A-2, B-10 and C-2

At the S/N ratios of 10 and 20 dB, System A-2 appears to generate more intelligible and higher quality speech than System B-10 or C-2. At the S/N ratio of 0 dB, System A-2 and B-10 perform better than System C-2. However, the choice between System A-2 and System B-10 is difficult, since System A-2 appears to have removed more random background noise but generated more "musical tone" like distortion which is quite pronounced at this S/N ratio. Despite this difficulty, we have chosen System A-2 to be compared to the conventional linear prediction analysis for a speech preference test discussed in the next section.

#### VIII.3.4 Evaluation of System A-2 Relative to Conventional LPC Method

In general, a fair evaluation of either a bandwidth compression system or speech enhancement system should be based on many factors such as intelligibility, speech quality, listener fatigue, etc. The main purpose of the subjective tests in this dissertation is a preliminary examination to determine whether or not the class

of systems developed in this thesis deserve further research efforts in terms of improving and evaluating them. With such a purpose in mind, we have taken a very limited point of view and performed a speech preference test with a small amount of test material. The test procedures and results are discussed in this section.

#### VIII.3.4.1 Test Material and Procedures

The test material consists of the five English sentences described in Section VIII.3.1. The S/N ratios considered in the test are 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB.

Two sentences were constructed for each of the five English sentences and five S/N ratios based on the analysis/synthesis system discussed in Section VIII.3.2. One of the two sentences corresponded to System A-2 and the other sentence corresponded to the conventional linear prediction analysis. Therefore, a total of fifty sentences were constructed.

The test consisted of three sessions: one practice session and two main sessions, Session I and Session II. The practice session was intended primarily to acquaint the listeners with the test procedures. Session I was devoted to evaluating System A-2 as a potential bandwidth compression system and Session II was devoted to evaluating System A-2 as a speech enhancement system.

The test materials, procedures and results of Session II will be presented in Section VIII.4 where we discuss System A-2 as a speech enhancement system.

Session I consisted of five parts, each part corresponding to one of the five S/N ratios. Each part consisted of five trials. Each of the five trials corresponded to one of the five English sentences. In each trial, two sentences were presented, one of which corresponded to System A-2 and the other corresponded to the conventional linear prediction analysis. The order of the presentation of the two sentences was randomized in each trial.

The listeners were asked to judge in each trial which of the two sentences was more preferable. It was explained to the listeners that "more preferable" could mean "more intelligible", "of higher quality", "less noisy", any combination of them, etc. and it was left entirely up to each individual listener to use his own interpretation of "more preferable". In each trial, the listeners were able to answer in five different discrete categories: the first sentence is definitely more preferable, the first sentence appears to be more preferable, no preference between the two sentences, the second sentence appears to be more preferable, and the second sentence is definitely more preferable. It was emphasized in the test that the judgement in each trial

should be made as independently as possible of all the previous trials.

#### VIII.3.4.2 Data Analysis and Results

Each response of a listener was converted to a numerical value in the following manner:

- 2: System A-2 is definitely more preferable
- 1: System A-2 appears to be more preferable
- 0: no preference
- 1: The conventional LPC analysis appears to be more preferable
- 2: The conventional LPC analysis is definitely more preferable

The numerical value assigned to each response was considered to represent the preference index of System A-2  $P(S_i, L_j, R_k)$  where  $S_i$  represents the  $i$ th English sentence and thus  $1 \leq i \leq 5$ ,  $L_j$  represents the  $j$ th listener and thus  $1 \leq j \leq 15$  since fifteen listeners participated in the test, and  $R_k$  represents the  $k$ th S/N ratio considered and thus  $1 \leq k \leq 5$  ( $k=1$  corresponding to S/N=0 dB,  $k=2$  corresponding to S/N=5 dB, etc.). From  $P(S_i, L_j, R_k)$ ,  $P(L_j, R_k)$  was obtained by

$$P(L_j, R_k) = \frac{1}{5} \sum_{i=1}^5 P(S_i, L_j, R_k) \quad (8-7)$$

From  $P(L_j, R_k)$  in equation (8-7),  $P_M(R_k)$  and  $P_{SD}(R_k)$  were

obtained by

$$P_M(R_k) = \frac{1}{15} \sum_{j=1}^{15} P(L_j, R_k) \quad (8-8a)$$

$$P_{SD}(R_k) = \left[ \frac{1}{15} \sum_{j=1}^{15} (P_M(R_k) - P(L_j, R_k))^2 \right]^{\frac{1}{2}} \quad (8-8b)$$

Therefore, a positive  $P_M(R_k)$  represents the preference of System A-2 over the conventional linear prediction analysis averaged over the five sentences used as test material and fifteen listeners. The highest number possible for  $P_M(R_k)$  is 2.  $P_{SD}(R_k)$  is the standard deviation of  $P(L_j, R_k)$  and represents the variability among the listeners in their responses.

$P_M(R_k)$  and  $P_{SD}(R_k)$  are tabulated in Table 8.3 and plotted in Figure 8.7. The solid line in Figure 8.7 corresponds to  $P_M(R_k)$  and the difference between the solid line and either the upper or lower dotted line corresponds to  $P_{SD}(R_k)$ . Even though the test was not performed at the S/N ratio of  $-\infty$  or  $+\infty$ , we can deduce the results from the theoretical considerations. At the S/N ratio of  $\infty$ , System A-2 is equivalent to the conventional linear prediction analysis and hence we would expect that  $P_M(S/N \text{ ratio} = \infty) = 0$ . At the S/N ratio of  $-\infty$ , the preference if any does not mean much.

Table 8.3

Results of the Speech Preference Test in which System A-2  
is Used as a Potential Bandwidth Compression System

S/N Ratio	$P_M(R_k)$	$P_{SD}(R_k)$
0 dB	1.413	0.529
5 dB	1.387	0.481
10 dB	1.040	0.662
15 dB	1.600	0.343
20 dB	1.293	0.473



