

26

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

OCT 29 1997

LIBRARIES

**Low Power Digital Filtering
Using Adaptive Approximate Processing**

by

Jeffrey Thomas Ludwig

S.B., Massachusetts Institute of Technology (1991)
S.M., Massachusetts Institute of Technology (1993)

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

Doctor of Philosophy

at the

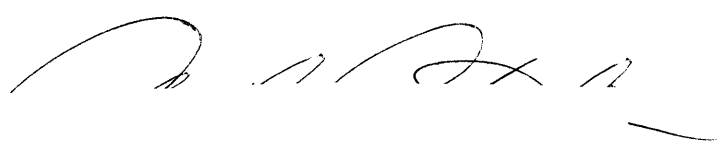
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1997

© 1997 Massachusetts Institute of Technology. All rights reserved.

Author: _____
Department of Electrical Engineering and Computer Science
September 2, 1997

Certified by: _____
S. Hamid Nawab
Associate Professor of Electrical and Computer Engineering
Thesis Supervisor



Accepted by: _____
A. C. Smith
Chair, Department Committee on Graduate Students

Low Power Digital Filtering Using Adaptive Approximate Processing

by

Jeffrey Thomas Ludwig

Submitted to the Department of Electrical Engineering and Computer Science
on September 2, 1997, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Techniques for reducing power consumption in digital circuits have become increasingly important because of the growing demand for portable multimedia devices. Digital filters, being ubiquitous in such devices, are a prime candidate for low power design. We present a new algorithmic approach to low power frequency-selective digital filtering which is based on the concepts of adaptive approximate processing. This approach is formalized by introducing the class of approximate filtering algorithms in which the order of a digital filter is dynamically varied to provide time-varying stopband attenuation in proportion to the time-varying signal-to-noise ratio (SNR) of the input signal, while maintaining a fixed SNR at the filter output. Since power consumption in digital filter implementations is proportional to the order of the filter, dynamically varying the filter order is a strategy which may be used to conserve power. We construct a framework to explore the statistical properties of approximate filtering algorithms and show that under certain assumptions, the performance of approximate filtering algorithms is asymptotically optimal. We investigate the transient effects of dynamically varying the order of a digital filter by developing deterministic and probabilistic frameworks for state transition error analysis. Approximate filter structures using FIR and IIR digital filter constituent elements are explored and shown to be an important element in the characterization of approximate filtering algorithms. Experiments involving the filtering of FDM speech signals are used to demonstrate the practical viability of approximate filtering algorithms in the context of low power signal processing.

Thesis Supervisor: S. Hamid Nawab

Title: Associate Professor of Electrical and Computer Engineering

Acknowledgments

I would like to express my sincere gratitude to Professor S. Hamid Nawab for his academic guidance, limitless support, and friendship during my four years in the Digital Signal Processing Group. It was a great honor and pleasure to have the opportunity to explore teaching, research, and the slippery surfaces of life together.

I am especially grateful to Professor Alan Oppenheim for his inspirational leadership, financial commitment, and faith in me. For his encouragement and rewarding collaboration, I thank Professor Anantha Chandrakasan. I also wish to acknowledge Professor Gregory Wornell for his valuable participation as a reader of this thesis. I thank Dr. Andrew Singer whom I have looked up to as a mentor and friend for 10 years. Andy's energy, endurance, and quest for excellence have served as inspiration to me for as long as I have known him. I thank Barry Blesser for being Barry Blesser. And I thank my parents for their uncommon love and support.

I would like to thank wholeheartedly James Ooi and Shawn Verbout, partners and best friends, for unforgettable hours of brainstorming and enthusiastic whiteboard sessions in a small room with no windows. Thank you for being strong enough to openly share with me your ideas and dreams and for honestly listening to mine. And finally, I thank you Christina for picking me up and carrying me on your back during some very confusing months of my life.

Contents

1	Introduction	15
1.1	Approximate Signal Processing	16
1.2	Low Power Design Methodologies	17
1.2.1	Motivation	18
1.2.2	Sources of Power Consumption	19
1.2.3	Summary of Previous Research	21
1.3	Outline of Thesis	23
2	Approximate Filtering Algorithms	27
2.1	Introduction	27
2.1.1	Overview	28
2.1.2	Approximate Filter Structures	29
2.2	Problem Statement	31
2.2.1	Summary of Low Power Approach	34
2.2.2	Summary of Maximum Likelihood Approach	36
2.3	Derivation of Low Power Solution	38
2.3.1	Low Power Estimation	44
2.3.2	Convergence Analysis	50
2.3.3	Numerical Example	55
2.4	Derivation of Maximum Likelihood Solution	57
2.4.1	Maximum Likelihood Estimation	58
2.4.2	Numerical Example	61
2.5	Adaptation for Non-Stationary Inputs	63
2.6	Summary	65

3	State Transition Error Analysis	67
3.1	Deterministic Analysis	70
3.1.1	Derivation of Deterministic Bound	72
3.1.2	Simulations	85
3.2	Probabilistic Analysis	89
3.2.1	Preliminaries	91
3.2.2	Derivation of Probabilistic Bounds	96
3.3	Considerations for Truncation Filter Structures	103
3.3.1	Deterministic Analysis	103
3.3.2	Probabilistic Analysis	112
3.4	Summary	120
4	Approximate Filter Structures	121
4.1	Replacement Filter Structures	128
4.1.1	Type FR Filter Structures	128
4.1.2	Type IR Filter Structures	135
4.2	Truncation Filter Structures	138
4.2.1	Type FT Filter Structures	140
4.2.2	Type IT Filter Structures	141
4.3	Summary and Future Directions	147
5	Experiments and Applications	151
5.1	Speech Signal Processing	151
5.2	Interpolation and Decimation	155
5.3	Future Directions	156
6	Conclusion	159
A	All-Pole Filter Matrix	161
B	Bound on Vector Norm	163
C	Autoregressive Parameter Values	165

List of Figures

2-1	An overview of approximate filtering. The adaptation strategy for updating the filter order after each new set of L output samples is defined by the decision module D	29
2-2	Magnitude-squared frequency responses for truncations of a 20th-order Butterworth filter with 3, 5, 7, 9, and 10 second-order sections. The half-power frequency of the Butterworth filters is $\pi/2$	31
2-3	Performance profile for the Parks-McLellan FIR replacement filter structure. The stopband is defined as $\omega \in [5\pi/8, \pi]$	47
2-4	Performance profile for the Butterworth IIR truncation filter structure. The stopband is defined as $\omega \in [5\pi/8, \pi]$	48
2-5	Solid curves represent input SNR estimates as a function of L and N_0 . The actual SNR for the input signal is 0.07. The straight dotted lines indicate the partitioning of the input SNR space by optimal values for the number of filter sections to use in order to obtain an output SNR of at least 1000. . .	52
2-6	Power spectral density of the 30th-order AR process which is used as the input signal in the numerical example.	55
2-7	Histograms of the LP input SNR estimates for $L = 5000, 1000, 100,$ and 50 . Each histogram represents the results of 100 Monte Carlo simulations. . . .	57
2-8	An overview of the concept of approximate filtering. The adaptation strategy for updating the filter order after each new set of L output samples is defined by the decision module D	63

3-1	Comparison of the approximate filter output $y_{N_1N_2}[n]$ and the fixed filter output $y_{N_2N_2}[n]$. The bottom plot depicts the absolute value of the state transition error $ y_{tr}[n] = y_{N_1N_2}[n] - y_{N_2N_2}[n] $. The filter order switch (state transition) occurred at time $n = 0$ in this case. In this illustrative example we used $N_1 = 2$ and $N_2 = 4$. The output signals were generated using the replacement IIR Butterworth filter structure.	70
3-2	A plot of the deterministic bound $B_{N_1N_2}$ vs. N_1 and N_2 for the replacement Butterworth filter structure. Refer to Eq. (3.12) for the definition of $B_{N_1N_2}$.	74
3-3	A plot of the deterministic bound $B_{N_1N_2}$ vs. N_1 and N_2 for the truncated Butterworth filter structure. Refer to Eq. (3.12) for the definition of $B_{N_1N_2}$.	75
3-4	A plot of the maximum value of the state transition error vs. N_1 and N_2 for $n \geq 0$ using the replacement Butterworth filter structure with half-power frequency $\pi/2$. In this example $B_X = 1$, or, equivalently, $ x[n] \leq 1$ for all n .	86
3-5	A plot of the maximum value of the state transition error vs. N_1 and N_2 for $n \geq 0$ using the truncated Butterworth filter structure with half-power frequency $\pi/2$. In this example $B_X = 1$, or, equivalently, $ x[n] \leq 1$ for all n .	87
3-6	A plot of the maximum value of the state transition error vs. N_1 and N_2 for $n \geq 0$ using the replacement Butterworth filter structure with half-power frequency $\pi/2$ and zero initial conditions. In this example $B_X = 1$, or, equivalently, $ x[n] \leq 1$ for all n	88
3-7	A plot of the maximum value of the state transition error vs. N_1 and N_2 for $n \geq 0$ using the truncated Butterworth filter structure with half-power frequency $\pi/2$ and zero initial conditions. In this example $B_X = 1$, or, equivalently, $ x[n] \leq 1$ for all n	89
3-8	Logarithmic plot of the normalized squared STE bound B_{tr}^2 vs. post-transition filter order N_2 . The curves were generated using the replacement Butterworth filter structure with $N_1 = 12$. The half-power frequency is $\pi/2$	90
3-9	Logarithmic plot of the normalized squared STE vs. post transition sample number for $N_2 = 2, 4, 6, 8,$ and 10 . The curves were generated using the replacement Butterworth filter structure with $N_1 = 12$. The half-power frequency is $\pi/2$	91

3-10	A plot of the normalized squared STE bound (plotted with '+'), along with the average (plotted with 'x'), maximum (plotted with '-'), and minimum (plotted with '-') values of the normalized squared STE computed by generating 1000 Monte Carlo simulations. The experiment used the replacement Butterworth filter structure, and parameter values $N_1 = 2$, $N_2 = 4$, and $B_X = 1$. The half-power frequency is $\pi/2$	92
3-11	A plot of the output power noise-to-signal ratio (OPNSR) for various values of the power window length L . Actual values of the OPNSR (denoted in the plot by 'x') based on 2000 Monte Carlo simulations are plotted along with the predicted theoretical values (denoted in the plot by '-'). The predicted theoretical values of the OPNSR are given by Eq. (3.89).	99
3-12	A plot of the function $\text{ONSR}[n]$ given by Eq. (3.90). Actual experimental values of the time series $\text{ONSR}[n]$ (denoted in the figure by 'x') based on 2000 Monte Carlo simulations are plotted along with the predicted theoretical values (denoted in the figure by 'o').	101
3-13	Truncation approximate filter structure for an IIR digital filter. The annotated variables are used in the STE analysis.	105
4-1	Frequency response magnitudes for rectangularly-windowed ideal FIR filters of orders $N = 20, 80$ and 140	124
4-2	An overview of the FIR approximate filter structure.	125
4-3	Conceptual diagram of the IIR replacement filter structure.	126
4-4	Conceptual diagrams of the IIR truncation filter structure: (a) the signal flow graph, and (b) the clocked shift register block diagram.	127
4-5	Performance profile for the eigenfilter FIR replacement filter structure.	132
4-6	Performance profile for the Parks-McLellan FIR replacement filter structure.	133
4-7	Frequency response magnitude-squared plots for the Butterworth IIR replacement filter structure.	136
4-8	Frequency response magnitude-squared plots for the Chebyshev IIR replacement filter structure.	137
4-9	Frequency response magnitude-squared plots for the inverse Chebyshev IIR replacement filter structure.	138

4-10	Frequency response magnitude-squared plots for the IIR elliptic replacement filter structure.	139
4-11	Comparison of the performance profiles for the Butterworth, Chebyshev, inverse Chebyshev, and elliptic IIR replacement filter structures.	140
4-12	Comparison of the OPNSR for the Butterworth, Chebyshev, inverse Chebyshev, and elliptic IIR replacement filter structures.	141
4-13	A comparison of the performance profiles for the Parks-McLellan FIR truncation and replacement filter structures. The truncated filters of orders $3 \leq N \leq 63$ were obtained by symmetrically truncating the coefficients of the order-64 Parks-McLellan filter.	142
4-14	A comparison of the performance profiles for the FIR truncation and replacement filter structures. The truncated filters of orders $3 \leq N \leq 63$ were obtained by symmetrically truncating the coefficients of the order-64 rectangularly-windowed ideal filter.	143
4-15	A comparison of the performance profiles for the eigenfilter FIR truncation and replacement filter structures. The truncated filters of orders $3 \leq N \leq 63$ were obtained by symmetrically truncating the coefficients of the order-64 eigenfilter.	144
4-16	Performance profile for truncations of a 20th-order Butterworth filter with half-power frequency $\pi/2$	145
4-17	Magnitude-squared frequency responses for truncations of a 6th-order Butterworth filter with 1, 2, and 3 second-order sections, for each of the possible distinct truncation filter structures defined in Eqs. (4.35)–(4.47). The optimal truncation filter structure is $\mathcal{H}_T^* = \mathcal{H}_T^2$	148
5-1	Demultiplexing of FDM speech using low power frequency selective filtering. (a) Passband speech, (b) stopband speech, and (c) number of filter sections as a function of sample number.	152
5-2	Evolution of filter order for the FDM speech signal processing example. Overlays of the approximate filter order and the stopband power in the input signal over time are shown. The approximate filter order clearly traces the envelope of the stopband power in the input signal.	154
5-3	Filter performance vs. percentage silence in stopband signal.	155

5-4	A conventional fixed-order FIR filter bank and associated spectral decomposition of the filter bank output signals.	157
5-5	An FIR filter bank which has incorporated approximate filters, resulting in an <i>approximate filter bank</i> . The filter order evolution for the first lowpass filter has been enclosed in the dashed box. The filter order can be seen to follow the energy in the input's stopband component $x_s[n]$, which is shown in the bottom plot.	158

List of Tables

- 2.1 Summary of the results of the numerical example using the LP estimator in which the true value of $N^* = 10$ and the true input SNR is 0.4831. The results are tabulated for power window lengths of $L = 50, 100, 500, 1000,$ and 5000. 56
- 2.2 Summary of the results of the numerical example using the ML estimator in which the true value of $N^* = 10$ and the true input SNR is 0.4831. The results are tabulated for power window lengths of $L = 50, 100, 500, 1000,$ and 5000. 62
- 3.1 Numerical values for the deterministic bound $B_{N_1 N_2}$ for the replacement Butterworth filter structure. Refer to Eq. (3.12) for the definition of $B_{N_1 N_2}$. 76
- 3.2 Numerical values for the deterministic bound $B_{N_1 N_2}$ for the truncated Butterworth filter structure. Refer to Eq. (3.12) for the definition of $B_{N_1 N_2}$. . . 77
- 4.1 Summary of the four types of approximate filter structures. 123
- 4.2 Numerical values for $J(\mathcal{H}_T^1) \cdots J_T(\mathcal{H}_T^6)$ for the Butterworth optimal truncation filter structure. Note that $J_T^*(\mathcal{H}_T^*) = J_T(\mathcal{H}_T^2)$ 147
- 5.1 Summary of the approximate filtering performance for demodulating FDM speech. 153
- C.1 The pole locations of the 30th-order autoregressive random process used in the numerical example of Chapter 2. 166

Chapter 1

Introduction

Living in the information age, we are all accustomed to having the luxury of sophisticated communications and computation systems at our fingertips. It is a rarity to go through the day and not witness evidence of the explosive popularity of portable electronic devices such as cellular telephones, camcorders, laptop computers, and even wristwatch pagers which are now capable of remotely delivering real-time stock quotes with the touch of a button. The signal processing demands of such devices have increased dramatically in the last decade as products continue to shrink in size and require increasing computational speed. Due to the nature of portability, these increased processing demands are accompanied by definite constraints on power consumption since rechargeable batteries must be used. Consequently, the important task of designing low power, computationally powerful processors has emerged and spurred great interest and activity in signal processing research.

Digital filters represent a fundamental signal processing element which is found in all of the portable systems already mentioned and many others. Motivated by the growing demand for low power digital signal processing techniques for use in portable multimedia devices, in this thesis we formulate a new algorithmic approach to low power frequency-selective digital filtering. We demonstrate that significant power savings may be achieved in digital filtering applications when the order of a digital filter is dynamically varied to provide time-varying stopband attenuation in proportion to the time-varying signal-to-noise ratio (SNR) of the input signal, while maintaining a fixed SNR at the filter output. In addition to providing the capability to dynamically conserve a limited resource such as battery power, the class of algorithms we develop provides the foundation for the development of other algorithms which have the ability to intelligently respond to dynamic changes in the

availability of other resources such as processor cycles in a shared environment.

The problem considered in this thesis originated by observing overlap in the research domains of approximate signal processing and low power digital circuit design. While there exist a whole host of methods for reducing power consumption in digital electronic circuits [9, 50], one strategy is to abstract and incorporate low power constraints into the algorithm design. This strategy is called algorithmic-based low power design. The concepts of approximate processing, which have been formalized in [1, 40, 42], are inherently well-suited for algorithmic-based low power design [15, 31]. Thus, the motivation for investigating low power digital filtering using adaptive approximate processing spawned naturally from interdisciplinary collaboration in the areas of low power digital circuit design and approximate signal processing. In the remainder of this chapter we highlight the fundamental concepts in approximate signal processing and low power digital circuit design methodologies, and then present a brief outline of the thesis.

1.1 Approximate Signal Processing

Computational efficiency is of paramount importance for a broad class of signal processing algorithms designed to operate in an environment with resource limitations or other real-time constraints. A traditional approach to reducing computational complexity has been to find approximations to the *signals* involved in the processing prior to the application of a particular algorithm. Reducing the number of parameters or bits required to adequately represent a signal usually will reduce the amount of computation required to process the signal. For example, certain applications involving highly-correlated signals such as speech, sound, or images use various source coding methods to strip away redundancy from the signals before further processing or transmission. Examples of well-established methods for signal approximation include: linear transform coding methods such as the discrete cosine, wavelet, wavelet packet, or Karhunen-Loeve transforms, subband coding methods, linear predictive coding methods, and vector quantization methods [24]. Given the amount of successful research and development that has been accomplished in approximating signals to enhance processing, it is sensible to consider the parallel problem of approximating the *algorithms* which are used to process these signals. Indeed, it is logical that the same cost

vs. quality tradeoffs¹ that are used when determining a signal approximation could be incorporated into the design of signal processing algorithms, in the spirit of maximizing the computational efficiency of complete systems with real-time resource constraints.

In this thesis we pursue the goal of dynamically reducing computational cost while maintaining a desired level of output quality in the context of frequency-selective digital filtering. More specifically, our optimization criterion is to minimize average power consumption subject to the constraint that a desired SNR at the output of a frequency-selective digital filter is maintained. This type of objective has been formally studied in the field of approximate processing in computer science [31]. Approximate processing is needed for applications in which it is desirable to dynamically adjust the quality of signal processing results to the availability of resources, such as time, bandwidth, memory, and power [41, 75]. An early example of an approximate signal processing algorithm is the approximate discrete short-time Fourier transform [15]. More recently, excellent research has been accomplished in the area of incremental refinement structures for approximate signal processing in the context of sinusoidal detection using the fast Fourier transform and in the context of image decoding using the discrete cosine transform [75]. While approximate processing concepts may be used to describe a variety of existing techniques in digital signal processing (DSP), communications, and other areas, there has recently been progress in formally using these concepts to develop new DSP techniques [1, 40, 42]. This thesis introduces and explores low power digital filtering using adaptive approximate processing, or more concisely *approximate filtering*.

1.2 Low Power Design Methodologies

Techniques for reducing power consumption in digital circuits have become increasingly important due to the growing demand for portable multimedia devices. Low power, low throughput products such as wristwatches and pocket calculators have historically been the focus of research for portable digital electronics. Recently more complex systems which merge computation with efficient signal transmission across complex communication networks call for low power, high throughput devices [8, 9, 16]. Examples of low power, high throughput portable devices include camcorders, cellular telephones, laptop computers,

¹such as the rate vs. distortion criterion which is popular in the source coding domain

paggers, and portable Global Positioning System receivers. Proposals for state-of-the-art personal communication service applications describe systems which are to provide multimedia access with full-motion digital video and speech recognition capabilities [9]. These systems pose perhaps the greatest challenge for digital circuit designers. Eventually these systems will require computational capabilities in excess of those demanded by the fixed workstations of today with the additional constraint of having to be powered by batteries [9].

A significant number of low power digital systems involve frequency-selective digital filtering, in which the goal is to reject signal components in one or more frequency bands while keeping the remaining portions of the input signal spectrum largely unaltered. Digital filtering is necessary to perform functions such as lowpass filtering for signal upsampling and downsampling, bandpass filtering for subband coding, and bandstop filtering for frequency-division multiplexing and demultiplexing [44]. Because of the demand for portability in computation and communication devices, the exploration of low power digital filtering is of significant interest, and a tremendous amount of related research is being pursued [28, 29].

1.2.1 Motivation

The designers of microprocessors have traditionally centered their efforts on increasing the processor clock rate, treating power dissipation as a design issue of secondary importance. This trend has resulted in single chip power levels in excess of 30 Watts [9]. More recently an enormous demand for low power design in complementary metal-oxide silicon (CMOS) devices has emerged for the following reasons:

- the demand for portable multimedia devices with high throughput and limited rechargeable battery weight and volume has sky-rocketed with the widespread use of cellular phones, laptop computers, and video conferencing systems
- as the density and size of the chips and systems continues to increase, the design of adequate cooling systems is ever more challenging and important
- given that personal computers presently account for 5% of commercial electricity consumption and are estimated to account for 10% by the year 2000 [9], the demand for fixed workstations with low power consumption for the purpose of reducing electricity costs will thrive.

In summary, the issues of portability, heat dissipation, and the economics of commercial electricity consumption all serve as practical motivation for the design of low power computation and communication systems.

1.2.2 Sources of Power Consumption

Reducing both peak power and average power are important priorities in low power digital circuit design. Reducing the peak power levels is important mainly for reliability and proper circuit operation. The required battery weight and size in a portable system is proportional to the time-averaged power consumption. Methods which reduce the average power consumption offer the added benefit of reducing peak power consumption and thus improve reliability [9].

There are four components to the time-averaged power consumption in digital circuits using CMOS technology: switching power, short-circuit power, leakage power, and static power [9]. The total time-averaged power consumption is the sum of these four individual components

$$P_{\text{average}} = P_{\text{switching}} + P_{\text{short-circuit}} + P_{\text{leakage}} + P_{\text{static}}. \quad (1.1)$$

The switching power component dominates and typically accounts for more than 90% of the total average power consumption. This makes the switching power component the primary target for power reduction. In addition, the switching power is the most signal-dependent and algorithm-dependent component, making the switching power the primary focus for algorithmic-based approaches to low power design. We now give a brief summary of the four components of power consumption.

1. **Short-circuit power.** When there is a direct conducting path from the voltage supply to ground, the short-circuit power component is present. This component of power consumption is defined to be

$$P_{\text{short-circuit}} = I_{\text{short-circuit}} V_{dd}, \quad (1.2)$$

where $I_{\text{short-circuit}}$ is the short-circuit current and V_{dd} is the supply voltage. Through

proper choice of transistor sizes, the short-circuit power can be kept below 10% of the total power consumption [9].

2. **Static power.** Circuits that have a constant source of current between their power supplies are subject to power dissipation due to the resulting static currents. The static component of power consumption is defined to be

$$P_{\text{static}} = I_{\text{static}}V_{dd}, \quad (1.3)$$

where I_{static} is the static current and V_{dd} is the supply voltage. In SRAM amplifiers, pulsed circuits may be used to minimize static currents. However, algorithmic-based methods for reducing power consumption can have little or no effect on the static power component.

3. **Leakage power.** The two types of leakage currents are reverse-bias diode leakage at the transistor drains and sub-threshold leakage through the channel of an “off” device. The leakage component of power consumption is defined to be

$$P_{\text{leakage}} = I_{\text{leakage}}V_{dd}, \quad (1.4)$$

where I_{leakage} is the total leakage current and V_{dd} is the supply voltage. The magnitude of both components of the leakage current is set predominantly by the processing technology; thus, algorithmic-based methods for reducing power consumption will have little or no effect on the leakage power component.

4. **Switching power.** The switching component of power for a CMOS gate with load capacitor C_L is given by

$$P_{\text{switching}} = \alpha C_L V_{dd}^2 f, \quad (1.5)$$

where α is the node transition activity factor, C_L is the physical load capacitance, V_{dd} is the supply voltage, and f is the operating frequency. The two components of the

node transition activity are transitions due to the static behavior of the circuit and transitions that occur due to the dynamic nature of the circuit. The node transition activity factor α is a function of the logic function being implemented, the logic style, the circuit topology, the input signal statistics, and the sequencing of operations. A system level approach which involves optimizing algorithms, architectures, logic design, circuit design, and physical design can be used to minimize the switched capacitance and thus, in turn, minimize the switching component of power.

1.2.3 Summary of Previous Research

In this section we give a brief overview of existing methods for reducing power consumption in CMOS devices, and highlight the context in which the contributions of this thesis fit in with respect to other methods. To first order, the average switching power consumption $P_{\text{switching}}$ in Eq. (1.5) may be expanded as

$$P_{\text{switching}} = \sum_i N_i C_i V_{dd}^2 f_s, \quad (1.6)$$

where C_i is the average capacitance switched per operation of type i corresponding to addition, multiplication, storage, or bus access, N_i is the number of operations of type i performed per output sample, V_{dd} is the operating supply voltage, and f_s is the sampling frequency.

Since the dominating switching component power consumption in CMOS devices is proportional to the square of the supply voltage V_{dd} , it is clear that supply voltage reduction will have a significant impact on the average switching power consumption. Indeed, reducing the supply voltage is the key to low power operation, even after taking into account the modifications to the system architecture which are required to maintain the computational throughput.

When the supply voltage is reduced by a factor k , the power consumption is reduced by a quadratic factor k^2 . Unfortunately this power reduction comes at a price. When we reduce V_{dd} , we encounter a corresponding decrease in throughput. An empirical model for the relationship between V_{dd} and circuit delay is $T_d = \frac{K_d}{V_{dd}}$, where K_d is a constant determined experimentally and T_d is the circuit processing delay [9]. Thus, while reducing the supply voltage is an excellent way to reduce power consumption, there is an associated penalty

to pay in decreased throughput. Typically this decrease in throughput is compensated for by introducing parallelism in the circuitry, which increases the required chip area. In this context we may trade a decrease in power consumption for an increase in chip area.

The supply voltage scaling approach to low power design achieves a reduction in average power consumption by scaling down the supply voltage at the expense of reducing throughput or increasing the required chip area. An alternative approach to low power design is to reduce the switching activity to the minimal level required to perform a given computation, since *CMOS circuits do not dissipate power if they are not switching*. For this purpose we may formulate optimization problems for signal processing algorithms to minimize the circuit switching activity. Minimizing the number of multiplications and additions required to perform a given function is one critical element in reducing the overall circuit switching activity. The framework we have developed for analyzing approximate filtering algorithms was developed for the purpose of reducing the average circuit switching activity via reducing the average number of operations required per output sample in a frequency-selective digital filter. Thus, our attention in this thesis is focused on *minimizing the switched capacitance in a CMOS circuit by dynamically minimizing the number of operations required to perform frequency-selective digital filtering, subject to output quality constraints*.

Real-time digital filtering is an example of a class of applications in which there is no advantage in exceeding a bounded computation rate. For such applications, an architecture-driven voltage scaling approach has previously been developed in which parallel and pipelined architectures can be used to compensate for increased delays at reduced voltages [9]. This strategy can result in supply voltages in the 1 to 1.5 V range by using conventional CMOS technology. Power supply voltages can be further scaled using reduced threshold devices. Circuits operating at power supply voltages as low as 70mV (at a temperature of 300K) and 27mV (at a temperature of 77K) have been demonstrated [6, 19].

Once the power supply voltage is scaled to the lowest possible level, the design goal is to minimize the switched capacitance at all levels of the design abstraction. At the logic level, for example, modules can be simply shut down at a very low level based on signal values [2]. Arithmetic structures such as ripple carry or carry select can also be optimized to reduce transition activity [7]. Architectural techniques include optimizing the sequencing of operations to minimize transition activity, avoiding time-multiplexed architectures which destroy signal correlations, and using balanced paths to minimize glitching transitions. At the algorithmic level, the computational complexity or the data representation can be

optimized for low power [9].

Another approach to reducing the switched capacitance and thus saving power is to lower N_i in Eq. (1.6). Efforts have been made to minimize N_i by intelligent choice of algorithm, given a particular signal processing task [19]. In digital filtering applications, the parameter N_i is approximately linearly proportional to the filter order. In the case of conventional filter design, the filter order in a particular application is typically fixed based on worst case signal statistics. This is inefficient if the worst case seldom occurs. More flexibility may be incorporated by using adaptive filtering algorithms, which are characterized by their ability to dynamically adjust the processing to the data by employing feedback mechanisms. In this thesis, we illustrate how adaptive filtering concepts may be exploited to develop low power implementations for digital filtering by lowering N_i and thus reducing the switched capacitance and saving power.

Adaptive filtering algorithms have traditionally been used to dynamically change the *values* of the filter coefficients based on an adaptation law, while maintaining a fixed filter order [20]. In contrast, in our adaptive approach to low power filtering we show how to dynamically adjust the filter *order*. This approach leads to filtering solutions in which the SNR of the filter output may be kept above a specified threshold while using as small a filter order as possible. Since power consumption, according to Eq. (1.6), is linearly proportional to N_i , which in turn is linearly proportional to the filter order, our approach achieves power reduction with respect to a fixed-order filter whose output is similarly guaranteed to have the output SNR above the specified threshold. Maximum power reduction is achieved by dynamically minimizing the order of the digital filter.

1.3 Outline of Thesis

We begin in Chapter 2 by considering the problem of conserving power by dynamically reducing the order of a frequency-selective digital filter while maintaining a desired level of output quality. The key is to vary the filter order over time to provide time-varying stopband attenuation in proportion to the time-varying SNR of the input signal, while maintaining a fixed SNR at the filter output. The order of the filter is varied by defining a control strategy in tandem with an approximate filter structure, thus producing an *approximate filtering algorithm*. An approximate filter structure is defined by a set of related filters with different orders. The control strategy produces a dynamic estimate of the best filter

order to use from those available in the approximate filter structure, based on real-time measurements of the input signal statistics.

From the practical concept of dynamically varying the order of a digital filter, we abstract an intimately related theoretical problem. This theoretical problem involves the determination of an *optimal filter order* based on observations of the input data and a set of concrete statistical assumptions. Two solutions to this theoretical problem are presented in Chapter 2. One solution is guided by a low power approach and achieves suboptimal performance with an extremely low computational cost. A second solution is guided by a maximum likelihood objective and provides superior performance while requiring much more computation. While computationally impractical, the maximum likelihood approach provides valuable insight as well as a performance benchmark for comparison with the low power solution. The key theoretical results are used to interpret the entire class of approximate filtering algorithms.

In an approximate filtering algorithm we begin filtering a given input signal with a frequency-selective digital filter of some nominal order, taken from an approximate filter structure. This filter has well-defined passband and stopband regions in frequency. After a number L of output samples have been produced, we use the most recent block of L input and output samples to form an easily computable estimate of the current input SNR, defined as the ratio of the input signal power in the passband of the filter to the input signal power in the stopband of the filter. This estimate of the input SNR is then used to update the filter order to the *minimum* value for which the output SNR will be greater than or equal to a pre-specified minimum tolerable value. The updated filter is then used to produce a second block of L output values, and the filter order update process is repeated. In Chapter 2 we develop an underlying theory to describe approximate filtering algorithms, based on the concepts of approximate signal processing. We construct a framework to explore the statistical properties of this theory, and show that under certain assumptions the performance of approximate filtering algorithms is asymptotically optimal.

In Chapter 3 we consider the transient effects of dynamically changing the filter order in approximate filtering. For this purpose, the output of an approximate filter is related to the output of a fixed digital filter by introducing the concept of *state transition error*. We statistically analyze the corruptive effects of the state transition error on: 1) the approximate filter output sequence, 2) the L -point approximate filter output power measurement, and 3) the optimal filter order estimate determined by the approximate filtering algorithm.

We have mentioned that the order of the approximate filter is varied over time by using an approximate filter structure, defined by a set of filters with similar spectral properties with different orders. In Chapter 4, a framework for analyzing approximate filter structures is presented. An approximate filter structure is a collection of frequency-selective digital filters, one for each filter order N in a given range $N_{\min} \leq N \leq N_{\max}$. We demonstrate that approximate filter structures represent a critical element in the characterization of approximate filtering algorithms. Two classes of approximate filter structures, *truncation* and *replacement* filter structures, are introduced and used as a basis for classifying all approximate filter structures into one of four types.

A replacement filter structure is characterized by the relationship between the coefficients of filters of different orders being completely unconstrained; the coefficients of each individual filter may be selected or *replaced* independently. In a truncation filter structure this is not allowed. For a truncation filter structure with FIR constituent filter elements, the coefficients defining the lower order filters are constrained to be subsets of the coefficients defining the filter with maximum order N_{\max} . Similarly, in a truncation filter structure with IIR constituent elements, the pole/zero pairs defining the lower order filters are constrained to be subsets of the pole/zero pairs defining the filter with maximum order N_{\max} . Thus, the lower order constituent elements in a truncation filter structure are *truncated* versions of the higher order constituent elements. It is clear that truncation filter structures may be described with fewer independent filter coefficients than replacement filter structures. Associated with this property is the fact that approximate filtering using a truncation filter structure requires less memory, chip area, and bus accesses than approximate filtering using a replacement filter structure.

A metric for evaluating the performance of approximate filter structures is presented in Chapter 4 based on the results of the analyses in Chapter 2 and Chapter 3. We show that generally the class of truncation filter structures offers better potential for power reduction in approximate filtering than the class of replacement filter structures. While this power efficiency of truncation filter structures is highlighted, we also show that replacement filter structures lead to approximate filtering algorithms with superior performance. Thus, the decision to use a truncation or replacement filter structure depends on the application as well as the associated power and performance specifications.

In Chapter 5 the results of computer simulation experiments involving speech signals are used to demonstrate the practical viability of approximate filtering for low power signal pro-

cessing. We demonstrate that an order of magnitude reduction in power consumption over fixed-order filters is possible using approximate filtering algorithms. Applications involving DSP functions found in portable multimedia devices are highlighted.

Finally, in Chapter 6 we provide a recapitulation of the main contributions of this thesis. We summarize the main contribution of this thesis as the development of a framework for the design and implementation of approximate filters using signal-dependent algorithms which meet fixed performance specifications while dynamically minimizing power consumption.

Chapter 2

Approximate Filtering Algorithms

2.1 Introduction

In this chapter we consider the practical problem of dynamically reducing the order of a frequency-selective digital filter to conserve power. We will demonstrate that it is possible to dynamically vary the stopband attenuation provided by a digital filter to obtain the minimum amount of attenuation needed to continuously maintain a given output signal-to-noise ratio (SNR), and show that approximate filtering algorithms significantly reduce the required average power consumption relative to that of conventional fixed-order filtering algorithms. From this practical problem we abstract a theoretical problem which involves the determination of an optimal filter order based on observations of the input data and a set of concrete assumptions on the statistics of the input signal. Two solutions to this theoretical problem will be presented, and the key results will be used to interpret the solution to the practical low power filtering problem.

An underlying theory for approximate filtering is developed. We construct a framework to explore the statistical properties of approximate filtering algorithms, and show that under certain assumptions the performance of approximate filtering algorithms is asymptotically optimal. The focus of the algorithm development is on applications involving frequency-selective digital filtering in which the goal is to reject one or more frequency bands while keeping the remaining portions of the input spectrum largely unaltered. Examples of such applications include lowpass filtering for signal upsampling and downsampling, bandpass filtering for subband coding, and lowpass filtering for frequency-division multiplexing and demultiplexing. In addition, approximate filtering algorithms appear to be useful in other

domains in which digital filters are used such as prediction, smoothing, echo cancellation, or equalization.

2.1.1 Overview

A brief summary of approximate filtering is now given. The basic idea is to begin filtering a given input signal with a frequency-selective digital filter of some nominal order, as shown in Fig. 2-1. This filter has well-defined passband and stopband regions in frequency. After L output samples have been produced, we use the most recent block of L input and output samples to form an easily computable *low power estimate* of the current input SNR, defined as the ratio of the input signal power in the passband of the filter to the input signal power in the stopband of the filter. In Fig. 2-1 the decision module D uses the signal power estimates \hat{P}_x and \hat{P}_y to form an estimate of the temporally local input SNR. This estimate of the input SNR is then used to update the filter order to be the *minimum* value which guarantees that the output SNR, defined as the ratio of the output signal power in the passband of the filter to the output signal power in the stopband of the filter, will be greater than or equal to a pre-specified minimum tolerable output SNR. This filter order is then used to produce another block of L output samples, and the filter order update process is repeated.

A key issue addressed in this chapter is how well the low power estimate of the filter order converges to the theoretical minimum order for situations satisfying certain statistical assumptions which are made in the derivation of the underlying theoretical framework for approximate filtering. Computer simulations are used to verify analytical results which we obtain in this chapter that show that convergence to the correct filter order depends upon: 1) the number L of input and output samples used in estimating the input SNR, 2) the nominal order of the filter applied in generating the output samples that are used in estimating the input SNR, and 3) the proximity of the true input SNR to the boundaries in the input SNR space corresponding to changes in the optimal choice of filter order [35].

The adaptation mechanism used with an approximate filtering structure is designed to determine and use the filter with the smallest order while ensuring that the approximate filter output meets a pre-specified quality constraint. Minimization of the filter order used at any given time is desirable because of the resulting savings in power consumption by the underlying hardware [36]. The output quality criterion we use is designed to keep the output SNR (the ratio of the passband power to the stopband power in the filter output) above a

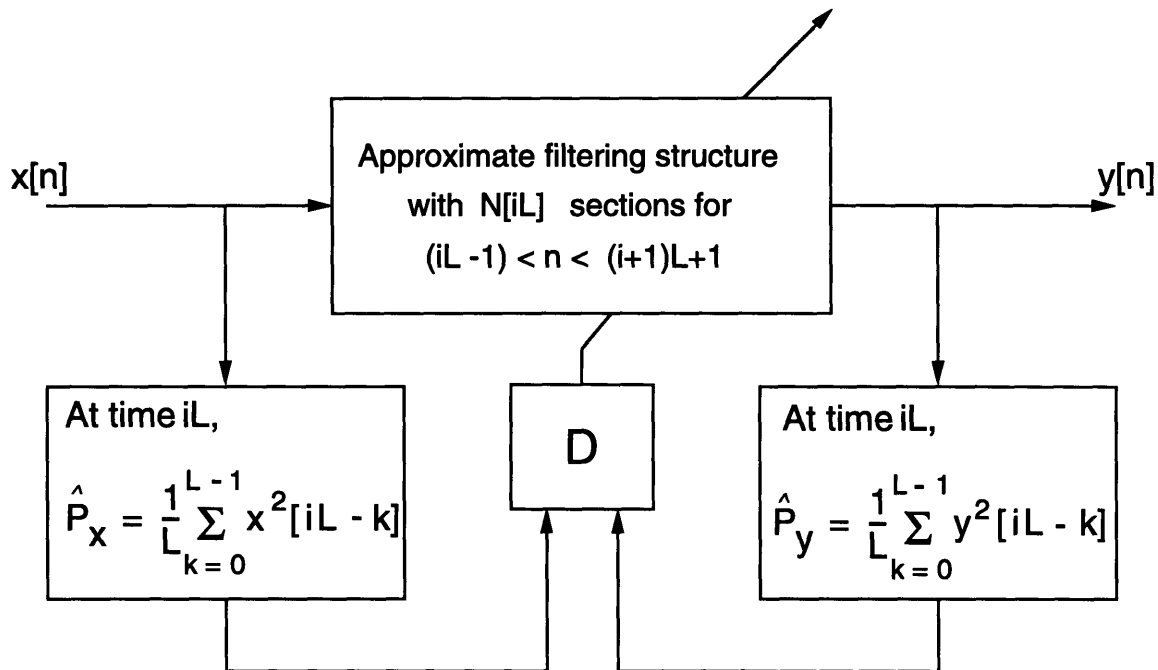


Figure 2-1: An overview of approximate filtering. The adaptation strategy for updating the filter order after each new set of L output samples is defined by the decision module D .

specified level. Other output quality constraints could easily be incorporated with minor modifications. One possible alternative output quality constraint is to keep the output signal power in the stopband of the filter below some pre-specified level. This strategy has been successfully investigated in [36].

2.1.2 Approximate Filter Structures

In approximate filtering algorithms the order of a frequency-selective digital filter is varied in a way defined by a control strategy and an approximate filter structure. A collection of frequency-selective digital filters, one for each filter order N in a given range $N_{\min} \leq N \leq N_{\max}$, constitutes an approximate filter structure \mathcal{H} . Each filter structure \mathcal{H} must possess the property that its progressively higher order filters have progressively increased average attenuation in the stopband region(s) while maintaining close to unity gain in the passband region(s).