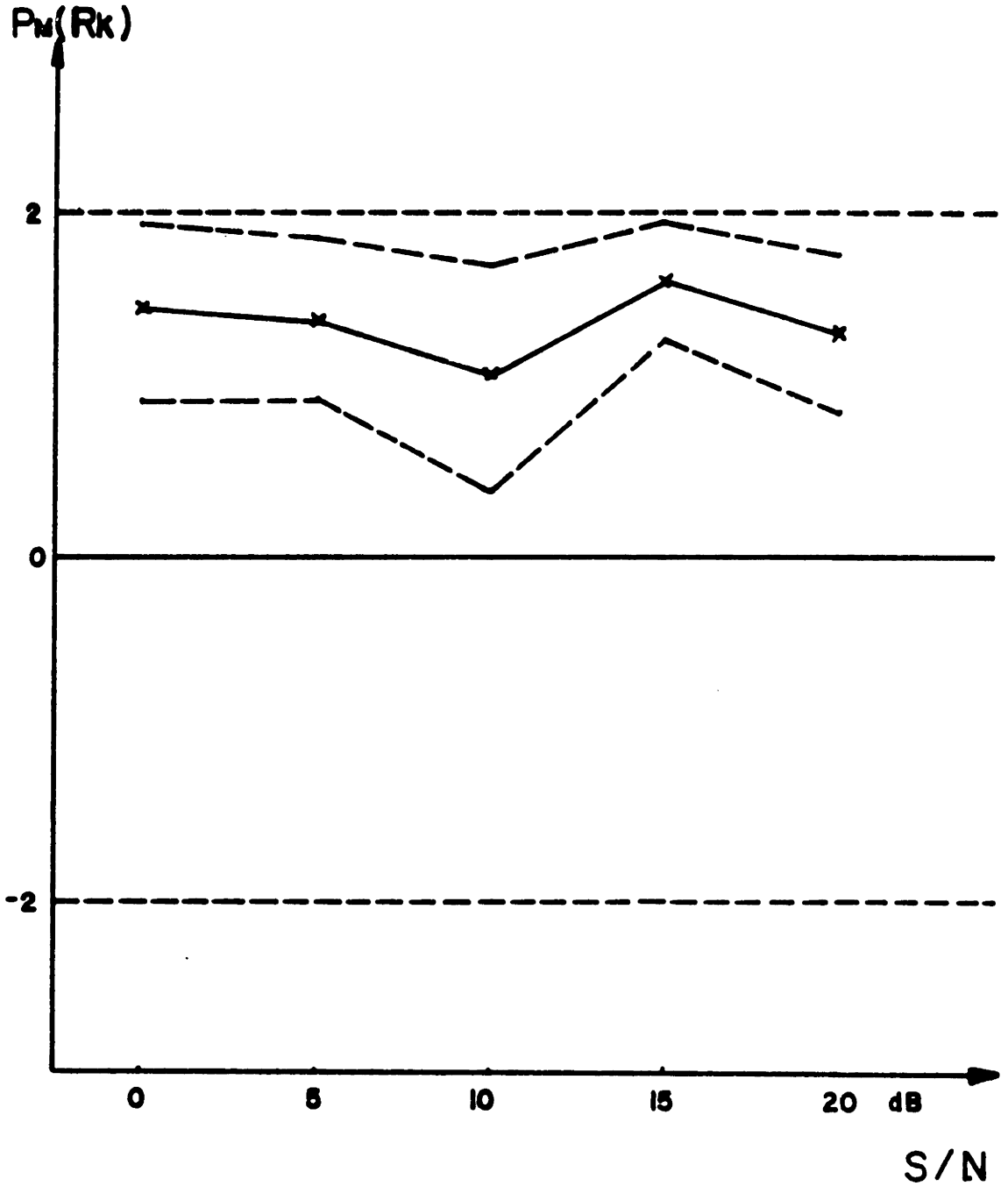


Figure 8.7 (a) Results of the speech preference test in which System A-2 is used as a potential bandwidth compression system. The solid line represents  $P_M(R_k)$ , and the distance between the solid line and the dotted line represents  $P_{SD}(R_k)$

### VIII.3.4.3 Discussions

From the results in Figure 8.7 it is clear that System A-2 is preferred over the conventional linear prediction analysis at all the five S/N ratios that have been considered. We conclude that these results are sufficiently encouraging to devote further research efforts in improving and evaluating a class of systems developed in this dissertation.

### VIII.4 Subjective Evaluation: Speech Enhancement Systems

As was discussed before, the systems that we developed in Chapter V and Chapter VI can be used not only as bandwidth compression systems but also as speech enhancement systems. There are two ways that the systems developed in this thesis can be used for speech enhancement. One of them is to use the estimated speech  $\hat{s}_w(n)$  as enhanced speech. An alternative way is to use the analysis/synthesis system as a speech enhancement system. Since a complete analysis/synthesis system requires the estimation of source information, the evaluation of the systems as speech enhancement systems in this section are restricted to the case in which the estimated speech  $\hat{s}_w(n)$  is used as enhanced speech. Some discussions on using a complete analysis/synthesis system for speech enhancement are given in Section VIII.5.

In Section VIII.4.1, the speech enhancement systems

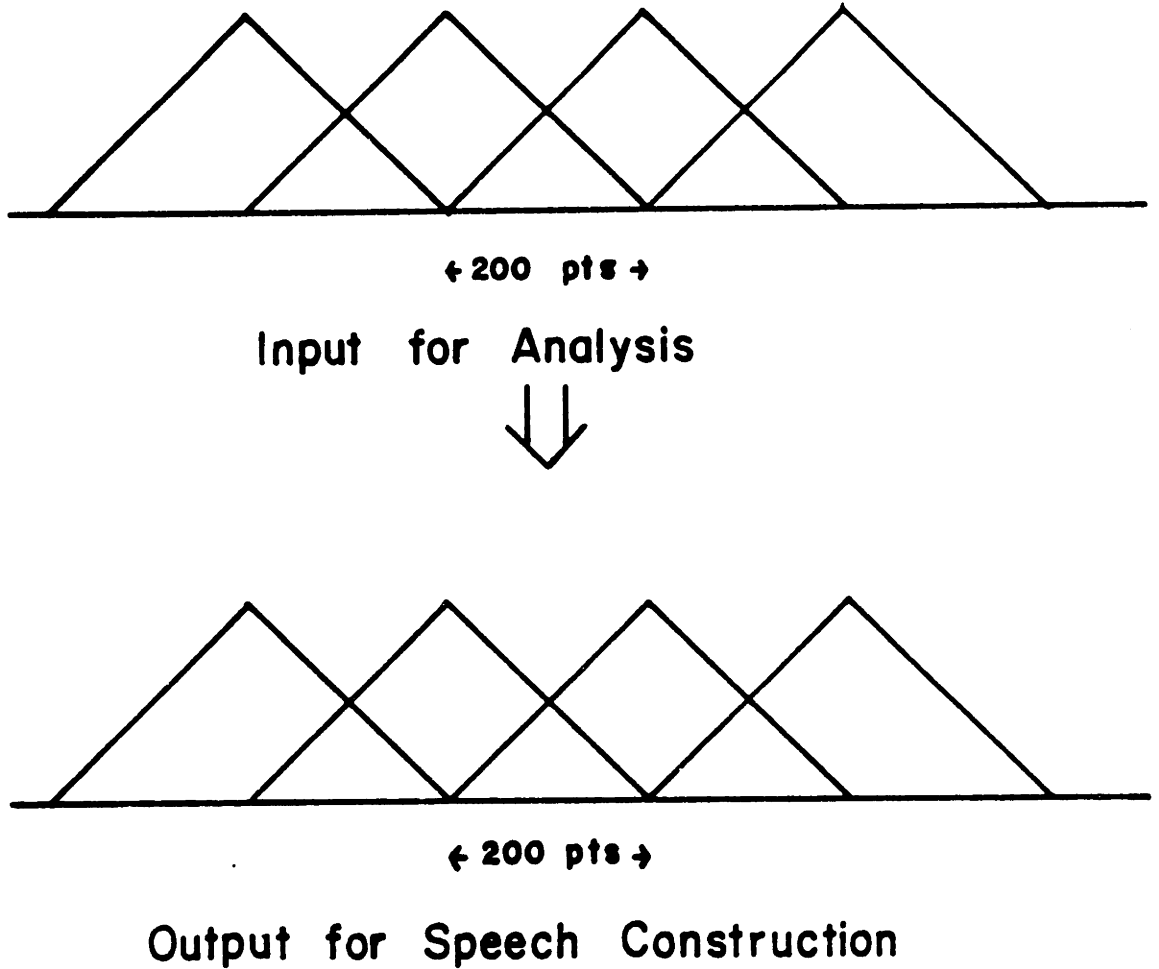
that have been used for evaluation are specified. In Section VIII.4.2, we remark briefly on the relative performance of various systems listed in Table 8.1 as speech enhancement systems. In Section VIII.4.3, the performance of System A-2 is evaluated by a speech preference test.

#### VIII.4.1 Speech Enhancement Systems

The speech enhancement systems are based on the estimated  $\hat{s}_w(n)$ . In System A,  $\hat{s}_w(n)$  is obtained in Step 2. In System B,  $\hat{s}_w(n)$  is obtained in Step 2B. In System C,  $\hat{s}_w(n)$  is obtained in Step 2. The analysis is again based on a tenth order all pole system with a 10 kHz sampling rate. In the analysis, a triangular window of 400 points was used with a frame rate of 200 points per frame. The estimated  $\hat{s}_w(n)$  is added back together in the same way it has been analyzed as is shown in Figure 8.8.

#### VIII.4.2 Preliminary Comparison

The differences in performance among various speech enhancement systems are very similar to the differences in performance among various potential bandwidth compression systems discussed in Section VIII.3.3. Therefore, the discussions in Section VIII.3.3 also apply to the three systems as speech enhancement systems.



**Figure 8.8** Data segmentation for the analysis and construction of speech in a speech enhancement system based on System A-2

### VIII.4.3 Evaluation of System A-2 as a Speech Enhancement System

All aspects of the evaluation of System A-2 as a speech enhancement system are identical to its evaluation as a potential bandwidth compression system discussed in Section VIII.3 with the following two differences. One difference is that the comparison was made between noisy speech and speech enhanced by System A-2 rather than between synthesized speech by the conventional LPC method and System A-2. Another difference is that System A-2 as a bandwidth compression system was evaluated in Session I as was discussed in Section VIII.3, while System A-2 as a speech enhancement system was evaluated in Session II of the speech preference test. The responses obtained in Session II of the speech preference test were analyzed in the same manner as those obtained in Session I. To differentiate the results of Session II from Session I, we use  $Q(S_i, L_j, R_k)$ ,  $Q(L_j, R_k)$ ,  $Q_M(R_k)$ ,  $Q_{SD}(R_k)$  in place of  $P(S_i, L_j, R_k)$ ,  $P(L_j, R_k)$ ,  $P_M(R_k)$ ,  $P_{SD}(R_k)$  to denote the preference index obtained from the responses in Session II. Therefore,  $Q(S_i, L_j, R_k)$  denotes the preference index as a function of the  $i$ th English sentence,  $j$ th listener and  $k$ th S/N ratio. The equations parallel to equations (8-7) and (8-8) are

$$Q(L_j, R_k) = \frac{1}{5} \sum_{i=1}^5 Q(S_i, L_j, R_k) \quad (8-9a)$$

$$Q_M(R_k) = \frac{1}{15} \sum_{j=1}^{15} Q(L_j, R_k) \quad (8-9b)$$

$$Q_{SD}(R_k) = \left[ \frac{1}{15} \sum_{j=1}^{15} (Q_M(R_k) - Q(L_j, R_k))^2 \right]^{\frac{1}{2}} \quad (8-9c)$$

Like  $P_M(R_k)$ , a positive  $Q_M(R_k)$  represents the preference of enhanced speech by System A-2 over the noisy speech averaged over the five sentences used as test material and fifteen listeners. The highest value possible for  $Q_M(R_k)$  is 2.  $Q_{SD}(R_k)$  is the standard deviation of  $Q(L_j, R_k)$  and represents the variability among the listeners in their responses.

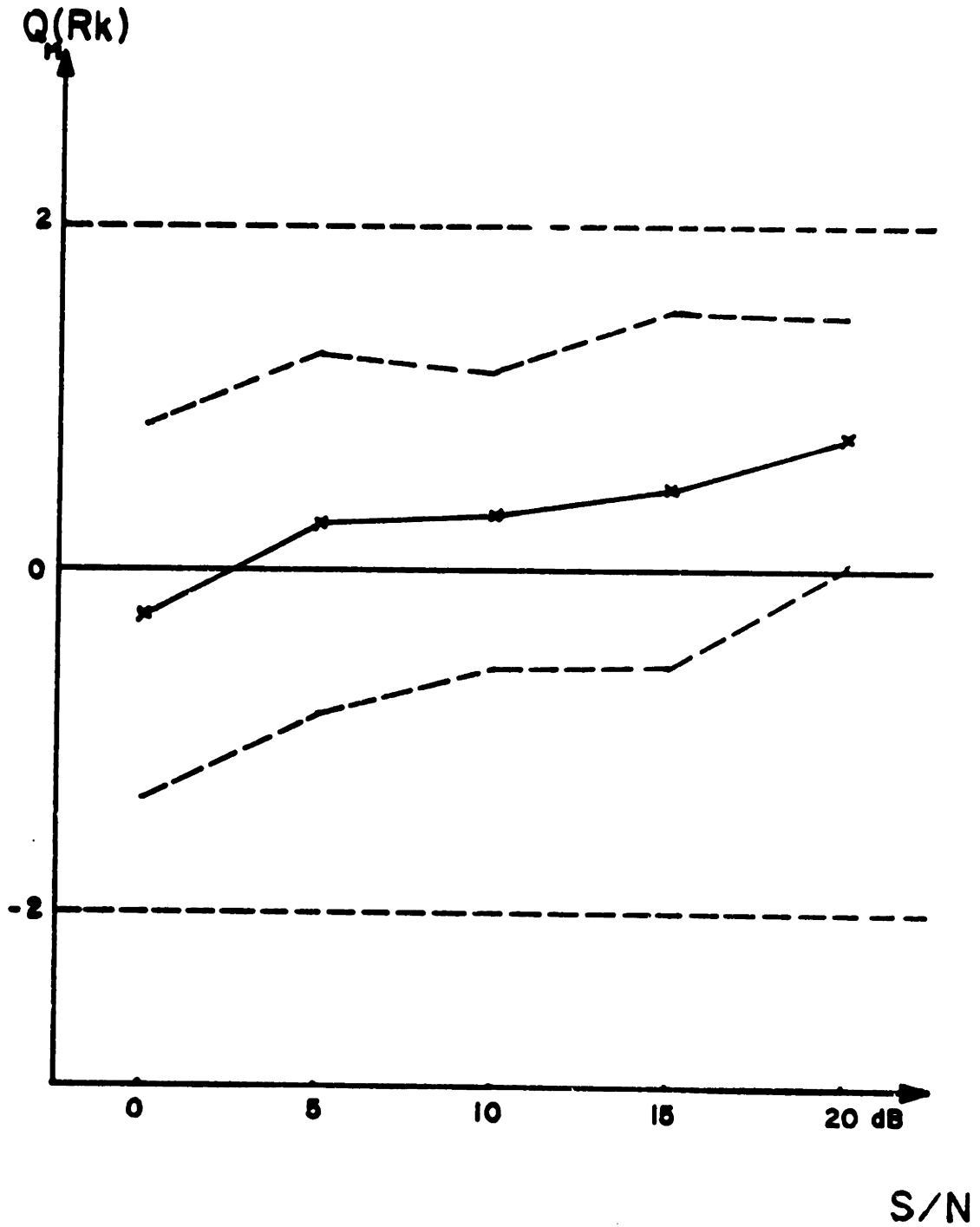
$Q_M(R_k)$  and  $Q_{SD}(R_k)$  are tabulated in Table 8.4 and plotted in Figure 8.9. The solid line in Figure 8.9 corresponds to  $Q_M(R_k)$  and the difference between the solid line and either the upper or lower dotted line corresponds to  $Q_{SD}(R_k)$ . For the same reasons discussed in Section VII.3,  $Q_M(S/N=\infty)$  would be zero and  $Q_M(S/N=-\infty)$  does not mean much.