Approximate filter structures represent an important element in the characterization of approximate filtering algorithms. Two classes of approximate filter structures, *truncation* and *replacement* filter structures, are important and used as a basis for classifying all

approximate filter structures into one of four types in Chapter 4. A replacement filter structure is characterized by the relationship between the coefficients of filters of different orders being completely unconstrained; the coefficients of each individual filter may be selected or *replaced* independently. In a truncation filter structure this is not allowed. For a truncation filter structure with FIR constituent filter elements, the coefficients defining the lower order filters are constrained to be subsets of the coefficients defining the filter with maximum order N_{\max} . Similarly, in a truncation filter structure with IIR constituent elements, the pole/zero pairs defining the lower order filters are constrained to be subsets of the pole/zero pairs defining the filter with maximum order. Thus the lower order constituent elements in a truncation filter structure are *truncated* versions of the higher order constituent elements. It is clear that truncation filter structures may be described with fewer independent filter coefficients than replacement filter structures. Associated with this property is the fact that truncation filter structures require less memory, chip area, and bus accesses than replacement filter structures.

The frequency response magnitudes of the filters drawn from an exemplary approximate filter structure based on truncations of an IIR Butterworth filter are shown in Fig. 2-2. The half-power frequency of the Butterworth filters is $\pi/2$. In this figure we show the magnitude-squared frequency responses for truncations of a 20th-order Butterworth filter with 3, 5, 7, 9, and 10 second-order sections. The key feature of an approximate filter structure is that the higher-order filters provide higher average stopband attenuation and thus have lower stopband power than the lower-order filters. This feature is clearly illustrated in Fig. 2-2, and allows us incorporate a tradeoff between filter quality and filter cost into approximate filtering algorithms. The filter quality is measured by the average stopband attenuation, while the filter cost is measured by the required power consumption or equivalently the required filter order.

The passband PB , stopband SB , and transition band TB regions for all filters in the approximate filter structure \mathcal{H} are identical. The passband and stopband regions must be explicitly specified in the definition of an approximate filter structure, and by default the transition band is defined to span the remaining portions of the spectrum $\omega \in [-\pi, \pi]$ which are not included in the passband or stopband regions.

Each of the individual filters which make up the constituent elements of the approximate filter structure \mathcal{H} must be properly normalized. Possible normalizations include a unit energy normalization or a unity DC (zero frequency) gain normalization. Other important

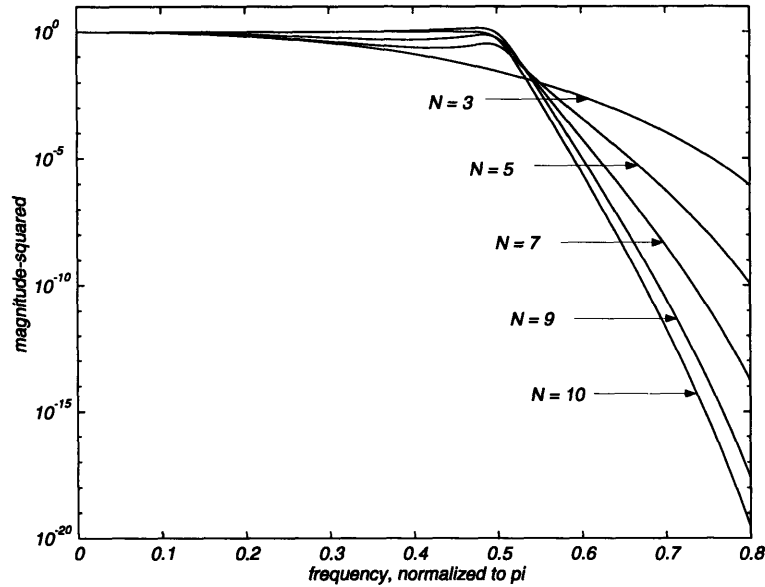


Figure 2-2: Magnitude-squared frequency responses for truncations of a 20th-order Butterworth filter with 3, 5, 7, 9, and 10 second-order sections. The half-power frequency of the Butterworth filters is $\pi/2$

characteristics of approximate filter structures will be studied in Chapter 4.

We have stated our intention to study the problem of dynamically reducing the order of a frequency-selective digital filter to conserve power while maintaining a desired level of output quality. We have stressed that the key is use vary the filter order over time to provide time-varying stopband attenuation in proportion to the time-varying SNR of the input signal, while maintaining a fixed SNR at the filter output. From the practical problem of dynamically reducing the order of a frequency-selective digital filter to conserve power, in the next section we abstract and explore an intimately related theoretical problem. The solutions to this theoretical problem will provide a basis for understanding and analyzing approximate filtering algorithms.

2.2 Problem Statement

In this section we introduce a theoretical problem, termed *the approximate filtering problem*, which involves the determination of an *optimal filter order* based on observations of input data and a set of concrete assumptions on the statistics of the input signal. Two solutions

to this theoretical problem are presented. One solution is guided by a low power approach and achieves suboptimal performance with an extremely low computational cost. A second solution is guided by a maximum likelihood objective and provides superior performance while requiring much more computation. While computationally impractical, the maximum likelihood approach provides valuable insight as well as a performance benchmark for comparison with the low power solution. The key theoretical results are used to interpret the entire class of approximate filtering algorithms.

The fundamental theoretical problem we address in this thesis is

Given

- a set of L input samples $x[0] \cdots x[L-1]$ from a wide sense stationary (WSS) Gaussian random process $x[n]$ with power spectral density $S_x(\omega)$
- a filter order set $\mathcal{N} = \{N_{\min} \cdots N_{\max}\}$ containing M elements
- a filter structure $\mathcal{H} = \{h_{N_{\min}}[n] \cdots h_{N_{\max}}[n]\}$, containing M frequency-selective filter elements, all having passband $\omega \in PB$ and stopband $\omega \in SB$,
- a passband region defined as $\omega \in PB$ for each filter in the filter structure \mathcal{H}
- a stopband region defined as $\omega \in SB$ for each filter in the filter structure \mathcal{H}
- a minimum tolerable output signal-to-noise ratio OSNR_{tol} ,

Determine

the optimal filter order N^* , defined as the *minimum* order $N \in \mathcal{N}$ of the frequency-selective filter $h_N[n] \in \mathcal{H}$ which provides sufficient stopband attenuation to assure

$$\text{OSNR}[N] \geq \text{OSNR}_{\text{tol}}, \quad (2.1)$$

where the output signal-to-noise ratio (SNR) is defined as

$$\text{OSNR}[N] \triangleq \frac{P_y^{PB}[N]}{P_y^{SB}[N]}. \quad (2.2)$$

The output power spectral density is

$$S_y(\omega) = |H_N(\omega)|^2 S_x(\omega), \quad (2.3)$$

the output power in the passband is

$$\begin{aligned} P_y^{PB}[N] &= \frac{1}{2\pi} \int_{PB} S_y(\omega) d\omega \\ &= \frac{1}{2\pi} \int_{PB} S_x(\omega) |H_N(\omega)|^2 d\omega, \end{aligned} \quad (2.4)$$

and the output power in the stopband is

$$\begin{aligned} P_y^{SB}[N] &= \frac{1}{2\pi} \int_{SB} S_y(\omega) d\omega \\ &= \frac{1}{2\pi} \int_{SB} S_x(\omega) |H_N(\omega)|^2 d\omega. \end{aligned} \quad (2.5)$$

We refer to the problem in Eqs. (2.1)–(2.5), as the *approximate filtering problem*. Any particular solution to the problem involves defining a method for reliably estimating the optimal filter order N^* based on observations of the input sequence $x[0] \cdots x[L-1]$. We measure the performance of a particular solution in terms of how accurately, on average, the solution determines the correct value for N^* . In this chapter we will present two distinct solutions to the approximate filtering problem. The first is guided by a low power consumption constraint and thus produces estimates of N^* which require very low computational overhead. We call this the *low power (LP) solution to the approximate filtering problem*. The second solution we present is motivated by a maximum likelihood formulation and is therefore termed *the maximum likelihood (ML) solution to the approximate filtering problem*. We will show that the ML solution requires much more computation and offers slightly better performance than the LP solution in certain circumstances. While computationally prohibitive for practical use, the ML solution is conceptually insightful as its performance may be used as a meaningful benchmark for comparison with the performance of the low power solution.

2.2.1 Summary of Low Power Approach

The low power (LP) approach to finding N^* is computationally simple and will be shown to perform almost equally as well as the computationally prohibitive ML-based approach, even for signals statistically tailored to favor the ML-based approach. In Chapter 5 we will see that for practical processing of speech signals, the LP solution to the approximate filtering problem provides an effective, reliable method for reducing power consumption by an order of magnitude or more over conventional filtering methods.

The LP approach computes an estimate of the output SNR based on an estimate of the total *power difference* between the input and output of a digital frequency-selective filter. Conceptually, if the frequency-selective filter were an ideal piecewise-constant filter

$$H_{\text{ideal}}(\omega) \begin{cases} 1 & \omega \in PB \\ 0 & \omega \in SB \end{cases}, \quad (2.6)$$

then the total power difference between the input and output signals would be equal to the exact input signal power in the stopband of the ideal filter. This is true since an ideal filter perfectly eliminates all the components of the input signal in its stopband, and passes with unity gain all the components of the input signal in its passband. Thus, ideally the output signal contains the exact passband components of the input signal, and nothing else. The input signal obviously contains both its passband and stopband signal components. Therefore the difference in input and output signals powers ideally gives the exact power in the stopband of the input signal. From this we may form a LP estimate $\widehat{\text{OSNR}}_{\text{LP}}[N, N_0]$ of the output SNR, and proceed to estimate N^* via Eq. (2.1) with $\text{OSNR}[N]$ replaced by $\widehat{\text{OSNR}}_{\text{LP}}[N, N_0]$.

Ideal filters are not practically realizable. However, if we use a non-ideal filter which approximates an ideal filter, our LP estimate of the output SNR based on the difference in input and output signal powers will approximate the true output SNR. To the extent that this is a good approximation, the LP approach to computing N^* using Eq. (2.1) will be effective. The detailed derivation of the LP solution to the approximate filtering problem will be given in Section 2.3. A summary of the final result is presented now.

The LP estimate \hat{N}_{LP}^* of the optimal filter order N^* is determined by searching for the minimum value of $N \in \mathcal{N}$ satisfying

$$\widehat{\text{OSNR}}_{\text{LP}}[N, N_0] \geq \text{OSNR}_{\text{tol}}, \quad (2.7)$$

where the LP estimate of the output SNR is defined as

$$\widehat{\text{OSNR}}_{\text{LP}}[N, N_0] = \left[\frac{\mathbf{y}^T \mathbf{y}}{\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y}} \right] \left(\frac{\int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega}{\int_{SB} |H_N(\omega)|^2 d\omega} \right) - \left(\frac{\int_{SB} |H_{N_0}(\omega)|^2 d\omega}{\int_{SB} |H_N(\omega)|^2 d\omega} \right). \quad (2.8)$$

The expression in Eq. (2.8) may be rearranged with simple algebraic manipulations and substitutions to simplify the decision rule for selecting \hat{N}_{LP}^* to be the *minimum* filter order $N \in \mathcal{N}$ satisfying

$$R \geq R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}], \quad (2.9)$$

where R is the ratio

$$R = \left[\frac{\mathbf{y}^T \mathbf{y}}{\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y}} \right], \quad (2.10)$$

and $R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}]$ is a function of N parameterized by N_0 and OSNR_{tol}

$$R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}] = \left\{ \text{OSNR}_{\text{tol}} + \left(\frac{\int_{SB} |H_{N_0}(\omega)|^2 d\omega}{\int_{SB} |H_N(\omega)|^2 d\omega} \right) \right\} \frac{\int_{SB} |H_N(\omega)|^2 d\omega}{\int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega}. \quad (2.11)$$

The signal vectors are $\mathbf{x} = [x[0] \ x[1] \ \dots \ x[L-1]]^T$ and $\mathbf{y} = [y[0] \ y[1] \ \dots \ y[L-1]]^T$, where the *power window length* L is the number of output samples which are produce by the filter $h_{N_0}[n] \in \mathcal{H}$ before \hat{N}_{LP}^* is computed. The nominal filter order N_0 may be chosen to be equal to any filter order $N \in \mathcal{N}$; however, as we shall see later in this chapter, the choice $N_0 = N_{\text{max}}$ produces the best results.

The LP solution to the approximate filtering problem takes advantage of the fact that during filtering L samples of the output $y[n]$ of the filter $h_{N_0}[n]$ are available, without additional computational cost. These samples are used to form the vector \mathbf{y} and the ratio R . The function $R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}]$ is a function of N assuming N_0 and OSNR_{tol} have been fixed, and its values may be easily computed and stored in advance. The rule for computing \hat{N}_{LP}^* is to search among the stored values of $R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}]$ to find amongst all those which satisfy $R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}] \leq R$ the unique one which corresponds to the minimum value of N . This value of N is defined to be \hat{N}_{LP}^* .

We now consider a second solution to the approximate filtering problem, using an ML-based approach. This solution does not use the available output signal $y[n]$ to compute its estimate of N^* . In certain situations the ML solution will achieve better estimates of N^* than the LP solution, but this performance advantage comes at the expense of requiring more computation.

2.2.2 Summary of Maximum Likelihood Approach

An alternative strategy for determining the *optimal filter order* N^* defined in Eq. (2.1) involves computing an estimate $\hat{S}_x(\omega)$ of the input power spectral density (PSD) $S_x(\omega)$ from observations $x[0] \cdots x[L-1]$ of the WSS input random process $x[n]$, and using this PSD estimate to compute an estimate of the optimal filter order N^* . Because the PSD estimate $\hat{S}_x(\omega)$ is based on ML estimates of the all-pole parameters of the underlying random process which is assumed to be autoregressive (AR), we call this the ML approach. A summary of the final result of the ML approach to the approximate filtering problem is now presented, with the details of the derivation presented in to Section 2.4.

Assuming that $S_x(\omega)$ corresponds to a p th-order AR random process, the ML-based estimate $\hat{S}_x(\omega)$ is defined as

$$\hat{S}_x(\omega) = \hat{\sigma}_u^2 \left| 1 + \sum_{m=1}^p \hat{a}_m e^{-j\omega m} \right|^{-2}, \quad (2.12)$$

where the ML-based estimates $\hat{\mathbf{a}} = [\hat{a}_1 \hat{a}_2 \cdots \hat{a}_p]^T$ and $\hat{\sigma}_u^2$ for the parameters $\mathbf{a} = [a_1 a_2 \cdots a_p]^T$ and σ_u^2 are computed via the well-known *Yule-Walker equations*. This method for AR parameter estimation is well known as the *autocorrelation method*. The

final result is that an ML estimate \hat{N}_{ML}^* of N^* may be computed via selecting the minimum value of $N \in \mathcal{N}$ satisfying

$$\widehat{\text{OSNR}}_{\text{ML}}[N] \geq \text{OSNR}_{\text{tol}}, \quad (2.13)$$

where the ML estimate of the output SNR is defined as

$$\widehat{\text{OSNR}}_{\text{ML}}[N] \triangleq \frac{[\hat{P}_y^{PB}[N]]_{\text{ML}}}{[\hat{P}_y^{SB}[N]]_{\text{ML}}}, \quad (2.14)$$

the ML estimate of the output power in the passband is defined as

$$\begin{aligned} [\hat{P}_y^{PB}[N]]_{\text{ML}} &= \frac{1}{2\pi} \int_{PB} [\hat{S}_y(\omega)]_{\text{ML}} d\omega \\ &= \frac{1}{2\pi} \int_{PB} [\hat{S}_x(\omega)]_{\text{ML}} |H_N(\omega)|^2 d\omega, \end{aligned} \quad (2.15)$$

and the ML estimate of the output power in the stopband is defined as

$$\begin{aligned} [\hat{P}_y^{SB}[N]]_{\text{ML}} &= \frac{1}{2\pi} \int_{SB} [\hat{S}_y(\omega)]_{\text{ML}} d\omega \\ &= \frac{1}{2\pi} \int_{SB} [\hat{S}_x(\omega)]_{\text{ML}} |H_N(\omega)|^2 d\omega, \end{aligned} \quad (2.16)$$

Once the ML estimate $\hat{S}_x(\omega)$ of $S_x(\omega)$ has been computed, the quantities $[\hat{P}_y^{PB}[N]]_{\text{ML}}$, $[\hat{P}_y^{SB}[N]]_{\text{ML}}$, and $\widehat{\text{OSNR}}_{\text{ML}}[N]$ may be evaluated for each value of $N \in \mathcal{N}$ and frequency-selective filter $h_N[n] \in \mathcal{H}$, and \hat{N}_{ML}^* may be determined via Eq. (2.13). This approach will be shown to experimentally produce excellent estimates of \hat{N}^* , especially when the input signal is synthetically generated to be a true AR WSS random process. This is an intuitively natural result. Unfortunately the ML approach is not practically viable due to its excessive computational requirements. However it will serve as a meaningful benchmark for performance comparison with the LP approach which is the focal point of this thesis.

Having now formally presented the statement of the approximate filtering problem as well as overviews of the two solutions which are developed in this chapter, we move on

to the detailed derivations of each solution. First we formulate the LP estimate \hat{N}_{LP}^* in Section 2.3, and then we formulate the ML estimate \hat{N}_{ML}^* in Section 2.4. The results of the LP solution will be directly used in our presentation of the approximate filtering algorithms in Section 2.5. The remarkable capabilities of this algorithm for reducing power consumption in digital filtering applications will be demonstrated in Chapter 5.

2.3 Derivation of Low Power Solution

In this section we develop the low power (LP) solution to the approximate filtering problem summarized in Eq. (2.1). The LP solution provides a method for computing the estimate \hat{N}_{LP}^* of the optimal filter order N^* based on low power operations. The LP method is necessarily computationally simple, and thus it has the advantage of requiring significantly less average power than the ML-based solution which will be given in Section 2.4. After presenting some underlying assumptions, our approach to deriving an expression for \hat{N}_{LP}^* begins with determining a LP estimate $\widehat{\text{ISNR}}_{\text{LP}}[N_0]$ of the input SNR based on the *difference* between the input power and the output power of a frequency-selective filter $h_{N_0}[n] \in \mathcal{H}$ with nominal order $N_0 \in \mathcal{N}$. We use our estimate $\widehat{\text{ISNR}}_{\text{LP}}[N_0]$ to produce an expression for a LP estimate $\widehat{\text{OSNR}}_{\text{LP}}[N, N_0]$ of the output SNR, which can be substituted into Eq. (2.1) for $\text{OSNR}[N]$ to determine the LP solution \hat{N}_{LP}^* to the approximate filtering problem.

To begin we suppose that a discrete-time WSS random process¹ $x[n]$ with power spectrum $S_x(\omega)$ is filtered using a digital frequency-selective filter with impulse response $h_{N_0}[n] \in \mathcal{H}$ and order $N_0 \in \mathcal{N}$ to obtain an output signal $y[n]$. We assume that the filter $h_{N_0}[n]$ is taken from an approximate filter structure \mathcal{H} and thus has a well-defined spectral passband

¹we assume that all random processes discussed in this thesis are ergodic

PB , stopband SB , and transition band TB . We make the following key assumptions:

Assumption 1

$S_x(\omega)$ is equal to an unknown constant σ_{SB}^2 in the stopband region of \mathcal{H}

$$S_x(\omega) = \sigma_{SB}^2 \quad \omega \in SB \quad (2.17)$$

Assumption 2

For frequencies in the passband of \mathcal{H} , $|H_N(\omega)|^2$ is approximately equal to unity

$$|H_N(\omega)|^2 \approx 1 \quad \omega \in PB. \quad (2.18)$$

This is assumed to be true for all $N \in \mathcal{N}$, and thus for all $H_N(\omega) \in \mathcal{H}$.

Assumption 3

$S_x(\omega)$ is negligible in the transition band of \mathcal{H}

$$S_x(\omega) \approx 0 \quad \omega \in TB. \quad (2.19)$$

Furthermore, this implies that

$$\int_{TB} S_x(\omega) f(\omega) d\omega \approx 0 \quad (2.20)$$

for any finite, continuous, differentiable function $f(\omega)$.

We note that no assumption is made about the shape of the function $S_x(\omega)$ within the passband. $S_x(\omega)$ in the passband is finite but otherwise *arbitrary*. Assumption 2 states that $S_x(\omega)$ is negligible in the transition band. This is reasonable for situations in which the input stopband and passband components are separated by a *guard band*, as is the case in a whole host of communications applications [30].

As mentioned earlier, our first step is to determine a LP estimate of the input SNR

under the stated assumptions. We define the signal-to-noise ratio (SNR) as the ratio of the signal power in the passband of \mathcal{H} to the signal power in the stopband of \mathcal{H} . The input SNR may be expressed as

$$\text{ISNR} \triangleq \frac{P_x^{PB}}{P_x^{SB}}, \quad (2.21)$$

where

$$P_x^{PB} = \frac{1}{2\pi} \int_{PB} S_x(\omega) d\omega, \quad (2.22)$$

and

$$P_x^{SB} = \frac{1}{2\pi} \int_{SB} S_x(\omega) d\omega. \quad (2.23)$$

By invoking the Assumption 1 which states that $S_x(\omega)$ is equal to an unknown constant σ_{SB}^2 in the stopband, it follows that

$$\begin{aligned} \text{ISNR} &= \frac{\frac{1}{2\pi} \int_{PB} S_x(\omega) d\omega}{\frac{1}{2\pi} \int_{SB} \sigma_{SB}^2 d\omega} \\ &= \frac{\frac{1}{2\pi} \int_{PB} S_x(\omega) d\omega}{\frac{1}{2\pi} \sigma_{SB}^2 \Delta_{SB}}, \end{aligned} \quad (2.24)$$

where Δ_{SB} is the spectral width of the stopband. For example, if the stopband is defined as $\pi/2 \leq |\omega| \leq \pi$, then $\Delta_{SB} = \pi$. Assumption 2 states that $|H_N(\omega)|^2 \approx 1$ for $\omega \in PB$. Since $S_y(\omega) = S_x(\omega)|H_N(\omega)|^2$, Assumption 2 implies that $S_y(\omega) \approx S_x(\omega)$ for $\omega \in PB$, and Eq. (2.24) becomes

$$\begin{aligned} \text{ISNR} \approx \text{ISNR}[N_0] &= \frac{\frac{1}{2\pi} \int_{PB} S_y(\omega) d\omega}{\frac{1}{2\pi} \sigma_{SB}^2 \Delta_{SB}} \\ &= \frac{P_y^{PB}[N_0]}{\frac{1}{2\pi} \sigma_{SB}^2 \Delta_{SB}}, \end{aligned} \quad (2.25)$$

where $P_y^{PB}[N_0]$ is the output power in the passband which was previously defined in Eq. (2.4). We note that the approximate expression in Eq. (2.25) for the input SNR is a function of N_0 due to its dependence on $P_y^{PB}[N_0]$. We now proceed to find an expression for $P_y^{PB}[N_0]$. First note that the total output power

$$\begin{aligned} P_y[N_0] &= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_y(\omega) d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) |H_{N_0}(\omega)|^2 d\omega. \end{aligned} \quad (2.26)$$

may be written as the sum of three spectrally-disjoint components

$$P_y[N_0] = P_y^{PB}[N_0] + P_y^{SB}[N_0] + P_y^{TB}[N_0], \quad (2.27)$$

where $P_y^{PB}[N_0]$ is given in Eq. (2.4), $P_y^{SB}[N_0]$ is given in Eq. (2.5), and

$$\begin{aligned} P_y^{TB}[N_0] &= \frac{1}{2\pi} \int_{TB} S_y(\omega) d\omega \\ &= \frac{1}{2\pi} \int_{TB} S_x(\omega) |H_N(\omega)|^2 d\omega. \end{aligned} \quad (2.28)$$

If we now invoke the Assumption 3 which states that $S_x(\omega)$ is negligible in the transition band, then $P_y^{TB}[N_0] \approx 0$, and rearranging Eq. (2.27) produces

$$P_y^{PB}[N_0] \approx P_y[N_0] - P_y^{SB}[N_0]. \quad (2.29)$$

We now examine the term $P_y^{SB}[N_0]$. Combining the definition of $P_y^{SB}[N_0]$ in Eq. (2.5) with Assumption 1 which states that $S_x(\omega) = \sigma_{SB}^2$ for $\omega \in SB$, we obtain

$$\begin{aligned} P_y^{SB}[N_0] &= \frac{1}{2\pi} \int_{SB} S_x(\omega) |H_N(\omega)|^2 d\omega \\ &= \frac{1}{2\pi} \sigma_{SB}^2 \int_{SB} |H_N(\omega)|^2 d\omega. \end{aligned} \quad (2.30)$$

In order to obtain a more elementary expression for $P_y^{SB}[N_0]$, it is apparent from Eq. (2.30) that an expression for the unknown parameter σ_{SB}^2 is needed. For this purpose we consider the difference in input and output signal power

$$P_x - P_y[N_0] = \frac{1}{2\pi} \int_{-\pi}^{\pi} [S_x(\omega) - S_y(\omega)] d\omega, \quad (2.31)$$

where the total input signal power P_x is defined as

$$P_x = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) d\omega. \quad (2.32)$$

As was shown for the total output power $P_y[N_0]$ in Eq. (2.27), the difference in input and output signal power may similarly be broken up into its spectrally disjoint passband, stopband, and transition band components

$$P_x - P_y[N_0] = \frac{1}{2\pi} \int_{PB} [S_x(\omega) - S_y(\omega)] d\omega + \frac{1}{2\pi} \int_{SB} [S_x(\omega) - S_y(\omega)] d\omega + \frac{1}{2\pi} \int_{TB} [S_x(\omega) - S_y(\omega)] d\omega. \quad (2.33)$$

We expand this and incorporate Assumption 1 to produce

$$P_x - P_y[N_0] = \frac{1}{2\pi} \int_{PB} S_x(\omega) [1 - |H_{N_0}(\omega)|^2] d\omega + \sigma_{SB}^2 \frac{1}{2\pi} \int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega + \frac{1}{2\pi} \int_{TB} S_x(\omega) [1 - |H_{N_0}(\omega)|^2] d\omega. \quad (2.34)$$

We now recall Assumption 2 which states that $|H_{N_0}(\omega)|^2 \approx 1$ for $\omega \in PB$. Under this assumption the first addend in the right-hand side of Eq. (2.34) is approximately zero. Furthermore, Assumption 3 which states that $S_x(\omega)$ is negligible in the transition band, we may argue that the third addend in the right-hand side of Eq. (2.34) is approximately zero. We therefore obtain

$$P_x - P_y[N_0] \approx \sigma_{SB}^2 \frac{1}{2\pi} \int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega, \quad (2.35)$$

which may be rearranged to yield

$$\sigma_{SB}^2 \approx (P_x - P_y[N_0]) \left(\frac{1}{2\pi} \int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega \right)^{-1}. \quad (2.36)$$

Substituting this approximate expression for σ_{SB}^2 into (2.30) we obtain

$$P_y^{SB}[N_0] = \frac{1}{2\pi} \sigma_{SB}^2 \int_{SB} |H_{N_0}(\omega)|^2 d\omega \quad (2.37)$$

$$\approx (P_x - P_y[N_0]) P_h^{SB}[N_0], \quad (2.38)$$

where

$$P_h^{SB}[N_0] = \left(\frac{1}{2\pi} \int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega \right)^{-1} \left(\frac{1}{2\pi} \int_{SB} |H_{N_0}(\omega)|^2 d\omega \right) \quad (2.39)$$

is a particularly relevant measure of the spectral quality of the filter $H_{N_0}(\omega)$. Note that $P_h^{SB}[N_0] = 0$ in the case of an ideal filter which was defined in Eq. (2.6). We now incorporate Eq. (2.29) into Eq. (2.25) and obtain

$$\text{ISNR}[N_0] = \frac{P_y^{PB}[N_0]}{\frac{1}{2\pi} \sigma_{SB}^2 \Delta_{SB}} \quad (2.40)$$

$$\approx \frac{P_y[N_0] - P_y^{SB}[N_0]}{\frac{1}{2\pi} \sigma_{SB}^2 \Delta_{SB}}. \quad (2.41)$$

Substituting in the approximate expression in Eq. (2.38) for $P_y^{SB}[N_0]$ produces

$$\text{ISNR}[N_0] \approx \frac{P_y[N_0] - (P_x - P_y[N_0]) P_h^{SB}[N_0]}{\frac{1}{2\pi} \sigma_{SB}^2 \Delta_{SB}}. \quad (2.42)$$

Plugging in our approximate expression in Eq. (2.36) for σ_{SB}^2 produces

$$\begin{aligned} \text{ISNR}[N_0] \approx & \left[\frac{P_y[N_0]}{P_x - P_y[N_0]} \right] \left(\frac{1}{\Delta_{SB}} \int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega \right) - \left(\frac{1}{\Delta_{SB}} \int_{SB} |H_{N_0}(\omega)|^2 d\omega \right). \end{aligned} \quad (2.43)$$

Armed with this expression for $\text{ISNR}[N_0]$ which is valid under the stated assumptions, we now turn to the problem of computing low power estimates of the signal-related quantities in Eq. (2.43): the total input power P_x and the total output power $P_y[N_0]$.

2.3.1 Low Power Estimation

To obtain the LP estimate $\widehat{\text{ISNR}}_{\text{LP}}[N_0]$ of the input SNR based on Eq. (2.43), suppose that we have applied a filter of order N_0 to the input $x[n]$ and have obtained L output samples prior to and including time n . We may then obtain the following estimates

$$\begin{aligned} \hat{P}_x &= \frac{1}{L} \sum_{k=0}^{L-1} x^2[n-k] \\ &= \mathbf{x}^T \mathbf{x}, \end{aligned} \quad (2.44)$$

and

$$\begin{aligned} \hat{P}_y[N_0] &= \frac{1}{L} \sum_{k=0}^{L-1} y^2[n-k] \\ &= \mathbf{y}^T \mathbf{y}, \end{aligned} \quad (2.45)$$

where the $L \times 1$ signal vectors are defined as $\mathbf{x} = [x[n-L+1] \cdots x[n-1] x[n]]^T$, $\mathbf{y} = [y[n-L+1] \cdots y[n-1] y[n]]^T$. We note that the explicit dependence of $\hat{P}_y[N_0]$ and

\hat{P}_x on n and L is omitted for notational simplicity. By incorporating the estimates $\hat{P}_y[N_0]$ and \hat{P}_x into Eq. (2.43) in place of $P_y[N_0]$ and P_x , respectively, we obtain the LP estimate $\widehat{\text{ISNR}}_{\text{LP}}[N_0]$ for the input SNR

$$\widehat{\text{ISNR}}_{\text{LP}}[N_0] = \left[\frac{\mathbf{y}^T \mathbf{y}}{\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y}} \right] \left(\frac{1}{\Delta_{SB}} \int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega \right) - \left(\frac{1}{\Delta_{SB}} \int_{SB} |H_{N_0}(\omega)|^2 d\omega \right), \quad (2.46)$$

This is our final expression for $\widehat{\text{ISNR}}_{\text{LP}}[N_0]$. We observe that $\widehat{\text{ISNR}}_{\text{LP}}[N_0]$ is easily computable from the signal-dependent quantities $\mathbf{y}^T \mathbf{y}$ and $\mathbf{x}^T \mathbf{x}$ and the integrals of the filter $h_{N_0}[n]$ that appear in Eq. (2.46).

In order to compute \hat{N}_{LP}^* an estimate $\widehat{\text{OSNR}}_{\text{LP}}[N, N_0]$ of the output SNR is required. To proceed, we define $\text{SNRI}[N]$, the signal-to-noise ratio improvement factor, as the multiplicative factor by which the input SNR is multiplied by to obtain the output SNR. This will enable us to easily obtain our estimate $\widehat{\text{OSNR}}_{\text{LP}}[N, N_0]$ of the output SNR from our estimate $\widehat{\text{ISNR}}_{\text{LP}}[N_0]$ of the input SNR via a simple multiplication by $\text{SNRI}[N]$. The SNR improvement factor is clearly a function of the filter order $N \in \mathcal{N}$. The signal-to-noise ratio improvement factor satisfies the relationship

$$\text{ISNR} \cdot \text{SNRI}[N] = \text{OSNR}[N], \quad (2.47)$$

which we may rearrange with substitution of the definitions for ISNR and $\text{OSNR}[N]$ from Eq. (2.21) and Eq. (2.2), respectively, to produce

$$\begin{aligned} \text{SNRI}[N] &= \frac{\text{OSNR}[N]}{\text{ISNR}} \\ &= \frac{P_y^{PB}[N]}{P_y^{SB}[N]} \cdot \frac{P_x^{SB}}{P_x^{PB}}. \end{aligned} \quad (2.48)$$

If we substitute in the definitions of $P_y^{PB}[N]$, $P_y^{SB}[N]$, P_x^{SB} , and P_x^{PB} , we arrive at

$$\text{SNRI}[N] = \left(\frac{\int_{PB} S_x(\omega) |H_N(\omega)|^2 d\omega}{\int_{SB} S_x(\omega) |H_N(\omega)|^2 d\omega} \right) \left(\frac{\int_{SB} S_x(\omega) d\omega}{\int_{PB} S_x(\omega) d\omega} \right). \quad (2.49)$$

If we now invoke the Assumption 1 and Assumption 2 which were stated at the beginning of Section 2.3, our expression for $\text{SNRI}[N]$ reduces to

$$\text{SNRI}[N] \approx \frac{\Delta_{SB}}{\int_{SB} |H_N(\omega)|^2 d\omega}. \quad (2.50)$$

Thus, the SNR improvement factor is inversely proportional to the power in the stopband of the filter $H_N(\omega)$. We noted earlier in Section 2.1.2 that the frequency selective filters in \mathcal{H} possess the property that as the filter order N increases, the average stopband attenuation also increases, and consequently the total stopband power decreases. Thus it is clear from Eq. (2.50) that the SNR improvement factor increases as the filter order N increases. The function $\text{SNRI}[N]$ is plotted vs. the filter order N for the Parks-McLellan FIR replacement filter structure in Fig. 2-3 and for the Butterworth IIR truncation filter structure in Fig. 2-4. In each case the stopband is defined as $\omega \in [5\pi/8, \pi]$. These two filter structures are discussed extensively in Chapter 4. As is clear from Fig. 2-3 and Fig. 2-4, the function $\text{SNRI}[N]$ monotonically increases with N . We refer to plots of the function $\text{SNRI}[N]$ as the *performance profile* for a given approximate filter structure \mathcal{H} .

If the input SNR is relatively low we must select a relatively high filter order to obtain a sufficiently large SNR improvement factor to assure that the output SNR is maintained above the minimum tolerable level OSNR_{tol} . Conversely, when the input SNR is relatively high we will be able to select a relatively low filter order which will provide an SNR improvement factor that will assure $\text{OSNR}[N] \geq \text{OSNR}_{\text{tol}}$.

To determine the LP solution to estimating N^* , we replace the exact ISNR in Eq. (2.47)

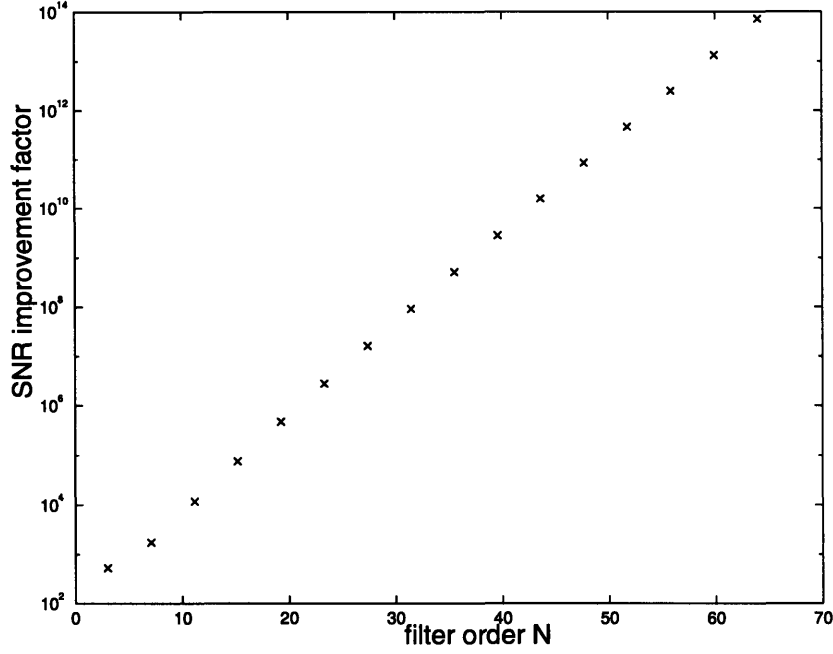


Figure 2-3: Performance profile for the Parks-McLellan FIR replacement filter structure. The stopband is defined as $\omega \in [5\pi/8, \pi]$.

with the LP estimate $\widehat{\text{ISNR}}_{\text{LP}}[N_0]$ given in Eq. (2.46), to obtain

$$\begin{aligned}
 \widehat{\text{OSNR}}_{\text{LP}}[N, N_0] &= \widehat{\text{ISNR}}_{\text{LP}}[N_0] \cdot \text{SNRI}[N] \\
 &= \left[\frac{\mathbf{y}^T \mathbf{y}}{\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y}} \right] \left(\frac{\int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega}{\int_{SB} |H_N(\omega)|^2 d\omega} \right) - \left(\frac{\int_{SB} |H_{N_0}(\omega)|^2 d\omega}{\int_{SB} |H_N(\omega)|^2 d\omega} \right), \quad (2.51)
 \end{aligned}$$

which we may compare to OSNR_{tol} to determine the low power estimate \hat{N}_{LP}^* for the optimal filter order N^* as the *minimum* filter order $N \in \mathcal{N}$ satisfying

$$\widehat{\text{OSNR}}_{\text{LP}}[N, N_0] \geq \text{OSNR}_{\text{tol}}. \quad (2.52)$$

The expression in Eq. (2.51) may be rearranged with simple algebraic manipulations and substitutions to produce

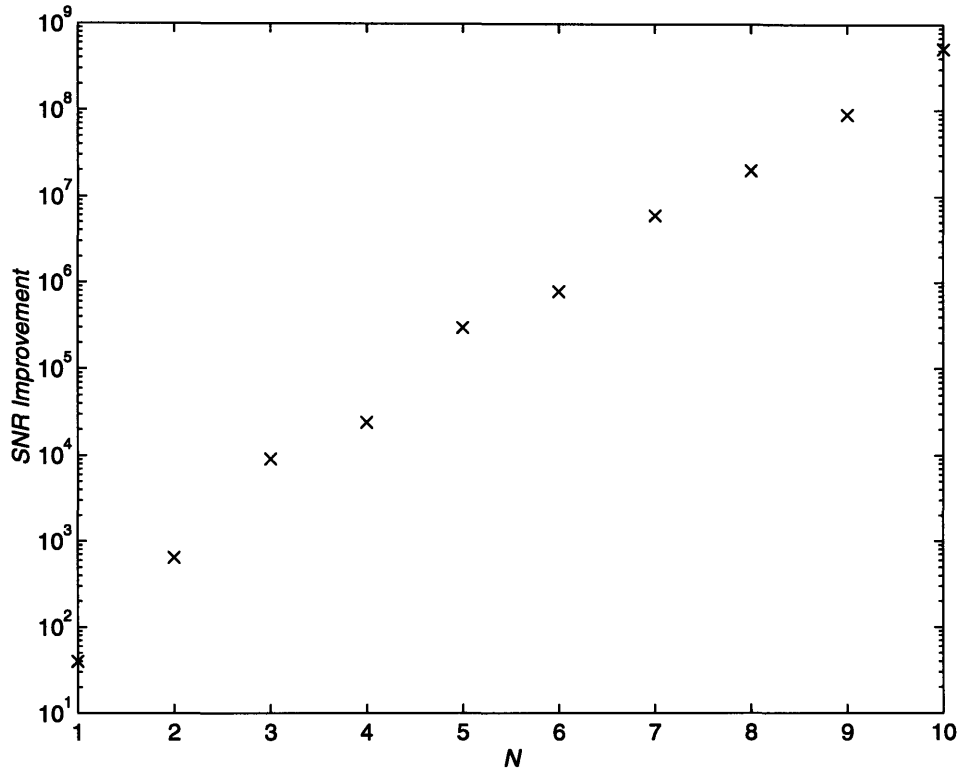


Figure 2-4: Performance profile for the Butterworth IIR truncation filter structure. The stopband is defined as $\omega \in [5\pi/8, \pi]$.

$$\left[\frac{\mathbf{y}^T \mathbf{y}}{\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y}} \right] \geq \left\{ \text{OSNR}_{\text{tol}} + \left(\frac{\int_{SB} |H_{N_0}(\omega)|^2 d\omega}{\int_{SB} |H_N(\omega)|^2 d\omega} \right) \right\} \frac{\int_{SB} |H_N(\omega)|^2 d\omega}{\int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega}. \quad (2.53)$$

By defining the ratio of quadratic forms in the above expression as

$$R = \left[\frac{\mathbf{y}^T \mathbf{y}}{\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y}} \right], \quad (2.54)$$

and the function $R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}]$ as

$$R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}] = \left\{ \text{OSNR}_{\text{tol}} + \left(\frac{\int_{SB} |H_{N_0}(\omega)|^2 d\omega}{\int_{SB} |H_N(\omega)|^2 d\omega} \right) \right\} \frac{\int_{SB} |H_N(\omega)|^2 d\omega}{\int_{SB} [1 - |H_{N_0}(\omega)|^2] d\omega}, \quad (2.55)$$

the decision rule for selecting \hat{N}_{LP}^* simplifies to choosing the *minimum* filter order $N \in \mathcal{N}$ satisfying

$$R \geq R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}]. \quad (2.56)$$

The notation $R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}]$ has been used to emphasize that $R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}]$ is a function of the filter order N and is parameterized by the nominal filter order N_0 and the minimum tolerable output SNR OSNR_{tol} . This enforces the fact that OSNR_{tol} and N_0 are application-specific parameters which are to be fixed in advance, leaving $R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}]$ a monotonic function of the single variable N . Note that the only signal-dependent quantity in Eq. (2.56) is the ratio of quadratic forms R , which was defined in Eq. (2.54).