Unlike the results of System A-2 as a potential bandwidth compression system, enhanced speech processed by System A-2 is preferred only at relatively high S/N ratios. At lower S/N ratios, the "musical tone" like background noise which arises primarily from the discontinuities of the upper formant frequencies in a frame by frame analysis scheme is sufficiently noticeable that the

Table 8.4

Results of the Speech Preference Test in Which  
System A-2 is Used as a Speech Enhancement System

S/N Ratio	$Q_M(R_k)$	$Q_{SD}(R_k)$
0 dB	-0.240	1.079
5 dB	0.240	1.023
10 dB	0.293	0.867
15 dB	0.467	1.042
20 dB	0.747	0.728



**Figure 8.9** Results of the speech preference test in which System A-2 is used as a speech enhancement system. The solid line represents  $Q_M(R_k)$ , and the distance between the solid line and the dotted line represents  $Q_{SD}(R_k)$ .



noise reduction by System A-2 does not sufficiently offset the speech degradation for some listeners. The responses of the listeners also indicate that some listeners have strong preference for processed speech while some other listeners have strong preference for unprocessed noisy speech. This is reflected by the large standard deviation shown in Figure 8.9.

In the context of this thesis, there are several methods that may be used to eliminate or mask the "musical tone" like background noise and they will be discussed in Chapter IX where various improvements are suggested for the class of systems developed in this thesis.

## VIII.5 Additional Studies

### VIII.5.1 Speech Enhancement by a Complete Analysis/Synthesis System

In the context of the work in this thesis, speech enhancement may be achieved by a complete analysis/synthesis system. To consider the feasibility of such a scheme, the speech material synthesized based on System A-2 in Section VIII.3 were compared with the enhanced speech obtained in Section VIII.4. Above the S/N ratio of about 10 dB, the enhanced speech in Section VIII.4 appeared to sound better, while below the S/N ratio of about 10 dB the opposite appeared to be true. It is difficult to

interpret this result for several reasons. The source information used in the synthesis of speech in Section VIII.3 was obtained from noise-free speech. Such an accurate source information is not available in practice. On the other hand, the source model (random noise or a train of pulses) used is a very simplified one and a more sophisticated excitation source such as voice excitation may improve the quality/intelligibility of the synthesized speech. Without further study in this area, the informal listening results imply that with the simple source model and System A-2, the approach to use the estimated  $\hat{s}_w(n)$  as enhanced speech is better than the approach to use an LPC analysis/synthesis scheme above the S/N ratio of 10 dB.

#### VIII.5.2 System A-2 as a Pre-processor for Other Bandwidth Compression Systems

As has been discussed in Chapter III, the fact that  $\underline{s}_0$  is estimated in addition to  $\underline{a}$  is important in the context of bandwidth compression of noisy speech as well as speech enhancement. This is because if we estimate only  $\underline{a}$ , then we are limited to a class of vocoding systems known as "LPC" vocoders.

As an example of using the class of systems developed in this dissertation as pre-processors for other vocoding systems, enhanced speech by System A-2 was processed by a

real time channel vocoder at Lincoln Laboratories and was compared to speech processed by the same vocoder with the unprocessed noisy speech as input. Based on informal listening, it appears that the improvement made by System A-2 is comparable to the improvement discussed in Section VIII.3 where System A-2 as a potential bandwidth compression system was compared to the conventional linear prediction analysis.

#### VIII.6 Summary

In this chapter, the three systems developed in Chapter VI have been evaluated under both an objective and subjective criteria. Under the objective criterion with the selection of the test material discussed in Section VIII.2, we conclude that all the three systems developed in Chapter VI with a proper choice of the parameters perform better than the conventional linear prediction analysis above -10 dB of the S/N ratio. Below -20 dB of the S/N ratio, none of the three systems performs any better than the conventional linear prediction analysis. Among the class of systems implemented in this dissertation, System A after two iterations performs best under the objective criterion at various S/N ratios of practical interest.

As a preliminary examination to determine whether or not the class of systems developed in this thesis have

potential to be used as bandwidth compression and speech enhancement systems of noisy speech, System A has been evaluated by a speech preference test. The results of the test indicate that System A is clearly preferred over the conventional linear prediction analysis as a potential bandwidth compression system. In the context of using System A as a speech enhancement system, the results are not as positive. However, there are a number of improvements that can be made as we will discuss in Chapter IX. Based on the evaluation performed in this chapter, we conclude that the results obtained are sufficiently encouraging to invest further research efforts in improving and evaluating the class of systems developed in this dissertation.

## CHAPTER IX FUTURE RESEARCH

### IX.1 Introduction

In this chapter, we discuss a number of areas for future research that are related to this dissertation. The areas of future research can be broadly classified into three different categories. The first category is improving the systems implemented in this thesis and is discussed in Section IX.2. The second category is issues related to adapting the systems to real world situations and is discussed in Section IX.3. The third category is the theoretical issues and systems for theoretical interest and is discussed in Section IX.4.

### IX.2 Improvements

A serious attempt has not been made in this dissertation to improve the performance of the systems implemented in this thesis. A few simple modifications may improve the performance of the systems developed. In this section, such modifications are discussed.

To indicate some potential areas in which some improvement can be made, three spectrograms are shown in Figures 9.1, 9.2 and 9.3. Figure 9.1 represents the spectrogram of noise-free speech that corresponds to "Line up at the screen door". Figure 9.2 represents the spectrogram of synthesized speech by the conventional LPC method at the S/N ratio of 0 dB. Figure 9.3 represents



Figure 9.1 Spectrogram of an English sentence "Line up at the screen door" spoken by a male speaker. The spacing between two consecutive lines on the vertical axis corresponds to 1 kHz.

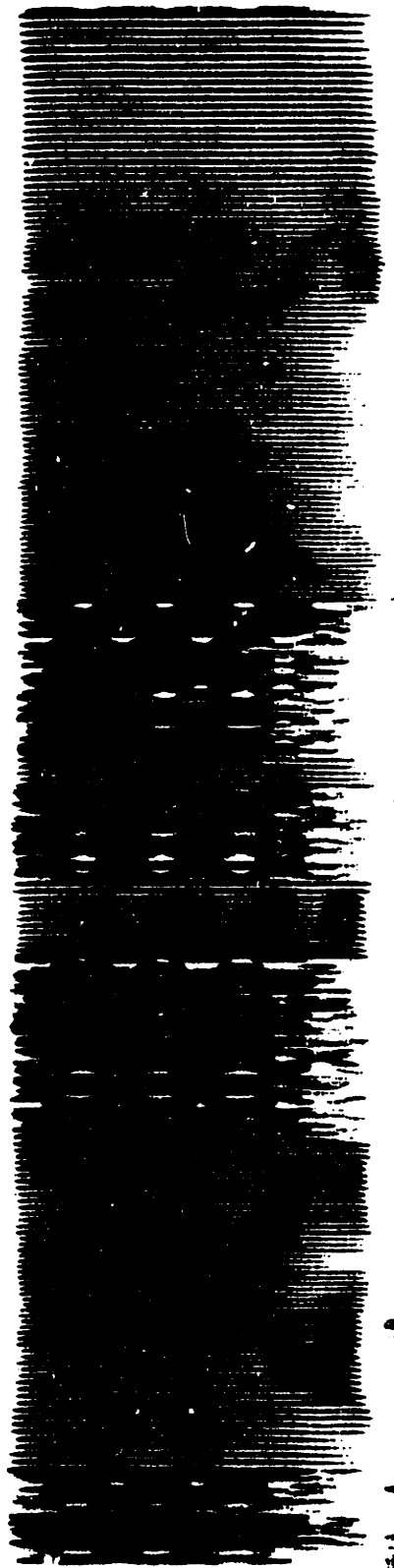


Figure 9.2 Spectrogram of speech synthesized by the conventional LPC analysis/synthesis system at the  $S/N = 0$  dB for the same English sentence shown in Figure 9.1. The spacing between two consecutive lines on the vertical axis corresponds to 1 kHz.

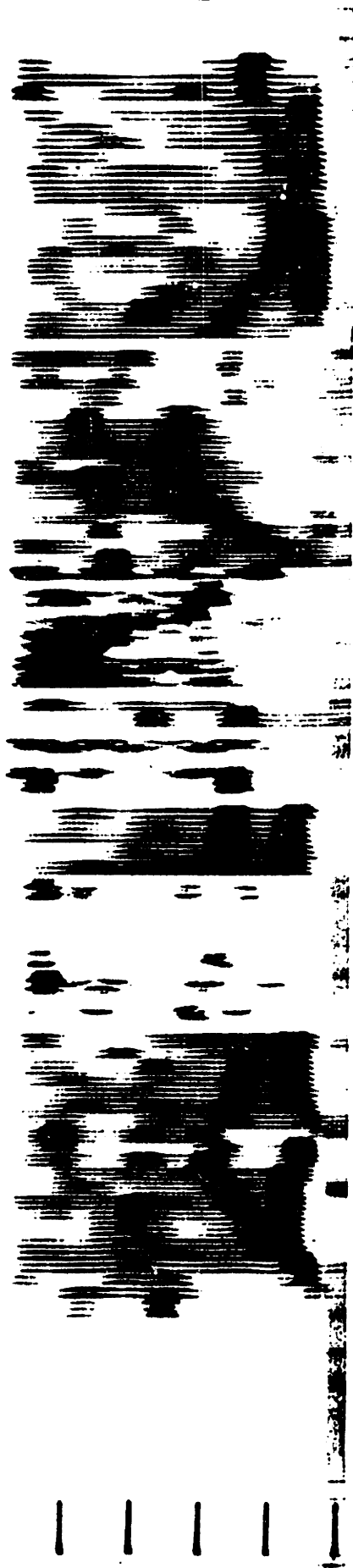


Figure 9.3 Spectrogram of speech synthesized by System A as a potential bandwidth compression system at the  $S/N = 0$  dB for the same English sentence shown in Figure 9.1. The spacing between two consecutive lines on the vertical axis corresponds to 1 kHz.



the spectrogram of the synthesized speech by System A-2 as a potential bandwidth compression system at the S/N ratio of 0 dB. Comparing Figures 9.1 and 9.3, it is clear that there are at least two main problems that cause speech degradation in the process of reducing the background noise. One of them is the non-smooth formant transitions. This problem occurs when the formant frequencies of speech change relatively fast and the frame rate is low in a frame by frame analysis environment. Such a problem may cause some speech degradation and can be solved by a higher frame rate with some optimization of the analysis window length, window type, or an interpolation scheme between frames in the synthesis. The second problem which is more serious arises due to the errors made by System A in estimating the formant frequencies. Such errors cause discontinuities in the formant frequencies, and occur more often in the higher formants where the local S/N ratio is relatively low. Such formant discontinuities are probably the primary cause of the "musical tone" like background noise discussed in Chapter VIII. In the remainder of this section, several ways that may solve or reduce the effect of the formant discontinuity problem are discussed.

#### IX.2.1 Incorporation of A Priori Information

In the theoretical results developed in this disserta-

tion, it is possible to incorporate a priori information of  $\underline{a}$ . One potential source from which some a priori information can be obtained is from the nearby analysis frames. Since the human vocal tract can not move arbitrarily fast, the results of one analysis frame are in some sense correlated with the results of the next analysis frame except at rapid onset or change. One way to incorporate the results of the past analysis frames in the analysis of the current frame is to determine  $p(\underline{a})$ , the a priori density of  $\underline{a}$ , in terms of the results of the previous analysis frame.

Some very preliminary experiment in which  $p(\underline{a})$  is assumed to be  $N(\bar{\underline{a}}, P_0)$  where  $\bar{\underline{a}}$  is the estimated  $\hat{\underline{a}}$  in the previous analysis frame and  $P_0$  is  $\sigma^2 \cdot I$  for some  $\sigma^2$  indicates that adding some a priori information from the previous analysis frame to the current analysis frame can reduce the "musical tone" like background noise. Some optimization in the choice of  $\bar{\underline{a}}$  and  $P_0$  may lead to some noticeable improvement.

### IX.2.2 Smoothing Formant Frequencies

One effect of adding some a priori information in a manner discussed in Section IX.2.1 is smoothing the estimated all pole coefficients  $\underline{a}$  of the individual analysis frames. Even though such a method to some extent leads to indirectly smoothing the formant frequencies and

thus eliminating the formant discontinuities, a more direct way would be to smooth the formant frequencies themselves. Such a direct procedure has the additional advantage that the formant frequencies can be smoothed discriminately. More specifically, in the white background noise environment the upper formant frequencies are degraded more often than the lower formant frequencies and therefore it may be desirable to smooth only the upper formant frequencies.

Such a smoothing procedure can eliminate the discontinuities in the formant frequencies and thus may reduce the "musical tone" like background noise. Furthermore, when the S/N ratio is relatively high such that the errors in the estimation of the formant frequencies do not occur often, the smoothed formant frequencies can in fact correspond to the true formant frequencies.

### IX.2.3 Masking with Random Noise

As we discussed in Section II.2.6, in a recent study, Schwartz, et al. [19], considered a system which is a modification of System C discussed in this thesis for speech enhancement. In the process of eliminating the effect of the background noise, System C creates some artificial speech degradation. Schwartz, et al. hypothesized that such a degradation arises due to setting the estimate of  $|S_w(\omega)|^2$  to zero when  $|Y_w(\omega)|^2$  is less than

$k \cdot E[|D(\omega)|^2]$ . Therefore, in their speech enhancement system, in place of zero  $|\hat{S}_w(\omega)|^2$  was set to  $\beta \cdot E[|D_w(\omega)|^2]$  for a very small value of  $\beta$  if  $|Y_w(\omega)|^2$  is less than  $(k+\beta) \cdot E[|D_w(\omega)|^2]$ . When such a modification is made, Schwartz, et al. has found that some speech degradation due to processing which is uncomfortable to listen to disappeared.