As a final note, if we desire to avoid the power hungry division involved in computing R , we may use an alternative form for the decision rule for selecting \hat{N}_{LP}^* . The resulting simplified decision rule for selecting \hat{N}_{LP}^* is to choose the *minimum* filter order $N \in \mathcal{N}$ satisfying

$$\mathbf{y}^T \mathbf{y} \geq (\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y}) R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}]. \quad (2.57)$$

The low power decision rule is now summarized.

Summary of Method for Determining \hat{N}_{LP}^*

1. Fix the values of the application-specific parameters N_0 , L , and OSNR_{tol}
2. Compute R using Eq. (2.54) and the signal vectors \mathbf{x} and \mathbf{y} defined in Eq. (2.44) and Eq. (2.45), respectively
3. Determine \hat{N}_{LP}^* as the minimum value of N for which

$$R \geq R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}],$$

as described in Eq. (2.56) or Eq. (2.57).

In summary, the LP solution to the approximate filtering problem invokes three explicit assumptions and relies on the signal-dependent estimates \hat{P}_x and $\hat{P}_y[N_0]$ given in Eq. (2.44) and Eq. (2.45), respectively. For situations in which the three assumptions are valid and in which \hat{P}_x and $\hat{P}_y[N_0]$ are good estimates of P_x and $P_y[N_0]$, respectively, we expect excellent estimator performance using the LP estimate \hat{N}_{LP}^* for N^* . This issue is explored in Section 2.3.2. The function $\text{SNRI}[N]$ and the nature of its dependence on N and \mathcal{H} are explored in Chapter 4 in which we study approximate filter structures.

2.3.2 Convergence Analysis

It is of interest to determine the degree to which the low power filter order estimate \hat{N}_{LP}^* converges to the theoretically optimal filter order N^* for input signals that satisfy the assumptions underlying the derivation of \hat{N}_{LP}^* . In this section, we illustrate empirically that better convergence is obtained as the duration L over which \hat{P}_x and $\hat{P}_y[N_0]$ are computed is made longer, and also as the nominal filter order N_0 is made larger. We also observe and discuss the fact that since optimal filter order selections partition the range of possible input SNR values, the relation of the actual input SNR to the boundaries in this partitioning is an

important factor in determining whether or not the truly optimal filter order N^* is exactly determined by the LP estimation method.

For our convergence analysis, we assume that the input signal satisfies the same conditions that were stipulated in the derivation of our expression for \hat{N}_{LP}^* in Section 2.3. This means that we assume the input signal $x[n]$ is a WSS random process. When L consecutive samples of the output $y[n]$ are produced using a filter of order N_0 , it follows that these output samples also belong to a WSS random process. We conclude that \hat{P}_x and $\hat{P}_y[N_0]$ as defined in Eq. (2.44) and Eq. (2.45), respectively, represent estimates of the zero-lag autocorrelation values of $x[n]$ and $y[n]$, respectively. These well-known estimators converge to the true values of the zero-lag autocorrelations as L is made larger. Since \hat{P}_x and $\hat{P}_y[N_0]$ are the only signal-related quantities used in obtaining the input SNR estimate in accordance with Eq. (2.46), we expect the input SNR estimate $\widehat{\text{ISNR}}_{LP}[N_0]$ to converge to the true input SNR as L and N_0 are made larger.

To verify the influence of the estimation interval L on the input SNR estimate, we applied the LP estimation method of Eq. (2.56) to a synthetically generated random signal $x[n]$. This signal was designed to have a flat spectrum in the passband $|\omega| \in [0, 3\pi/8]$, a flat spectrum in the stopband $|\omega| \in [5\pi/8, \pi]$, negligible energy in the transition band $|\omega| \in [3\pi/8, 5\pi/8]$, and a fixed SNR throughout its 10,000 point duration. The signal was filtered using an order- N_0 digital Butterworth filter. The L consecutive input and output samples (starting from the 1000th sample to avoid filter startup transient effects) were used to obtain the LP estimate $\widehat{\text{ISNR}}_{LP}[N_0]$ of the input SNR. For a case where the true SNR of the input signal $x[n]$ was set to 0.07, in Fig. 2-5 we show the LP estimates of the input SNR obtained for different values of L in the range $1 \leq L \leq 4000$ and N_0 in the range $4 \leq N_0 \leq 10$. It is clear from Fig. 2-5 that as L and N_0 increase, the LP estimate of the input SNR visually converges to the true input SNR of 0.07. It should be noted that lower values of N_0 correspond to frequency response shapes which violate the underlying assumptions to a greater degree. We must keep in mind that unless the filter $h_{N_0}[n]$ is ideal as in Eq. (2.6), the estimate $\widehat{\text{ISNR}}_{LP}[N_0]$ will never truly converge to the true input SNR, no matter how large L is made.

In Fig. 2-5, we have also indicated the partitioning of the input SNR space in accordance with the corresponding optimal filter order N^* which should be used to ensure a minimum tolerable output SNR of 1000. Except for very small values of L , it is seen that the LP estimate of the input SNR leads to $\hat{N}_{LP}^* = N^*$. This result is dependent on the fact that

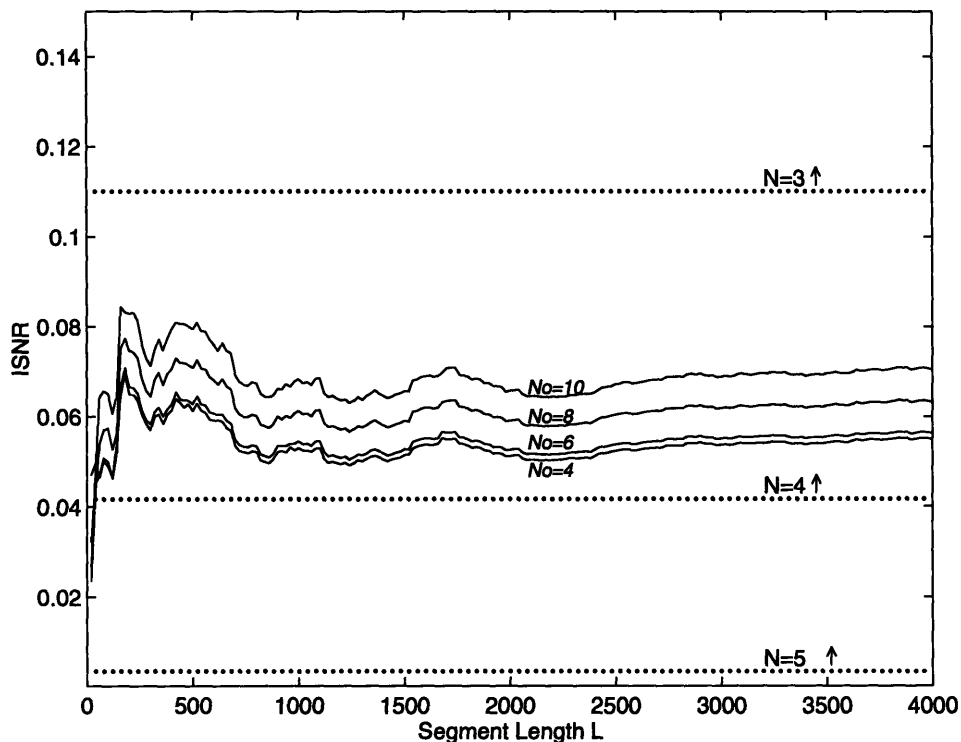


Figure 2-5: Solid curves represent input SNR estimates as a function of L and N_0 . The actual SNR for the input signal is 0.07. The straight dotted lines indicate the partitioning of the input SNR space by optimal values for the number of filter sections to use in order to obtain an output SNR of at least 1000.

an input SNR of .07 happens to fall near the middle of the input SNR space corresponding to $N^* = 4$. For example, if the actual input SNR had been made equal to 0.04, the LP estimates of the input SNR may have crossed into the incorrect $N^* = 5$ region for a larger set of values of L .

To close this section, we offer two remarks regarding the convergence properties of the LP optimal filter order estimate \hat{N}_{LP}^* . First, we note that

$$\lim_{N_0 \rightarrow \infty} \lim_{L \rightarrow \infty} \widehat{\text{ISNR}}_{LP}[N_0] = \text{ISNR}. \quad (2.58)$$

This statement simply elucidates the fact that as $N_0 \rightarrow \infty$, the filter $h_{N_0}[n]$ becomes ideal with no transition band. In the limiting case of Eq. (2.58) the power difference $P_x - P_y[N_0]$

converges to the true value of P_x^{PB} , and consequently the ratio $R = \mathbf{y}^T \mathbf{y} / (\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y})$ converges to the true input SNR. While in practice we will never be able to realize the conditions of this limiting case, the result is nevertheless insightful. Secondly, we observe that under Assumption 1

$$\lim_{N \rightarrow \infty} \text{SNRI}[N] = \frac{\Delta_{SB}}{\int_{SB} |H_N(\omega)|^2 d\omega}, \quad (2.59)$$

which is the same expression we get for $\text{SNRI}[N]$ when we invoke the assumptions presented in the derivation of the LP estimation method. Remarkably, then, we may conclude that

$$\lim_{N \rightarrow \infty} \lim_{N_0 \rightarrow \infty} \lim_{L \rightarrow \infty} \hat{N}_{LP}^* = N^*, \quad (2.60)$$

which implies that if we use sufficiently large values of L and N_0 to compute $\widehat{\text{ISNR}}_{LP}[N_0]$, then we can expect *asymptotically optimal* performance as the filter orders $N \in \mathcal{N}$ that we search over to compute \hat{N}_{LP}^* increase without bound.

A second note we make in closing is that by introducing the $L \times L$ convolution matrix

$$\mathbf{H}[N_0] = \begin{bmatrix} h_{N_0}[0] & 0 & 0 & \cdots & 0 \\ h_{N_0}[1] & h_{N_0}[0] & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{N_0}[L-1] & h_{N_0}[L-2] & h_{N_0}[L-3] & \cdots & h[0] \end{bmatrix}, \quad (2.61)$$

we may express the vector \mathbf{y} as

$$\mathbf{y} = \mathbf{H}[N_0] \mathbf{x}, \quad (2.62)$$

and thus the expression for R in Eq. (2.54) simplifies to

$$R = \frac{\mathbf{x}^T \mathbf{A}[N_0] \mathbf{x}}{\mathbf{x}^T \mathbf{B}[N_0] \mathbf{x}} \quad (2.63)$$

where the $L \times L$ matrices \mathbf{A} and \mathbf{B} satisfy

$$\mathbf{A} = \mathbf{H}^T[N_0] \mathbf{H}[N_0] \quad (2.64)$$

$$\mathbf{B} = (\mathbf{I} - \mathbf{H}[N_0])^T (\mathbf{I} - \mathbf{H}[N_0]). \quad (2.65)$$

In this formulation R has the form of a ratio of two quadratic forms in the random vector \mathbf{x} . It is interesting to note that if the vector \mathbf{x} is a multivariate Gaussian random vector, an extremely complicated nevertheless computable expression may be obtained for the variance of R as a function of the filter coefficients in \mathbf{H} and the power window length L . Various forms of the variance of the random variable R may be found in [18, 32, 61, 62]. A future direction of this research is to analytically evaluate this variance and compare it to the sample variance computed in computer simulations. Furthermore, the problem of *designing* the filter $h_N[n]$ which appears in the matrix $\mathbf{H}[N]$ to produce filter structures \mathcal{H} which *minimize the variance* of R provides an exciting and challenging future avenue to pursue in the area of approximate filtering algorithms.

In addition, expressions for the mean and variance of $\mathbf{x}^T \mathbf{A}[N_0] \mathbf{x}$ and $\mathbf{x}^T \mathbf{B}[N_0] \mathbf{x}$ for a multivariate Gaussian random vector \mathbf{x} with zero mean and covariance matrix Σ may be obtained in closed form [54]. Specifically, they are given by

$$E \left(\mathbf{x}^T \mathbf{A}[N_0] \mathbf{x} \right) = \text{trace} (\Sigma \mathbf{A}[N_0]) \quad (2.66)$$

$$\text{VAR} \left(\mathbf{x}^T \mathbf{A}[N_0] \mathbf{x} \right) = 2 \text{trace} (\Sigma \mathbf{A}[N_0])^2, \quad (2.67)$$

and

$$E \left(\mathbf{x}^T \mathbf{B}[N_0] \mathbf{x} \right) = \text{trace} (\Sigma \mathbf{B}[N_0]) \quad (2.68)$$

$$\text{VAR} \left(\mathbf{x}^T \mathbf{B}[N_0] \mathbf{x} \right) = 2 \text{trace} (\Sigma \mathbf{B}[N_0])^2. \quad (2.69)$$

While these expressions do not give the true mean and variance of the random variable R or \hat{N}_{LP}^* , they do offer insight into the statistical properties of two signal-dependent quantities $\mathbf{y}^T \mathbf{y} = \mathbf{x}^T \mathbf{A}[N_0] \mathbf{x}$ and $(\mathbf{y}^T \mathbf{y} - \mathbf{x}^T \mathbf{x}) = \mathbf{x}^T \mathbf{B}[N_0] \mathbf{x}$ involved in the simplified division-free decision rule for determining \hat{N}_{LP}^* given in Eq. (2.57), and thus are worth mentioning here.

2.3.3 Numerical Example

In this section we present a simple numerical example to demonstrate the efficacy of the LP solution to the approximate filtering problem. We first synthetically generated a random driving noise signal which consisted of independent and identically distributed samples distributed according to a unit variance, zero mean Gaussian probability density function (PDF). This driving noise sequence was then filtered with an 30th-order all-pole filter to create a WSS Gaussian random process. A plot of the PSD of this random process is shown in Fig. 2-6.

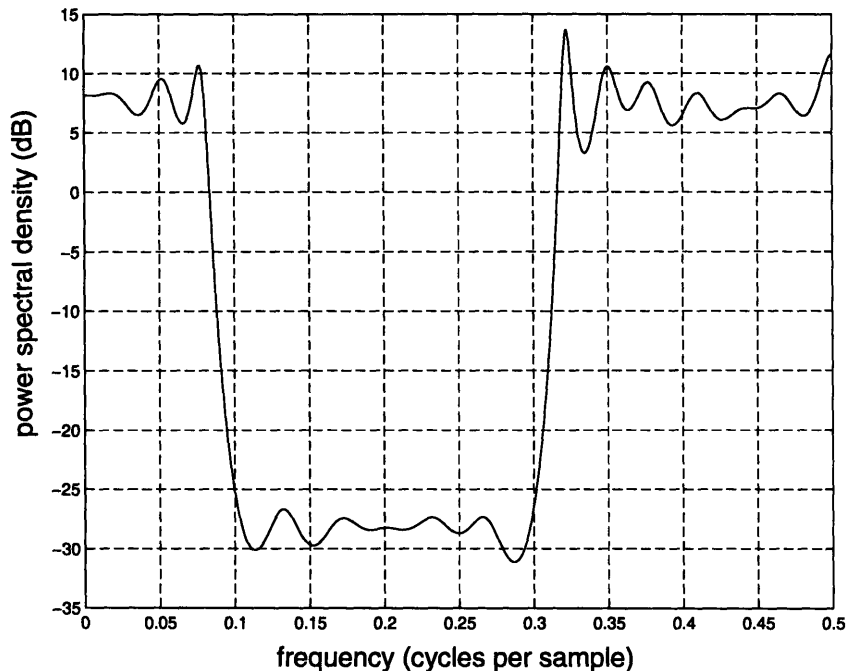


Figure 2-6: Power spectral density of the 30th-order AR process which is used as the input signal in the numerical example.

The all-pole filter parameters were selected to assure the power spectral density $S_x(\omega)$ was negligible in the transition band, defined in this example to be $3\pi/8 \leq |\omega| \leq 5\pi/8$.

The exact all-pole parameters that were used in this example are given in Appendix C. The passband was defined to be $0 \leq |\omega| \leq 3\pi/8$ and the stopband was defined as $5\pi/8 \leq |\omega| \leq \pi$. As can be seen from the spectral shape of $S_x(\omega)$, the input power spectral density is negligible in the transition band and relatively flat in the stopband in accordance with the assumptions underlying the development of the LP solution to the approximate filtering problem.

Table 2.1: Summary of the results of the numerical example using the LP estimator in which the true value of $N^* = 10$ and the true input SNR is 0.4831. The results are tabulated for power window lengths of $L = 50, 100, 500, 1000$, and 5000.

L	\hat{N}_{LP}^*		\widehat{ISNR}_{LP}	
	<i>Sample Mean</i>	<i>Sample STD</i>	<i>Sample Mean</i>	<i>Sample STD</i>
50	7.5700	2.0013	0.6243	0.6540
100	7.7100	1.5718	0.5009	0.2320
500	7.2300	1.4829	0.5046	0.0987
1000	7.2000	1.4771	0.4998	0.0678
5000	7.6800	1.4967	0.4809	0.0260

In Table 2.1 we summarize the LP estimation results after 100 Monte Carlo trials were performed for each of the values of $L = 50, 100, 500, 1000$, and 5000. The sample mean and sample standard deviation (STD) are listed in Table 2.1. Clearly as L increases, the quality of our LP estimates \hat{N}_{LP}^* and \widehat{ISNR}_{LP} improve. This is evident from the fact that the sample standard deviations decrease as L increases. In addition, the low power estimate \widehat{ISNR}_{LP} of the input SNR converges towards the true value of 0.4831 as L increases. In this example the Parks-McLellan FIR replacement approximate filter structure was used with $N_0 = N_{\max} = 64$ and $N_{\min} = 3$.

In Fig. 2-7 we have plotted four histograms of the actual LP estimates of the input SNR for the same 100 Monte Carlo trials. Each histogram corresponds the estimates of the input SNR for different values of L . As L increases the estimates “tighten up” around their means. From the entries in Table 2.1 it is clear that while \widehat{ISNR}_{LP} converges towards the true input SNR as L increases, the estimate \hat{N}_{LP}^* does not coverage to N^* in this example. This is a consequence of the fact that theoretically \widehat{ISNR}_{LP} converges to $ISNR$ as L and N_0

increase without bound, while the convergence of \hat{N}_{LP}^* to N^* requires L , N_0 , and N_{\max} to increase without bound, as was discussed in Section 2.3.2.

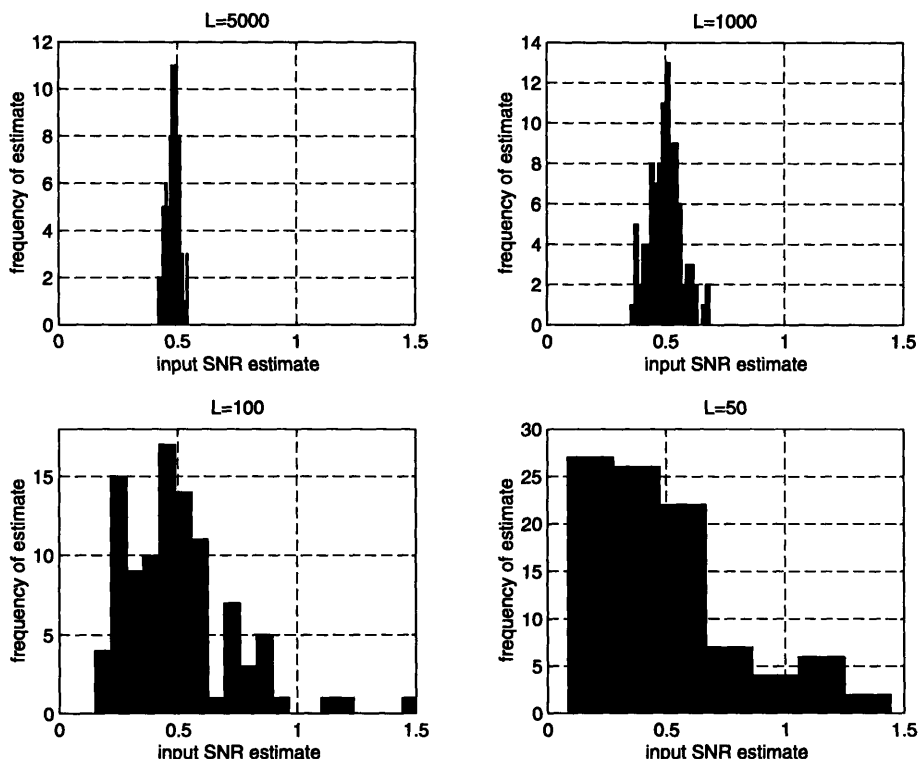


Figure 2-7: Histograms of the LP input SNR estimates for $L = 5000, 1000, 100,$ and 50 . Each histogram represents the results of 100 Monte Carlo simulations.

We shall revisit this same numerical example in Section 2.4.2 and evaluate the performance of the ML solution for comparison to the results given here.

2.4 Derivation of Maximum Likelihood Solution

In this section we assume $S_x(\omega)$ is the PSD of an autoregressive (AR) Gaussian random process, and we use a maximum likelihood (ML) objective to find estimates for the parameters defining this process. As we shall see the PSD $S_x(\omega)$ is a function of these parameters, and this function is one-to-one (invertible). Therefore we may easily obtain the ML estimate $[\hat{S}_x(\omega)]_{ML}$ by invoking the well-known invariance property of the ML estimator [26].

By inspecting Eq. (2.1) we observe that the optimal filter order N^* is *not* in one-to-one correspondence with $S_x(\omega)$. This is true since many different functions $\hat{S}_x(\omega)$ could result in the same ratio of integrals which define $\text{OSNR}[N]$, and thus many different functions $\hat{S}_x(\omega)$ could result in the same N^* . Nevertheless, we shall still be able to find a maximum *modified* likelihood estimate, which we shall denote \hat{N}_{ML}^* , for the optimal filter order N^* which is based on the maximum likelihood estimate $[\hat{S}_x(\omega)]_{\text{ML}}$. We will show that the estimate \hat{N}_{ML}^* , although not the estimate which truly maximizes the associated likelihood function, instead maximizes a modified likelihood function. Thus the performance of \hat{N}_{ML}^* will serve as a meaningful benchmark for comparison with the LP estimate \hat{N}_{LP}^* of the optimal filter order which was presented in Section 2.3.

2.4.1 Maximum Likelihood Estimation

In this section we first present an expression for the asymptotic ML estimates of the AR parameters of a Gaussian random process. Directly following the insightful presentation in [26], we consider the random process generated as the output $x[n]$ of a stable, causal all-pole filter

$$H(z) = \frac{1}{A(z)} \quad (2.70)$$

excited at the input by a zero-mean white Gaussian noise sequence $u[n]$. The p th-order polynomial function $A(z)$ is defined by the AR filter parameters $[a_1 \ a_2 \ \dots \ a_p]$ as

$$A(z) = 1 + \sum_{k=1}^p a_k z^{-k}. \quad (2.71)$$

If the all-pole filter $H(z)$ is stable, the excitation noise sequence $u[n]$ assures that the output $x[n]$ is a WSS random process. The effect of the filter is to color the white noise sequence $u[n]$. The AR model is capable of producing a wide variety of PSD functions, depending on the choice of the AR filter parameters $[a_1 \ a_2 \ \dots \ a_p]$ and excitation noise variance σ_u^2 . The problem is to estimate the parameters $[a_1 \ a_2 \ \dots \ a_p]$ and σ_u^2 based on the observed data sequence $x[0] \ \dots \ x[L-1]$. Once the parameter estimates $[\hat{a}_1 \ \hat{a}_2 \ \dots \ \hat{a}_p]$ and $\hat{\sigma}_u^2$ are computed, the PSD is estimated as

$$\hat{S}_x(\omega) = \hat{\sigma}_u^2 \left| 1 + \sum_{m=1}^p \hat{a}_m e^{-j\omega m} \right|^{-2} \quad (2.72)$$

We now given expressions for the asymptotic ML estimates for the parameters $[a_1 a_2 \cdots a_p]$ and σ_u^2 . First, the estimated autocorrelation function is

$$\hat{r}_{xx}[k] = \begin{cases} \frac{1}{L} \sum_{n=0}^{L-1-|k|} x[n]x[n+|k|] & |k| \leq L-1 \\ 0 & |k| \geq L \end{cases} . \quad (2.73)$$

The set of equations to be solved for the asymptotic ML estimate of the AR filter parameters \mathbf{a} is

$$\sum_{l=1}^p \hat{r}_{xx}[k-l] = -\hat{r}_{xx}[k] \quad k = 1, 2, \dots, p \quad (2.74)$$

which can be rewritten in matrix form as

$$\begin{bmatrix} \hat{r}_{xx}[0] & \hat{r}_{xx}[1] & \cdots & \hat{r}_{xx}[p-1] \\ \hat{r}_{xx}[1] & \hat{r}_{xx}[0] & \cdots & \hat{r}_{xx}[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{xx}[p-1] & \hat{r}_{xx}[p-2] & \cdots & \hat{r}_{xx}[0] \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = - \begin{bmatrix} \hat{r}_{xx}[1] \\ \hat{r}_{xx}[2] \\ \vdots \\ \hat{r}_{xx}[p] \end{bmatrix} . \quad (2.75)$$

These are the well-known estimated *Yule-Walker equations*, which may be recursively solved using the Levinson recursion algorithm [25]. What is left is to solve for the asymptotic ML estimate σ_u^2 . The asymptotic ML estimate is given by

$$\hat{\sigma}_u^2 = \hat{r}_{xx}[0] + \sum_{k=1}^p \hat{a}_k \hat{r}_{xx}[k]. \quad (2.76)$$

Thus, the asymptotic ML estimates $\hat{\mathbf{a}}$ and $\hat{\sigma}_u^2$ for the parameters \mathbf{a} and σ_u^2 are given in

Eq. (2.75) and Eq. (2.76), respectively. These estimates converge to the true ML estimates as $L \rightarrow \infty$, and yield reasonable estimates for sufficiently large finite values L .

We recall that our ML estimate of the PSD is

$$[\hat{S}_x(\omega)]_{\text{ML}} = \hat{\sigma}_u^2 \left| 1 + \sum_{m=1}^p \hat{a}_m e^{-j\omega m} \right|^{-2}. \quad (2.77)$$

This estimate may be used to determine an ML estimate \hat{N}_{ML}^* of the optimal filter order N^* . This ML-based estimate is produced by choosing \hat{N}_{ML}^* to be the *minimum* order $N \in \mathcal{N}$ of the frequency-selective filter $h_N[n] \in \mathcal{H}$ which provides sufficient stopband attenuation to assure

$$\widehat{\text{OSNR}}_{\text{ML}}[N] \geq \text{OSNR}_{\text{tol}}, \quad (2.78)$$

where the ML-based estimate of the output SNR is defined as

$$\widehat{\text{OSNR}}_{\text{ML}}[N] \triangleq \frac{[\hat{P}_y^{PB}[N]]_{\text{ML}}}{[\hat{P}_y^{SB}[N]]_{\text{ML}}}, \quad (2.79)$$

the ML-based estimate of the output power in the passband is defined as

$$\begin{aligned} [\hat{P}_y^{PB}[N]]_{\text{ML}} &= \frac{1}{2\pi} \int_{PB} [\hat{S}_y(\omega)]_{\text{ML}} d\omega \\ &= \frac{1}{2\pi} \int_{PB} [\hat{S}_x(\omega)]_{\text{ML}} |H_N(\omega)|^2 d\omega, \end{aligned} \quad (2.80)$$

and the ML-based estimate of the output power in the stopband is defined as

$$\begin{aligned} [\hat{P}_y^{SB}[N]]_{\text{ML}} &= \frac{1}{2\pi} \int_{SB} [\hat{S}_y(\omega)]_{\text{ML}} d\omega \\ &= \frac{1}{2\pi} \int_{SB} [\hat{S}_x(\omega)]_{\text{ML}} |H_N(\omega)|^2 d\omega. \end{aligned} \quad (2.81)$$

Consequently, the estimate \hat{N}_{ML}^* maximizes the modified likelihood function which is related

to the true likelihood function as explained in [26].

Summary of Method for Determining \hat{N}_{ML}^*

1. *Given observations $x[0] \cdots x[L-1]$, compute the ML estimates of $\hat{\mathbf{a}}$ and $\hat{\sigma}_u^2$ using Eq. (2.75) and Eq. (2.76), respectively*
2. *Compute $[\hat{S}_x(\omega)]_{ML}$ via Eq. (2.77)*
3. *Determine \hat{N}_{ML}^* according to Eq. (2.78).*

2.4.2 Numerical Example

In this section we present the results of using the ML estimation method for determining N^* on the same numerical example that was presented in Section 2.3.3. Numerical results are given to demonstrate the performance of the ML method. Recall from Section 2.3.3 that this example involves a random driving noise signal consisting of independent and identically distributed samples distributed according to a unit variance, zero mean Gaussian PDF. This driving noise signal was filtered with an 30th-order all-pole filter to create a WSS Gaussian random process with a PSD which was shown previously in Fig. 2-6.

In Table 2.2 we summarize the ML estimation results after 100 Monte Carlo trials were performed for each of the values of $L = 50, 100, 500, 1000,$ and 5000 . The sample mean and sample standard deviation (STD) are listed Table 2.2. Clearly as L increases, the quality of the ML optimal filter order estimate \hat{N}_{ML}^* improves since its standard deviation decreases as L increases. In addition, the ML estimate of the optimal filter order converges towards the true value of $N^* = 10$ as L becomes larger.

In the simulations we used the Yule-Walker equations and the Levinson recursion to solve for the ML AR coefficients which give the ML power spectrum estimate and thus the ML estimate of the optimal filter order N_{ML}^* . While the Yule-Walker equations give the asymptotic ML estimates of the AR parameters which converge to the true ML estimates as $L \rightarrow \infty$, it is well known that another method produces better estimates for finite data

Table 2.2: Summary of the results of the numerical example using the ML estimator in which the true value of $N^* = 10$ and the true input SNR is 0.4831. The results are tabulated for power window lengths of $L = 50, 100, 500, 1000$, and 5000.

L	\hat{N}_{ML}^*	
	<i>Sample Mean</i>	<i>Sample STD</i>
50	11.0300	1.5983
100	11.2000	1.5176
500	10.6000	1.2060
1000	10.7200	1.2877
5000	10.2400	0.8180

records [49]. This method of AR parameter estimation is known as the *forward-backward least-squares method*. Using the forward-backward least-squares method to compute the AR parameter estimates would probably improve the performance of the estimator \hat{N}_{ML}^* . In our simulations here we used the Yule-Walker AR parameter estimates since we were guided by the mathematical optimality of the ML approach. Small errors in the AR parameter estimates should not have a significant effect on the estimator \hat{N}_{ML}^* . This is true since \hat{N}_{ML}^* is based on a ratio of integrals of the power spectrum. Small errors in the AR parameter estimates will be integrated out when computing \hat{N}_{ML}^* .

As a final note, we observe that in determining the ML estimates for the numerical example we assumed the order of the AR process was known to be equal to 30. This introduces an element of unfairness when comparing the performance of the ML and LP approaches. In practice the AR order would not be known exactly, and would thus have to be estimated. Furthermore, this example was specifically tailored for the ML approach, since the input signal was synthetically generated to represent a true AR random process. This provides a second reason why we would expect the ML estimates to outperform the LP estimates. In Chapter 5 we will see that when processing real speech signals, the LP estimation method is reliably effective at producing good estimates for N^* .

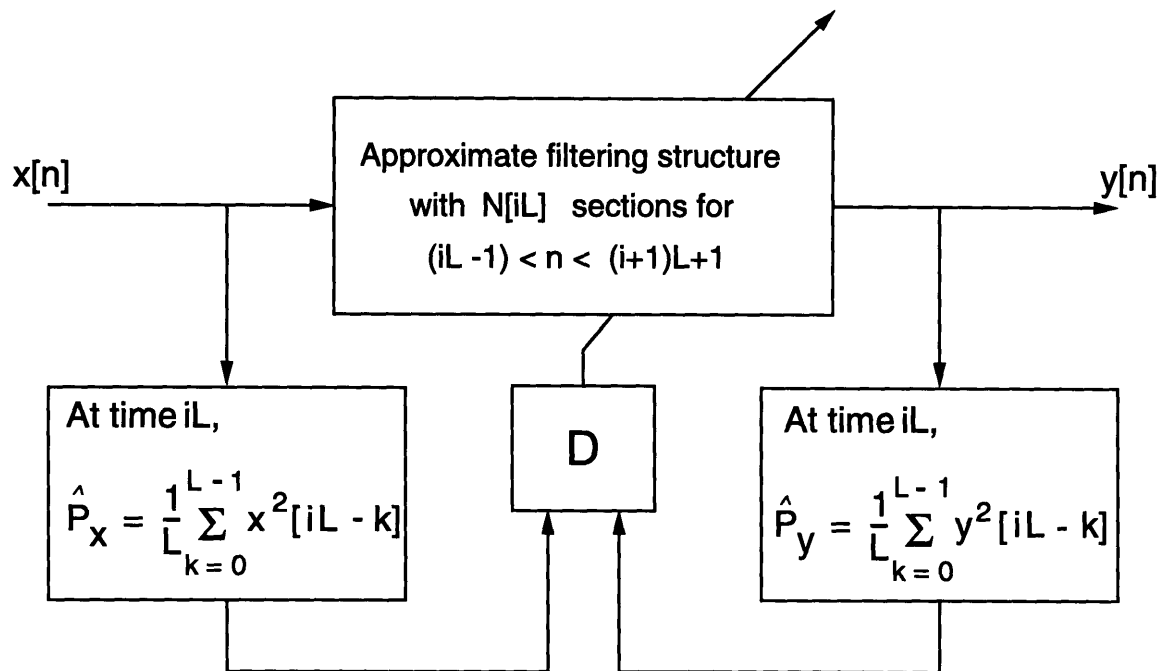


Figure 2-8: An overview of the concept of approximate filtering. The adaptation strategy for updating the filter order after each new set of L output samples is defined by the decision module D .

2.5 Adaptation for Non-Stationary Inputs

In this section we return our focus to the LP solution to the approximate filtering problem. We consider how we may use the previous results which assumed the input was a WSS random process to develop a method for accommodating non-stationary input signals. This situation arises in a whole host of filtering contexts in which the input signal is not modeled well by a WSS random process.

If the input is non-stationary, the output SNR variable $OSNR[N]$ in Eq. (2.1) will be time varying, and consequently the optimal filter order will change over time. The LP estimate \hat{N}_{LP}^* of the optimal filter order N^* must therefore change over time to dynamically minimize power consumption. Of course, this requires an adaptation framework whose overhead is low relative to the expected savings in power consumption.

One adaptation strategy is illustrated in Fig. 2-8. The order of the filter is updated for every new set of L output samples. The update procedure involves the calculation of input and output signal power estimates followed by the application of the decision module D shown in Fig. 2-8.

This module uses the signal power estimates \hat{P}_x and $\hat{P}_y[N_0]$ to form an estimate of the temporally local input SNR. This input SNR estimate is then used as the basis for selecting the filter order to be applied in computing the next set of L output samples.

Assuming that M values of the function $R_{\text{tol}}[N; N_0, \text{OSNR}_{\text{tol}}]$ in Eq. (2.54) are pre-stored, the search for \hat{N}_{LP}^* in Eq. (2.78) requires $O(M)$ operations—at most M table look-up operations, two subtractions, at most $M_0 + 1$ multiplications and at most M comparisons—where M is the number of filter elements in \mathcal{H} . Specifically, $M = (N_{\text{max}} - N_{\text{min}} + 1)$. The estimated filter order \hat{N}_{LP}^* becomes the new order of the filter used in the approximate filtering structure to produce the next L output samples. The process continues in this way, updating the filter every L samples. Thus, for every new set of L output samples, the adaptation overhead involves $O(M)$ operations for the decision module D, and on the order of $2L$ multiplications and $2L$ additions for calculating the estimates \hat{P}_x and $\hat{P}_y[N_0]$. Assuming that L is much greater than N_{max} , the overhead is approximately two multiplications and two additions per output sample.

As discussed throughout this chapter, the goal of approximate filtering algorithms is to dynamically adjust the order of a digital filter such that the minimum possible filter order is used while assuring that the output SNR is maintained above a given threshold. For this purpose we may choose to estimate the optimal filter order, given a block of L output samples, and use our estimate of the optimal filter order to filter the next block of input samples. This is the method we have presented thus far and is termed the *optimal control strategy*. Alternatively, we may instead decide to use a simpler filter order update method. This simpler update method involves estimating, from a given block of L output samples, *in which direction* the filter order should be incremented (up or down) by a fixed amount. The newly incremented filter order is then used to filter the next block of input samples. We call this simpler method the *incremental control strategy*. These are two possible *control strategies* that may be used in conjunction with an approximate filter structure \mathcal{H} to incorporate flexibility in meeting required processing specifications.

In closing, we summarize what is needed to define an approximate filtering algorithm. The three definitive elements of an approximate filtering algorithm are

- the power window length L and nominal filter order N_0
- the filter structure \mathcal{H}
- the control strategy for dynamically updating the filter order based on \hat{N}_{LP}^*

In this chapter we have discussed the issues involved in choosing the power window length L and nominal filter order N_0 , which both have an important influence on the convergence properties of the LP estimators $\widehat{\text{ISNR}}_{\text{LP}}[N_0]$ and \hat{N}_{LP}^* . We have deferred the important considerations that must be made with respect to the choice of the filter structure \mathcal{H} to Chapter 4. Two possibilities for the control strategy were presented in this chapter, namely the *optimal control strategy* and the *incremental control strategy*. Other control strategies are certainly possible. It should be clear at this point that approximate filtering algorithms are flexible; we may fine tune them via the choice of the three elements listed above to meet the requirements of a particular application.

2.6 Summary

We have considered the practical problem of dynamically reducing the order of a frequency-selective digital filter to reduce average power consumption, and presented the class of approximate filtering algorithms for which this is accomplished. Approximate filtering algorithms were developed by abstracting a theory from the practical low power filtering problem. The theory centered on the problem of determining an optimal filter order based on observations of the input data and a set of concrete assumptions on the statistics of the input signal. We explored the statistical properties of this theory, and showed that under certain assumptions the class of approximate filtering algorithms is asymptotically optimal. The theory served the purpose of aiding us in understanding interpreting the performance of approximate filtering algorithms.

Chapter 3

State Transition Error Analysis

The crux of the approximate filtering algorithms presented in Chapter 2 is that the order of a frequency selective digital filter may be dynamically varied based on the input signal statistics to reduce the time-averaged power consumption of the filtering operation while maintaining a fixed level of output quality. This possibility offers attractive potential for low power signal processing applications. A formula for how to cheaply compute the low power (LP) estimate \hat{N}_{LP}^* of the optimal filter order N^* was presented in Chapter 2 for applications involving wide sense stationary input signals, and an adaptation mechanism for computing the time-varying estimate \hat{N}_{LP}^* of the short-time optimal filter order N^* was established to accommodate nonstationary input signals. In approximate filtering the instantaneous change in the order of the digital frequency-selective filter has a corruptive effect on the output of the filter. This is due to the effective *state transition*. In this chapter we develop a model for this corruption as an additive noise term in the approximate filter output signal. We refer to this additive noise term as the *state transition error*.

Minimization of the state transition error (STE) in the approximate filter output is important for two reasons. First, in some applications we would like to produce the exact time series corresponding to the output of a fixed-order filter. For example, this may be the case if the temporal features of the output of a fixed filter are relevant, as is the case, for example, in heart rate monitoring and seismic signal processing applications. Instantaneously switching the filter order corrupts the fixed-order filter output and may partially or completely destroy the desired temporal features, rendering the approximate filter output ineffectual. Quantifying the corruption due to the STE as an additive noise term is the first step in trying to preserve some or all of the desired temporal features of the approximate

filter output. Secondly, for optimal low power performance in approximate filtering, it is important that the low power estimate \hat{N}_{LP}^* of the optimal filter order N^* is not vitiated by the STE component of the filter output. We will directly address both of these issues in this chapter via the development of deterministic and probabilistic frameworks for STE analysis. Further, in Chapter 4 we will use the results from this chapter for investigating how to choose approximate filter structures in order to *minimize the effect of the STE*.