One explanation that such a thresholding method can reduce some perceptually undesirable speech degradation is that it is a way of masking the speech degradation. Based on this explanation, then, an alternative way to mask the speech degradation which is easier to implement than the threshold method is to simply add some random noise to the processed speech. The concept of masking the artificial speech degradation is not limited to System C but can be applied to any system which generates some perceptually undesirable speech degradation. The amount of noise necessary to mask the speech degradation depends on the level of the speech degradation that is to be masked. In a very preliminary experiment, the processed speech by System A has been added with some white random noise. The reasonable level of noise added to mask the "musical tone" like background noise is about 15 dB below the original background noise level. If the processing suggested in Sections IX.2.1 or IX.2.2 is carried out successfully and thus reduce the level of the "musical tone"

like background noise, then it is expected that even a lower level random noise than 15 dB below the original noise level may be able to mask the perceptually unpleasant speech degradation due to processing by System A-2. Further, if the speech degradation occurs primarily in the higher frequency regions in which the local S/N ratio is relatively low, then adding high pass filtered noise may be more desirable. A further study should be carried out in determining the proper noise level and the type of noise necessary to mask the speech degradation that occurs by processing noisy speech with the class of systems developed in this dissertation.

### IX.3 Adaptation to Practical Problems

There are many issues which require further study in implementing the class of systems developed in this thesis in practical environments. In this section, we discuss some of these issues.

#### IX.3.1 Estimation of $P_d(\omega)$

In the systems discussed in this thesis, the power spectrum of the background noise  $P_d(\omega)$  is assumed to be known. In practice,  $P_d(\omega)$  has to be estimated from the noisy speech  $y(n)$ . If the silence intervals are to be used for the estimation of  $P_d(\omega)$ , a silence detector from the noisy speech has to be incorporated in the overall system.

A related study to the estimation of  $P_d(\omega)$  is to determine the sensitivity of the performance of the systems developed in this thesis to a possible incorrect estimation of  $P_d(\omega)$ . A system which performs well when  $P_d(\omega)$  is correctly estimated may degrade quickly as the estimated  $P_d(\omega)$  differs from the true  $P_d(\omega)$ . The sensitivity issue is an important area to be investigated.

### IX.3.2 Estimation of Source Information

To develop a complete analysis/synthesis system based on the theoretical results developed in this thesis, it is necessary to develop an algorithm that estimates the source parameters. In the context of this dissertation, we may simply apply existing pitch detectors [40,41,42] to the estimated  $\hat{s}_w(n)$ . Alternatively, there may be a more optimum way of obtaining the source information that accounts for the presence of background noise. The estimation of the source parameters from the noisy speech is an important area for future research in developing a complete analysis/synthesis system.

### IX.3.3 Evaluation of Systems

After some further study on the system improvement, it is important to evaluate the systems in terms of their performance in improving speech intelligibility, quality, etc. The choice of the system may depend on the specific

background noise environment, cost of implementation, etc.

#### IX.4 Further Theoretical Study and Related Work

##### IX.4.1 Implementation of Other Systems

In this dissertation, we considered estimating  $\underline{a}$  by maximizing  $p(\underline{a}|\underline{y}_0)$ . Since maximizing  $p(\underline{a}|\underline{y}_0)$  is a non-linear problem, we considered "sub-optimal" procedures in which  $p(\underline{a}, \underline{s}_0|\underline{y}_0)$  is maximized. An attempt to maximize  $p(\underline{a}, \underline{s}_0|\underline{y}_0)$  led to the LMAP and RLMAP algorithms which require solving only sets of linear equations in an iterative manner. Further approximations of these algorithms led to System A and System B which were implemented.

An important area of future research from a theoretical point of view is a theoretical understanding of the relations and properties of the MAP, LMAP and RLMAP algorithms, and their implementations. As we discussed in Section V.6, a theoretical study to understand the relations and properties of the three algorithms is currently in progress. The implementation of the MAP algorithm is important since the results obtained by maximizing  $p(\underline{a}|\underline{y}_0)$  are the optimum that can be achieved if we follow the philosophy that is taken in this research. The implementation of the LMAP and RLMAP algorithms is important since it allows us to understand the performance degradation due to the approximations made in developing System A and System B from the LMAP and RLMAP algorithms. It also allows us to understand

the effect of changing the problem from maximizing  $p(\underline{a}|\underline{y}_0)$  to  $p(\underline{a}, \underline{s}_0|\underline{y}_0)$ . A comparison of the MAP, LMAP, RLMAP methods, System A and System B in terms of their performances can be a basis for determining the extent of further research efforts in developing a different approximation method to the true MAP estimation procedure.

#### IX.4.2 Different Initial Estimates of $\underline{a}$

In the LMAP, RLMAP algorithms, System A and System B, we begin from some initial estimate of  $\underline{a}$ . In the systems that were implemented, the initial estimate was obtained by simply applying the correlation method of linear prediction analysis to the noisy speech. Since the LMAP and RLMAP algorithms are not guaranteed to give the global maximum of  $p(\underline{a}|\underline{y}_0)$  or  $p(\underline{a}, \underline{s}_0|\underline{y}_0)$ , other initial estimates of  $\underline{a}$  may lead to different estimates of  $\underline{a}$ .

Beginning from other initial estimates of  $\underline{a}$  can be useful in at least two different ways. First, they may lead to better estimates of  $\underline{a}$ . Second, the primary disadvantage of System B relative to System A is its slow convergence to a reasonable solution. If we begin from some other initial estimates of  $\underline{a}$ , System B may converge to a solution more quickly. This is an area for further study.



#### IX.4.3 Incorporation of A Priori Information

There are many levels in incorporating a priori information based on the knowledge that the noisy signal we deal with is speech plus noise. In one extreme, we could add some a priori information in a manner similar to the discussions in Section IX.3. In the other extreme, we may want to capitalize more fully on the physiological constraints imposed by the human vocal mechanism and even the linguistic constraints imposed by the rules of the language. Since any accurate extra information added in estimating the speech parameters can potentially lead to a better estimate, such additional knowledge may be important in dealing with the noisy speech. To understand what knowledge of speech we can capitalize on and how such knowledge can be used to estimate the speech parameters better is an important area for future research in many areas of speech processing.

#### IX.4.4 Excitation by a Train of Pulses

In the theoretical development in this dissertation, various systems were developed based on the assumption that the excitation is white Gaussian noise and we simply applied the same systems to both unvoiced and voiced sounds. If we estimate the system parameters of voiced speech based on the assumption that the excitation is a train of pulses, then a better estimate of the speech

parameters may perhaps be obtained. Since a majority of speech sounds are voiced and the voiced sounds are very important in the perception of speech, an attempt to estimate the speech parameters of voiced sounds more accurately appears attractive. The notion to capitalize on the periodicity of voiced sounds is also related to the incorporation of more knowledge of speech in estimating the speech parameters.

#### IX.4.5 Pole-Zero Modelling

In the theoretical development in this thesis, we have assumed an all pole transfer function in the underlying speech model. In a stationary background noise environment, the low energy speech segments such as unvoiced speech degrade more quickly due to the relatively low S/N ratio and thus are probably an important factor in decreasing speech intelligibility. Since unvoiced speech can be better modelled by a pole-zero than an all pole transfer function, the approach to use a pole-zero system may lead to a better performance and it is an important area for future research.

## CHAPTER X CONCLUSION

In this thesis, the problem of enhancement and bandwidth compression of noisy speech was formulated as a parameter estimation problem, in which we attempted to estimate the parameters of an underlying speech model from the noisy speech based on the MAP estimation procedure. Such an approach led to two algorithms which require solving sets of linear equations in an iterative manner. Some approximations of the two algorithms led to two systems which are computationally simpler than the two algorithms by taking advantage of a high speed FFT algorithm.