Although we are generally interested in exploring the properties of the STE due to periodic switching of the filter order, for the purpose of theoretical simplicity in this chapter we consider the output of an approximate filter with a single state transition (filter order switch) from filter order N_1 to filter order N_2 at time $n = 0$. This is a simplification of the periodic switching which occurs in the approximate filtering algorithm for non-stationary inputs. This simplification will adequately serve our purposes in this chapter, since the STE will be shown to decay exponentially and become negligible a few samples after the filter order is switched. In order to analyze and understand the STE, we view the output $y_{N_1 N_2}[n]$ of an approximate filter with a single state transition as the sum of the output $y_{N_2 N_2}[n]$ of a filter with fixed order N_2 and the STE denoted by $y_{tr}[n]$. This relationship may be expressed as

$$y_{N_1 N_2}[n] = y_{N_2 N_2}[n] + y_{tr}[n]. \quad (3.1)$$

The notation for the approximate filter output $y_{N_1 N_2}[n]$ is used to emphasize that the filter order switches once from order N_1 to order N_2 at time $n = 0$. In this chapter the notation for the fixed filter output $y_{N_2 N_2}[n]$ is used to emphasize that in this case the filter order *does not switch* at time $n = 0$. We now develop deterministic and probabilistic frameworks for analyzing the STE. In the deterministic framework we determine a bound on the magnitude of the STE at any post-transition sample number. In the probabilistic framework we explore the impact of the STE on: 1) the approximate filter output $y_{N_1 N_2}[n]$, 2) the approximate filter output power estimate

$$\hat{P}_y[N_0] = \sum_{k=0}^{L-1} y_{N_1 N_2}^2[k], \quad (3.2)$$

and 3) the low power filter order estimate \hat{N}_{LP}^* which is based on the approximate filter output $y_{N_1 N_2}[n]$.

By considering the impact of the STE on these quantities, we demonstrate in the probabilistic framework that the STE is *essentially negligible*. Formally, we accomplish this by evaluating the following statistical metrics:

1. $E\{y_{tr}^2[n]\}$, the expected value of the squared STE,
2. $E\{P_{tr}[n]\}$, the expected value of the total additive corruption induced into the output power estimate $\hat{P}_y[N_0]$ due to the STE, and
3. $\text{Prob}(\hat{N}_k \geq \hat{N}_{LP}^* \mid y_{N_1 N_2}[n] = y_{N_2 N_2}[n] + y_{tr}[n])$, the probability that the STE-corrupted filter order estimate \hat{N}_k is greater than or equal to the ideal STE-free filter order estimate \hat{N}_{LP}^* . This is a statistical measure of the effect of the STE on the LP estimate of the optimal filter order. We use this probabilistic *excedance* measure since the nature of our problem is that it is much better to have $\hat{N}_k \geq \hat{N}_{LP}^*$, in which case we are assured that the output SNR is greater than or equal to the minimum tolerable level, than it is to have $\hat{N}_k < \hat{N}_{LP}^*$, in which case the output SNR can be less than the minimum tolerable level.

It is important to point out immediately that *the STE is always zero when we use an FIR approximate filter structure*. This is true since in FIR filtering each output sample does not depend directly on previous output samples. Therefore we may change filter orders transparently with FIR filters and the STE is always zero. This issue will be revisited in Chapter 4 in our study of approximate filter structures. Thus, it is only when we use an IIR approximate filter structure that we need to be concerned with the STE.

We now make a final note before proceeding to establish the deterministic and probabilistic frameworks for STE analysis. Our present purpose is to motivate our investigation of the STE by considering Fig. 3-1 in which we show a comparison of the approximate filter output, denoted by $y_{N_1 N_2}[n]$, and the output of a fixed filter of order N_2 , denoted by $y_{N_2 N_2}[n]$. The bottom plot in Fig. 3-1 depicts the absolute value of the STE $|y_{tr}[n]| = |y_{N_1 N_2}[n] - y_{N_2 N_2}[n]|$, which is visually small compared to $|y_{N_1 N_2}[n]|$ or $|y_{N_2 N_2}[n]|$. In addition, the STE signal appears to decay rapidly. In light of this observation, we are motivated to establish a bound on the magnitude of the STE as well as an expression for the expected value of the squared

STE. Our goal is to establish that we may view the output of the approximate filter $y_{N_1 N_2}[n]$ as being approximately equal to the fixed filter output $y_{N_2 N_2}[n]$.

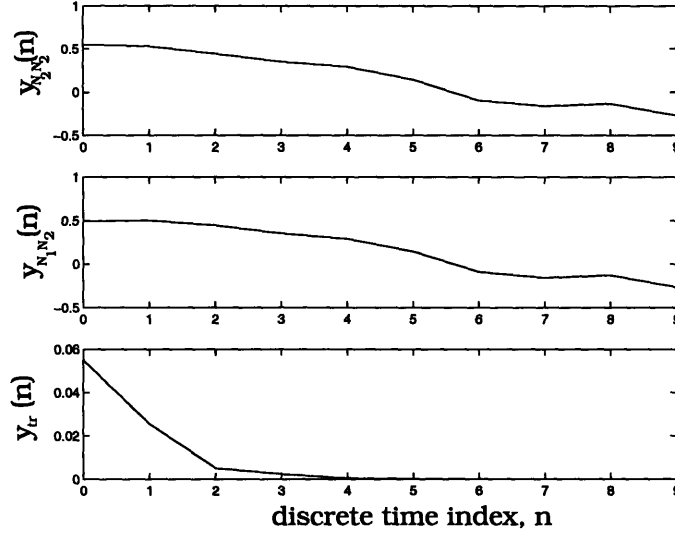


Figure 3-1: Comparison of the approximate filter output $y_{N_1 N_2}[n]$ and the fixed filter output $y_{N_2 N_2}[n]$. The bottom plot depicts the absolute value of the state transition error $|y_{tr}[n]| = |y_{N_1 N_2}[n] - y_{N_2 N_2}[n]|$. The filter order switch (state transition) occurred at time $n = 0$ in this case. In this illustrative example we used $N_1 = 2$ and $N_2 = 4$. The output signals were generated using the replacement IIR Butterworth filter structure.

3.1 Deterministic Analysis

We now address the problem of deriving a deterministic upper bound B_{tr} on the absolute value of the STE $y_{tr}[n]$ such that $|y_{tr}[n]| < B_{tr}$ for $n \geq 0$. To review, the output $y_{N_1 N_2}[n]$ of the approximate filter is produced by a filter which instantaneously switches its order from N_1 to N_2 at time $n = 0$. The STE, denoted $y_{tr}[n]$, is defined to be the difference between the approximate filter output $y_{N_1 N_2}[n]$ and the output $y_{N_2 N_2}[n]$ which is obtained by filtering the same input sequence with a filter of fixed order N_2 for all time. That is,

$$y_{tr}[n] = y_{N_1 N_2}[n] - y_{N_2 N_2}[n], \quad n \geq 0, \quad (3.3)$$

where $y_{N_2 N_2}[n] = \sum_{k=-\infty}^{\infty} x[n-k]h_{N_2}[k]$ and $y_{N_1 N_2}[n]$ is the approximate filter output. We will demonstrate in this section that the STE decays exponentially after the filter switch. Before proceeding with the details of the derivation of the bound B_{tr} on the STE, we present a brief summary of the final result.

Deterministic Bound on the STE: B_{tr}

The deterministic bound on the magnitude of the STE is B_{tr} , and the STE satisfies

$$|y_{\text{tr}}[n]| = |y_{N_2 N_2}[n] - y_{N_1 N_2}[n]| \quad (3.4)$$

$$\leq B_{\text{tr}} \cdot |z_{\text{max}}|^n, \quad (3.5)$$

where

$$B_{\text{tr}} = N_2 \cdot B_{\mathbf{y}_{\text{tr}}} \cdot \lambda_{\min}^{-1/2}(\mathbf{Z}^T \mathbf{Z}), \quad (3.6)$$

N_2 is the post-transition order of the approximate filter, $B_{\mathbf{y}_{\text{tr}}}$ is a bound on the norm of the $L \times 1$ vector of the first L post-transition samples of the STE, $\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})$ is the minimum eigenvalue of the matrix $\mathbf{Z}^T \mathbf{Z}$, and \mathbf{Z} is the Vandermonde matrix corresponding to the filter $h_{N_2}[n]$. Note that the parameters N_2 , $B_{\mathbf{y}_{\text{tr}}}$, and $\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})$ in Eq. (3.6) depend only on our choice of the impulse responses $h_{N_1}[n]$ and $h_{N_2}[n]$ and the input signal bound B_X , which is defined such that $|x[n]| \leq B_X$ for all n .

3.1.1 Derivation of Deterministic Bound

We begin our deterministic analysis of the STE by deriving a bound $B_X \cdot B_{N_1 N_2}$ on the outputs of two distinct filters of orders N_1 and N_2 with the same bounded input signal satisfying $|x[n]| \leq B_X$. This result will be used to find an expression for B_{tr} . Consider the outputs of two distinct filters with impulse responses $h_{N_1}[n]$ and $h_{N_2}[n]$ and orders N_1 and N_2 , respectively. Suppose the two filters $h_{N_1}[n]$ and $h_{N_2}[n]$ belong to the same approximate filter structure \mathcal{H} , and have outputs $y_{N_1 N_1}[n]$ and $y_{N_2 N_2}[n]$, respectively. If the signal $x[n]$ is the input to both of the filters, the output signals may be written in terms of their convolution sums as

$$y_{N_i N_i}[n] = \sum_{k=-\infty}^{\infty} x[n-k] h_{N_i}[k], \quad i = 1, 2. \quad (3.7)$$

The difference $\gamma_{N_1 N_2}[n]$ between the output signals $y_{N_1 N_1}[n]$ and $y_{N_2 N_2}[n]$ is

$$\gamma_{N_1 N_2}[n] = y_{N_1 N_1}[n] - y_{N_2 N_2}[n] = \sum_{k=-\infty}^{\infty} x[n-k] [h_{N_1}[k] - h_{N_2}[k]]. \quad (3.8)$$

Applying the Cauchy-Schwartz inequality, we bound the magnitude of $\gamma_{N_1 N_2}[n]$ as

$$\begin{aligned} |\gamma_{N_1 N_2}[n]| &= \left| \sum_{k=-\infty}^{\infty} x[n-k] [h_{N_1}[k] - h_{N_2}[k]] \right| \\ &\leq \sum_{k=-\infty}^{\infty} |x[n-k]| |h_{N_1}[k] - h_{N_2}[k]|. \end{aligned} \quad (3.9)$$

If we assume that each of the values in the input signal $x[n]$ is bounded such that for all n

$$|x[n]| \leq B_X, \quad (3.10)$$

for some finite, positive number B_X , then a bound on the magnitude of $\gamma_{N_1 N_2}[n]$ is given by

$$\begin{aligned}
|\gamma_{N_1 N_2}[n]| &\leq B_X \sum_{k=-\infty}^{\infty} |h_{N_1}[k] - h_{N_2}[k]| \\
&\leq B_X \cdot B_{N_1 N_2},
\end{aligned} \tag{3.11}$$

where

$$B_{N_1 N_2} = \sum_{k=-\infty}^{\infty} |h_{N_1}[k] - h_{N_2}[k]|. \tag{3.12}$$

Thus, for input signals satisfying $|x[n]| \leq B_X$, the two fixed filter outputs $y_{N_1 N_1}[n]$ and $y_{N_2 N_2}[n]$ differ in absolute value by at most $B_X \cdot B_{N_1 N_2}$. Note that this bound is *not* a function of n . It is worthy to note that $(M^2 - M)/2$ distinct values of $B_{N_1 N_2}$ may be computed for an approximate filter structure \mathcal{H} having M distinct filter elements. These values constitute one measure the STE performance of the approximate filter structure \mathcal{H} , and may be taken into consideration in the design of approximate filter structures.

In Fig. 3-2 and Fig. 3-3 we show plots of the bound $B_{N_1 N_2}$ as a function of N_1 and N_2 for the replacement Butterworth filter structure and the truncated Butterworth filter structure, respectively. Some important properties of these filter structures were discussed in Chapter 1, and they will be further investigated in Chapter 4. We note that along the line $N_1 = N_2$ the bound $B_{N_1 N_2}$ is zero since the STE is zero when $N_1 = N_2$ which is equivalent to having no filter order switch occur. We also note that as $|N_1 - N_2|$ increases, so does $B_{N_1 N_2}$ symmetrically about the line $N_1 = N_2$. This is what we expect since as $|N_1 - N_2|$ increases the spectral differences in $h_{N_1}[n]$ and $h_{N_2}[n]$ become more profound. The shape of the curve for the replacement Butterworth filter structure in Fig. 3-2 is smoother than that of the truncation Butterworth filter structure in Fig. 3-3 due to the especially abrupt changes in spectral magnitude shape between filters of different orders in the truncation Butterworth filter structure which are not as prominent in the replacement Butterworth filter structure.

In addition, in Table 3.1 and Table 3.2 we have provided the numerical values of the bound $B_{N_1 N_2}$ as a function of N_1 and N_2 for the replacement Butterworth filter structure and the truncated Butterworth approximate filter structure, respectively. The half-power

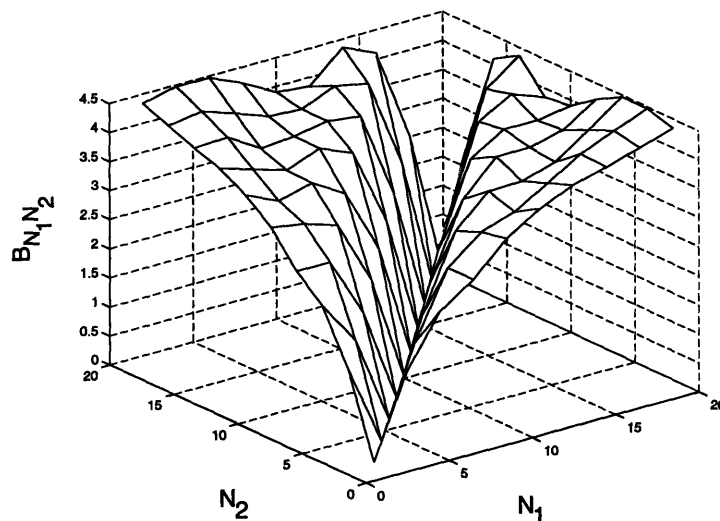


Figure 3-2: A plot of the deterministic bound $B_{N_1 N_2}$ vs. N_1 and N_2 for the replacement Butterworth filter structure. Refer to Eq. (3.12) for the definition of $B_{N_1 N_2}$.

frequency in each case is $\pi/2$.

We now return our attention to deriving the deterministic bound B_{tr} on the STE, which we recall is defined such that

$$|y_{\text{tr}}[n]| = |y_{N_1 N_2}[n] - y_{N_2 N_2}[n]| \leq B_{\text{tr}}, \quad n \geq 0. \quad (3.13)$$

We will show that such a bound exists, and that the STE decays exponentially in time. The STE induced by switching from filter $h_{N_1}[n]$ to $h_{N_2}[n]$ at time $n = 0$ may be written in the form

$$|y_{\text{tr}}[n]| = B_{\text{tr}} \cdot |\alpha|^n, \quad n \geq 0 \quad (3.14)$$

for some α such that $|\alpha| < 1$ and for some positive real number B_{tr} . We will find an expression for B_{tr} and α in terms of the filter coefficients $h_{N_1}[n]$ and $h_{N_2}[n]$ and the input signal bound B_X which is chosen such that the input signal satisfies $x[n] \leq B_X$. To begin, consider the order- N_2 difference equation

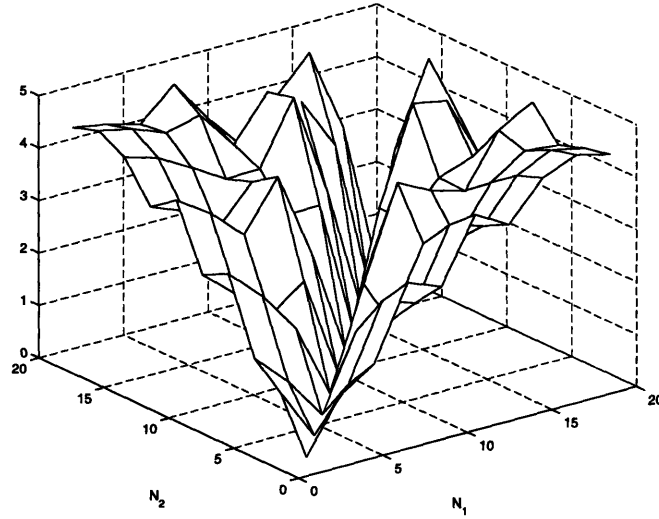


Figure 3-3: A plot of the deterministic bound $B_{N_1 N_2}$ vs. N_1 and N_2 for the truncated Butterworth filter structure. Refer to Eq. (3.12) for the definition of $B_{N_1 N_2}$.

$$y_{N_2 N_2}[n] = \sum_{k=1}^{N_2} a_{k N_2} y_{N_2 N_2}[n-k] + \sum_{k=0}^{N_2} b_{k N_2} x[n-k], \quad (3.15)$$

where $a_{k N_2}$ and $b_{k N_2}$ are the k th coefficients in the order- N_2 polynomials (ordered in ascending powers of z^{-1}) whose roots define the poles and zeros of the order- N_2 filter $h_{N_2}[n]$, respectively. Fundamental linear system theory dictates that the solution to the difference equation in Eq. (3.15) may be expressed as the sum of a zero-input response (ZIR) $y_{N_2 N_2}^{\text{ZIR}}[n]$ and a zero-state response (ZSR) $y_{N_2 N_2}^{\text{ZSR}}[n]$ which satisfy

$$y_{N_2 N_2}[n] = y_{N_2 N_2}^{\text{ZSR}}[n] + y_{N_2 N_2}^{\text{ZIR}}[n], \quad (3.16)$$

where

$$y_{N_2 N_2}^{\text{ZSR}}[n] = \sum_{k=0}^{N_2} b_{k N_2} x[n-k], \quad (3.17)$$

and

Table 3.1: Numerical values for the deterministic bound $B_{N_1N_2}$ for the replacement Butterworth filter structure. Refer to Eq. (3.12) for the definition of $B_{N_1N_2}$.

Filter Order N_1	Filter Order N_2									
	3	5	7	9	11	13	15	17	19	21
3	0	1.41	2.30	2.69	3.23	3.58	3.84	3.99	4.16	4.34
5	1.41	0	1.54	2.68	3.13	3.32	3.66	4.01	4.25	4.47
7	2.30	1.54	0	1.66	2.95	3.48	3.49	3.66	4.06	4.47
9	2.69	2.68	1.66	0	1.86	3.13	3.71	3.71	3.85	4.22
11	3.23	3.13	2.95	1.86	0	2.03	3.47	3.82	3.81	3.90
13	3.58	3.32	3.48	3.13	2.03	0	2.14	3.74	4.05	3.84
15	3.84	3.66	3.49	3.71	3.47	2.14	0	2.23	3.94	4.36
17	3.99	4.01	3.66	3.71	3.82	3.74	2.23	0	2.41	4.10
19	4.16	4.25	4.06	3.85	3.81	4.05	3.94	2.41	0	2.52
21	4.34	4.47	4.47	4.22	3.90	3.84	4.36	4.10	2.52	0

$$y_{N_2N_2}^{\text{ZIR}}[n] = \sum_{k=1}^{N_2} a_{kN_2} y_{N_2N_2}[n-k], \quad (3.18)$$

with initial conditions

$$y_{N_2N_2}[-N_2] \cdots y_{N_2N_2}[-1] = \bar{y}_{N_2N_2}[-N_2] \cdots \bar{y}_{N_2N_2}[-1]. \quad (3.19)$$

The initial condition values $\bar{y}_{N_2N_2}[-N_2] \cdots \bar{y}_{N_2N_2}[-1]$ in Eq. (3.19) are produced by filtering the input signal $x[n]$ with the filter $h_{N_2}[n]$ on the interval $-\infty < n \leq -1$.

The approximate filter output signal $y_{N_1N_2}[n]$ satisfies a set of equations similar to Eqs. (3.16)–(3.18) given by

$$y_{N_1N_2}[n] = y_{N_1N_2}^{\text{ZSR}}[n] + y_{N_1N_2}^{\text{ZIR}}[n], \quad (3.20)$$

where

Table 3.2: Numerical values for the deterministic bound $B_{N_1 N_2}$ for the truncated Butterworth filter structure. Refer to Eq. (3.12) for the definition of $B_{N_1 N_2}$.

Filter Order N_1	Filter Order N_2									
	3	5	7	9	11	13	15	17	19	21
3	0	1.00	1.42	2.80	2.56	3.75	3.39	4.11	4.38	4.21
5	1.00	0	0.80	2.21	2.01	3.32	3.13	3.92	4.24	4.10
7	1.42	0.80	0	1.41	1.43	2.74	2.64	3.56	3.99	3.93
9	2.80	2.21	1.41	0	1.92	3.67	3.35	3.17	4.15	4.52
11	2.56	2.01	1.43	1.92	0	2.04	2.54	3.19	3.42	3.69
13	3.75	3.32	2.74	3.67	2.04	0	2.55	4.37	4.25	3.00
15	3.39	3.13	2.64	3.35	2.54	2.55	0	2.41	3.89	3.74
17	4.11	3.92	3.56	3.17	3.19	4.37	2.41	0	2.89	4.42
19	4.38	4.24	3.99	4.15	3.42	4.25	3.89	2.89	0	2.83
21	4.21	4.10	3.93	4.52	3.69	3.00	3.74	4.42	2.83	0

$$y_{N_1 N_2}^{\text{ZSR}}[n] = \sum_{k=0}^{N_2} b_{k N_2} x[n-k], \quad (3.21)$$

and

$$y_{N_1 N_2}^{\text{ZIR}}[n] = \sum_{k=1}^{N_2} a_{k N_2} y_{N_1 N_2}[n-k], \quad (3.22)$$

with *different* initial conditions

$$y_{N_1 N_2}[-N_2] \cdots y_{N_1 N_2}[-1] = \bar{y}_{N_1 N_1}[-N_2] \cdots \bar{y}_{N_1 N_1}[-1] \quad (3.23)$$

The initial condition values $\bar{y}_{N_1 N_1}[-N_2] \cdots \bar{y}_{N_1 N_1}[-1]$ in Eq. (3.23) are produced by filtering the input $x[n]$ with the filter $h_{N_1}[n]$ on the interval $-\infty \leq n < -1$.

The STE $y_{\text{tr}}[n]$ is defined on the interval $-N_2 \leq n < \infty$ as

$$y_{\text{tr}}[n] = y_{N_1 N_2}[n] - y_{N_2 N_2}[n], \quad n \geq -N_2, \quad (3.24)$$

which may be expanded to produce

$$y_{\text{tr}}[n] = y_{N_1 N_2}^{\text{ZSR}}[n] + y_{N_1 N_2}^{\text{ZIR}}[n] - y_{N_2 N_2}^{\text{ZSR}}[n] - y_{N_2 N_2}^{\text{ZIR}}[n]. \quad (3.25)$$

This may be expanded yet again to obtain

$$\begin{aligned} y_{\text{tr}}[n] &= \sum_{k=0}^{N_2} b_{k N_2} x[n-k] + \sum_{k=1}^{N_2} a_{k N_2} y_{N_1 N_2}[n-k] \\ &\quad - \sum_{k=0}^{N_2} b_{k N_2} x[n-k] - \sum_{k=1}^{N_2} a_{k N_2} y_{N_2 N_2}[n-k], \end{aligned} \quad (3.26)$$

which simplifies to

$$y_{\text{tr}}[n] = \underbrace{\sum_{k=1}^{N_2} a_{k N_2} y_{N_1 N_2}[n-k]}_{y_{N_1 N_2}^{\text{ZIR}}[n]} - \underbrace{\sum_{k=1}^{N_2} a_{k N_2} y_{N_2 N_2}[n-k]}_{y_{N_2 N_2}^{\text{ZIR}}[n]}, \quad (3.27)$$

and thus

$$y_{\text{tr}}[n] = y_{N_1 N_2}^{\text{ZIR}}[n] - y_{N_2 N_2}^{\text{ZIR}}[n]. \quad (3.28)$$

This implies that the STE satisfies the recursion relation

$$y_{\text{tr}}[n] = \sum_{k=1}^{N_2} a_{k N_2} [y_{N_1 N_2}[n-k] - y_{N_2 N_2}[n-k]], \quad (3.29)$$

or equivalently satisfies

$$y_{\text{tr}}[n] = \sum_{k=1}^{N_2} a_{k N_2} y_{\text{tr}}[n-k], \quad (3.30)$$

with initial conditions

$$y_{\text{tr}}[-N_2] \cdots y_{\text{tr}}[-1] = [\bar{y}_{N_1 N_1}[-N_2] - \bar{y}_{N_2 N_2}[-N_2]] \cdots [\bar{y}_{N_1 N_1}[-1] - \bar{y}_{N_2 N_2}[-1]]. \quad (3.31)$$

We note that each of the N_2 initial conditions $y_{\text{tr}}[-N_2] \cdots y_{\text{tr}}[-1]$ is of the form of $\gamma_{N_1 N_2}[n]$ given in Eq. (3.8). Therefore each initial condition in this set has a magnitude satisfying

$$|y_{\text{tr}}[n]| \leq B_X \cdot B_{N_1 N_2} \quad -N_2 \leq n \leq -1 \quad (3.32)$$

This result will be invoked later to establish the bound B_{tr} .

For now we consider the expression in Eq. (3.28) for the STE for $n \geq 0$. This expression identifies the STE as a pure ZIR for $n \geq 0$. We note that the STE does not have the form of a pure ZIR on the interval $-N_2 \leq n \leq -1$. Nevertheless, for $n \geq 0$ the STE may be expressed as a weighted sum of exponential signals [54],

$$y_{\text{tr}}[n] = \sum_{k=1}^{N_2} c_k z_k^n, \quad (3.33)$$

where the z_k are the N_2 pole locations of the filter $h_{N_2}[n]$ and the c_k are scalar coefficients. Note that if the filter $h_{N_2}[n]$ is stable, then $|z_k| < 1$ for $1 \leq k \leq N_2$. We may write Eq. (3.33) in matrix form as

$$\underbrace{\begin{bmatrix} y_{\text{tr}}[0] \\ y_{\text{tr}}[1] \\ \vdots \\ y_{\text{tr}}[L-1] \end{bmatrix}}_{\mathbf{y}_{\text{tr}}} = \underbrace{\begin{bmatrix} z_1^0 & z_2^0 & \cdots & z_{N_2}^0 \\ z_1^1 & z_2^1 & \cdots & z_{N_2}^1 \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{L-1} & z_2^{L-1} & \cdots & z_{N_2}^{L-1} \end{bmatrix}}_{\mathbf{Z}} \underbrace{\begin{bmatrix} c_1 \\ \vdots \\ c_{N_2} \end{bmatrix}}_{\mathbf{c}}, \quad (3.34)$$

which can be more compactly written as

$$\mathbf{y}_{\text{tr}} = \mathbf{Z}\mathbf{c}, \quad (3.35)$$

for the $L \times 1$ vector \mathbf{y}_{tr} , the $L \times N_2$ Vandermonde matrix \mathbf{Z} , and the $N_2 \times 1$ vector \mathbf{c} . Each of the N_2 columns of \mathbf{Z} is made up of the one of the N_2 poles of $h_{N_2}[n]$ raised to successively higher powers from zero to $L - 1$ in rows 1 to L , as shown in Eq. (3.34). Recall that by definition of the induced matrix 2-norm [67]

$$\|\mathbf{Z}\| = \max_{\mathbf{c} \neq 0} \frac{\|\mathbf{Z}\mathbf{c}\|}{\|\mathbf{c}\|}. \quad (3.36)$$

This implies that

$$\|\mathbf{Z}\mathbf{c}\| \leq \|\mathbf{Z}\| \cdot \|\mathbf{c}\|. \quad (3.37)$$

To determine an expression for $\|\mathbf{Z}\|$, following Strang [67] we observe

$$\|\mathbf{Z}\|^2 = \max_{\mathbf{c} \neq 0} \frac{\|\mathbf{Z}\mathbf{c}\|^2}{\|\mathbf{c}\|^2} = \max_{\mathbf{c} \neq 0} \frac{\mathbf{c}^T \mathbf{Z}^T \mathbf{Z} \mathbf{c}}{\mathbf{c}^T \mathbf{c}}. \quad (3.38)$$

But

$$\mathbf{Z}^T \mathbf{Z} \mathbf{c} = \lambda \mathbf{c} \quad (3.39)$$

if λ is an eigenvalue of the matrix $\mathbf{Z}^T \mathbf{Z}$, so

$$\max_{\mathbf{c} \neq 0} \frac{\mathbf{c}^T (\mathbf{Z}^T \mathbf{Z} \mathbf{c})}{\mathbf{c}^T \mathbf{c}} = \max_{\mathbf{c} \neq 0} \frac{\mathbf{c}^T (\lambda \mathbf{c})}{\mathbf{c}^T \mathbf{c}} = \max_{\mathbf{c} \neq 0} \lambda \frac{\mathbf{c}^T \mathbf{c}}{\mathbf{c}^T \mathbf{c}} = \max_{\mathbf{c} \neq 0} \lambda = \lambda_{\max}(\mathbf{Z}^T \mathbf{Z}). \quad (3.40)$$

Therefore, combining Eq. (3.38) and Eq. (3.40) produces

$$\|\mathbf{Z}\|^2 = \lambda_{\max}(\mathbf{Z}^T \mathbf{Z}), \quad (3.41)$$

and

$$\|\mathbf{Z}\mathbf{c}\|^2 \leq \lambda_{\max}(\mathbf{Z}^T \mathbf{Z}) \|\mathbf{c}\|^2, \quad (3.42)$$

where $\lambda_{\max}(\mathbf{Z}^T \mathbf{Z})$ is the largest eigenvalue of the matrix $\mathbf{Z}^T \mathbf{Z}$. A similar argument can be made [67] to show that

$$\|\mathbf{Z}\mathbf{c}\|^2 \geq \lambda_{\min}(\mathbf{Z}^T \mathbf{Z}) \|\mathbf{c}\|^2, \quad (3.43)$$

where $\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})$ is the smallest eigenvalue of the matrix $\mathbf{Z}^T \mathbf{Z}$. We now proceed to bound $\|\mathbf{c}\|$ using Eq. (3.43). Since $\mathbf{y}_{\text{tr}} = \mathbf{Z}\mathbf{c}$, we know from Eq. (3.43) that

$$\|\mathbf{y}_{\text{tr}}\|^2 \geq \lambda_{\min}(\mathbf{Z}^T \mathbf{Z}) \|\mathbf{c}\|^2. \quad (3.44)$$

Rearranging, the inequality in Eq. (3.44) becomes

$$\begin{aligned} \|\mathbf{c}\| &\leq \frac{\|\mathbf{y}_{\text{tr}}\|}{\sqrt{\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})}} \\ &\leq \frac{B_{\mathbf{y}_{\text{tr}}}}{\sqrt{\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})}}, \end{aligned} \quad (3.45)$$

where $B_{\mathbf{y}_{\text{tr}}}$ is defined such that $\|\mathbf{y}_{\text{tr}}\| \leq B_{\mathbf{y}_{\text{tr}}}$. To determine $B_{\mathbf{y}_{\text{tr}}}$, we determine bounds on the first two individual elements $y_{\text{tr}}[0]$ and $y_{\text{tr}}[1]$ of the vector \mathbf{y}_{tr} , and then recursively compute bounds on the remaining elements $y_{\text{tr}}[2] \cdots y_{\text{tr}}[L-1]$. For the first element of \mathbf{y}_{tr} we have

$$\begin{aligned}
y_{\text{tr}}[0] &= \left| \sum_{k=1}^{N_2} a_{kN_2} y_{\text{tr}}[-k] \right| \\
&\leq B_X B_{N_1 N_2} \sum_{k=1}^{N_2} |a_{kN_2}| = Y_{\text{tr}}[0]
\end{aligned} \tag{3.46}$$

and for the second element of \mathbf{y}_{tr} we have

$$\begin{aligned}
y_{\text{tr}}[1] &= \left| \sum_{k=1}^{N_2} a_{kN_2} y_{\text{tr}}[1-k] \right| \\
&= \left| \sum_{k=2}^{N_2} a_{kN_2} y_{\text{tr}}[1-k] + a_{1N_2} y_{\text{tr}}[0] \right| \\
&\leq \left| \sum_{k=2}^{N_2} a_{kN_2} y_{\text{tr}}[1-k] \right| + |a_{1N_2} y_{\text{tr}}[0]| \\
&\leq \left| B_X B_{N_1 N_2} \sum_{k=1}^{N_2} |a_{kN_2}| \right| + \left| B_X B_{N_1 N_2} a_{1N_2} \sum_{k=1}^{N_2} |a_{kN_2}| \right| \\
&\leq B_X B_{N_1 N_2} [1 + a_{1N_2}] \sum_{k=1}^{N_2} |a_{kN_2}| = Y_{\text{tr}}[1].
\end{aligned} \tag{3.47}$$

Building on the form of Eq. (3.46) and Eq. (3.47), we now present a recursive expression for the bounds on each of the elements of the vector \mathbf{y}_{tr}

$$y_{\text{tr}}[l] \leq Y_{\text{tr}}[l] \quad l \geq 0, \tag{3.48}$$

where

$$Y_{\text{tr}}[l] = B_X B_{N_1 N_2} \sum_{k=1}^{N_2} |a_{kN_2}| + \sum_{k=0}^{l-1} |a_{l-k, N_2}| Y_{\text{tr}}[k], \quad l > 0, \tag{3.49}$$

with initial condition

$$Y_{\text{tr}}[0] = B_X B_{N_1 N_2} \sum_{k=1}^{N_2} |a_{k N_2}|. \quad (3.50)$$

Using Eq. (3.49) and Eq. (3.50) we may generate each of the bounds $Y_{\text{tr}}[l]$ for $0 \leq l \leq L-1$, and then determine $B_{y_{\text{tr}}}$ using

$$\begin{aligned} \|\mathbf{y}_{\text{tr}}\|^2 &= \sum_{k=0}^{L-1} |y_{\text{tr}}[k]|^2 \\ &\leq \sum_{k=0}^{L-1} Y_{\text{tr}}^2[k] = B_{y_{\text{tr}}}^2. \end{aligned} \quad (3.51)$$

We shall use this result in the forthcoming derivation of B_{tr} . We start by recalling that the post-transition STE is a pure ZIR and thus may be expressed as

$$y_{\text{tr}}[n] = \sum_{k=1}^{N_2} c_k z_k^n, \quad n \geq 0, \quad (3.52)$$

so that

$$\begin{aligned} |y_{\text{tr}}[n]| &= \left| \sum_{k=1}^{N_2} c_k z_k^n \right| \\ &\leq \sum_{k=1}^{N_2} |c_k| \cdot |z_k^n| \\ &\leq |z_{\max}|^n \sum_{k=1}^{N_2} |c_k| \\ &\leq N_2 \cdot \|\mathbf{c}\| \cdot |z_{\max}|^n, \end{aligned} \quad (3.53)$$

where z_{\max} is defined to be the pole location of the filter $h_{N_2}[n]$ with the largest magnitude. We have used the fact that

$$\frac{1}{N_2} \sum_{k=1}^{N_2} |c_k| \leq \|\mathbf{c}\|, \quad (3.54)$$

where $\|\mathbf{c}\|$ is the conventional 2-norm of the vector \mathbf{c}

$$\|\mathbf{c}\| = \sqrt{\sum_{k=1}^{N_2} c_k^2}. \quad (3.55)$$

A proof of Eq. (3.55) is given in Appendix B. Substituting in the bound for $\|\mathbf{c}\|$ from Eq. (3.45) into Eq. (3.53) produces for $n \geq 0$

$$\begin{aligned} |y_{\text{tr}}[n]| &\leq N_2 \cdot \frac{\|\mathbf{y}_{\text{tr}}\|}{\sqrt{\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})}} \cdot |z_{\max}|^n \\ &\leq N_2 \cdot \frac{B_{\mathbf{y}_{\text{tr}}}}{\sqrt{\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})}} \cdot |z_{\max}|^n \\ &\leq B_{\text{tr}} \cdot |z_{\max}|^n, \end{aligned} \quad (3.56)$$

where

$$B_{\text{tr}} = N_2 \cdot B_{\mathbf{y}_{\text{tr}}} \cdot \lambda_{\min}^{-1/2}(\mathbf{Z}^T \mathbf{Z}). \quad (3.57)$$

To obtain the bound B_{tr} above we have used the result given in Eq. (3.51) that $\|\mathbf{y}_{\text{tr}}\| \leq B_{\mathbf{y}_{\text{tr}}}$. In Eq. (3.56) we have given a bound for the STE in terms of the bound $B_{\mathbf{y}_{\text{tr}}}$, the filter order N_2 , the pole location z_{\max} , and the eigenvalue $\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})$ of the matrix $\mathbf{Z}^T \mathbf{Z}$ defined in Eq. (3.34). Thus, our expression for the bound B_{tr} is now complete. A summary of the steps which are needed to obtain a numerical value for B_{tr} , given B_X , $h_{N_1}[n]$, and $h_{N_2}[n]$, is now given.

Summary of How to Compute B_{tr}

1. Given B_X , $h_{N_1}[n]$, and $h_{N_2}[n]$, compute $B_{N_1N_2}$ using Eq. (3.12)
2. Compute $\lambda_{\min}(\mathbf{Z}^T\mathbf{Z})$ as the minimum singular value of the matrix \mathbf{Z} defined in Eq. (3.34)
3. Compute $B_{y_{tr}}$ using Eqs. (3.48)–(3.51)
4. Compute B_{tr} using Eq. (3.57)

3.1.2 Simulations

The expression for the deterministic bound B_{tr} on the STE given in Eq. (3.57) is complicated; its evaluation requires the recursive computation of the bound $B_{y_{tr}}$ using Eq. (3.48) through Eq. (3.51) as well as the computation of the minimum singular value of the matrix \mathbf{Z} in Eq. (3.34). In order to gain more insight into the nature of the bound B_{tr} and its dependence on B_X , $h_{N_1}[n]$, and $h_{N_2}[n]$, in this section we provide a numerical evaluation of the STE and its bound B_{tr} . We compare the relative magnitudes of the STE and its deterministic bound B_{tr} using computer simulations.

To begin, we generated a random input sequence bounded by $B_X = 1$ whose samples were independent and identically distributed according to a uniform probability distribution. We filtered this input sequence with a filter taken from an IIR Butterworth approximate filter structure with half-power frequency $\pi/2$ and order N_1 . At time $n = 0$ the filter order was switched to N_2 and the maximum value of the STE was measured and recorded. This experiment was repeated for all pairs of N_1, N_2 in the range $2 \leq N_1, N_2 \leq 20$. The results are presented in Fig. 3-4 for the replacement Butterworth filter structure and in Fig. 3-5 for the truncation Butterworth filter structure.

It is clear from Fig. 3-4 and Fig. 3-5 that the maximum value of the STE is directly proportional to: 1) the absolute difference $|N_2 - N_1|$ between the filter orders, and 2) the value of N_2 . For large values of $|N_2 - N_1|$, the spectral magnitude responses of the two filters in the approximate filter structure differ significantly so the spectral components of the input signal which pass through the filter and the spectral components of the input

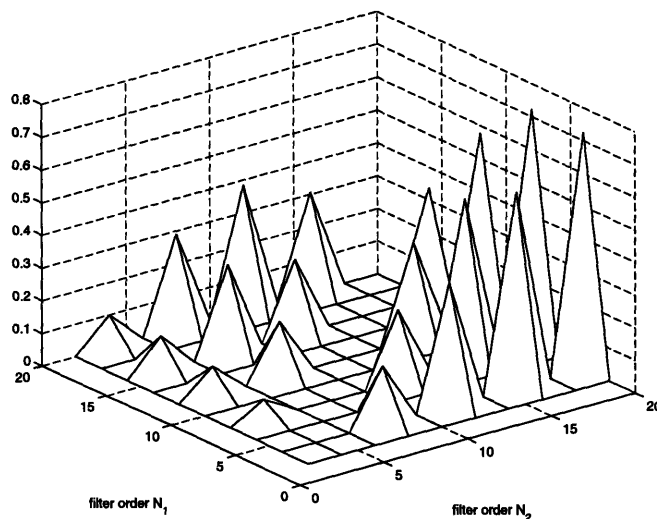


Figure 3-4: A plot of the maximum value of the state transition error vs. N_1 and N_2 for $n \geq 0$ using the replacement Butterworth filter structure with half-power frequency $\pi/2$. In this example $B_X = 1$, or, equivalently, $|x[n]| \leq 1$ for all n .

signal which are rejected by each filter differ significantly. This large spectral difference in filters $h_{N_1}[n]$ and $h_{N_2}[n]$ translates into a large maximum STE. In addition, the number of initial conditions which are weighted and summed to produce the STE at each post-transition time is equal to N_2 . Since in the case of an independent, identically distributed WSS random input signal the expected value of each of the N_2 initial conditions is the same, the more initial conditions there are (that is, the higher the value of N_2), the larger maximum value of the STE we would expect. These two dependencies are clearly visible in the mesh plots of the maximum value of the STE given in Fig. 3-4 and Fig. 3-5. We note that no matter how large the maximum STE is, over time the STE always decays exponentially according to Eq. (3.56). The speed of the STE exponential decay depends on how close the pole location z_{\max} is to the unit circle. For example, with $|z_{\max}| \ll 1$, the decay will be very fast, whereas for $|z_{\max}| \approx 1$, the decay of the STE will be much slower.

For the purpose of gaining further insight into the dependence of the STE on N_1 and N_2 , we consider the effect at time $n = 0$ of replacing the initial conditions leftover from using the filter $h_{N_1}[n]$ with *zero initial conditions* to produce the post-transition output samples. The results are presented in Fig. 3-6 for the replacement Butterworth filter structure and in Fig. 3-7 for the truncation Butterworth filter structure. In this case we can see clearly that the maximum value of the STE no longer depends on $|N_2 - N_1|$. Using the zero

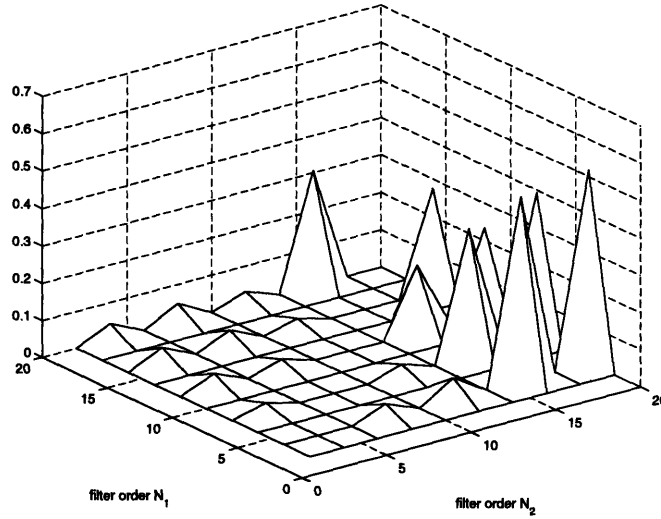


Figure 3-5: A plot of the maximum value of the state transition error vs. N_1 and N_2 for $n \geq 0$ using the truncated Butterworth filter structure with half-power frequency $\pi/2$. In this example $B_X = 1$, or, equivalently, $|x[n]| \leq 1$ for all n .

initial conditions is equivalent to using a filter order of $N_1 = 0$ to generate the STE initial conditions, regardless of the actual value of N_1 used to produce the approximate filter output. This explains why the STE bound using zero initial conditions does not depend on the filter order N_1 . The direct dependence of the maximum value of the STE on the filter order N_2 is still strongly evident.

Proceeding with our deterministic analysis, we now define the normalized STE as

$$y_{\text{nttr}}^2[n] = \frac{y_{\text{tr}}^2[n]}{B_X \cdot Y_{N_1} \cdot Y_{N_2}}, \quad (3.58)$$

where

$$Y_{N_k} = \sum_{m=-\infty}^{\infty} h_{N_k}[m] \quad k = 1, 2. \quad (3.59)$$

These normalization factors are defined such that $B_X \cdot Y_{N_k}$ is the maximum value that the output of a fixed filter of order N_k could attain for $k = 1, 2$.

Fig. 3-8 depicts a logarithmic plot of the normalized squared STE bound $B_{\text{tr}}^2 / (B_X \cdot Y_{N_1} \cdot Y_{N_2})$

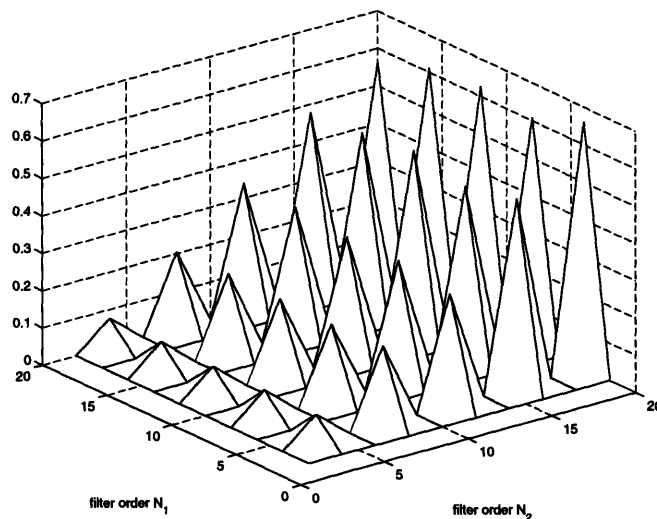


Figure 3-6: A plot of the maximum value of the state transition error vs. N_1 and N_2 for $n \geq 0$ using the replacement Butterworth filter structure with half-power frequency $\pi/2$ and zero initial conditions. In this example $B_X = 1$, or, equivalently, $|x[n]| \leq 1$ for all n .

vs. post-transition filter order N_2 for $N_2 = 2, 4, 6, 8$, and 10 . The points plotted in Fig. 3-8 were generated using the replacement Butterworth filter structure with $N_1 = 12$, using the expression for B_{tr} in Eq. (3.57). We see that as the post-transition filter order N_2 increases, so does the normalized squared STE bound $B_{tr}^2 / (B_X \cdot Y_{N_1} \cdot Y_{N_2})$.

Fig. 3-9 shows a logarithmic plot of the normalized squared STE vs. post transition sample number for $N_2 = 2, 4, 6, 8$, and 10 . The curves were generated for the replacement Butterworth filter structure with $N_1 = 12$, using the bounded time series from Eq. (3.56). The values for the bounded time series at $n = 0$ are equal to the normalized squared STE bound B_{tr}^2 shown in Fig. 3-8, and then exponentially decay with $|z_{max}|^n$.

Finally we plot the normalized squared STE in Fig. 3-10 found by averaging over 1000 Monte Carlo simulations. The plot includes the normalized squared STE bound B_{tr}^2 for $N_1 = 2$, $N_2 = 4$, and $B_X = 1$, using again the replacement Butterworth filter structure. It is evident from the plot that the average STE is approximately 5 orders of magnitude less than the deterministic bound. This suggests that the bound B_{tr}^2 is valid but very conservative. We are thus motivated to pursue the development of a probabilistic framework for STE analysis in which a more useful characterization of the STE may be obtained.

We now offer a closing comment before proceeding to the development of a probabilistic framework for STE analysis in the next section. First we notice that the normalized squared

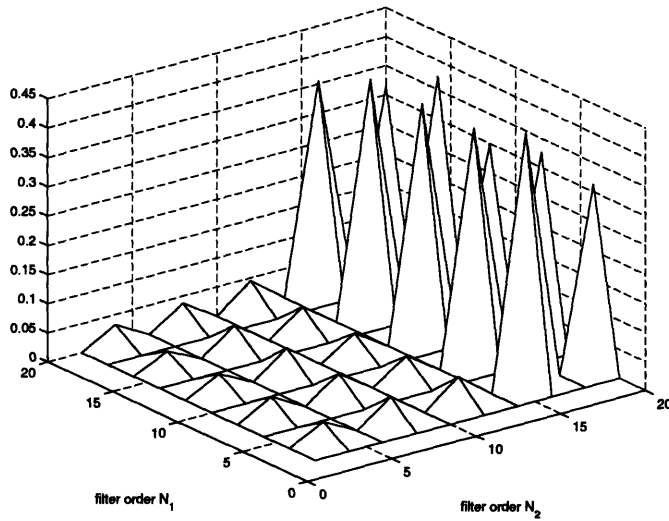


Figure 3-7: A plot of the maximum value of the state transition error vs. N_1 and N_2 for $n \geq 0$ using the truncated Butterworth filter structure with half-power frequency $\pi/2$ and zero initial conditions. In this example $B_X = 1$, or, equivalently, $|x[n]| \leq 1$ for all n .

STE in Fig. 3-10 appears to have a periodic ripple component. This is due to the fact that since $N_2 = 4$, the STE for $N \geq 0$ is a pure sum of four complex exponentials, with the weight of one of the complex exponentials much larger than the other three. Thus one mode of the pure ZIR response is dominant. We hypothesize that when N_2 is large, the STE looks more like an exponential decay since on average the STE is a sum of a number sinusoidal modes. The smaller the value of N_2 , the more purely sinusoidal the ripple on the STE appears.

3.2 Probabilistic Analysis

Our deterministic analysis of the STE resulted in the derivation of a deterministic bound on the magnitude of the STE which depended on the approximate filter structure being used, the pre-transition filter order N_1 , the post-transition filter order N_2 , and the input signal bound B_X . The deterministic bound serves the purpose of proving that the STE is bounded by an exponential decay. Although interesting and insightful, this deterministic bound is too conservative to demonstrate that on average the STE is negligible. Indeed, in our probabilistic analysis we will find that the expected STE is much less than its deterministic bound, and is for practical purposes *essentially negligible*. This is not at all obvious from

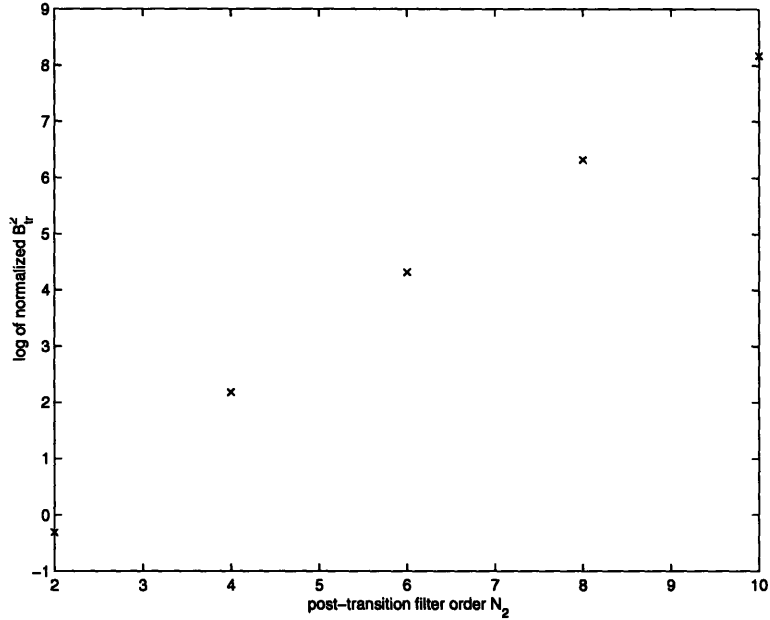


Figure 3-8: Logarithmic plot of the normalized squared STE bound B_{tr}^2 vs. post-transition filter order N_2 . The curves were generated using the replacement Butterworth filter structure with $N_1 = 12$. The half-power frequency is $\pi/2$.

the deterministic analysis, which motivates us to adapt a probabilistic framework for STE analysis. Formally, we consider the following statistical metrics

1. $E\{y_{tr}^2[n]\}$, the expected value of the squared STE,
2. $E\{P_{tr}[n]\}$, the expected value of the total additive corruption induced into the output power estimate $\hat{P}_y[N_0]$ due to the STE, and
3. $\text{Prob}(\hat{N}_k \geq \hat{N}_{LP}^* \mid y_{N_1 N_2}[n] = y_{N_2 N_2}[n] + y_{tr}[n])$, the probability that the STE-corrupted filter order estimate \hat{N}_k is greater than or equal to the ideal STE-free filter order estimate \hat{N}_{LP}^* . This is a statistical measure of the effect of the STE on the LP estimate of the optimal filter order. We use this probabilistic *excedance* measure since the nature of our problem is that it is much better to have $\hat{N}_k \geq \hat{N}_{LP}^*$, in which case we are assured that the output SNR is greater than or equal to the minimum tolerable level, than it is to have $\hat{N}_k < \hat{N}_{LP}^*$, in which case the output SNR can be less than the minimum tolerable level.

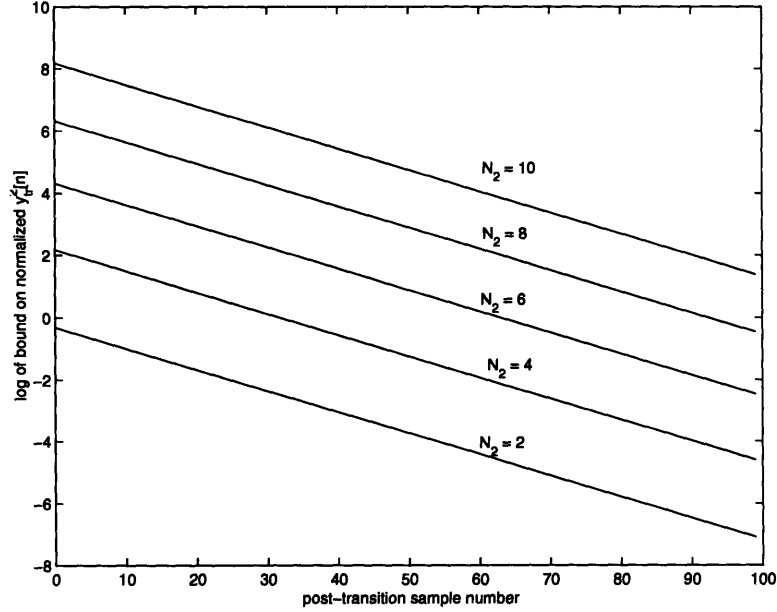


Figure 3-9: Logarithmic plot of the normalized squared STE vs. post transition sample number for $N_2 = 2, 4, 6, 8,$ and 10 . The curves were generated using the replacement Butterworth filter structure with $N_1 = 12$. The half-power frequency is $\pi/2$.

3.2.1 Preliminaries

In this section we adopt a vector-based formulation. We begin by assuming that $x[n]$ is a random input signal whose samples are independent and identically distributed. The infinite input vector is

$$\mathbf{x} = [x[L-1] \ x[L-2] \ x[L-3] \ \dots]^T. \quad (3.60)$$

The random vector \mathbf{x} has mean $\mathbf{m}_\mathbf{x} = E\{\mathbf{x}\}$ which we hereafter assume to be zero, and autocovariance matrix $\Lambda_\mathbf{x} = E\{\mathbf{x}\mathbf{x}^T\}$. We define

$$\mathbf{y}_0 = [y_{\text{tr}}[-1] \ y_{\text{tr}}[-2] \ \dots \ y_{\text{tr}}[-N_2]]^T, \quad (3.61)$$

which is an $N_2 \times 1$ vector of initial conditions for the STE. Since for $n < 0$ the STE is simply the difference between two fixed filter outputs, we may alternatively define \mathbf{y}_0 as

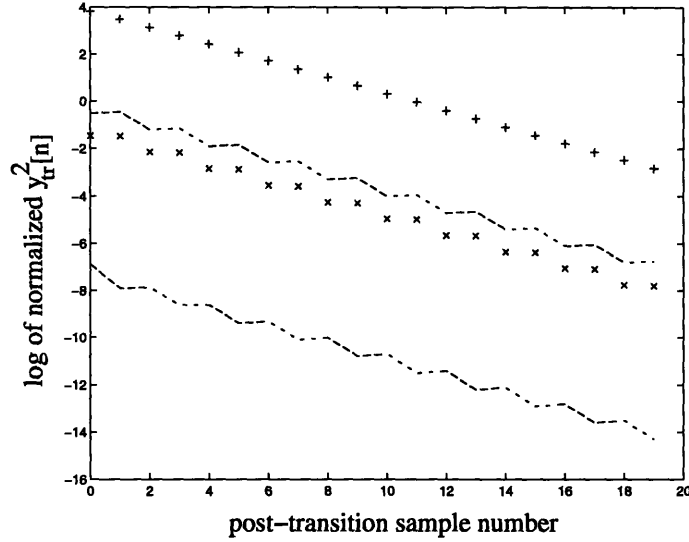


Figure 3-10: A plot of the normalized squared STE bound (plotted with '+'), along with the average (plotted with 'x'), maximum (plotted with '-'), and minimum (plotted with '-') values of the normalized squared STE computed by generating 1000 Monte Carlo simulations. The experiment used the replacement Butterworth filter structure, and parameter values $N_1 = 2$, $N_2 = 4$, and $B_X = 1$. The half-power frequency is $\pi/2$.

$$\mathbf{y}_0 = [(y_{N_1 N_1}[-1] - y_{N_2 N_2}[-1]) \cdots (y_{N_1 N_1}[-N_2] - y_{N_2 N_2}[-N_2])]^T. \quad (3.62)$$

The expected values of the fixed filtered outputs are

$$\mathbb{E} \{ y_{N_i N_i}[n] \} = \mathbb{E} \left\{ \sum_{k=-\infty}^{\infty} x[n-k] h_{N_i}[k] \right\} \quad i = 1, 2, \quad (3.63)$$

so that

$$\mathbb{E} \{ y_{N_i N_i}[n] \} = \mathbb{E} \{ x[n] \} \sum_{k=-\infty}^{\infty} h_{N_i}[k] \quad i = 1, 2. \quad (3.64)$$

We have assumed that $\mathbb{E} \{ \mathbf{x} \} = \mathbf{0}$, where the $\mathbf{0}$ is the vector of all zeros with appropriate dimension. This implies that the random vector \mathbf{y}_0 also has zero mean

$$\mathbf{E}\{\mathbf{y}_0\} = \mathbf{0}. \quad (3.65)$$

The autocovariance matrix of the vector \mathbf{y}_0 is $\Lambda_{\mathbf{y}_0} = \mathbf{E}\{\mathbf{y}_0\mathbf{y}_0^T\}$. Proceeding, we now define the fixed order- N_2 filter output vector

$$\mathbf{y}_{N_2N_2} = [y_{N_2N_2}[0] \ y_{N_2N_2}[1] \ \cdots \ y_{N_2N_2}[L-1]]^T \quad (3.66)$$

with mean $\mathbf{E}\{\mathbf{y}_{N_2N_2}\} = \mathbf{0}$ and autocovariance matrix $\Lambda_{\mathbf{y}_{N_2N_2}} = \mathbf{E}\{\mathbf{y}_{N_2N_2}\mathbf{y}_{N_2N_2}^T\}$, the approximate filter output vector

$$\mathbf{y}_{N_1N_2} = [y_{N_1N_2}[0] \ y_{N_1N_2}[1] \ \cdots \ y_{N_1N_2}[L-1]]^T \quad (3.67)$$

with mean $\mathbf{E}\{\mathbf{y}_{N_1N_2}\} = \mathbf{0}$ and autocovariance matrix $\Lambda_{\mathbf{y}_{N_1N_2}} = \mathbf{E}\{\mathbf{y}_{N_1N_2}\mathbf{y}_{N_1N_2}^T\}$, and the STE vector

$$\mathbf{y}_{\text{tr}} = [y_{\text{tr}}[0] \ y_{\text{tr}}[1] \ \cdots \ y_{\text{tr}}[L-1]]^T \quad (3.68)$$

with autocovariance matrix $\Lambda_{\mathbf{y}_{\text{tr}}} = \mathbf{E}\{\mathbf{y}_{\text{tr}}\mathbf{y}_{\text{tr}}^T\}$. With this formulation the STE vector has been defined for $n \geq 0$ such that

$$\mathbf{y}_{N_1N_2} = \mathbf{y}_{N_2N_2} + \mathbf{y}_{\text{tr}}. \quad (3.69)$$

We now take a moment to mention a few key points. The STE has been modeled as an additive noise term in the output of the approximate filter. Ideally we desire the fixed-order filter output vector $\mathbf{y}_{N_2N_2}$, but due to the filter order switching (state transition) our approximate filter produces $\mathbf{y}_{N_1N_2} = \mathbf{y}_{N_2N_2} + \mathbf{y}_{\text{tr}}$. Our purpose in this section is to statistically evaluate the consequences of the noise term \mathbf{y}_{tr} on the desired filter output $\mathbf{y}_{N_2N_2}$, the desired L -point output power measurement $\hat{P}_{\mathbf{y}_{N_2N_2}}$, and the low power filter order estimate \hat{N}_{LP}^* . It should be noted that we have defined the desired output to be

$\mathbf{y}_{N_2 N_2}$, a natural but not unique choice. Alternatively, we could have defined $\mathbf{y}_{N_1 N_1}$ or $\mathbf{y}_{N_k N_k}$ to be the desired output, for some application-specific value of k . In any case our statistical analysis would follow a similar development.

Continuing with our probabilistic analysis, if \mathbf{H}_d is the convolution matrix for the IIR impulse response $h_d[n] = h_{N_1}[n] - h_{N_2}[n]$ with N_2 rows and an infinite number of columns,

$$\mathbf{H}_d = \begin{bmatrix} 0 & \cdots & 0 & h_d[0] & h_d[1] & h_d[2] & h_d[3] & \cdots \\ 0 & \cdots & 0 & 0 & h_d[0] & h_d[1] & h_d[2] & \cdots \\ \vdots & \vdots & \vdots & 0 & 0 & h_d[0] & h_d[1] & \cdots \\ 0 & \cdots & 0 & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}, \quad (3.70)$$

in which the first L columns are all zeros. The matrix \mathbf{H}_d operates on the infinite input vector $\mathbf{x} = [x[L-1] \ x[L-2] \ x[L-3] \ \cdots]^T$ to produce

$$\mathbf{y}_0 = \mathbf{H}_d \mathbf{x}. \quad (3.71)$$

With this construction, the autocovariance matrix for the random vector \mathbf{y}_0 may be expressed as

$$\begin{aligned} \Lambda_{\mathbf{y}_0} &= \mathbf{E}\{\mathbf{H}_d \mathbf{x} (\mathbf{H}_d \mathbf{x})^T\} \\ &= \mathbf{E}\{\mathbf{H}_d \mathbf{x} \mathbf{x}^T \mathbf{H}_d^T\} \\ &= \mathbf{H}_d \mathbf{E}\{\mathbf{x} \mathbf{x}^T\} \mathbf{H}_d^T \\ &= \mathbf{H}_d \Lambda_{\mathbf{x}} \mathbf{H}_d^T, \end{aligned} \quad (3.72)$$

Now if \mathbf{H}_{N_2} is the convolution matrix for the IIR impulse response $h_{N_2}[n]$ with L rows and an infinite number of columns,

$$\mathbf{H}_{N_2} = \begin{bmatrix} 0 & \cdots & 0 & h_{N_2}[0] & h_{N_2}[1] & h_{N_2}[2] & \cdots \\ \vdots & & \cdots & & \cdots & & \\ 0 & h_{N_2}[0] & h_{N_2}[1] & h_{N_2}[2] & h_{N_2}[3] & \cdots & \\ h_{N_2}[0] & h_{N_2}[1] & h_{N_2}[2] & h_{N_2}[3] & \cdots & & \end{bmatrix}, \quad (3.73)$$

in which the first $L - 1$ columns of the first row are all zeros. The matrix \mathbf{H}_{N_2} operates on the infinite input vector $\mathbf{x} = [x[L - 1] \ x[L - 2] \ x[L - 2] \ \cdots]^T$ to produce

$$\mathbf{y}_{N_2 N_2} = \mathbf{H}_{N_2} \mathbf{x}. \quad (3.74)$$

With this construction, the autocovariance matrix for the random vector $\mathbf{y}_{N_2 N_2}$ may be expressed as

$$\begin{aligned} \mathbf{\Lambda}_{\mathbf{y}_{N_2 N_2}} &= E\{\mathbf{H}_{N_2} \mathbf{x} (\mathbf{H}_{N_2} \mathbf{x})^T\} \\ &= E\{\mathbf{H}_{N_2} \mathbf{x} \mathbf{x}^T \mathbf{H}_{N_2}^T\} \\ &= \mathbf{H}_{N_2} E\{\mathbf{x} \mathbf{x}^T\} \mathbf{H}_{N_2}^T \\ &= \mathbf{H}_{N_2} \mathbf{\Lambda}_{\mathbf{x}} \mathbf{H}_{N_2}^T, \end{aligned} \quad (3.75)$$

From our deterministic analysis in Section 3.1, we know that the STE satisfies the difference equation

$$y_{\text{tr}}[n] = \sum_{k=1}^{N_2} a_{k N_2} y_{\text{tr}}[n - k]. \quad (3.76)$$

By inspection of the summation in Eq. (3.76) we see that the first element $y_{\text{tr}}[0]$ of the vector \mathbf{y}_{tr} defined in Eq. (3.68) is a weighted sum of the elements $y_{\text{tr}}[-1] \ \cdots \ y_{\text{tr}}[-N_2]$ in the vector \mathbf{y}_0 defined in Eq. (3.61). The second element $y_{\text{tr}}[1]$ of the vector \mathbf{y}_{tr} is also a weighted sum of all but one of the elements in the vector \mathbf{y}_0 , plus the first element $y_{\text{tr}}[0]$ of the vector \mathbf{y}_{tr} multiplied by $a_{1 N_2}$. This is equivalent to a weighted sum of the elements in

the vector \mathbf{y}_0 . In fact all of the elements of the vector \mathbf{y}_{tr} depend on a linear combination of the elements of the vector \mathbf{y}_0 , so that the vector \mathbf{y}_{tr} may be computed via the matrix product

$$\begin{aligned}\mathbf{y}_{\text{tr}} &= \mathbf{Q}_{N_2} \mathbf{y}_0 \\ &= \mathbf{Q}_{N_2} \mathbf{H}_d \mathbf{x},\end{aligned}\tag{3.77}$$

where \mathbf{Q}_{N_2} is a $L \times N_2$ matrix whose elements are defined in terms of the polynomial coefficients a_{kN_2} whose roots are the pole locations of the filter $h_{N_2}[n]$. Although its elements are easily computed, the analytical form of the matrix \mathbf{Q}_{N_2} is exceedingly complicated. For this reason the formula for the elements of the matrix \mathbf{Q}_{N_2} in terms of the coefficients a_{kN_2} has been relegated to Appendix A.

3.2.2 Derivation of Probabilistic Bounds

Proceeding, we now consider the total output power estimate $\hat{P}_{\mathbf{y}_{N_2 N_2}}$ based on L samples of the output signal $y_{N_2 N_2}[n]$, defined to be

$$\hat{P}_{\mathbf{y}_{N_2 N_2}} = \sum_{k=0}^{L-1} y_{N_2 N_2}^2[k].\tag{3.78}$$

The expected value of $\hat{P}_{\mathbf{y}_{N_2 N_2}}$ may be written as

$$\mathbb{E} \left\{ \hat{P}_{\mathbf{y}_{N_2 N_2}} \right\} = \mathbb{E} \left\{ \mathbf{y}_{N_2 N_2}^T \mathbf{y}_{N_2 N_2} \right\}\tag{3.79}$$

or alternatively as

$$\mathbb{E} \left\{ \hat{P}_{\mathbf{y}_{N_2 N_2}} \right\} = \text{trace} \left(\mathbb{E} \left\{ \mathbf{y}_{N_2 N_2} \mathbf{y}_{N_2 N_2}^T \right\} \right).\tag{3.80}$$

Using the simple substitution $\mathbf{y}_{N_2 N_2} = \mathbf{H}_{N_2} \mathbf{x}$, we arrive at

$$\begin{aligned}
\mathbb{E} \left\{ \hat{P}_{y_{N_2 N_2}} \right\} &= \text{trace} \left(\mathbb{E} \left\{ \mathbf{H}_{N_2} \mathbf{x} (\mathbf{H}_{N_2} \mathbf{x})^T \right\} \right) \\
&= \text{trace} \left(\mathbf{H}_{N_2} \mathbf{\Lambda}_x \mathbf{H}_{N_2}^T \right)
\end{aligned} \tag{3.81}$$

As discussed previously, when using an approximate filter, the output signal $y_{N_2 N_2}[n]$ is not available. Consequently we must use the approximate filter output $y_{N_1 N_2}[n]$ to compute the total output power estimate $\hat{P}_{y_{N_1 N_2}}$ based on L samples of the approximate filter output signal $y_{N_1 N_2}[n]$, instead of output signal $y_{N_2 N_2}[n]$. This estimate is defined to be

$$\hat{P}_{y_{N_1 N_2}} = \sum_{k=0}^{L-1} y_{N_1 N_2}^2[k] \tag{3.82}$$

which can be expanded to produce

$$\begin{aligned}
\hat{P}_{y_{N_1 N_2}} &= \sum_{k=0}^{L-1} (y_{N_2 N_2}[k] + y_{\text{tr}}[k])^2 \\
&= \hat{P}_{y_{N_2 N_2}} + \underbrace{\sum_{k=0}^{L-1} 2y_{N_2 N_2}[k]y_{\text{tr}}[k] + y_{\text{tr}}^2[k]}_{P_{\text{tr}}},
\end{aligned} \tag{3.83}$$

where P_{tr} is defined as noted in Eq. (3.83). The term P_{tr} clearly represents the total additive corruption induced into $\hat{P}_{y_{N_2 N_2}}$ due to the STE. Using our vector notation we observe that the expected value of P_{tr} may be written as

$$\mathbb{E} \{ P_{\text{tr}} \} = \mathbb{E} \left\{ \sum_{k=0}^{L-1} 2y_{N_2 N_2}[k]y_{\text{tr}}[k] + y_{\text{tr}}^2[k] \right\} \tag{3.84}$$

or alternatively as

$$\mathbb{E} \{ P_{\text{tr}} \} = \text{trace} \left(\mathbb{E} \left\{ \mathbf{y}_{N_2 N_2} \mathbf{y}_{\text{tr}}^T \right\} \right) + \text{trace} \left(\mathbb{E} \left\{ \mathbf{y}_{\text{tr}} \mathbf{y}_{\text{tr}}^T \right\} \right). \tag{3.85}$$

Using the simple substitutions $\mathbf{y}_{N_2 N_2} = \mathbf{H}_{N_2} \mathbf{x}$ and $\mathbf{y}_{\text{tr}} = \mathbf{Q}_{N_2} \mathbf{H}_d \mathbf{x}$, we arrive at

$$\begin{aligned}
E\{P_{\text{tr}}\} &= \text{trace} \left(E \left\{ \mathbf{H}_{N_2} \mathbf{x} (\mathbf{Q}_{N_2} \mathbf{H}_d \mathbf{x})^T \right\} \right) + \text{trace} \left(E \left\{ \mathbf{Q}_{N_2} \mathbf{H}_d \mathbf{x} (\mathbf{Q}_{N_2} \mathbf{H}_d \mathbf{x})^T \right\} \right) \\
&= \text{trace} \left(2\mathbf{H}_{N_2} \mathbf{\Lambda}_x \mathbf{H}_d^T \mathbf{Q}_{N_2}^T \right) + \text{trace} \left(\mathbf{Q}_{N_2} \mathbf{H}_d \mathbf{\Lambda}_x \mathbf{H}_d^T \mathbf{Q}_{N_2}^T \right) \quad (3.86)
\end{aligned}$$

We have stated that one goal of our probabilistic analysis is to quantify the average error induced into our output power estimate $\hat{P}_{y_{N_2 N_2}}$ due to the STE. For this purpose we now define the output power noise-to-signal ratio (OPNSR) as

$$\begin{aligned}
\text{OPNSR} &= \frac{E\{P_{\text{tr}}\}}{E\{\hat{P}_{y_{N_2 N_2}}\}} \\
&= \frac{E\left\{ \sum_{k=0}^{L-1} 2y_{\text{tr}}[k]y_{N_2 N_2}[k] + y_{\text{tr}}^2[k] \right\}}{E\left\{ \sum_{k=0}^{L-1} y_{N_2 N_2}^2[k] \right\}}. \quad (3.87)
\end{aligned}$$

Equivalently, the expected output power noise-to-signal ratio (OPNSR) may be written as

$$\text{OPNSR} = \frac{\text{trace} \left(2\mathbf{H}_{N_2} \mathbf{\Lambda}_x \mathbf{H}_d^T \mathbf{Q}_{N_2}^T \right) + \text{trace} \left(\mathbf{Q}_{N_2} \mathbf{H}_d \mathbf{\Lambda}_x \mathbf{H}_d^T \mathbf{Q}_{N_2}^T \right)}{\text{trace} \left(\mathbf{H}_{N_2} \mathbf{\Lambda}_x \mathbf{H}_d^T \right)} \quad (3.88)$$

Denoting $\sigma_{y_{N_2 N_2}}^2$ as the variance of the output signal $y_{N_2 N_2}[n]$, the above equation may be simplified to

$$\text{OPNSR} = \frac{\text{trace} \left(2\mathbf{H}_{N_2} \mathbf{\Lambda}_x \mathbf{H}_d^T \mathbf{Q}_{N_2}^T \right) + \text{trace} \left(\mathbf{Q}_{N_2} \mathbf{H}_d \mathbf{\Lambda}_x \mathbf{H}_d^T \mathbf{Q}_{N_2}^T \right)}{L\sigma_{y_{N_2 N_2}}^2} \quad (3.89)$$

For the purpose of experimentally confirming the analytical result we have just obtained, consider the plot of the function OPNSR for various values of the power window length L given in Fig. 3-11. We recall that the power window length was introduced in Chapter 2

as the number of output samples generated before an output power estimate was formed. Note that as the power window length L increases, the ratio OPNSR decreases. In the computer simulations used to generate the results shown in Fig. 3-11 we used the truncated Butterworth filter structure with half-power frequency $\pi/2$ and $N_1 = 2$, $N_2 = 4$, and $B_X = 1$. Two thousand Monte Carlo simulations were used to generate the average OPNSR for each value of L , shown in the plot with 'x'. The theoretical OPNSR is also shown in the plot which was calculated using the ratio of matrix products in Eq. (3.89).

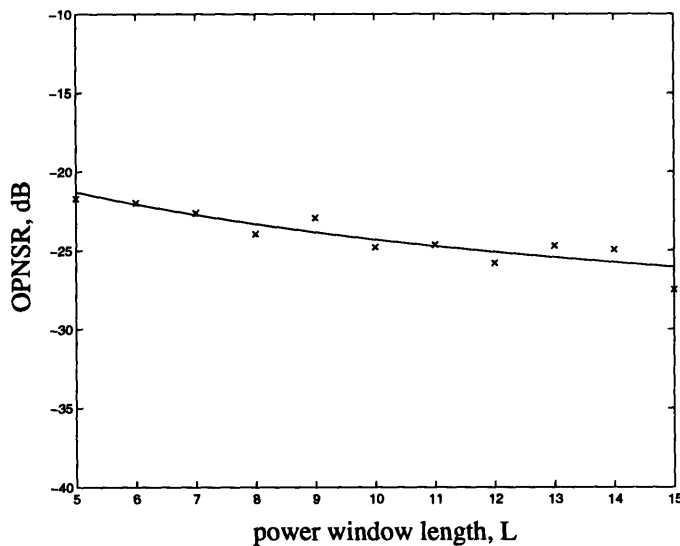


Figure 3-11: A plot of the output power noise-to-signal ratio (OPNSR) for various values of the power window length L . Actual values of the OPNSR (denoted in the plot by 'x') based on 2000 Monte Carlo simulations are plotted along with the predicted theoretical values (denoted in the plot by '-'). The predicted theoretical values of the OPNSR are given by Eq. (3.89).

Another goal of our probabilistic analysis is to quantify the mean-squared STE, $E\{y_{tr}^2[n]\}$. This quantity is conveniently represented by the diagonal elements of a matrix we have already formulated. We are not particularly interested in the absolute value of the time series $y_{tr}^2[n]$ but rather its relative magnitude compared to the desired output $y_{N_2 N_2}^2[n]$. For this purpose we define the output noise-to-signal ratio vector, **ONSR**, whose n th element is defined by

$$\text{ONSR}[n] = \frac{E\{y_{tr}^2[n]\}}{E\{y_{N_2 N_2}^2[n]\}}, \quad 0 \leq n \leq L - 1. \quad (3.90)$$

The vector **ONSR** has dimensions $L \times 1$, and may be more compactly written as

$$\mathbf{ONSR} = \frac{\text{diag} \left(\mathbf{Q}_{N_2} \mathbf{H}_d \mathbf{\Lambda}_x \mathbf{H}_d^T \mathbf{Q}_{N_2}^T \right)}{\text{diag} \left(\mathbf{H}_{N_2} \mathbf{\Lambda}_x \mathbf{H}_{N_2}^T \right)} \quad (3.91)$$

For the purpose of experimentally validating the expression for $\text{ONSR}[n]$ in Eq. (3.90), consider the plot of the $\text{ONSR}[n]$ given in Fig. 3-12. The plot shows the actual experimental averages based on 2000 Monte Carlo simulations and the theoretical prediction given by Eq. (3.90). Note that as the post-transition sample number increases, the ratio $\text{ONSR}[n]$ exponentially decays. In the computer simulations used to generate the results shown in Fig. 3-11 we used the replacement Butterworth filter structure with $N_1 = 2$, $N_2 = 4$, and $B_X = 1$.

The final goal of the probabilistic analysis is to examine the effect of the STE on our low power optimal filter order estimate \hat{N}_{LP}^* , based on the approximate filter output $y_{N_1 N_2}[n]$. For this purpose we assume the input random vector elements are statistically independent and identically distributed according to a uniform distribution on the interval $[-B_X, B_X]$, as before, and consider the conditional probability distribution

$$p_{\mathbf{y}_{N_2 N_2} | \mathbf{y}_{N_1 N_2} = \mathbf{y}_{N_2 N_2} + \mathbf{y}_{tr}}(\mathbf{Y}_{N_2 N_2} | \mathbf{Y}_{N_1 N_2} = \mathbf{Y}_{N_2 N_2} + \mathbf{Y}_{tr}). \quad (3.92)$$

This probability density summarizes the probability that the random vector $\mathbf{y}_{N_2 N_2} = \mathbf{Y}_{N_2 N_2}$ given that the random vector $\mathbf{y}_{N_1 N_2} = \mathbf{Y}_{N_1 N_2}$. Because the filter outputs $\mathbf{y}_{N_2 N_2}$ and $\mathbf{y}_{N_1 N_2}$ are filtered versions of the uniformly-distributed random input vector, we may assume by the central limit theorem that the random vectors $\mathbf{y}_{N_2 N_2}$ and $\mathbf{y}_{N_1 N_2}$ are approximately jointly Gaussian. Under this assumption the conditional density in Eq. (3.92) is multivariate Gaussian in the random vector $\mathbf{y}_{N_2 N_2}$, and thus completely characterized by its autocovariance matrix

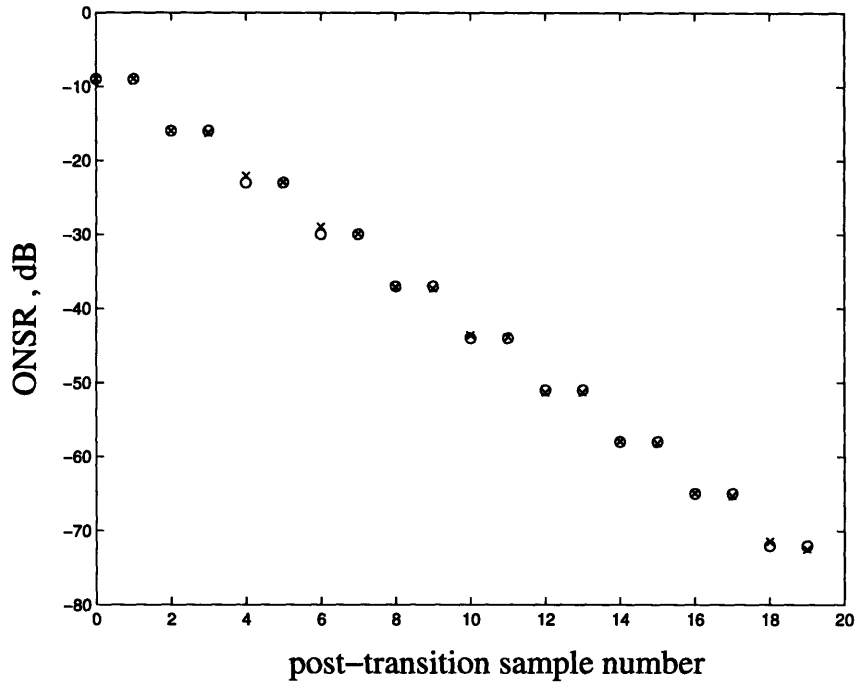


Figure 3-12: A plot of the function $\text{ONSR}[n]$ given by Eq. (3.90). Actual experimental values of the time series $\text{ONSR}[n]$ (denoted in the figure by ‘x’) based on 2000 Monte Carlo simulations are plotted along with the predicted theoretical values (denoted in the figure by ‘o’).

$$\Lambda_c = \Lambda_{\mathbf{y}_{N_2 N_2}} - \Lambda_{\mathbf{y}_{N_2 N_2} \mathbf{y}_{N_1 N_2}} \Lambda_{\mathbf{y}_{N_1 N_2}}^{-1} \Lambda_{\mathbf{y}_{N_2 N_2} \mathbf{y}_{N_1 N_2}}^T \quad (3.93)$$

and its mean

$$\mu_c = \Lambda_{\mathbf{y}_{N_2 N_2} \mathbf{y}_{N_1 N_2}} \Lambda_{\mathbf{y}_{N_1 N_2}}^{-1} \Lambda_{\mathbf{y}_{N_1 N_2}}, \quad (3.94)$$

where

$$\Lambda_{\mathbf{y}_{N_2 N_2} \mathbf{y}_{N_1 N_2}} = \Lambda_{\mathbf{y}_{N_2 N_2}} + \mathbf{H}_{N_2} \Lambda_x \mathbf{H}_d^T \mathbf{Q}^T, \quad (3.95)$$

and

$$\Lambda_{\mathbf{y}_{N_1 N_2}} = \Lambda_{\mathbf{y}_{N_2 N_2}} + \mathbf{H}_{N_2} \Lambda_{\mathbf{x}} \mathbf{H}_d^T \mathbf{Q}^T + \mathbf{Q} \mathbf{H}_d \Lambda_{\mathbf{x}}^T \mathbf{H}_{N_2}^T + \mathbf{Q} \mathbf{H}_d \Lambda_{\mathbf{x}} \mathbf{H}_d^T \mathbf{Q}^T. \quad (3.96)$$

These matrices are easily found by observing the fact that the vectors $\mathbf{y}_{N_2 N_2}$ and $\mathbf{y}_{N_1 N_2}$ are jointly Gaussian. Note that Λ_c does not depend on the actual value of the vector $\mathbf{y}_{N_1 N_2}$ while μ_c depends linearly on the observed approximate filter output vector $\mathbf{y}_{N_1 N_2}$.