As a preliminary investigation into the performance of the two systems developed in this thesis, the two systems were implemented and applied to both real and synthetic speech data. An objective and informal subjective evaluation indicate that the systems implemented perform well as enhancement and potential bandwidth compression systems of noisy speech.

A number of studies were suggested for future research in this thesis. They include various improvements and further evaluation of the systems implemented in this thesis, implementation and evaluation of other systems developed but have not been implemented in this thesis and development of new systems by incorporating more knowledge of speech.

REFERENCES

- [1] G. A. Miller, G. A. Heise, and W. Lichten, "The Intelligibility of Speech as a Function of the Context of the Test Materials," J. Exp. Psychol., vol. 41, pp. 329-335, 1951.
- [2] B. Gold, "Robust Speech Processing," M.I.T. Lincoln Lab., Technical Note 1976-6, DDC AD-A012P99/0, Jan. 27, 1976
- [3] M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise," IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-24, pp. 488-494 Dec. 1976.
- [4] Y. M. Perlmutter, L. D. Braida, R. H. Frazier, and A. V. Oppenheim, "Evaluation of a Speech Enhancement System," Proc. 1977 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 212-215, May 9-11, 1977.
- [5] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition," IEEE Trans. Acoustics, Speech and Signal Processing, to be published.
- [6] J. S. Lim, "Evaluation of Autocorrelation Subtraction Method for Enhancing Speech Degraded by Additive White Noise," submitted to IEEE Trans. Acoustics, Speech and Signal Processing.
- [7] J. D. Markel and A. H. Gray, Jr., Linear Prediction of Speech, Berlin, Heidelberg, New York: Springer-Verlag, 1976.
- [8] F. Itakura and S. Saito, "Analysis Synthesis Telephony Based on the Maximum Likelihood Method," Rep. 6th Int. Congr. Acoustics, Y. Kohasi, Ed., Paper C-5-5, pp. C17-C20, August 1968.
- [9] J. D. Gibson, J. L. Melsa and S. K. Jones, "Digital Speech Analysis Using Sequential Estimation Techniques," IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-23, pp. 362-369, August 1975.
- [10] V. C. Shields, Jr., "Separation of Added Speech Signals by Digital Comb Filtering," S.M. Thesis, Dept. of Elec. Eng., M.I.T., 1970.

- [11] R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim, "Enhancement of Speech by Additive Filtering," Proc. 1976 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 251-253, April 12-14, 1976.
- [12] T. W. Parsons, "Separation of Simultaneous Vocalic Utterances of Two Talkers," Ph.D. Thesis, Polytechnic Institute of New York, June 1975.
- [13] James L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd Edition, New York, Heidelberg, Berlin: Springer-Verlag, 1972.
- [14] D. T. Magill and C. K. Un, "Wide-band Noise Reduction of Noisy Speech," 92nd Meeting of Acoustical Soc. of Amer., paper # RR8, November 15-19, 1976.
- [15] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Study and Development of the INTEL Technique for Improving Speech Intelligibility," Nicolet Scientific Corp., Final Rep. NSC-FR/4023, Dec. 1974.
- [16] S. F. Boll, "Application of the SABER Method for Improved Spectral Analysis of Noisy Speech," Technical Report UUCS-77-107, Computer Science Dept., University of Utah, August 1977.
- [17] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," submitted to IEEE Trans. on Acoustics, Speech and Signal Processing, 1978.
- [18] J. S. Lim, "Estimation of a Speech Model Parameters from Speech Waveforms Degraded by Additive Random Noise," Ph.D. Thesis Proposal submitted to the Department of Elec. Eng. and Comp. Sci., M.I.T., August 1977.
- [19] R. Schwartz and M. Berouti, BBN, personal communications, May 1978.
- [20] J. Suzuki, "Speech Processing by Splicing of Autocorrelation Function," Proc. 1976 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 113-116, April 12-14, 1976.

- [21] J. Suzuki, "Speech Processing System by use of Short-time Crosscorrelation Function," Proc. 1977 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 24-27, May 9-11, 1977.
- [22] H. L. Van Trees, Detection, Estimation and Modulation Theory, New York: Wiley, 1968.
- [23] M. W. Callahan, "Acoustic Signal Processing Based on the Short-time Spectrum," Ph.D. dissertation, Computer Science Dept., University of Utah, March 1976.
- [24] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoustical Soc. Amer., vol. 50, pp. 637-655, 1971.
- [25] J. D. Markel and A. H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based upon the Autocorrelation Method," IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-22, pp. 124-134, April 1974.
- [26] A. V. Oppenheim, "Speech Analysis-Synthesis System Based on Homomorphic Filtering," J. Acoustical Soc. Amer., vol. 45, pp. 459-462, 1969.
- [27] J. S. Lim, "Spectral Root Homomorphic Deconvolution System," submitted to IEEE Trans. Acoustics, Speech and Signal Processing, 1978.
- [28] P. Eykhoff, System Identification: Parameter and State Estimation, New York: Wiley, 1974.
- [29] J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, vol. 63, pp. 561-580, April 1975.
- [30] N. Levinson, "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction," J. Math. Phys., vol. 25, pp. 261-278, 1947.
- [31] A. Gelb, et al., Applied Optimal Estimation, A. Gelb, Ed., Cambridge, Mass.: M.I.T. Press, 1974.
- [32] F. C. Schwappe, Uncertain Dynamic Systems, Englewood Cliffs, New Jersey: Prentice-Hall, 1973.
- [33] D. Q. Mayne, "A Solution of the Smoothing Problem for Linear Dynamic Systems," Automatica, vol. 4, pp. 73-92, 1966.

- [34] D. C. Fraser and J. E. Potter, "The Optimum Linear Smoother as a Combination of Two Optimum Linear Filters," IEEE Trans. Aut. Contr., August 1969.
- [35] R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization," Computer J., vol. 6, pp. 163-168, 1963.
- [36] R. Fletcher, "Function Minimization without Evaluating Derivatives--A Review," Computer J., vol. 8, pp. 31-41, 1965.
- [37] W. J. D. Powell, "A Survey of Numerical Methods for Unconstrained Optimization," SIAM Review, vol. 12, pp. 79-97, 1970.
- [38] B. Musicus, "An Iterative Technique on Maximum Likelihood Parameter Estimation on Noisy Data," S.M. Thesis, M.I.T., to be submitted, December, 1978.
- [39] A. V. Oppenheim and R. W. Schafer, Digital Signal Processing, Englewood Cliffs, New Jersey: Prentice-Hall, 1975.
- [40] A. M. Noll, "Cepstrum Pitch Determination," J. Acoustical Soc. Amer., vol. 41, pp. 293-309, 1967.
- [41] J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. Audio Electroacoust., vol. AU-20, pp. 367-377, 1972.
- [42] B. Gold and L. R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoustical Soc. Amer., vol. 46, pp. 442-448, 1969.

APPENDIX

In Appendix I, we summarize briefly the notations that have been used in the thesis. In Appendix II, a table of  $\overline{\text{LCSE}}$  and Normalized  $\overline{\text{LCSE}}$  which were discussed in Section VIII.2 is shown.