If we integrate the conditional density in Eq. (3.92) over an annulus defined by

$$P_{k-1} \leq \mathbf{y}_{N_2 N_2}^T \mathbf{y}_{N_2 N_2} \leq P_k, \quad (3.97)$$

where the region $[P_{k-1}, P_k]$ defines the partition the output power space corresponding to $\hat{N}_{\text{LP}}^* = N_k$, then the probability density

$$\text{Prob}(N_k = \hat{N}_{\text{LP}}^* \mid \mathbf{y}_{N_1 N_2} = \mathbf{y}_{N_2 N_2} + \mathbf{y}_{\text{tr}}) \quad (3.98)$$

is equal to the integral

$$\int_C p_{\mathbf{y}_{N_2 N_2} \mid \mathbf{y}_{N_2 N_2} + \mathbf{y}_{\text{tr}}}(\mathbf{Y}_{N_2 N_2} \mid \mathbf{Y}_{N_2 N_2} + \mathbf{Y}_{\text{tr}}) d\mathbf{y}_{N_2 N_2}, \quad (3.99)$$

where $C = P_{k-1} \leq \mathbf{y}_{N_2 N_2}^T \mathbf{y}_{N_2 N_2} \leq P_k$ is the k th annulus region. An expression for this integral may be obtained as an infinite series and thus computed with arbitrary precision [45, 56]. Graphically this integral corresponds to the integral of an L -dimensional multivariate Gaussian density with nonzero mean over an annulus centered at the origin. If N_k is the our low power estimate of the optimal filter order based the STE-corrupted output $\mathbf{y}_{N_1 N_2}$, then the probability that N_k is equal to the true value of \hat{N}_{LP}^* that we would have obtained using the desired fixed filter output vector $\mathbf{y}_{N_2 N_2}$ is given by the integral in Eq. (3.99). From this we may easily deduce the cumulative distribution, or *excedance function* $\text{Prob}(N_k \geq \hat{N}_{\text{LP}}^* \mid \mathbf{y}_{N_2 N_2} + \mathbf{y}_{\text{tr}})$.

In summary, we first recall that the STE has been modeled as an additive noise term in the desired fixed filter output $\mathbf{y}_{N_2 N_2}$, such that the approximate filter output $\mathbf{y}_{N_1 N_2} =$

$\mathbf{y}_{N_2N_2} + \mathbf{y}_{\text{tr}}$. The effect of the STE from this viewpoint is summarized in the output noise-to-signal ratio $\text{ONSR}[n]$.

Secondly, the STE also corrupts our output power estimate $\hat{P}_{\mathbf{y}_{N_2N_2}}$ with an additive noise term P_{tr} such that $\hat{P}_{\mathbf{y}_{N_2N_2}} = \hat{P}_{\mathbf{y}_{N_1N_2}} + P_{\text{tr}}$. The effect of the STE from this viewpoint is summarized in the output power noise-to-signal ratio OPNSR.

Finally, the additive corruption P_{tr} in our output power measurement induces an error into our low power optimal filter order estimate \hat{N}_{LP}^* . This effect is encapsulated in the probability density $\text{Prob}(\hat{N}_k = \hat{N}_{\text{LP}}^* \mid \mathbf{y}_{N_1N_2} = \mathbf{y}_{N_2N_2} + \mathbf{y}_{\text{tr}})$. This probability density function summarizes the effect of the STE on our low power filter order estimate \hat{N}_{LP}^* .

3.3 Considerations for Truncation Filter Structures

So far in this chapter we have developed deterministic and probabilistic frameworks for STE analysis assuming that a replacement approximate filter structure is used. The final section in this chapter is dedicated to exploring the adjustments to these frameworks which are required to analyze the effect of the STE in truncation approximate filter structures.

For the purposes of STE analysis, we restrict the truncation filter structure to increase the number of second-order sections used by *at most one* at each stage. The truncation filter structure may decrease the number of second-order sections used at each stage *arbitrarily* as long as at least one second-order section is used at all times. When a state transition from using filter order N_1 (or, equivalently, $S_1 = 2N_1$ second-order sections) to filter order N_2 (or, equivalently, $S_2 = 2N_2$ second-order sections) occurs such that $N_2 < N_1$, then *the STE is identical to that derived previously in Section 3.1 for replacement approximate filter structures*. When a state transition from filter order N_1 to N_2 occurs such that $N_2 > N_1$, then the new analysis presented in this section applies. As in Section 3.1, we develop deterministic and probabilistic frameworks to analyze the STE in truncation approximate filter structures.

3.3.1 Deterministic Analysis

We first address the problem of deriving a deterministic upper bound B_{tr} on the absolute value of the STE $y_{\text{tr}}[n]$ such that $|y_{\text{tr}}[n]| < B_{\text{tr}}$ for $n \geq 0$. The results in this section are obtained assuming that a truncation approximate filter structure with a cascade of second-order sections is used. The truncation approximate filter structure is shown in Fig. 3-13.

The output of the S_k th second-order section is labeled $y_{S_k S_k}[n]$, with $y_{00}[n] = x[n]$. The intermediate signals $e_{S_k S_k}[n]$ are labeled in Fig. 3-13 and will be used in the STE analysis. Before proceeding with the details of the derivation of the deterministic bound on the STE, we present a brief summary of the final result.

**Deterministic Bound on the STE: B_{tr}
Using a Truncation Filter Structure with $N_2 > N_1$**

Assume a truncation approximate filter structure is used with $N_2 > N_1$. The deterministic bound on the magnitude of the STE is B_{tr} , and the STE satisfies

$$|y_{\text{tr}}[n]| = |y_{N_2 N_2}[n] - y_{N_1 N_2}[n]| \quad (3.100)$$

$$\leq B_{\text{tr}} \cdot |z_{\text{max}}|^n, \quad (3.101)$$

where

$$B_{\text{tr}} = 2 \cdot B_{\mathbf{e}_{\text{tr}}} \cdot \lambda_{\min}^{-1/2}(\mathbf{Z}^T \mathbf{Z}) \cdot \left| \sum_{k=1}^2 b_{k S_2} \right|, \quad (3.102)$$

S_2 is the current number of second-order sections of the approximate filter, $B_{\mathbf{e}_{\text{tr}}}$ is a bound on the norm of the $L \times 1$ vector of the first L post-transition samples of the the signal $e_{\text{tr}}[n]$, $\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})$ is the minimum eigenvalue of the matrix $\mathbf{Z}^T \mathbf{Z}$, and \mathbf{Z} is the Vandermonde matrix corresponding to the S_2 th second-order section pole pair in the truncation filter structure. Note that the parameters N_2 , $B_{\mathbf{e}_{\text{tr}}}$, and $\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})$ in Eq. (3.102) depend only on our choice of the truncation filter structure, S_1 , S_2 , and the input signal bound B_X , which is defined such that $|x[n]| \leq B_X$ for all n .

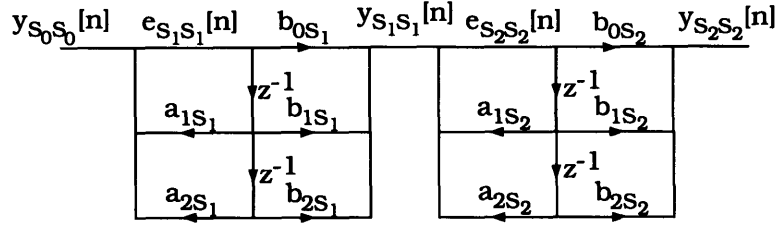


Figure 3-13: Truncation approximate filter structure for an IIR digital filter. The annotated variables are used in the STE analysis.

We now proceed with a detailed derivation of the result which is summarized above. The deterministic bound B_{tr} on the STE is defined such that

$$|y_{tr}[n]| = |y_{S_1 S_2}[n] - y_{S_2 S_2}[n]| \leq B_{tr}, \quad n \geq 0, \quad (3.103)$$

where $y_{S_1 S_2}[n]$ is the output of the approximate filter which switches from using S_1 second-order sections to S_2 second-order sections at time $n = 0$ and $y_{S_2 S_2}[n]$ is the fixed filter output produced by using S_2 second-order sections for all time. As noted before, we assume $S_2 > S_1$. For cases in which $S_2 < S_1$, the STE analysis presented earlier in Section 3.1 for replacement approximate filter structures applies. We further assume that $S_2 = S_1 + 1$, so that at most one second-order section is added to the truncation filter structure at each stage of the approximate filtering algorithm. The output $y_{S_2 S_2}[n]$ of the fixed filter with S_2 second-order sections may be expressed as

$$y_{S_2 S_2}[n] = \sum_{k=1}^2 b_{kS_2} e_{S_2 S_2}[n - k], \quad (3.104)$$

where b_{kS_2} is the k th coefficient in the second-order polynomial (ordered in ascending powers of z^{-1}) whose roots define the zeros of the S_2 th second-order section in the cascade filter structure. The signal $e_{S_2 S_2}[n]$, satisfies

$$e_{S_2 S_2}[n] = \underbrace{y_{S_1 S_1}[n]}_{e_{S_2 S_2}^{\text{ZSR}}[n]} + \underbrace{\sum_{k=1}^2 a_{k S_2} e_{S_2 S_2}[n-k]}_{e_{S_2 S_2}^{\text{ZIR}}[n]}. \quad (3.105)$$

In Eq. 3.105 we have labeled the ZIR and ZSR components of the signal $e_{S_2 S_2}[n]$. The two initial conditions for $e_{S_2 S_2}[n]$ are $e_{S_2 S_2}[-1] = z[-1]$ and $e_{S_2 S_2}[-2] = z[-2]$, where $z[n]$ is the output of an all-pole filter with a causal, stable impulse response with the system function

$$H_z(z) = \frac{1}{1 - \sum_{k=1}^2 a_{k S_2} z^{-k}}, \quad (3.106)$$

and input $y_{S_1 S_1}[n]$. Although not immediately apparent, these initial conditions naturally arise in the cascade filter structure in Fig. 3-13 if the number of second-order sections is fixed to be S_2 for all time.

The output $y_{S_1 S_2}[n]$ of the approximate filter using the truncation approximate filter structure that switches from using S_1 second-order sections to using $S_2 = S_1 + 1$ second-order sections at time $n = 0$ is defined as

$$y_{S_1 S_2}[n] = \sum_{k=1}^2 b_{k S_2} e_{S_1 S_2}[n-k], \quad (3.107)$$

where $b_{k S_2}$ is the k th coefficient in the second-order polynomial (ordered in ascending powers of z^{-1}) whose roots define the zeros of the S_2 th second-order section in the cascade filter structure. The signal $e_{S_1 S_2}[n]$, satisfies

$$e_{S_1 S_2}[n] = \underbrace{y_{S_1 S_1}[n]}_{e_{S_1 S_2}^{\text{ZSR}}[n]} + \underbrace{\sum_{k=1}^2 a_{k S_2} e_{S_1 S_2}[n-k]}_{e_{S_1 S_2}^{\text{ZIR}}[n]}. \quad (3.108)$$

In Eq. 3.108 we have labeled the ZIR and ZSR components of the signal $e_{S_1 S_2}[n]$. The two initial conditions for $e_{S_1 S_2}[n]$ are $e_{S_1 S_2}[-1] = y_{S_1 S_1}[-1]$ and $e_{S_1 S_2}[-2] = y_{S_1 S_1}[-2]$, where

$y_{S_1 S_1}[n]$ is the output of the cascade filter structure using a fixed number S_1 second-order sections for all time. These initial conditions naturally arise in the S_2 -section cascade filter structure when the number of second-order sections is increased from S_1 to S_2 at time $n = 0$. We now consider the *intermediate state transition error* $e_{\text{tr}}[n]$ in the truncation filter structure, defined as

$$\begin{aligned}
e_{\text{tr}}[n] &= e_{S_1 S_2}[n] - e_{S_2 S_2}[n] \\
&= e_{S_1 S_2}^{\text{ZSR}}[n] + e_{S_1 S_2}^{\text{ZIR}}[n] - e_{S_2 S_2}^{\text{ZSR}}[n] - e_{S_2 S_2}^{\text{ZIR}}[n] \\
&= e_{S_1 S_2}^{\text{ZIR}}[n] - e_{S_2 S_2}^{\text{ZIR}}[n],
\end{aligned} \tag{3.109}$$

which can be expanded with substitutions and rearrangement to yield

$$e_{\text{tr}}[n] = \sum_{k=1}^2 a_{k S_2} e_{\text{tr}}[n - k]. \tag{3.110}$$

with initial conditions $e_{\text{tr}}[-1] = y_{S_1 S_1}[-1] - z[-1]$ and $e_{\text{tr}}[-2] = y_{S_1 S_1}[-2] - z[-2]$. We note that each of the two initial conditions $e_{\text{tr}}[-2]$ and $e_{\text{tr}}[-1]$ is of the form of $\delta_{N_1 N_2}[n]$ from Eq. (3.8). Thus, from our earlier results, each initial condition in this set has a magnitude satisfying

$$|e_{\text{tr}}[n]| \leq B_{S_1} \cdot B_z \quad -2 \leq n \leq -1, \tag{3.111}$$

where

$$B_{S_1} = B_X \sum_{k=0}^{\infty} h_{S_1}[k] \tag{3.112}$$

is the bound on the fixed filter output signal $y_{S_1 S_1}[n]$, B_X is the input signal bound, $h_{S_1}[k]$ is the impulse response of the fixed filter obtained by using the first S_1 sections of the truncation filter structure, and B_z is the bound on the absolute difference between $y_{S_1 S_1}[n]$ and $z[n]$, given by

$$B_z = \sum_{k=0}^{\infty} (\delta[k] - h_{\text{ap}}[k]), \quad (3.113)$$

where $h_{\text{ap}}[n]$ is the causal, stable impulse response corresponding to the all-pole system function

$$H_{\text{ap}}(z) = \frac{1}{1 - \sum_{k=1}^2 a_k S_2 z^{-k}} \quad (3.114)$$

This result will be invoked later to establish the bound B_{tr} .

The expression for $e_{\text{tr}}[n]$ in Eq. (3.110) identifies the signal $e_{\text{tr}}[n]$ to be a pure ZIR for $n \geq 0$. We note that $e_{\text{tr}}[n]$ does not have the form of a pure ZIR on the interval $-2 \leq n \leq -1$. Nevertheless, for $n \geq 0$ the signal $e_{\text{tr}}[n]$ may be expressed as a weighted sum of exponential signals [54],

$$e_{\text{tr}}[n] = \sum_{k=1}^{N_2} c_k z_k^n, \quad (3.115)$$

where the z_k are the two pole locations of the second-order filter which includes the pole/zero pair of the S_2 th second-order section in the truncation filter structure. Note that if this section of the truncation filter structure is stable, then $|z_k| < 1$ for $1 \leq k \leq N_2$. We may write Eq. (3.115) in matrix form as

$$\underbrace{\begin{bmatrix} e_{\text{tr}}[0] \\ e_{\text{tr}}[1] \\ \vdots \\ e_{\text{tr}}[L-1] \end{bmatrix}}_{\mathbf{e}_{\text{tr}}} = \underbrace{\begin{bmatrix} z_1^0 & z_2^0 \\ z_1^1 & z_2^1 \\ \vdots & \vdots \\ z_1^{L-1} & z_2^{L-1} \end{bmatrix}}_{\mathbf{Z}} \underbrace{\begin{bmatrix} c_1 \\ c_2 \end{bmatrix}}_{\mathbf{c}}, \quad (3.116)$$

which can be more compactly expressed as

$$\mathbf{e}_{\text{tr}} = \mathbf{Z}\mathbf{c}, \quad (3.117)$$

for the $L \times 1$ vector \mathbf{y}_{tr} , the $L \times 2$ Vandermonde matrix \mathbf{Z} , and the 2×1 vector \mathbf{c} . Each of the 2 columns of \mathbf{Z} is made up of the one of the two poles of the S_2 th second-order section raised to successively higher powers from zero to $L - 1$ in rows 1 to L , as shown in Eq. (3.116).

Following the development of our presentation in Section 3.1 of the STE analysis using a replacement approximate filter structure, we now proceed to bound $\|\mathbf{c}\|$ in Eq. (3.117) using Eq. (3.43). Since $\mathbf{e}_{\text{tr}} = \mathbf{Z}\mathbf{c}$, we know from Eq. (3.43) that

$$\|\mathbf{e}_{\text{tr}}\|^2 \geq \lambda_{\min}(\mathbf{Z}^T\mathbf{Z}) \cdot \|\mathbf{c}\|^2. \quad (3.118)$$

Rearranging, the inequality in Eq. (3.118) becomes

$$\begin{aligned} \|\mathbf{c}\| &\leq \frac{\|\mathbf{e}_{\text{tr}}\|}{\sqrt{\lambda_{\min}(\mathbf{Z}^T\mathbf{Z})}} \\ &\leq \frac{B_{\mathbf{e}_{\text{tr}}}}{\sqrt{\lambda_{\min}(\mathbf{Z}^T\mathbf{Z})}}, \end{aligned} \quad (3.119)$$

where $B_{\mathbf{e}_{\text{tr}}}$ is defined such that $\|\mathbf{e}_{\text{tr}}\| \leq B_{\mathbf{e}_{\text{tr}}}$. To determine $B_{\mathbf{e}_{\text{tr}}}$, we use the results from our previous analysis Section 3.1 to produce

$$e_{\text{tr}}[l] \leq E_{\text{tr}}[l] \quad l > 0, \quad (3.120)$$

where

$$E_{\text{tr}}[l] = B_{S_1} B_z \sum_{k=1}^2 |a_{kS_2}| + \sum_{k=0}^{l-1} |a_{l-k,S_2}| e_{\text{tr}}[k], \quad (3.121)$$

with initial condition

$$E_{\text{tr}}[0] = B_{S_1} B_z \sum_{k=1}^2 |a_{kS_2}|. \quad (3.122)$$

Using Eq. (3.120) and Eq. (3.122) we may generate each of the bounds $E_{\text{tr}}[l]$ for $0 \leq l \leq L-1$, and then determine $B_{\text{e}_{\text{tr}}}$ using

$$\begin{aligned} \|\mathbf{e}_{\text{tr}}\|^2 &= \sum_{k=0}^{L-1} |e_{\text{tr}}[k]|^2 \\ &\leq \sum_{k=0}^{L-1} E_{\text{tr}}^2[k] = B_{\text{e}_{\text{tr}}}^2. \end{aligned} \quad (3.123)$$

We shall use this result in the forthcoming derivation of B_{tr} . We are now prepared to derive B_{tr} . We start by recalling that e_{tr} is a pure ZIR and thus may be expressed as

$$e_{\text{tr}}[n] = \sum_{k=1}^2 c_k z_k^n, \quad n \geq 0, \quad (3.124)$$

so that

$$\begin{aligned} |e_{\text{tr}}[n]| &= \left| \sum_{k=1}^2 c_k z_k^n \right| \\ &\leq \sum_{k=1}^2 |c_k| \cdot |z_k^n| \\ &\leq |z_{\max}|^n \sum_{k=1}^2 |c_k| \\ &\leq 2 \cdot \|\mathbf{c}\| \cdot |z_{\max}|^n, \end{aligned} \quad (3.125)$$

where z_{\max} is defined to be the pole location with the largest magnitude of the second-order filter in the section S_2 of the truncation filter structure.

Substituting in the previously derived bound for $\|\mathbf{c}\|$ from Eq. (3.119) into Eq. (3.125), produces for $n \geq 0$

$$\begin{aligned}
|e_{\text{tr}}[n]| &\leq 2 \cdot \frac{\|e_{\text{tr}}\|}{\sqrt{\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})}} \cdot |z_{\max}|^n \\
&\leq 2 \cdot \frac{B_{e_{\text{tr}}}}{\sqrt{\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})}} \cdot |z_{\max}|^n \\
&\leq B_e \cdot |z_{\max}|^n,
\end{aligned} \tag{3.126}$$

where

$$B_e = 2 \cdot B_{e_{\text{tr}}} \cdot \lambda_{\min}^{-1/2}(\mathbf{Z}^T \mathbf{Z}). \tag{3.127}$$

To obtain the bound B_e above we have used the result given in Eq. (3.123) that $\|e_{\text{tr}}\| \leq B_{e_{\text{tr}}}$. In Eq. (3.126) we have given a bound for the signal $e_{\text{tr}}[n]$ in terms of the bound $B_{e_{\text{tr}}}$, the pole location z_{\max} , and the eigenvalue $\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})$ of the matrix $\mathbf{Z}^T \mathbf{Z}$ defined in Eq. (3.116).

By inspection of Fig. 3-13, it is clear that

$$y_{s_2 s_2}[n] = \sum_{k=0}^2 b_{k s_2} e_{s_2 s_2}[n - k], \tag{3.128}$$

and that

$$y_{s_1 s_2}[n] = \sum_{k=0}^2 b_{k s_2} e_{s_1 s_2}[n - k], \tag{3.129}$$

so that the STE for truncation approximate filter structures is related to the signal e_{tr} as

$$y_{\text{tr}}[n] = \sum_{k=0}^2 b_{k s_2} e_{\text{tr}}[n - k]. \tag{3.130}$$

Substituting in the bound we have derived for $e_{\text{tr}}[n]$ produces for $n \geq 0$,

$$\begin{aligned}
|y_{\text{tr}}[n]| &= \left| \sum_{k=0}^2 b_{kS_2} e_{\text{tr}}[n-k] \right| \\
&\leq B_e \left| \sum_{k=0}^2 b_{kS_2} \right| \\
&\leq B_{\text{tr}},
\end{aligned} \tag{3.131}$$

where $B_{\text{tr}} = B_e \left| \sum_{k=0}^2 b_{kS_2} \right|$. Thus, our explicit derivation of the bound B_{tr} is now complete. A summary of the steps which are needed to obtain a numerical value for B_{tr} , given B_X , S_1 , S_2 , and a truncation approximate filter structure is now given.

**Summary of How to Compute B_{tr}
Using a Truncation Filter Structure with $N_2 > N_1$**

1. Given B_X , S_1 , S_2 , and a truncation approximate filter structure, compute the bounds B_{S_1} using Eq. (3.112) and B_z using Eq. (3.113).
2. Compute $\lambda_{\min}(\mathbf{Z}^T \mathbf{Z})$ as the minimum singular value of the matrix \mathbf{Z} defined in Eq. (3.116).
3. Compute B_{etr} using Eqs. (3.120)–(3.123)
4. Compute B_e using Eq. (3.127)
5. Compute B_{tr} using Eq. (3.131)

3.3.2 Probabilistic Analysis

In this section we focus on exploring the STE in truncation filter structures with a probabilistic framework similar to that developed for probabilistic analysis of the STE for replacement filter structures in Section 3.2. We consider the following statistical metrics:

1. $E\{y_{\text{tr}}^2[n]\}$, the expected value of the squared STE,

2. $E\{P_{\text{tr}}[n]\}$, the expected value of the total additive corruption induced into $\hat{P}_y[N_0]$ due to the STE, and

We again adopt a vector-based formulation for notational simplicity. We begin by assuming that $x[n]$ is a random input signal whose samples are independent and identically distributed. The infinite input vector is

$$\mathbf{x} = [x[L-1] \ x[L-2] \ x[L-3] \ \cdots]^T. \quad (3.132)$$

The random vector \mathbf{x} has mean $\mathbf{m}_x = E\{\mathbf{x}\}$ which we hereafter assume to be zero, and autocovariance matrix $\Lambda_x = E\{\mathbf{x}\mathbf{x}^T\}$. We define

$$\mathbf{e}_0 = [e_{\text{tr}}[-1] \ e_{\text{tr}}[-2]]^T, \quad (3.133)$$

which is a 2×1 vector of initial conditions for the intermediate STE. Recall that the intermediate STE was defined in Eq. (3.109). The vector \mathbf{e}_0 has autocovariance matrix $\Lambda_{\mathbf{e}_0} = E\{\mathbf{e}_0\mathbf{e}_0^T\}$. Since for $n < 0$ the STE is simply the difference between two fixed filter outputs, we may alternatively define \mathbf{e}_0 as

$$\mathbf{e}_0 = [(y_{S_1 S_1}[-1] - z[-1]) \ (y_{S_1 S_1}[-2] - z[-2])]^T. \quad (3.134)$$

We define the following intermediate output vector

$$\mathbf{e}_{S_1 S_2} = [e_{S_1 S_2}[0] \ e_{S_1 S_2}[1] \ \cdots \ e_{S_1 S_2}[L-1]]^T \quad (3.135)$$

with autocovariance matrix $\Lambda_{\mathbf{e}_{S_1 S_2}} = E\{\mathbf{e}_{S_1 S_2}\mathbf{e}_{S_1 S_2}^T\}$, and the intermediate state transition error vector

$$\mathbf{e}_{\text{tr}} = [e_{\text{tr}}[0] \ e_{\text{tr}}[1] \ \cdots \ e_{\text{tr}}[L-1]]^T \quad (3.136)$$

with autocovariance matrix $\Lambda_{\mathbf{e}_{\text{tr}}} = \text{E}\{\mathbf{e}_{\text{tr}}\mathbf{e}_{\text{tr}}^T\}$.

We also define the following output vectors

$$\mathbf{y}_{S_2S_2} = [y_{S_2S_2}[0] \ y_{S_2S_2}[1] \ \cdots \ y_{S_2S_2}[L-1]]^T \quad (3.137)$$

with autocovariance matrix $\Lambda_{\mathbf{y}_{S_2S_2}} = \text{E}\{\mathbf{y}_{S_2S_2}\mathbf{y}_{S_2S_2}^T\}$,

$$\mathbf{y}_{S_1S_2} = [y_{S_1S_2}[0] \ y_{S_1S_2}[1] \ \cdots \ y_{S_1S_2}[L-1]]^T \quad (3.138)$$

with autocovariance matrix $\Lambda_{\mathbf{y}_{S_1S_2}} = \text{E}\{\mathbf{y}_{S_1S_2}\mathbf{y}_{S_1S_2}^T\}$, and the STE vector

$$\mathbf{y}_{\text{tr}} = [y_{\text{tr}}[0] \ y_{\text{tr}}[1] \ \cdots \ y_{\text{tr}}[L-1]]^T \quad (3.139)$$

with autocovariance matrix $\Lambda_{\mathbf{y}_{\text{tr}}} = \text{E}\{\mathbf{y}_{\text{tr}}\mathbf{y}_{\text{tr}}^T\}$. The approximate filter output vector satisfies

$$\mathbf{y}_{S_1S_2} = \mathbf{y}_{S_2S_2} + \mathbf{y}_{\text{tr}}. \quad (3.140)$$

We now take a moment to mention a few key points. The STE vector \mathbf{y}_{tr} has been modeled as an additive noise term in the output of the approximate filter output vector $\mathbf{y}_{S_1S_2}$. Ideally we would like to produce vector $\mathbf{y}_{S_2S_2}$, but due to the filter order switching (state transition) our approximate filter produces $\mathbf{y}_{S_1S_2} = \mathbf{y}_{S_2S_2} + \mathbf{y}_{\text{tr}}$. Our purpose in this section is to statistically evaluate the consequences of the noise term \mathbf{y}_{tr} on the desired output vector $\mathbf{y}_{S_2S_2}$. We also want to examine the effect of \mathbf{y}_{tr} on the desired L -point output power measurement $\hat{P}_{\mathbf{y}_{S_2S_2}}$, as well as the low power filter order estimate \hat{N}_{LP}^* .

We define the matrix \mathbf{H}_e , with 2 rows and an infinite number of columns, as the convolution matrix for the IIR impulse response with system function

$$H_e(z) = H_{2S_1}(z) \cdot \left[1 - \frac{1}{1 - \sum_{k=1}^2 a_{kS_2} z^{-k}} \right], \quad (3.141)$$

where $H_{2S_1}(z)$ is the system function of the order- $2S_1$ fixed filter of the first S_1 sections of the truncation filter structure. The matrix \mathbf{H}_e has the form

$$\mathbf{H}_e = \begin{bmatrix} 0 & \cdots & 0 & h_e[0] & h_e[1] & h_e[2] & h_e[3] & \cdots \\ 0 & \cdots & 0 & 0 & h_e[0] & h_e[1] & h_e[2] & \cdots \\ \vdots & \vdots & \vdots & 0 & 0 & h_e[0] & h_e[1] & \cdots \\ 0 & \cdots & 0 & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}, \quad (3.142)$$

in which the first L columns are all zeros. The matrix \mathbf{H}_e operates on the infinite input vector $\mathbf{x} = [x[L-1] \ x[L-2] \ x[L-3] \ \cdots]^T$ to produce

$$\mathbf{e}_0 = \mathbf{H}_e \mathbf{x}. \quad (3.143)$$

With this construction, the autocovariance matrix for the random vector \mathbf{e}_0 may be expressed as

$$\begin{aligned} \Lambda_{\mathbf{e}_0} &= \mathbf{E}\{\mathbf{H}_e \mathbf{x} (\mathbf{H}_e \mathbf{x})^T\} \\ &= \mathbf{E}\{\mathbf{H}_e \mathbf{x} \mathbf{x}^T \mathbf{H}_e^T\} \\ &= \mathbf{H}_e \mathbf{E}\{\mathbf{x} \mathbf{x}^T\} \mathbf{H}_e^T \\ &= \mathbf{H}_e \Lambda_{\mathbf{x}} \mathbf{H}_e^T, \end{aligned} \quad (3.144)$$

Now if \mathbf{H}_{2S_2} is the convolution matrix for the fixed order- $2S_2$ filter with S_2 second-order sections, having L rows and an infinite number of columns,

$$\mathbf{H}_{2S_2} = \begin{bmatrix} 0 & \cdots & 0 & h_{2S_2}[0] & h_{2S_2}[1] & h_{2S_2}[2] & \cdots \\ \vdots & & \cdots & & \cdots & & \\ 0 & h_{2S_2}[0] & h_{2S_2}[1] & h_{2S_2}[2] & h_{2S_2}[3] & \cdots & \\ h_{2S_2}[0] & h_{2S_2}[1] & h_{2S_2}[2] & h_{2S_2}[3] & \cdots & & \end{bmatrix}, \quad (3.145)$$

in which the first $L - 1$ columns of the first row are all zeros. The matrix \mathbf{H}_{2S_2} operates on the infinite input vector $\mathbf{x} = [x[L - 1] \ x[L - 2] \ x[L - 3] \ \cdots]^T$ to produce

$$\mathbf{y}_{S_2S_2} = \mathbf{H}_{2S_2}\mathbf{x}. \quad (3.146)$$

With this construction, the autocovariance matrix for the random vector $\mathbf{y}_{S_2S_2}$ may be expressed as

$$\begin{aligned} \mathbf{\Lambda}_{\mathbf{y}_{S_2S_2}} &= \mathbf{E}\{\mathbf{H}_{2S_2}\mathbf{x}(\mathbf{H}_{2S_2}\mathbf{x})^T\} \\ &= \mathbf{E}\{\mathbf{H}_{2S_2}\mathbf{x}\mathbf{x}^T\mathbf{H}_{2S_2}^T\} \\ &= \mathbf{H}_{2S_2}\mathbf{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{H}_{2S_2}^T \\ &= \mathbf{H}_{2S_2}\mathbf{\Lambda}_{\mathbf{x}}\mathbf{H}_{2S_2}^T, \end{aligned} \quad (3.147)$$

From our deterministic analysis in Section 3.3.1, we know that the intermediate STE satisfies the difference equation

$$e_{\text{tr}}[n] = \sum_{k=1}^2 a_{kS_2} e_{\text{tr}}[n - k]. \quad (3.148)$$

By inspection of the summation in Eq. (3.148) we see that the first element $e_{\text{tr}}[0]$ of the vector \mathbf{e}_{tr} defined is a weighted sum of the elements $e_{\text{tr}}[-1]$ and $e_{\text{tr}}[-2]$ in the vector \mathbf{e}_0 . The second element $e_{\text{tr}}[1]$ of the vector \mathbf{e}_{tr} is also a weighted sum of all but one of the elements in the vector \mathbf{e}_0 , plus the first element $e_{\text{tr}}[0]$ of the vector \mathbf{e}_{tr} multiplied by a_{1S_2} . This is equivalent to a weighted sum of the elements in the vector \mathbf{e}_0 . In fact all of the

elements of the vector \mathbf{e}_{tr} depend on a linear combination of the elements of the vector \mathbf{e}_0 , so that the vector \mathbf{e}_{tr} may be computed via the matrix product

$$\begin{aligned}\mathbf{e}_{\text{tr}} &= \mathbf{Q}_2 \mathbf{e}_0 \\ &= \mathbf{Q}_2 \mathbf{H}_e \mathbf{x},\end{aligned}\tag{3.149}$$

where \mathbf{Q}_2 is a $L \times 2$ matrix whose elements are defined in terms of the polynomial coefficients a_{kS_2} . The formula for the elements of the matrix \mathbf{Q}_2 in terms of the coefficients a_{kS_2} may be found in Appendix A.

Lastly, we note that the STE vector may be expressed as

$$\mathbf{y}_{\text{tr}} = \mathbf{B}_{S_2} \begin{bmatrix} \mathbf{e}_0 \\ \mathbf{e}_{\text{tr}} \end{bmatrix}\tag{3.150}$$

$$= \mathbf{B}_{S_2} \begin{bmatrix} \mathbf{I} \\ \mathbf{Q}_2 \end{bmatrix} \mathbf{H}_e \mathbf{x},\tag{3.151}$$

where the $L \times (L + 2)$ matrix \mathbf{B}_{S_2} is given by

$$\mathbf{B}_{2S_2} = \begin{bmatrix} b_{2S_2} & b_{1S_2} & b_{0S_2} & 0 & 0 & \cdots \\ 0 & b_{2S_2} & b_{1S_2} & b_{0S_2} & 0 & \cdots \\ 0 & 0 & b_{2S_2} & b_{1S_2} & b_{0S_2} & 0 & \cdots \\ \cdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.\tag{3.152}$$

We now consider the total output power estimate $\hat{P}_{y_{S_2 S_2}}$ based on L samples of the output signal $y_{S_2 S_2}[n]$, defined to be

$$\hat{P}_{y_{S_2 S_2}} = \sum_{k=0}^{L-1} y_{S_2 S_2}^2[k].\tag{3.153}$$

The expected value of $\hat{P}_{\mathbf{y}_{S_2 S_2}}$ may be written as

$$\mathbb{E} \left\{ \hat{P}_{\mathbf{y}_{S_2 S_2}} \right\} = \mathbb{E} \left\{ \mathbf{y}_{S_2 S_2}^T \mathbf{y}_{S_2 S_2} \right\} \quad (3.154)$$

or alternatively as

$$\mathbb{E} \left\{ \hat{P}_{\mathbf{y}_{S_2 S_2}} \right\} = \text{trace} \left(\mathbb{E} \left\{ \mathbf{y}_{S_2 S_2} \mathbf{y}_{S_2 S_2}^T \right\} \right). \quad (3.155)$$

Using the simple substitution $\mathbf{y}_{S_2 S_2} = \mathbf{H}_{2S_2} \mathbf{x}$, we arrive at

$$\begin{aligned} \mathbb{E} \left\{ \hat{P}_{\mathbf{y}_{S_2 S_2}} \right\} &= \text{trace} \left(\mathbb{E} \left\{ \mathbf{H}_{2S_2} \mathbf{x} (\mathbf{H}_{2S_2} \mathbf{x})^T \right\} \right) \\ &= \text{trace} \left(\mathbf{H}_{2S_2} \mathbf{\Lambda}_{\mathbf{x}} \mathbf{H}_{2S_2}^T \right) \end{aligned} \quad (3.156)$$

As discussed previously, when using an approximate filter, the output signal $y_{S_2 S_2}[n]$ is not available. Consequently we must instead use the approximate filter output $y_{S_1 S_2}[n]$ to compute the total output power estimate $\hat{P}_{\mathbf{y}_{S_1 S_2}}$ from L samples of the approximate filter output signal $y_{S_1 S_2}[n]$. This estimate is defined to be

$$\hat{P}_{\mathbf{y}_{S_1 S_2}} = \sum_{k=0}^{L-1} y_{S_1 S_2}^2[k] \quad (3.157)$$

which can be expanded to produce

$$\begin{aligned} \hat{P}_{\mathbf{y}_{S_1 S_2}} &= \sum_{k=0}^{L-1} (y_{S_2 S_2}[k] + y_{\text{tr}}[k])^2 \\ &= \hat{P}_{\mathbf{y}_{S_2 S_2}} + \underbrace{\sum_{k=0}^{L-1} 2y_{S_2 S_2}[k]y_{\text{tr}}[k] + y_{\text{tr}}^2[k]}_{P_{\text{tr}}}, \end{aligned} \quad (3.158)$$

where P_{tr} is defined as noted in Eq. (3.158). The term P_{tr} clearly represents the total additive corruption induced into $\hat{P}_{\mathbf{y}_{S_2 S_2}}$ due to the STE. Using our vector notation we

observe that the expected value of P_{tr} may be written as

$$\mathbb{E}\{P_{\text{tr}}\} = \mathbb{E}\left\{\sum_{k=0}^{L-1} 2y_{S_2 S_2}[k]y_{\text{tr}}[k] + y_{\text{tr}}^2[k]\right\} \quad (3.159)$$

or alternatively as

$$\mathbb{E}\{P_{\text{tr}}\} = \text{trace}\left(\mathbb{E}\left\{\mathbf{y}_{S_2 S_2} \mathbf{y}_{\text{tr}}^T\right\}\right) + \text{trace}\left(\mathbb{E}\left\{\mathbf{y}_{\text{tr}} \mathbf{y}_{\text{tr}}^T\right\}\right). \quad (3.160)$$

Using the substitutions $\mathbf{y}_{S_2 S_2} = \mathbf{H}_{2S_2} \mathbf{x}$ and $\mathbf{y}_{\text{tr}} = \mathbf{B}_{S_2} \begin{bmatrix} \mathbf{I} \\ \mathbf{Q}_2 \end{bmatrix} \mathbf{H}_e \mathbf{x}$, we arrive at

$$\begin{aligned} \mathbb{E}\{P_{\text{tr}}\} &= \text{trace}\left(\mathbb{E}\left\{\mathbf{H}_{2S_2} \mathbf{x} (\mathbf{B}_{S_2} \bar{\mathbf{Q}}_2 \mathbf{H}_e \mathbf{x})^T\right\}\right) + \text{trace}\left(\mathbb{E}\left\{\mathbf{B}_{S_2} \bar{\mathbf{Q}}_2 \mathbf{H}_e \mathbf{x} (\mathbf{B}_{S_2} \bar{\mathbf{Q}}_2 \mathbf{H}_e \mathbf{x})^T\right\}\right) \\ &= \text{trace}\left(\mathbf{H}_{2S_2} \boldsymbol{\Lambda}_{\mathbf{x}} \mathbf{H}_e^T \bar{\mathbf{Q}}_2^T \mathbf{B}_{S_2}^T\right) + \text{trace}\left(\mathbf{B}_{S_2} \bar{\mathbf{Q}}_2 \mathbf{H}_e \boldsymbol{\Lambda}_{\mathbf{x}} \mathbf{H}_e^T \bar{\mathbf{Q}}_2^T \mathbf{B}_{S_2}^T\right), \end{aligned} \quad (3.161)$$

where the matrix $\bar{\mathbf{Q}}_2 = \begin{bmatrix} \mathbf{I} \\ \mathbf{Q}_2 \end{bmatrix}$. We define the output power noise-to-signal ratio (OPNSR) as

$$\text{OPNSR} = \frac{\mathbb{E}\{P_{\text{tr}}\}}{\mathbb{E}\{\hat{P}_{y_{S_2 S_2}}\}} \quad (3.162)$$

In the case of a truncation approximate filter structure with $S_2 > S_1$, the output power noise-to-signal ratio (OPNSR) is given by

$$\text{OPNSR} = \frac{\text{trace}\left(\mathbf{H}_{2S_2} \boldsymbol{\Lambda}_{\mathbf{x}} \mathbf{H}_e^T \bar{\mathbf{Q}}_2^T \mathbf{B}_{S_2}^T\right) + \text{trace}\left(\mathbf{B}_{S_2} \bar{\mathbf{Q}}_2 \mathbf{H}_e \boldsymbol{\Lambda}_{\mathbf{x}} \mathbf{H}_e^T \bar{\mathbf{Q}}_2^T \mathbf{B}_{S_2}^T\right)}{\text{trace}\left(\mathbf{H}_{2S_2} \boldsymbol{\Lambda}_{\mathbf{x}} \mathbf{H}_{2S_2}^T\right)} \quad (3.163)$$

The second goal of our probabilistic analysis for truncation approximate filter structures is to quantify the mean-squared STE, $\mathbb{E}\{y_{\text{tr}}^2[n]\}$. This quantity is conveniently represented by the diagonal elements of a matrix we have already derived. We are not particularly

interested in the absolute value of the time series $y_{\text{tr}}^2[n]$, but rather its relative magnitude compared to the desired output $y_{S_2 S_2}^2[n]$. For this purpose we define the output noise-to-signal ratio vector, **ONSR**, whose n th element is defined by

$$\text{ONSR}[n] = \frac{\text{E} \{y_{\text{tr}}^2[n]\}}{\text{E} \{y_{S_2 S_2}^2[n]\}}, \quad 0 \leq n \leq L - 1. \quad (3.164)$$

The vector **ONSR** has dimensions $L \times 1$, and may be more compactly written as

$$\text{ONSR} = \frac{\text{diag} \left(\mathbf{B}_{S_2} \bar{\mathbf{Q}}_2 \mathbf{H}_e \mathbf{\Lambda}_x \mathbf{H}_e^T \bar{\mathbf{Q}}_2^T \mathbf{B}_{S_2}^T \right)}{\text{diag} \left(\mathbf{H}_{2S_2} \mathbf{\Lambda}_x \mathbf{H}_{2S_2}^T \right)}. \quad (3.165)$$

We note that the effect of the STE on the low power optimal filter order estimate \hat{N}_{LP}^* , based on the approximate filter output $y_{S_1 S_2}[n]$, follows directly from our previous formulation for the STE in replacement approximate filter structures.