APPENDIX I SUMMARY OF NOTATIONS

$\underline{a}$ :  $(a_1, a_2, \dots, a_p)^T$ , an all pole coefficient vector,  
 T represents transpose of a matrix

$\bar{\underline{a}}$ : a priori mean of  $\underline{a}$

$\hat{\underline{a}}_i$ :  $i$ th estimate of  $\underline{a}$

$A(\omega)$ :  $F[\alpha(n)]$ , discrete time Fourier transform of  $\alpha(n)$

A:

$$\begin{bmatrix} 0 & a_1 & a_2 & \dots & a_p & 0 & & \\ & 0 & a_1 & a_2 & & a_p & 0 & \underline{0} \\ & & 0 & a_1 & & & & 0 \\ & & & 0 & a_1 & & & a_p \\ & & & & 0 & & & a_2 \\ & \underline{0} & & & & & & a_1 \end{bmatrix}$$

$A_I$ :

$$\begin{bmatrix} a_p & & \underline{0} & & \\ & a_{p-1} & & & \\ \vdots & & & & \\ a_1, \dots, a_{p-1}, & & & & a_p \end{bmatrix}$$

$B(\omega)$ :  $F[\beta(n)]$ , discrete time Fourier Transform of  $\beta(n)$

$d(n)$ : disturbance or background noise; assumed to be generated by a Gaussian random process

$d_w(n)$ :  $d(n) \cdot w_g(n)$ , windowed background noise

$\underline{d}(n_1, n_2)$ :  $(d(n_1), \dots, d(n_2))^T$

$\underline{d}_0$ :  $\underline{d}(N-1, 0)$ , a vector of background noise

DFT[x(n)]:  $X(k) = \sum_{n=0}^{M-1} x(n) \cdot e^{-j\frac{2\pi}{M} k \cdot n}$ , M point Discrete

Fourier Transform of x(n)

E[x]: expected value of x

$\epsilon_p$ : error function to be minimized

F[x(n)]:  $X(\omega) = \sum_{n=-\infty}^{\infty} x(n) \cdot e^{-j\omega n}$ , discrete time Fourier

Transform of x(n)

$F^{-1}[X(\omega)]$ :  $x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) \cdot e^{j\omega n} \cdot d\omega$ , the inverse discrete

time Fourier Transform of X(ω)

g: gain factor

H(z): z transform of the transfer function in the underlying speech model

IDFT[X(k)]:  $x(n) = \frac{1}{M} \sum_{k=0}^{M-1} X(k) \cdot e^{j\frac{2\pi}{M} k \cdot n}$ , the Inverse Discrete

Fourier Transform of X(k)

$\underline{k}(n)$ : Kalman filter gain

$\underline{m}_s$ :  $(I-A)^{-1} \cdot A_I \cdot \underline{s}_I$

$\underline{m}$ : mean of  $\underline{s}_0$  conditioned on  $\underline{a}$  and  $\underline{y}_0$

N(A,B): Gaussian with mean of A and covariance of B

$P_0$ : a priori covariance of  $\underline{a}$

$P_x(\omega)$ :  $\sum_{n=-\infty}^{\infty} R_x(n) \cdot e^{j\omega n}$ , the power spectral density of  $x(n)$

$p(A_0)$ : probability density function of  $A$ , or probability density function evaluated at  $A=A_0$ ; see footnote 2

$p(A_0|B_0)$ : analogous to  $p(A_0)$  with the conditional density function

$R_x(n)$ :  $E[x(k) \cdot x(k-n)]$  for a stationary signal  $x(n)$ , or correlation of  $x(n)$

$R_s$ :  $g^2 \cdot (I-A)^{-1} \cdot ((I-A)^{-1})^T$

$R_d$ :  $E[\underline{d}_0 \cdot \underline{d}_0^T]$

$s(n)$ : signal or speech

$s_w(n)$ :  $s(n) \cdot w_s(n)$ , windowed speech

$\underline{s}(n_1, n_2)$ :  $(s(n_1), \dots, s(n_2))^T$

$\underline{s}_0$ :  $\underline{s}(N-1, 0)$

$\hat{\underline{s}}_{0i}$ : the  $i$ th estimate of  $\underline{s}_0$

$\underline{s}_I$ :  $\underline{s}(-1, -p)$

$S(\omega)$ :  $\sum_{n=-\infty}^{\infty} s(n) \cdot e^{-j\omega n}$ , discrete time Fourier Transform of  $s(n)$

$|S(\omega)|$ : magnitude of  $S(\omega)$

$\angle S(\omega)$ : phase of  $S(\omega)$ , also denoted as  $\langle S(\omega)$

$u(n)$ : a pulse train or random noise excitation

$\underline{u}(n)$ : an excitation vector, typically zero mean white Gaussian noise

$\text{Var}[x]$ : variance of  $x$

$\underline{v}(n)$ : an observation vector, typically zero mean white Gaussian noise

$\underline{V}$ : covariance of  $\underline{s}_0$  conditioned on  $\underline{a}$  and  $\underline{y}_0$

$w(n)$ : zero mean white Gaussian noise with unit variance

$w_s(n)$ : a smooth window function

$\underline{x}(n)$ : a state vector

$\underline{x}(-1)$ : the initial state vector

$\hat{x}_{ML}$ : Maximum Likelihood estimate of  $x$

$\hat{x}_{MAP}$ : Maximum A posteriori estimate of  $x$

$\hat{x}_{MMSE}$ : Minimum Mean Square Error estimate of  $x$

$y(n)$ :  $s(n)+d(n)$ , noisy signal or noisy speech

$y_w(n)$ :  $y(n) \cdot w_s(n)$ , windowed noisy speech

$\underline{Y}(n_1, n_2)$ :  $(y(n_1), \dots, y(n_2))^T$

$\underline{Y}_0$ :  $\underline{Y}(N-1, 0)$

$Y(\omega)$ :  $\sum_{n=-\infty}^{\infty} y(n) \cdot e^{j\omega n}$ , the discrete time Fourier Transform of  $y(n)$

$|Y(\omega)|$ : magnitude of  $Y(\omega)$

$\angle Y(\omega)$ : phase of  $Y(\omega)$ , also denoted as  $\langle Y(\omega)$

$\underline{z}(n)$ : an observation vector

$\phi_x(n)$ :  $\sum_{k=-\infty}^{\infty} x_w(k) \cdot x_w(k-n)$ , the short time correlation of  $x(n)$

$\Phi_x(\omega)$ :  $F[\phi_x(n)]$

$$\phi_x^*(n) : \sum_{\ell=n_0}^{n_0+M-1} x(\ell) \cdot x(\ell-n), \text{ another definition of the}$$

short time correlation of  $x(n)$ ; note that

$$\phi_x(n) \neq \phi_x^*(n)$$

$$\theta(\omega) : F[\theta(n)]$$

$$\Gamma(\omega) : F[\gamma(n)]$$



BIOGRAPHICAL NOTE

Jae S. Lim was born on December 2nd, 1950 and raised until the 11th year of High School in Taegu, Korea. He came to the United States in 1967 and graduated from Mahwah High School, Mahwah, New Jersey, in 1968. After his sophomore year at M.I.T., he served in the Korean army for three years and returned to M.I.T. in 1973. He received the S.B., S.M., and E.E. degrees in Electrical Engineering and Computer Science from M.I.T. in 1974, 1975, and 1978 respectively.

From September, 1974 to June, 1975 and from June, 1976 to August, 1976, he was supported by the M.I.T. Research Laboratory of Electronics Industrial Fellowship. From September, 1975 to June, 1976, he was a Teaching Assistant in the courses "Signals and Systems" and "Digital Signal Processing". From September, 1976, he was supported by a Research Assistantship.

Mr. Lim is a member of Eta Kappa Nu and Sigma Xi.