This concludes our analysis of the STE in truncation approximate filter structures. As a final note, we mention that while the STE for replacement filter structures was shown to decay exponentially with the largest magnitude pole out of the N_2 poles of the post-transition filter $h_{N_2}[n]$, the STE for truncation filter structures with $N_2 > N_1$ was shown to decay exponentially with the pole of largest magnitude out of the 2 poles of the new second-order section. Since the pole of largest magnitude out of the 2 poles of the new second-order section will in all but one case be smaller than the pole of largest magnitude out of the N_2 poles of the entire cascade of $S_2 = N_2/2$ second-order sections, we may conclude that most of the time the STE will decay more rapidly using a truncation filter structure than when using a replacement filter structure.

3.4 Summary

In this chapter we defined and analyzed the state transition error (STE). We developed a model for this corruption as an additive noise term in the approximate filter output signal. By formulating deterministic and probabilistic frameworks, we analytically and empirically investigated the effects of the STE on the approximate filter output $y_{N_1 N_2}[n]$, the approximate filter output power estimate, and the low power filter order estimate \hat{N}_{LP}^* based on the approximate filter output $y_{N_1 N_2}[n]$.

Chapter 4

Approximate Filter Structures

In Chapter 2 we showed that approximate filtering algorithms dynamically reduce the order of a frequency-selective digital filter while maintaining a desired level of output quality and thus conserve power. The order of the filter is varied by defining a control strategy which works in conjunction with an approximate filter structure. The approximate filter structure is defined by a set of filters of different orders. The control strategy produces an estimate of the best filter order to use from those available in the approximate filter structure based on real-time measurements of the input signal statistics.

In this chapter we study approximate filter structures and investigate their important role in defining good approximate filtering algorithms. We introduce *replacement* and *truncation* filter structures, which represent two important classes of approximate filter structures, and analyze their respective advantages and shortcomings. These two classes are further broken down by the type of constituent filter elements they may use, either finite impulse response (FIR) or infinite impulse response (IIR). This decomposition results in four types of distinct approximate filter structures, for which we adopt the names Type FR, Type IR, Type FT, and Type IT. We will see that each of these four approximate filter types has its own design subtleties, and determining the best approximate filter depends heavily on the nature of the application and performance specifications. To assess the relative performance of approximate filter structures, we use the signal-to-noise ratio (SNR) improvement factor defined in Chapter 2 as well as the output power noise-to-signal ratio due to the state transition error (STE), defined in Chapter 3.

A collection of frequency-selective digital filters, one for each filter order N in a given range $N_{\min} \leq N \leq N_{\max}$ constitutes an approximate filter structure \mathcal{H} . Each filter struc-

ture \mathcal{H} possesses the property that its progressively higher order filters have progressively increased average attenuation in the stopband region(s) while maintaining close to unity gain in the passband region(s). The passband and stopband regions for all the filters in the filter structure \mathcal{H} are identical. In addition, each of the individual filters which make up the constituent elements of the filter structure \mathcal{H} must be properly normalized. Possible normalizations include a unit energy normalization or a unity DC (zero frequency) gain normalization. In the analysis and simulations of this chapter, a unity DC (zero frequency) gain normalization is used.

We stated earlier that an approximate filtering algorithm may be used in conjunction with a filter structure whose constituent elements are IIR or FIR digital filters. The choice of whether to use an IIR or FIR structure involves a tradeoff between filtering performance characteristics such as sharpness of the transition band, desired spectral magnitude, linearity of the phase response, processing delay, stability in the presence of filter coefficient quantization, and efficiency of implementation (for example, FFT, convolution, or direct forms).

The primary advantage of an IIR filter structure is that it can provide significantly better stopband attenuation and less delay than an FIR filter structure having the same number of coefficients. This is a consequence of the output feedback which generates an infinite impulse response with only a finite number of parameters [60]. FIR filter structures are desirable for their guaranteed stability even in the presence of coefficient quantization and for the possibility of an exact linear phase characteristic. However, it should be noted that some commercially available DSP chips can implement certain FIR filters more computationally efficiently than standard IIR filters because the chip architecture has been optimized for a particular FIR filter. In addition, there exist nonlinear phase FIR filters which can provide significantly better stopband attenuation than the linear phase FIR filters. Therefore, the statement that IIR filters are always more computationally efficient than FIR filters should not be made without careful consideration of the variables at hand.

In addition to the flexibility to choose IIR or FIR filter elements in an approximate filter structure, we are free to use either a replacement filter structure or a truncation filter structure. In a truncation filter structure with FIR constituent elements, the coefficients defining the lower order filters are constrained to be subsets of the coefficients defining the filter with maximum order N_{\max} . In a truncation filter structure with IIR constituent elements, the set of pole/zero pairs defining each lower order filter is similarly constrained

to subset of the pole/zero pairs defining the filter with maximum order N_{\max} . Thus the lower order constituent elements in a truncation filter structure are *truncated* versions of the higher order constituent elements.

In a replacement filter structure the relationship between the coefficients defining filters of different orders are not necessarily related in any way; the coefficients of each individual filter may be *replaced* independently. Given this, we expect the replacement filter structures with unconstrained filter coefficients to perform better than the truncation filter structures with constrained coefficients. This expectation will be confirmed in our analysis and simulations.

Truncation filter structures offer more power savings opportunities in a CMOS technology implementation than replacement filter structures. This is true since truncation filter structures may be described with fewer independent filter coefficients than replacement filter structures and thus require less memory, chip area, and bus accesses. In summary, while replacement filter structures have the advantage of offering better filtering performance, truncation filter structures are more power efficient.

We classify approximate filter structures as belonging to one of four possible approximate filter structure types. The definitions of the four approximate filter structure types are summarized in Table 4.1.

Table 4.1: Summary of the four types of approximate filter structures.

Constituent Filter Impulse Responses	Pruning Method	
	Replacement	Truncation
FIR	Type FR	Type FT
IIR	Type IR	Type IT

The implementation of approximate filter structures with FIR constituent elements (Type FR and Type FT) is conceptually simpler than that of approximate filter structures with IIR constituent elements (Type IR and Type IT), so we focus on FIR approximate filter structures first. In Fig. 4-1 we have plotted the frequency response magnitudes for rectangularly-windowed ideal FIR filters of orders $N = 20, 80$ and 140 . From these plot we observe that as the filter order increases, the average attenuation in the stopband of the filter also increases.

Now consider the conceptual diagram in Fig. 4-2. In order to instantaneously change

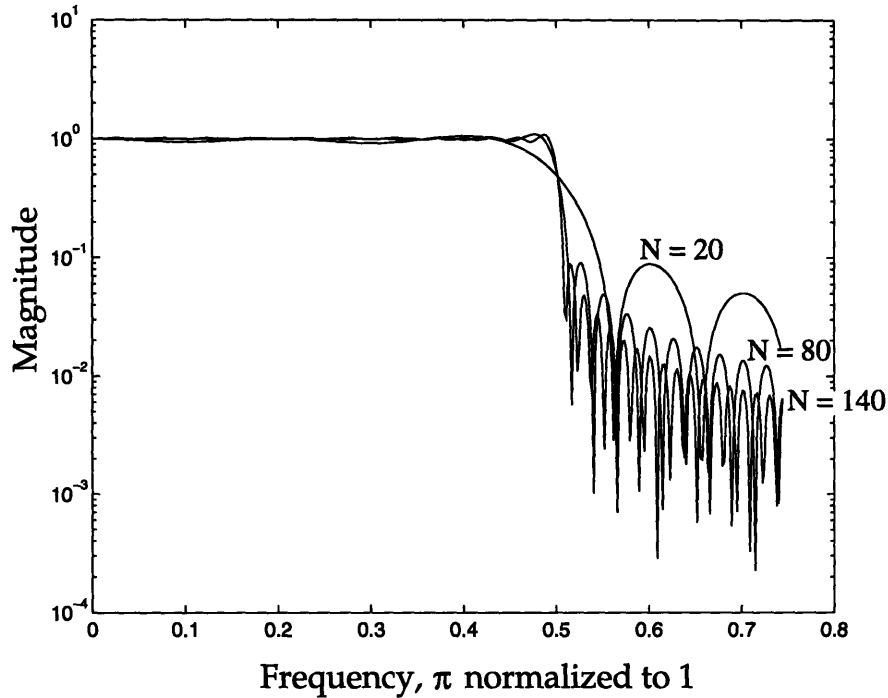


Figure 4-1: Frequency response magnitudes for rectangularly-windowed ideal FIR filters of orders $N = 20, 80$ and 140 .

the filter order using a Type FR or Type FT structure at a particular time, we select the desired filter order N , use the appropriate $(N + 1)$ FIR filter coefficients to weight the $(N + 1)$ past input samples, and then sum these weighted input values to produce the current output sample value. For a replacement FIR (Type FR) filter structure with $M = (N_{\max} - N_{\min} + 1)$ distinct orders, we must store $M(M + 1)/2$ distinct FIR filter coefficients. This is true since all the filter coefficients for the filters of distinct orders are unrelated, so each one must occupy a memory location. However for a truncation FIR (Type III) filter structure with $M = (N_{\max} - N_{\min} + 1)$ distinct orders, we must store only N_{\max} distinct FIR filter coefficients. This is true since all of the coefficients of each of the lower order filters are subsets of the order- N_{\max} filter coefficients, so only the N_{\max} coefficients of the order- N_{\max} filter need to be stored. As a final comment we note that in either Type FR or Type FT approximate filter structures, the STE is zero since the filters are all FIR.

We measure the performance of an FIR approximate filter structure by the signal-to-noise ratio (SNR) improvement factor given in Chapter 2. Recall that the SNR improvement

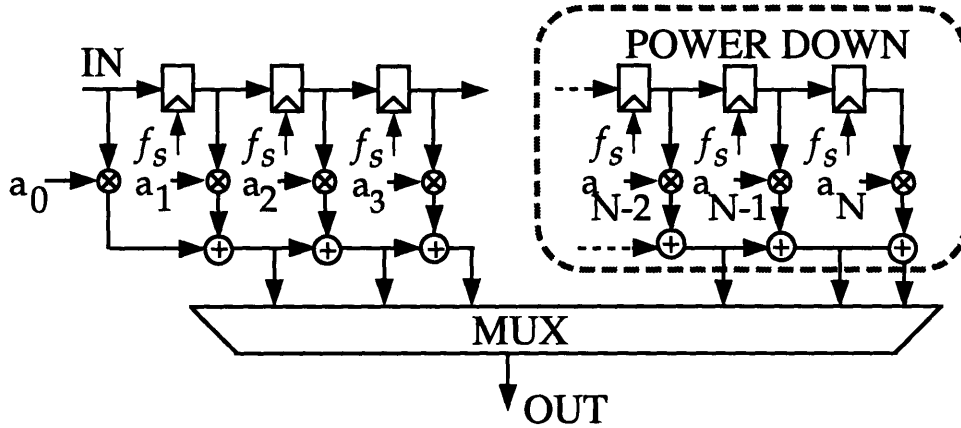


Figure 4-2: An overview of the FIR approximate filter structure.

factor for the k th-order filter in an approximate filter structure is

$$\text{SNRI}[k] = \frac{\Delta_{SB}}{\int_{SB} |H_N(\omega)|^2 d\omega} \quad (4.1)$$

For each fixed filter order in the range $N_{\min} \leq N \leq N_{\max}$, we would like to maximize the SNR improvement factor in Eq. (4.1) for optimal approximate filtering performance. From Eq. (4.1) we see that for filter order k this is equivalent to determining the k th-order FIR filter with *minimum stopband power*.

The exact solution to this optimization problem when the FIR filter coefficients are symmetric may be found and is easily computable. The resulting class of FIR filters are known as *eigenfilters* [71]. Because eigenfilters are the optimal constituent elements for Type FR or Type FT approximate filter structures, we will focus our attention on their derivation and properties in part of Section 4.1.1.

We now consider approximate filter structures with IIR constituent filter elements. A conceptual diagram of an IIR replacement (Type IR) filter structure is given in Fig. 4-3. As an example, consider that at time $n = n_0$, the order of the filter structure may be decreased by one from N_{\max} to $(N_{\max} - 1)$ by simply setting the coefficient pair $(a_{N_{\max}}, b_{N_{\max}})$ in Fig. 4-3 to zero. More generally at time $n = n_0$, the order of the approximate filter may be set to any order $N_0 < N_{\max}$ by simply setting the coefficient pairs $(a_{N_{\max}}, b_{N_{\max}}) \cdots (a_{N_0+1}, b_{N_0+1})$ to zero. It is important to note that the data stream in the middle of the replacement filter structure shown in Fig. 4-3 continues to shift through all of the N_{\max} vertical delay

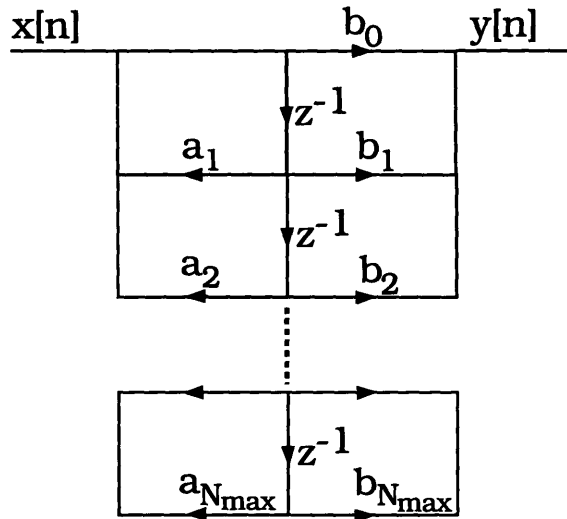


Figure 4-3: Conceptual diagram of the IIR replacement filter structure.

elements, regardless of which coefficients have been set to zero. In addition, we note that all coefficients are allowed to change their values at each instant in time, according to the elements in the filter structure \mathcal{H} .

Conceptual diagrams of the IIR truncation (Type IT) filter structure are given in Fig. 4-4. In Fig. 4-4(a) is a conceptual diagram of the signal flow graph, while in Fig. 4-4(b) we show a clocked shift register hardware block diagram. As an example, consider that at time $n = n_0$, the order of the filter structure may be decreased by two from $2M_0$ to $(2M_0 - 2)$ by simply truncating the last second-order section and taking the output to be $y_{M_0-1}[n]$. In general, at time $n = n_0$, the order of the truncation filter structure may be set to any even order $2M < 2M_0$ by truncating the last $(M_0 - M)$ second-order sections from the cascade structure shown in Fig. 4-4 and taking the output to be $y_M[n]$. If we desire to increase the order of the Type IT filter structure, second-order sections may be added to the truncation filter structure at any time.

One measure of the performance of an IIR approximate filter structure is the signal-to-noise ratio (SNR) improvement factor, originally given in Chapter 2, which was presented again in Eq. (4.1). This measure is used for FIR approximate filter structures as well. For approximate filter structures with IIR filter elements we also use the output power noise-to-signal ratio, defined in Eq. (3.89) in Chapter 3, as an additional performance metric to encapsulate the STE. By intelligently choosing an appropriate Type IR or Type IT approximate filter structure, it is possible to reduce the effect of the STE in approximate

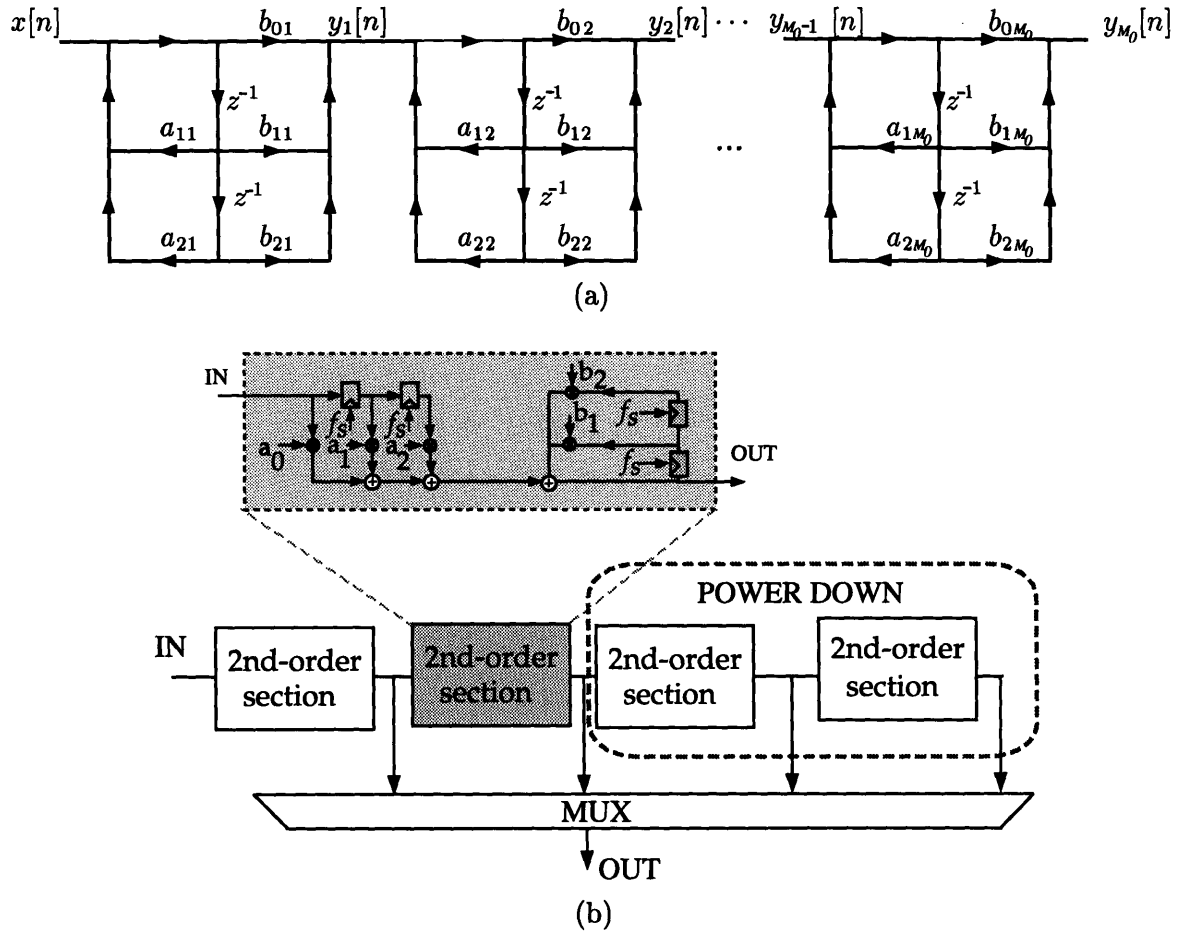


Figure 4-4: Conceptual diagrams of the IIR truncation filter structure: (a) the signal flow graph, and (b) the clocked shift register block diagram.

filtering.

Since finding optimal IIR filter structures designed according to these two performance metrics (SNRI and OPNSR) involves an unsolvable constrained nonlinear optimization problem, we do not pursue direct IIR filter structure design. Instead we evaluate the performance of IIR approximate filter structures using four classical IIR digital filter constituent elements, namely Butterworth, Chebyshev, inverse Chebyshev, and elliptic digital filters.

We have introduced the four approximate filter structure types, and presented one performance metric (SNRI) for FIR approximate filter structures and two performance metrics

(SNRI and OPNSR) for IIR approximate filter structures. In the next two sections of this chapter we will evaluate specific replacement and truncation filter structures, according to the appropriate performance metric(s), and discuss the advantages and shortcomings of each type of approximate filter structure.

4.1 Replacement Filter Structures

A replacement filter structure \mathcal{H}_R is defined by a set of $(N_{\max} - N_{\min} + 1)$ digital filters, one for each filter order N in a given range $N_{\min} \leq N \leq N_{\max}$. We denote this set by

$$\mathcal{H}_R = \{H_{N_{\max}}(\omega), H_{N_{\max}-1}(\omega), \dots, H_{N_{\min}}(\omega)\}. \quad (4.2)$$

We define the filter structure \mathcal{H}_R by defining its constituent filter elements. These filters must be either all FIR (Type FR approximate filter structure) or all IIR (Type IR approximate filter structure), and should possess similar spectral characteristics for obvious practical reasons. Some examples of Type FR approximate filter structures are those defined by individual FIR constituent elements which are Parks-McLellan equiripple filters [44], eigenfilters [71], prolate spheroidal windowed filters [70], or other windowed ideal filters [44]. Some examples of Type IR approximate filter structures are those defined by individual IIR constituent elements which are digital Butterworth, Chebyshev, inverse Chebyshev, or elliptic filters [44].

Since we are free to choose the constituent filters in a replacement filter structure arbitrarily, provided that all filters are uniformly IIR or FIR, there is opportunity for substantial filter structure design flexibility. In all our analyses and simulations we normalize each constituent filter element to have a unity DC gain.

4.1.1 Type FR Filter Structures

Type FR filter structures are characterized as replacement filter structures having FIR constituent filter elements. Because Type FR filter structures have the replacement quality, the coefficients of the filter of a particular order are unrelated to the coefficients of any other filter with a different order within the structure. The performance of a Type FR approximate filter structure is measured by the SNR improvement factor. By inspection

of Eq. (4.1), we see that minimizing the power in the stopband of each filter maximizes the SNR improvement factor, which motivates us to investigate the use of eigenfilters for approximate filtering with a Type FR filter structure.

An important family of symmetric FIR filters corresponds to the symmetric windowing of the impulse responses of corresponding ideal filters. For example, a lowpass filter of this type has an impulse response given by [44]:

$$h[n] = w[n] \frac{\sin \omega_c n}{\pi n}, \quad (4.3)$$

where $w[n]$ is a symmetric N -point window. In order to motivate the use of eigenfilters, we follow the reasoning presented in [70]. Consider the rectangularly-windowed FIR filter $h[n]$ with frequency response $H(\omega)$. This filter minimizes the squared-error between the ideal IIR filter $h_i[n]$ with frequency response $H_i(\omega)$ and the FIR filter $h[n]$. Equivalently, the rectangularly-windowed FIR filter coefficients minimize the integral

$$\int_{-\pi}^{\pi} |H_i(\omega) - H(\omega)|^2 d\omega \quad (4.4)$$

Unfortunately, the rectangularly-windowed FIR filters suffer from the well-known Gibbs phenomenon. They also minimize the squared error across the transition band, which by definition is a “don’t care” region and should be ignored. A class of FIR filters called eigenfilters [71] resolve these two issues. First, eigenfilters minimize the sum of the stopband and passband errors only, ignoring the transition band error. Second, the magnitude response of an eigenfilter does not demonstrate the Gibbs phenomenon. The eigenfilter coefficients may be easily computed as an eigenvector of an appropriate positive definite matrix which we shall now describe.

We provide a brief summary of the derivation of eigenfilters, and then proceed to demonstrate their utility in approximate filtering. Consider linear phase FIR filters [44] which are real-valued and have symmetric impulse responses which satisfy $h[n] = h[N - n]$, with odd length $N + 1$. The amplitude of the frequency response is

$$H_A(\omega) = \sum_{n=0}^M b_n \cos(\omega n) = \mathbf{b}^T \mathbf{c}(\omega), \quad (4.5)$$

where $M = N/2$,

$$\mathbf{b} = [h[0] \quad 2h[1] \quad 2h[2] \quad \cdots \quad 2h[N-2] \quad 2h[N-1]]^T, \quad (4.6)$$

$$\mathbf{c}(\omega) = [1 \quad \cos(\omega) \quad \cos(2\omega) \quad \cdots \quad \cos(M\omega)]^T. \quad (4.7)$$

and b_n is the n th element of the vector \mathbf{b} . The eigenfilter problem is to find the coefficients in the vector \mathbf{b} which minimize the sum of the passband and stopband squared errors. Since $H(\omega) = e^{-j\omega M} H_A(\omega)$, we have

$$|H(\omega)| = H_A^2(\omega) = \mathbf{b}^T \mathbf{c}(\omega) \mathbf{c}^T(\omega) \mathbf{b}. \quad (4.8)$$

The stopband power of the filter is

$$P_{SB}^h = \int_{\omega_s}^{\pi} |H(\omega)|^2 \frac{d\omega}{\pi} = \mathbf{b}^T \mathbf{S} \mathbf{b}, \quad (4.9)$$

where

$$\mathbf{S} = \int_{\omega_s}^{\pi} \mathbf{c}(\omega) \mathbf{c}^T(\omega) \frac{d\omega}{\pi}. \quad (4.10)$$

The (m, n) element of \mathbf{S} is

$$s_{m,n} = \int_{\omega_s}^{\pi} \cos(m\omega) \cos(n\omega) \frac{d\omega}{\pi}, \quad (4.11)$$

which can be evaluated in terms of ω_s , m , and n [71]. In order to incorporate the passband error into the eigenfilter design, we observe that the amplitude response at zero frequency $H_A(0)$ may be expressed as $H_A(0) = \mathbf{b}^T \mathbf{1}$, where $\mathbf{1}$ is the $N \times 1$ vector of all 1's. Taking $H_A(0)$ as a reference, the passband deviation at any frequency may be written as

$$\mathbf{b}^T \mathbf{1} - \mathbf{b}^T \mathbf{c}(\omega) = \mathbf{b}^T [\mathbf{1} - \mathbf{c}(\omega)], \quad (4.12)$$

so that the total passband error is

$$P_{PB}^h == \mathbf{b}^T \mathbf{P} \mathbf{b}, \quad (4.13)$$

where

$$\mathbf{P} = \int_0^{\omega_p} [1 - \mathbf{c}(\omega)][1 - \mathbf{c}(\omega)]^T \frac{d\omega}{\pi}. \quad (4.14)$$

We now define the objective function Φ that is to be minimized by the appropriate choice of filter coefficients as

$$\Phi = \alpha P_{SB}^h + (1 - \alpha) P_{PB}^h, \quad (4.15)$$

where $0 < \alpha < 1$. Here α is a parameter which dictates the tradeoff between passband and stopband error. We may simplify our expression for Φ to:

$$\Phi = \mathbf{b}^T \mathbf{R} \mathbf{b}, \quad (4.16)$$

where

$$\mathbf{R} = \alpha \mathbf{S} + (1 - \alpha) \mathbf{P} \quad (4.17)$$

It can easily be shown that \mathbf{R} is a real, symmetric, and positive definite matrix. The unit-

norm vector \mathbf{b} which minimizes the objective function, Φ , is the eigenvector corresponding to the minimum eigenvalue, λ_0 , of \mathbf{R} , and can be easily calculated using the well-known *power method* [70]. The elements of the vector \mathbf{b} then define the eigenfilter coefficients $h[n]$ according to the relationship in Eq. (4.6).

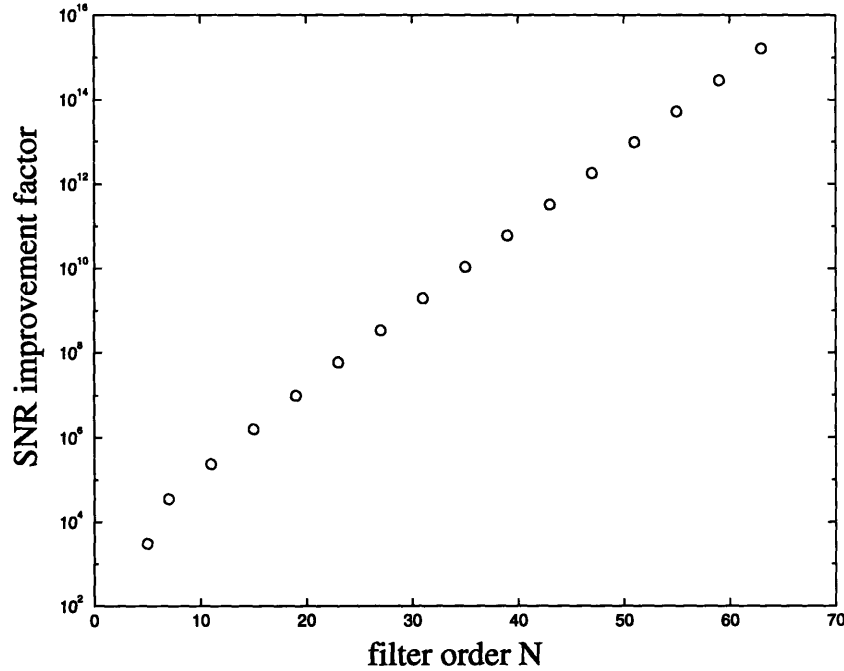


Figure 4-5: Performance profile for the eigenfilter FIR replacement filter structure.

By construction, eigenfilters are naturally the optimal Type FR approximate filter structure constituent filter elements in terms of providing the highest SNR improvement factor for a fixed filter order and fixed power in the passband ripple. In Fig. 4-5 the performance profile for the Type FR approximate filter structure using eigenfilters is shown. The eigenfilters were designed using a passband $|\omega| \in [0, 3\pi/8]$, a stopband $|\omega| \in [5\pi/8, \pi]$, and a weighting parameter $\alpha = 0.1$. The filter structure performance profile is a plot of the SNR improvement factor vs. filter order. In Fig. 4-6 we show the performance profile for the Type FR approximate filter structure using Parks-McLellan filters. For comparison purposes, the Parks-McLellan filters were designed to have approximately the same power in the passband ripple as the eigenfilters. By inspecting the performance profiles we observe that the eigenfilters are the best, as expected.

To end this section on Type FR approximate filter structures, we note that an analytical

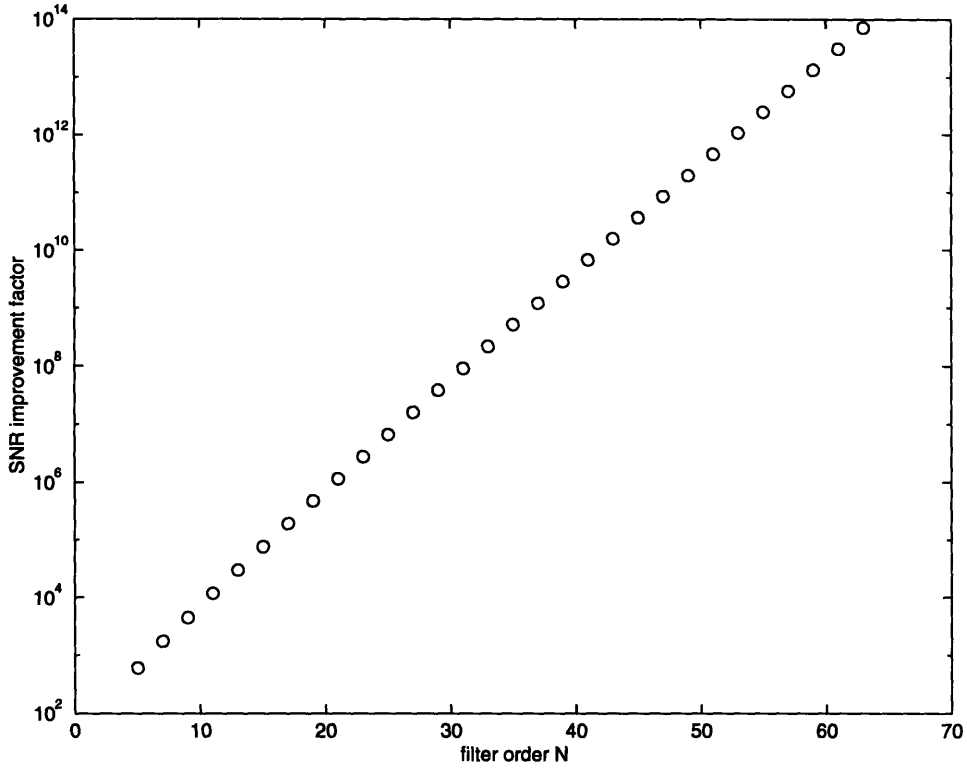


Figure 4-6: Performance profile for the Parks-McLellan FIR replacement filter structure.

expression may be found for the integral of the frequency response magnitude-squared of an FIR filter. This integral is needed to calculate the SNR improvement factors shown in Fig. 4-5 and Fig. 4-6. A derivation of this expression is provided here for the sake of completeness. The derivation presented here directly follows that presented in [70].

In order to determine the SNR improvement factor for a given FIR filter, we need to compute the total filter stopband power. The stopband power of a filter is defined as integral of the frequency response magnitude-squared. Assuming that the real-valued FIR filter $h[n]$ is causal with length $N + 1$, we know that

$$H_N(\omega) = \sum_{n=0}^N h[n]e^{-j\omega n}. \quad (4.18)$$

The filter stopband power is defined as

$$P_{SB}^h[N] = \frac{1}{2\pi} \int_{SB} |H_N(\omega)|^2 d\omega. \quad (4.19)$$

Defining the vectors

$$\mathbf{h} = [h[0] \ h[1] \ h[2] \ \dots \ h[N]]^T \quad (4.20)$$

and

$$\mathbf{e}(e^{j\omega}) = [1 \ e^{-j\omega} \ \dots \ e^{-j\omega N}]^T, \quad (4.21)$$

we have $H(\omega) = \mathbf{h}^T \mathbf{e}(e^{j\omega})$, so that

$$|H(\omega)|^2 = H(\omega)H^*(\omega) = \mathbf{h}^T \mathbf{e}(e^{j\omega}) \mathbf{e}^T(e^{j\omega}) \mathbf{h}. \quad (4.22)$$

Now we may rewrite the stopband power as

$$P_{SB}^h[N] = \mathbf{h}^T \left[\frac{1}{2\pi} \int_{SB} \mathbf{R}(\omega) d\omega \right] \mathbf{h} \quad (4.23)$$

where

$$\mathbf{R}(\omega) = \mathbf{e}(e^{j\omega}) \mathbf{e}^T(e^{j\omega}). \quad (4.24)$$

The element with index (m, n) in the $(N + 1) \times (N + 1)$ matrix $\mathbf{R}(\omega)$ is:

$$e^{-j(m-n)\omega} = \cos \omega(m - n) - j \sin \omega(m - n), \quad (4.25)$$

so that $\mathbf{R}(\omega) = \mathbf{I}(\omega) + j\mathbf{Q}(\omega)$ is a Hermitian matrix. Therefore $\mathbf{Q}(\omega)$ is antisymmetric, and

$\mathbf{h}^T \mathbf{Q}(\omega) \mathbf{h} = 0$. Thus, $P_{SB}^h[N]$ can be simplified to

$$P_{SB}^h[N] = \mathbf{h}^T \mathbf{P} \mathbf{h}, \quad (4.26)$$

where \mathbf{P} has (m, n) th entry

$$p_{mn} = \frac{1}{2\pi} \int_{SB} \cos \omega(m-n) d\omega, \quad 0 \leq m, n \leq N. \quad (4.27)$$

In the case of a lowpass filter with cutoff frequency ω_c , the stopband region is defined as $\omega_c \leq \omega \leq \pi$, and we have

$$p_{mn} = \frac{1}{\pi} \int_{\omega_c}^{\pi} \cos \omega(m-n) d\omega = -\frac{\sin \omega_c(m-n)}{\pi(m-n)} \quad 0 \leq m, n \leq N, \quad (4.28)$$

and $P_{SB}^h[N]$ may be computed directly using Eq. (4.26). The computation of the SNR improvement factor then follows directly using Eq. (4.1).

4.1.2 Type IR Filter Structures

Type IR approximate filter structures have IIR constituent filter elements and the replacement quality. Because Type IR approximate filter structures have the replacement quality, the coefficients of the filter of a particular order are unrelated to the coefficients of filters of different orders within the structure.

In Fig. 4-7 we have plotted the frequency responses magnitudes for the Butterworth IIR replacement filter structure for $N_{\min} = 2$ and $N_{\max} = 10$. In Fig. 4-8, Fig. 4-9, and Fig. 4-10 we show similar plots for the Chebyshev, inverse Chebyshev, and elliptic replacement filter structures. All the filters have been normalized such that the maximum ripple in the passband is equal to 0.01.

The performance of a Type IR approximate filter structure is measured by the SNR improvement factor, as was the case with Type FR filter structures. From inspection of the frequency response magnitude-squared plots, we observe that the elliptic replacement filter responses have visually the lowest stopband power, and thus we would expect the elliptic replacement filter structure to have the best performance profile. Indeed, this is

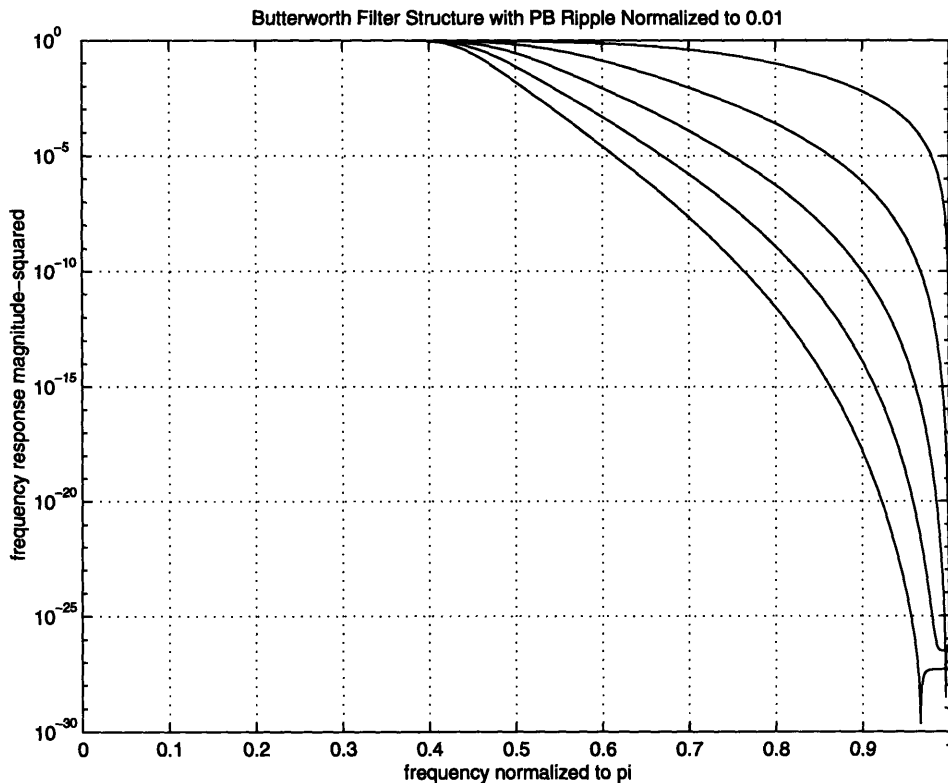


Figure 4-7: Frequency response magnitude-squared plots for the Butterworth IIR replacement filter structure.

confirmed in Fig. 4-11, where we show a comparison of the performance profiles for the four IIR replacement filter structures considered thus far.

To assess performance we also consider the STE performance metric which measures the corruptive effect of instantaneously switching from the *maximum filter* order to each of the other lower order filters in the approximate filter structure. The STE performance metric for a filter of order N is defined as the output power noise-to-signal ratio, $OPNSR[N]$, which was given in Chapter 3. An expression for $OPNSR[N]$ was given in Eq. (3.89). The output power noise-to-signal ratio $OPNSR[N_k]$ here is defined as that arising from instantaneously switching from the N_{max} -order filter to the N_k -order filter. In Fig. 4-12 we show a comparison of the STE performance metric for the same set of four IIR replacement filter structures. In this case the Chebyshev IIR replacement filter structure is the best in terms of STE performance.

In closing, we make a few notes on the relative number of operations required to implement different IIR filter types. In general an order- N IIR filter requires $2N$ additions and

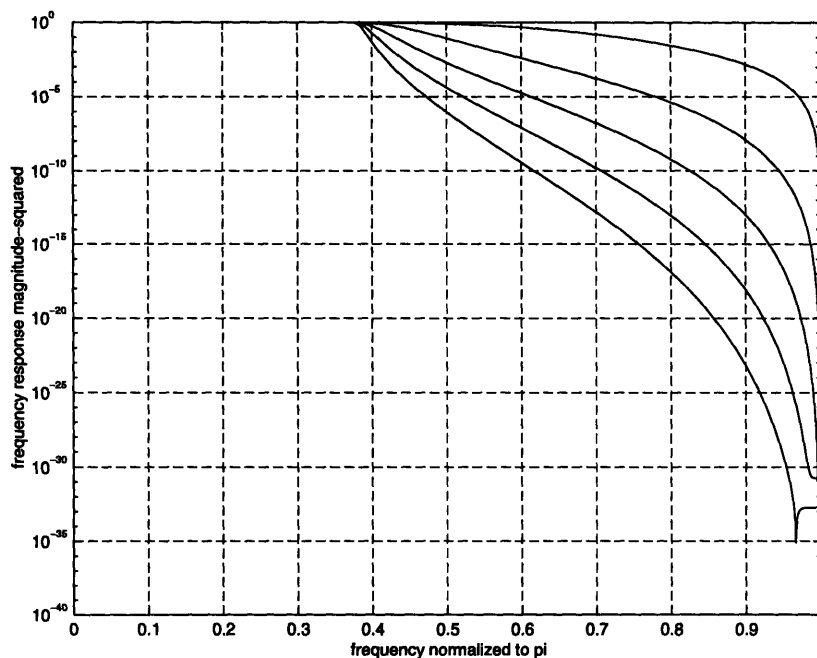


Figure 4-8: Frequency response magnitude-squared plots for the Chebyshev IIR replacement filter structure.

$2N$ multiplies per output sample to implement in direct form. However only $2N$ additions and $(N + 1)$ multiplies are needed to implement an N th-order Butterworth filter in direct form, due to the fact that all its zeros are at $z = -1$. This is true for Chebyshev and inverse Chebyshev IIR filters as well, which are both all-pole filters in the analog domain and thus transform to having all their zeros at $z = -1$ in the discrete-time domain via the bilinear transformation.

In [70] it is shown that an elliptic filter can be implemented as the sum of two allpass filters. With this special decomposition only N multiplies per output sample are required. Even without this allpass decomposition, an elliptic filter can be implemented with $1.5N$ multiplies due to the symmetry of its numerator polynomial coefficients. Thus, there is no great advantage to the Butterworth Chebyshev, and inverse Chebyshev filters having all their zeros at $z = -1$. Elliptic filters can be implemented equally as efficiently with a better SNR improvement factor performance profile. Because of the equiripple nature of both the passband and the stopband, the elliptic filter requires a much smaller order than that of a Butterworth or Chebyshev filter meeting the same specifications. Thus, the elliptic

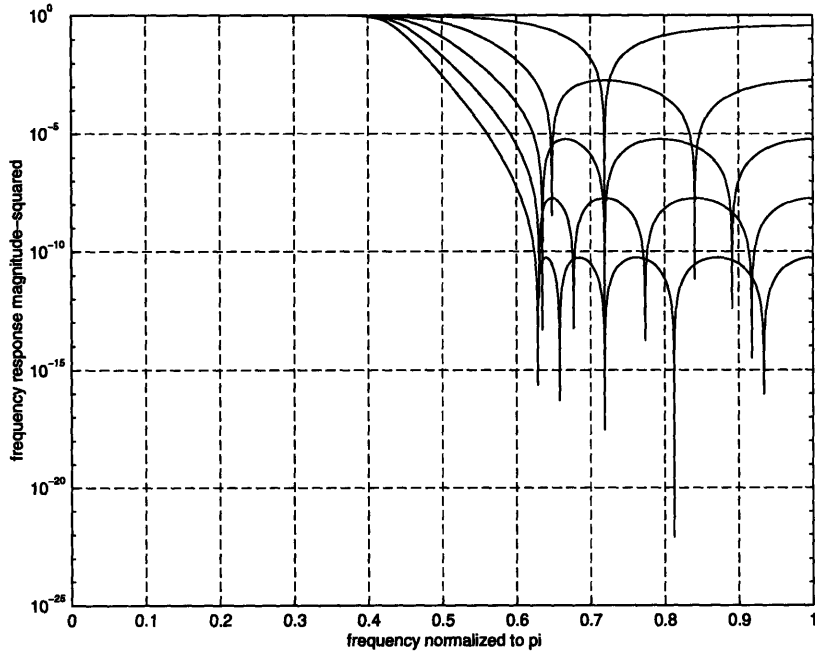


Figure 4-9: Frequency response magnitude-squared plots for the inverse Chebyshev IIR replacement filter structure.

replacement filter structure is in general a good choice for IIR approximate filtering.

4.2 Truncation Filter Structures

A truncation filter structure \mathcal{H}_T is defined by a set of $(N_{\max} - N_{\min} + 1)$ digital filters, one for each filter order N in a given range $N_{\min} \leq N \leq N_{\max}$. We denote this set by

$$\mathcal{H}_T = \{H_{N_{\max}}(\omega), H_{N_{\max}-1}(\omega), \dots, H_{N_{\min}}(\omega)\}. \quad (4.29)$$

Again we define the filter structure \mathcal{H}_T by defining its constituent filter elements. These filters must be either all FIR (Type FT approximate filter structure) or all IIR (Type IT approximate filter structure), and should possess similar spectral characteristics for obvious practical reasons. We impose an additional constraint on truncation filter structures. In the FIR truncation filter structure, the coefficients defining the lower order filters are constrained to be subsets of the coefficients defining the filter with maximum order N_{\max}

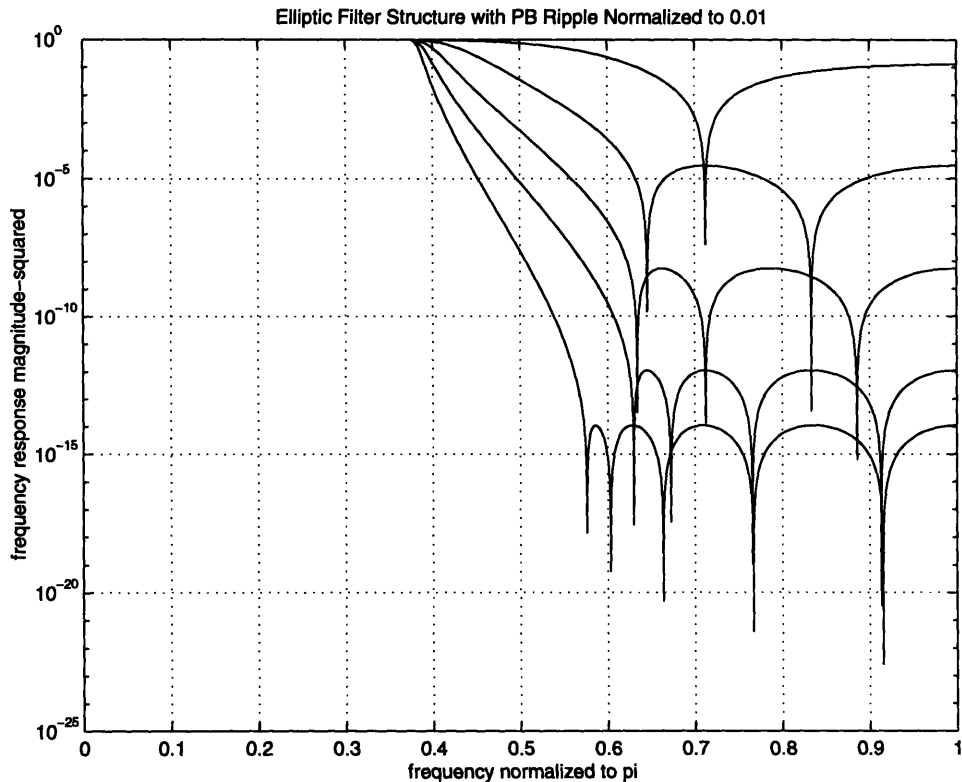


Figure 4-10: Frequency response magnitude-squared plots for the IIR elliptic replacement filter structure.

in the truncation filter structure. This constraint reduces the number of filter coefficients that must be stored and accessed and thus offers potential for reducing chip area and chip power consumption over approximate filtering using replacement filter structures. In an FIR truncation approximate filter structure the filter of maximum order $H_{N_{\max}}(\omega)$ may be chosen to be a Parks-McLellan filter, an eigenfilter, a prolate spheroidal windowed filter, or another windowed ideal filter.

To define an IIR truncation filter structure, we must first define the IIR constituent filter with maximum order, $H_{N_{\max}}$. Then the rest of the filters of lower orders are defined by a pruning sequence pole/zero pairs. In a Type IT approximate filter structure the filter of maximum order $H_{N_{\max}}(\omega)$ may be a digital Butterworth filter, a Chebyshev filter, an inverse Chebyshev filter, or an elliptic filter. In defining an IIR truncation filter structure there is freedom to choose the pruning sequence to meet desired performance specifications. This issue will be explored in Chapter 5. As a final note before proceeding to discuss Type FT filter structures, we mention that in all our analyses and simulations we normalize each

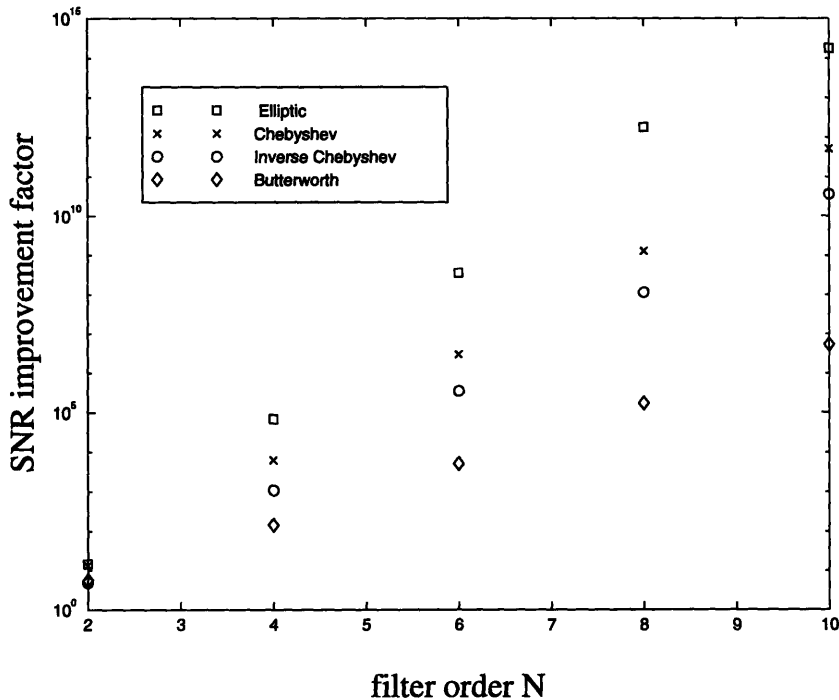


Figure 4-11: Comparison of the performance profiles for the Butterworth, Chebyshev, inverse Chebyshev, and elliptic IIR replacement filter structures.

constituent filter element to have a unity DC gain.

4.2.1 Type FT Filter Structures

In order to describe the FIR truncation (FT) type of approximate filter structures, we first recall that in a truncation filter structure the coefficients defining the lower order filters are constrained to be subsets of the coefficients defining the filter with maximum order in the truncation filter structure.

In Fig. 4-13 we compare the SNR improvement factor performance profiles of the replacement and truncation FIR filter structures using Parks-McLellan filters with $N_{\max} = 64$. In Fig. 4-14 we compare the SNR improvement factor performance profiles of the replacement and truncation FIR filter structures using rectangularly-windowed ideal filters again with $N_{\max} = 64$. Finally in Fig. 4-15 we compare the SNR improvement factor performance profiles of the replacement and truncation FIR filter structures using eigenfilters with $N_{\max} = 64$. As expected, in all three cases the unconstrained replacement filter structure outperforms the truncation filter structure in terms of having a higher SNR improvement

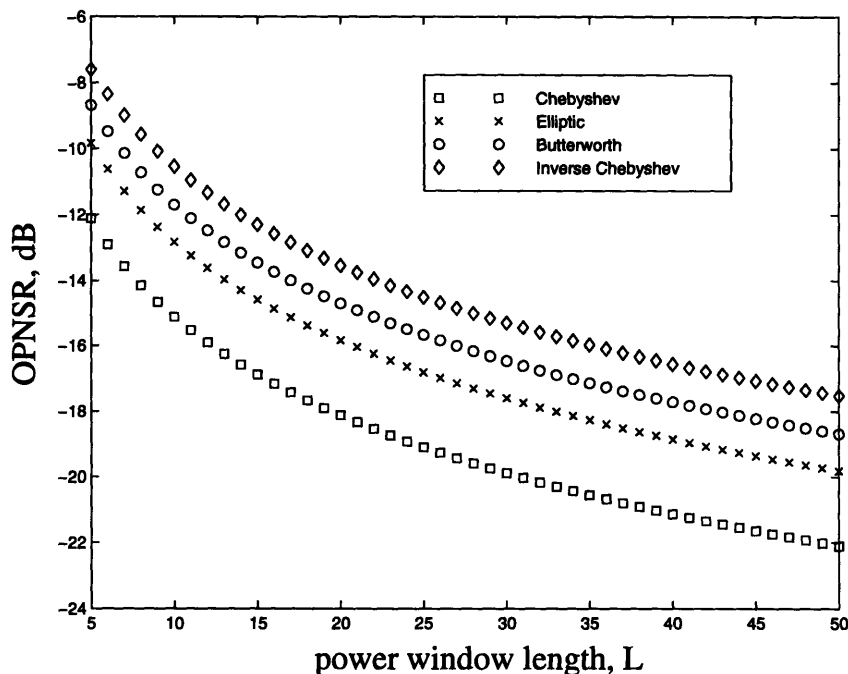


Figure 4-12: Comparison of the OPNSR for the Butterworth, Chebyshev, inverse Chebyshev, and elliptic IIR replacement filter structures.

factor for each filter order. We note that this is true at all but the lowest filter orders. The lowest order truncation filters may provide better SNR improvement than the corresponding replacement filters. Upon examination of the frequency response magnitudes of these low order truncated filters, however, we observe an especially poor passband characteristic which has not been taken into account in this analysis.

4.2.2 Type IT Filter Structures

The SNR improvement factor represents the factor by which the first set of N sections of a truncation filtering structure improves upon the input SNR. Fig. 4-16 shows the SNR improvement factor as a function of N for the case of truncations of a Butterworth filter with a half-power frequency of $\pi/2$ implemented as a cascade of ten second-order sections. The stopband in this case was defined to be $\omega \in [5\pi/8, \pi]$.

Truncation filter structures with IIR filter elements are called Type IT filter structures. To specify a Type IT filter structure, we begin by selecting the IIR filter of maximum order N_{\max} . The poles and zeros for the lower order filters are then defined, and the problem is

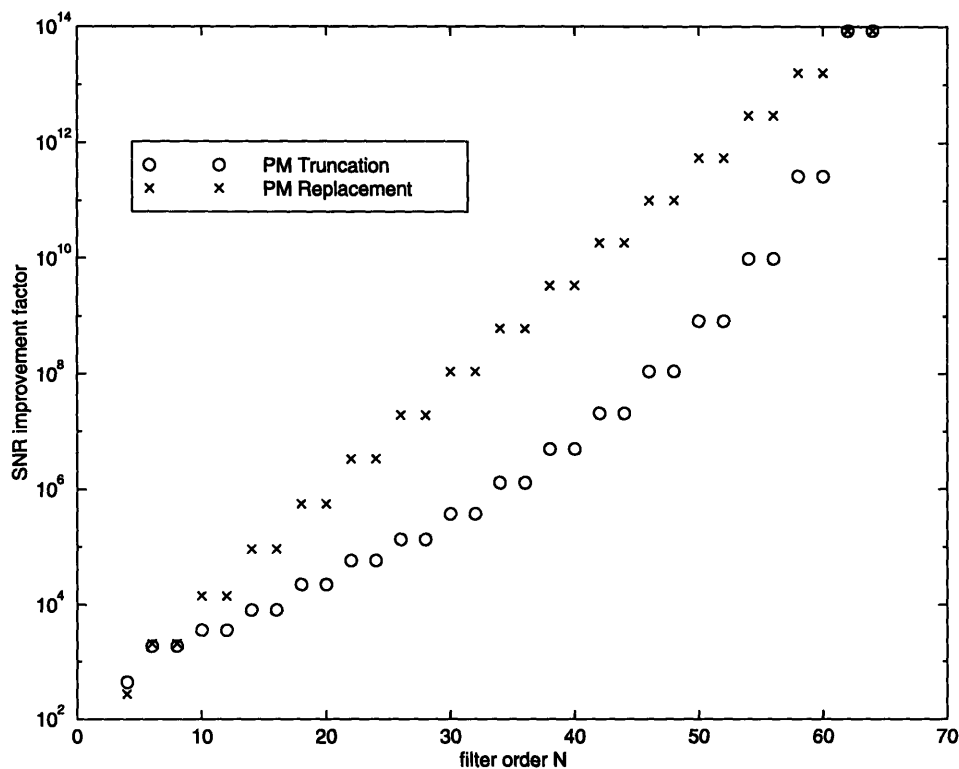


Figure 4-13: A comparison of the performance profiles for the Parks-McLellan FIR truncation and replacement filter structures. The truncated filters of orders $3 \leq N \leq 63$ were obtained by symmetrically truncating the coefficients of the order-64 Parks-McLellan filter.

to choose the best ordered sequence of poles and zeros to prune away from the given IIR filter with order N_{\max} in order to obtain each of the lower order filters. The combination of the filter of order N_{\max} with the order pole/zero pruning sequence defines a Type IT filter structure \mathcal{H}_T . To give some insight into the nature of Type IT filter structure specification, we present the following example.

Let us consider the case of a Butterworth filter of order $2M_0$. A cascade structure for this filter consists of a serial connection of M_0 second-order Direct-Form II sections, as was previously shown in Fig. 4-4. Each section corresponds to a pair of conjugate poles of the Butterworth filter and two zeros (both located at $z = -1$). Denoting the frequency response of the order- $2M_0$ Butterworth filter by $H_{M_0}(\omega)$, we may write

$$H_{M_0}(\omega) = G_1(\omega)G_2(\omega)G_3(\omega) \cdots G_{M_0}(\omega) \quad (4.30)$$

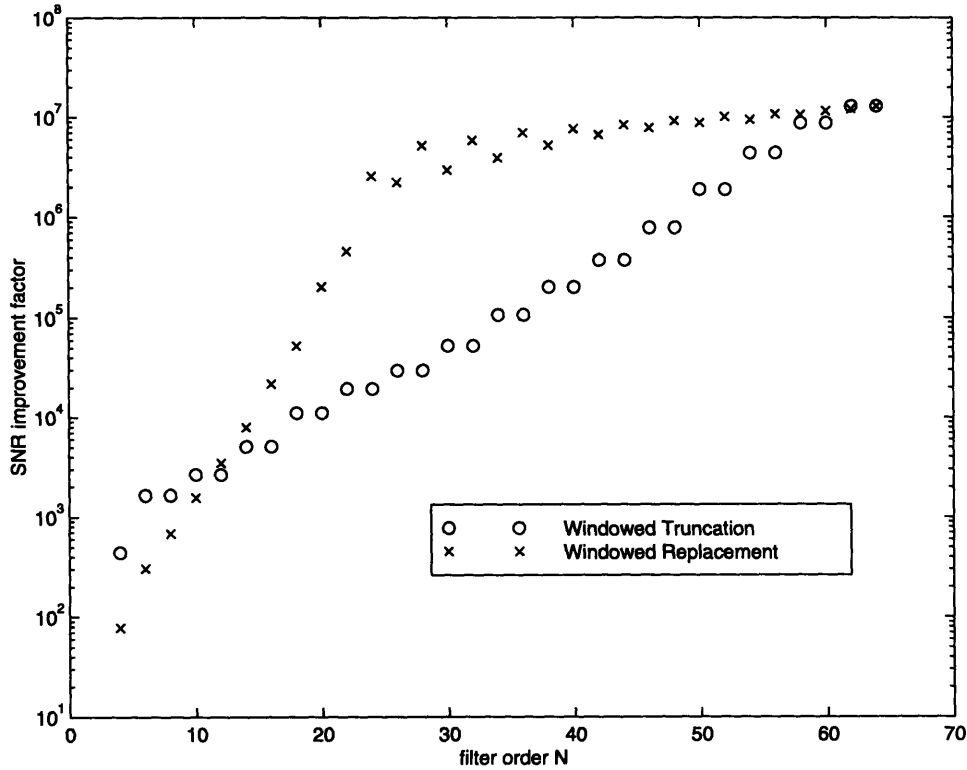


Figure 4-14: A comparison of the performance profiles for the FIR truncation and replacement filter structures. The truncated filters of orders $3 \leq N \leq 63$ were obtained by symmetrically truncating the coefficients of the order-64 rectangularly-windowed ideal filter.

where $G_i(\omega)$ denotes the frequency response of the i th second-order section in the cascade structure of Fig. 4-4. It can be furthermore assured that $G_i(0) = 1$. If only the first N sections ($N \leq M_0$) of the cascade structure in Fig. 4-4 are used, the resulting order- $2N$ *truncated* Butterworth filter has the frequency response $H_N(\omega)$, given by

$$H_N(\omega) = \prod_{k=1}^N G_k(\omega). \quad (4.31)$$

We are free to assign the Butterworth pole pairs to each of the second-order sections $G_k(\omega)$. It is desirable to make this assignment assure that as the number of second-order sections is increased, the average attenuation in the stopband of the filter also increases, while keeping the passband gain of each of the filters in \mathcal{H}_T close to unity. One strategy for

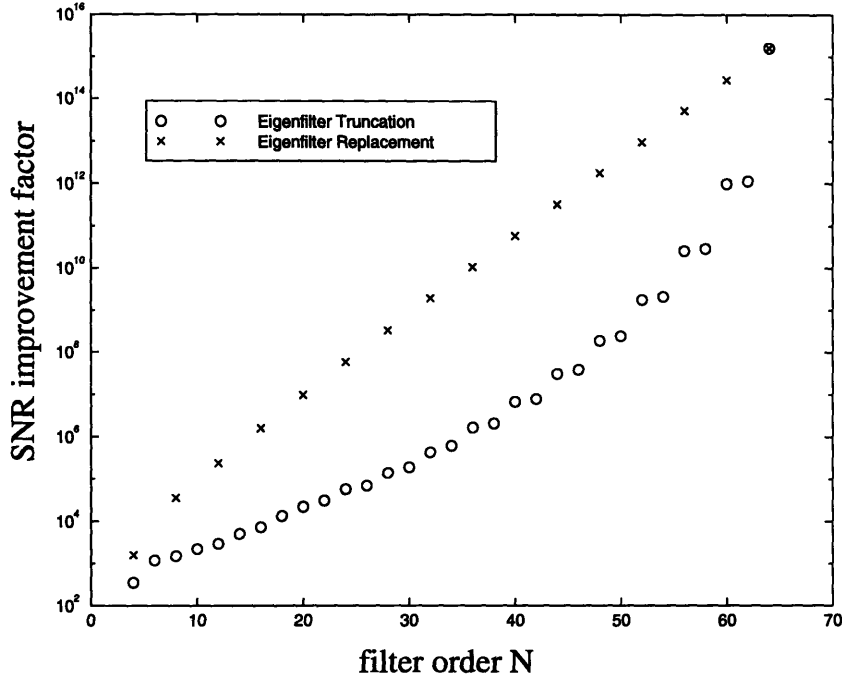


Figure 4-15: A comparison of the performance profiles for the eigenfilter FIR truncation and replacement filter structures. The truncated filters of orders $3 \leq N \leq 63$ were obtained by symmetrically truncating the coefficients of the order-64 eigenfilter.

making such a pole-pair assignment is as follows: the ordered set of second-order sections $G_1(\omega)G_2(\omega)G_3(\omega)\cdots G_{M_0}(\omega)$ is chosen from the $M_0!$ possible ordered sets to be the one which *minimizes* the objective function

$$J_T(G_1(\omega)\cdots G_{M_0}(\omega)) = \max_{1 \leq k \leq M_0} \left| |H_k(\omega)|^2 - 1 \right| \quad \omega \in PB, \quad (4.32)$$

where PB denotes the spectral region of support of the passband of $H_k(\omega)$. In other words, given the order- $2M_0$ filter $H_{M_0}(\omega)$, the problem is to determine the sequence of filters $G_1(\omega)G_2(\omega)\cdots G_{M_0}(\omega)$ such that $J_T(G_1(\omega)\cdots G_{M_0}(\omega))$ is minimized. The second-order sections $G_k(\omega)$ define the pole/zero truncation (pruning) sequence since $G_1(\omega)$ is truncated first to obtain the $(M_0 - 1)$ -section filter, $G_2(\omega)$ is truncated second to obtain the $(M_0 - 2)$ -section filter, and so on. Thus each ordered truncation sequence $G_1(\omega)G_2(\omega)\cdots G_{M_0}(\omega)$ defines a corresponding truncation filter structure

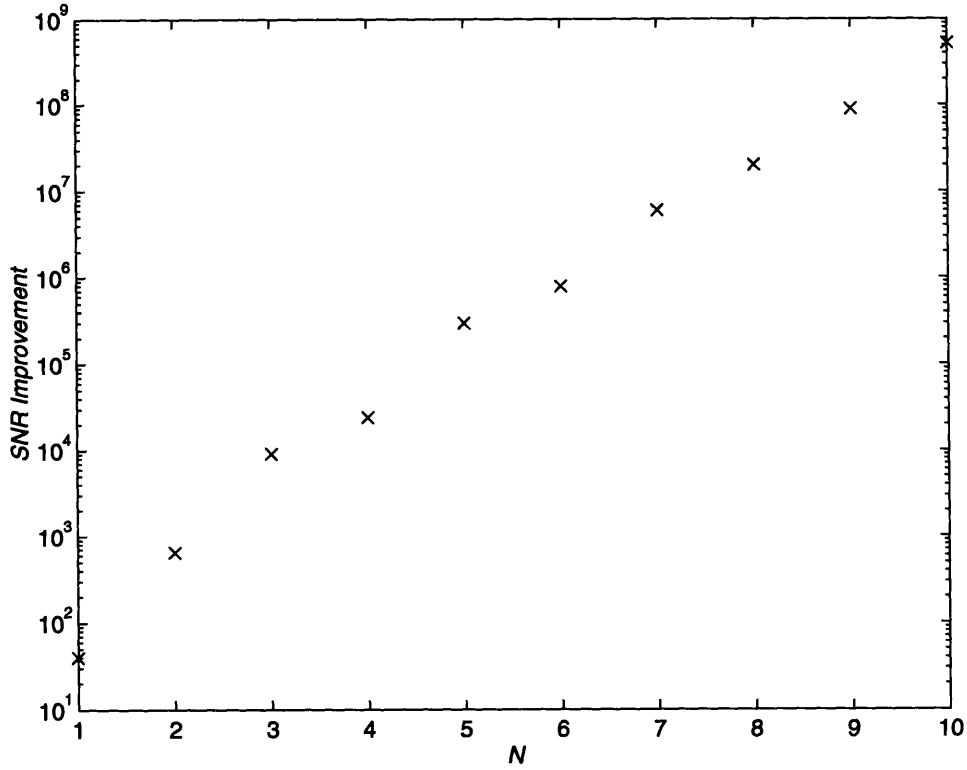


Figure 4-16: Performance profile for truncations of a 20th-order Butterworth filter with half-power frequency $\pi/2$.

$$\mathcal{H}_T = \{H_N(\omega), H_{N-1}(\omega), \dots, H_1(\omega)\}, \quad (4.33)$$

which can be used in the approximate filtering algorithm. We define the globally optimal truncation filter structure \mathcal{H}_T^* to be the particular truncation filter structure which minimizes $J_T(G_1(\omega) \cdots G_{M_0}(\omega))$. That is,

$$\mathcal{H}_T^* = \arg \min_{1 \leq k \leq (M_0!)^2} J_T(\mathcal{H}_T^k). \quad (4.34)$$

As indicated in Eq. (4.34), in order to find \mathcal{H}_T^* we must exhaustively search over $(M_0!)^2$ distinct filter structures \mathcal{H}_T^k and evaluate $J_T(\mathcal{H}_T^k)$ for each one. As defined earlier, \mathcal{H}_T^* is the truncation filter structure which results in the minimum value of $J_T(\mathcal{H}_T^k)$ over the range $1 \leq k \leq (M_0!)^2$. Since digital Butterworth filters have all their zeros at $z = -1$, the ordering

of the zeros in this truncation filter structure does not matter. All second-order section pole pairs are accompanied by two zeros at $z = -1$. For general IIR filters in which the zero locations are not all the same, the optimization must be done over all possible pole/zero pair combinations, resulting in $(M_0!)^2$ distinct filter structures to search over. There are $(M_0!)^2$ possibilities since for each of the $M_0!$ distinct pole pair orderings there exist $M_0!$ distinct possible zero pair orderings. In the case of the Butterworth filter structure, since the ordering of the zero pairs does not matter (all of the zeros are at $z = -1$), only $M_0!$ distinct filter structures exist.

To illustrate, consider the application of this strategy to a Butterworth filter with $M_0 = 3$ and a half-power frequency of $\pi/2$. As explained earlier, there are $M_0! = 3 \cdot 2 \cdot 1 = 6$ distinct filter structures $\mathcal{H}_T^1 \cdots \mathcal{H}_T^6$ to consider in determining the optimal truncation filter structure, \mathcal{H}_T^* , for which $J_T(\mathcal{H}_T^*)$ is minimum. For this case we have

$$H_N(\omega) = \prod_{k=1}^3 G_k(\omega) \quad (4.35)$$

with

$$\mathcal{H}_T^1 = \{H_6^1(\omega), H_4^1(\omega), H_2^1(\omega)\} \quad (4.36)$$

$$= \{G_1(\omega)G_2(\omega)G_3(\omega), G_1(\omega)G_2(\omega), G_1(\omega)\}, \quad (4.37)$$

$$\mathcal{H}_T^2 = \{H_6^2(\omega), H_4^2(\omega), H_2^2(\omega)\} \quad (4.38)$$

$$= \{G_1(\omega)G_3(\omega)G_2(\omega), G_1(\omega)G_3(\omega), G_1(\omega)\}, \quad (4.39)$$

$$\mathcal{H}_T^3 = \{H_6^3(\omega), H_4^3(\omega), H_2^3(\omega)\} \quad (4.40)$$

$$= \{G_2(\omega)G_1(\omega)G_3(\omega), G_2(\omega)G_1(\omega), G_2(\omega)\}, \quad (4.41)$$

$$\mathcal{H}_T^4 = \{H_6^4(\omega), H_4^4(\omega), H_2^4(\omega)\} \quad (4.42)$$

$$= \{G_2(\omega)G_3(\omega)G_1(\omega), G_2(\omega)G_3(\omega), G_2(\omega)\}, \quad (4.43)$$

$$\mathcal{H}_T^5 = \{H_6^5(\omega), H_4^5(\omega), H_2^5(\omega)\} \quad (4.44)$$

$$= \{G_3(\omega)G_1(\omega)G_2(\omega), G_3(\omega)G_1(\omega), G_3(\omega)\}, \quad (4.45)$$

and

$$\mathcal{H}_T^6 = \{H_6^6(\omega), H_4^6(\omega), H_2^6(\omega)\} \quad (4.46)$$

$$= \{G_3(\omega)G_2(\omega)G_1(\omega), G_3(\omega)G_2(\omega), G_3(\omega)\}, \quad (4.47)$$

Overlays of the frequency responses of the three filters in each of the above truncation filter structures \mathcal{H}_T^k above are shown in Fig. 4-17. It should be observed that as the number of sections (N) is increased, the average attenuation the stopband also increases. On the other hand, as can be seen in Fig. 4-17, the filter gain remains close to unity in most of the passband.

Table 4.2: Numerical values for $J(\mathcal{H}_T^1) \cdots J_T(\mathcal{H}_T^6)$ for the Butterworth optimal truncation filter structure. Note that $J_T^*(\mathcal{H}_T^*) = J_T(\mathcal{H}_T^2)$.

Filter Structure	Value of $J_T(\mathcal{H}_T^k)$
\mathcal{H}_T^1	0.5767
$\mathcal{H}_T^2 = \mathcal{H}_T^*$	0.4926
\mathcal{H}_T^3	0.5767
\mathcal{H}_T^4	0.9553
\mathcal{H}_T^5	1.3438
\mathcal{H}_T^6	1.3438

After evaluating $J_T(\mathcal{H}_T^k)$ over the range $1 \leq k \leq 6$, we determined empirically that $J_T(\mathcal{H}_T^2)$ is the minimum, and thus for this example $\mathcal{H}_T^* = \mathcal{H}_T^2$. All six of the values of $J_T(\mathcal{H}_T^k)$ for $1 \leq k \leq 6$, are tabulated in Table 2 for reference.

4.3 Summary and Future Directions

In this chapter we have defined four types of approximate filter structures and investigated their relative performance. The replacement structures were shown to perform the best in terms of maximizing the SNR improvement factor, while truncation filter structures were

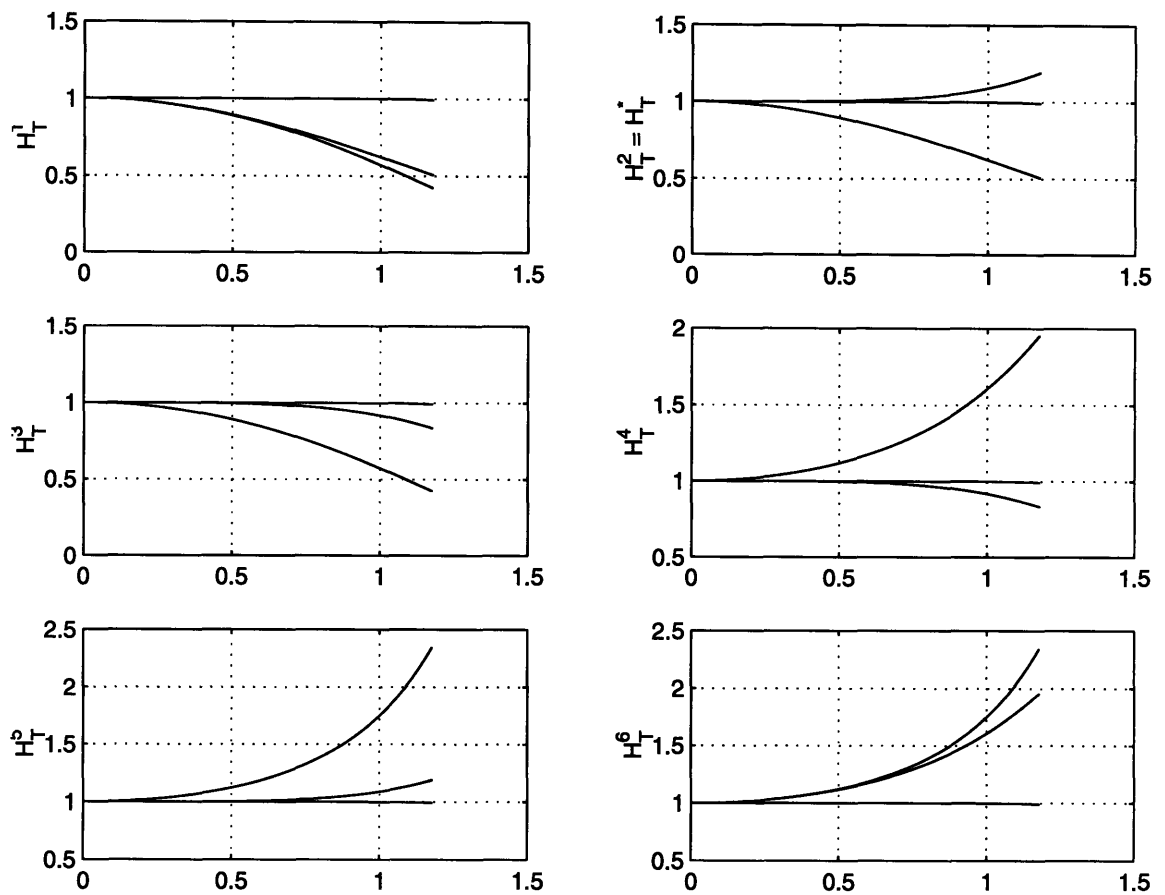


Figure 4-17: Magnitude-squared frequency responses for truncations of a 6th-order Butterworth filter with 1, 2, and 3 second-order sections, for each of the possible distinct truncation filter structures defined in Eqs. (4.35)–(4.47). The optimal truncation filter structure is $\mathcal{H}_T^* = \mathcal{H}_T^2$.

shown to have the advantage of less storage requirements which is equivalent to less power consumption in CMOS devices. FIR eigenfilters and IIR elliptic filters were shown to be excellent choices for constituent filter elements in approximate filter structures.

A future direction of research would be to explore how an additional performance metric could be incorporated into the choice of approximate filter structures. This performance metric is formulated to represent the variance of our estimate $\hat{P}_y[N_0]$ of the output power $P_y[N_0]$. An expression for the variance of the estimate $\hat{P}_y[N_0]$ was discussed in Chapter 2. Minimization of this variance is important since it is desirable to have our estimate of the output power as close to the actual output power as possible with high probability. Perhaps a composite objective function could be formed which encapsulates the STE, SNR improvement factor, and the variance of the estimate $\hat{P}_y[N_0]$. Designing filter structures to optimize such an objective function would provide a promising method for enhancing the performance of approximate filtering algorithms.

Chapter 5

Experiments and Applications

In the first four chapters of this thesis we have developed a theoretical basis for approximate filtering algorithms and explored the interplay between approximate filtering algorithms, approximate filter structures, and the state transition error. In this chapter we present computer simulations which show that significant power savings may be achieved when the order of a digital filter is dynamically varied to provide time-varying stopband attenuation in proportion to the time-varying signal-to-noise ratio (SNR) of the input signal, while maintaining a fixed level of output quality. We highlight experiments involving speech signals to demonstrate the practical viability of approximate filtering algorithms. An order of magnitude or more reduction in power consumption over fixed-order filters is achieved in the context of demultiplexing frequency-division multiplexed (FDM) speech signals. We also survey an actual DSP chip implementation of an approximate filtering algorithm which was developed at Stanford University. The results of this chip implementation solidify the effectiveness of approximate filtering algorithms in the context of interpolation and decimation for use in low power analog-to-digital and digital-to-analog data converters.

5.1 Speech Signal Processing

In this section we illustrate the potential of approximate filtering algorithms to reduce power consumption in speech processing. We use a Butterworth truncation filter structure with 10 second-order sections. This approximate filter structure and the adaptation control strategy described in Section 2.5 was applied to two speech signals which had been frequency-division multiplexed. The power window length was chosen to be $L = 100$ and the minimum

tolerable output SNR was set to $OSNR_{tol} = 1000$. The IIR filters in the Butterworth truncation filter structure each had a half-power frequency of $\pi/2$. The stopband was defined to be between $5\pi/8$ and π , while the passband was defined to be between 0 and $3\pi/8$. One speech signal was spectrally centered in the passband region of the lowpass filter and the other was modulated into the stopband region of the lowpass filter. The sampling rate for each of the speech signal was 16000 Hz. Fig. 5-1 shows the speech signal in the passband, the speech signal in the stopband, and the evolution of the number of filter sections used by the approximate filtering technique. Examination of the figure shows that as would be expected, the number of filter sections is large when the input SNR is small. Furthermore, the number of filter sections is small when the input SNR is high.

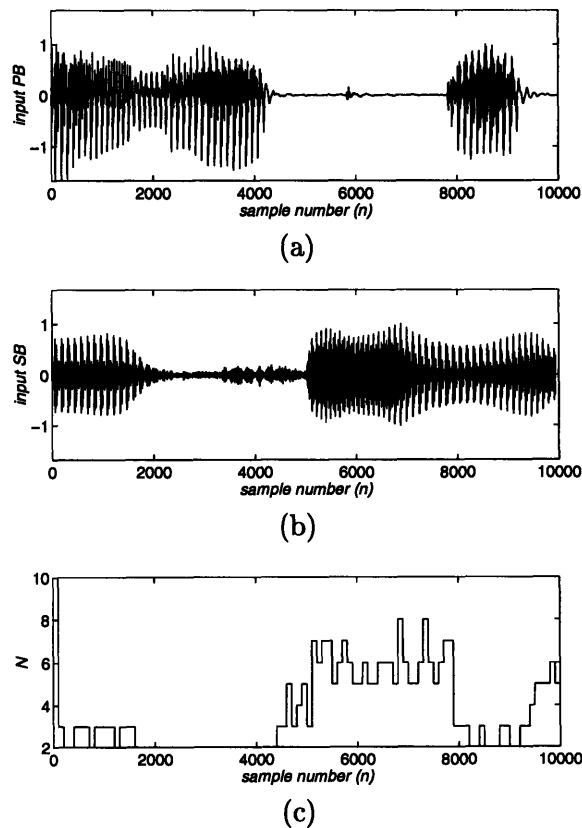


Figure 5-1: Demultiplexing of FDM speech using low power frequency selective filtering. (a) Passband speech, (b) stopband speech, and (c) number of filter sections as a function of sample number.

Table 5.1: Summary of the approximate filtering performance for demodulating FDM speech.

<i>Sentence Number</i>	<i>Minimum Order</i>	<i>Maximum Order</i>	<i>Power Consumption ($P_{\text{fixed}}/P_{\text{adaptive}}$)</i>
1	3	99	4.4
2	3	117	6.3
3	3	87	5.6
4	3	93	5.9
5	3	61	4.2
6	3	119	5.9
7	3	125	7.1
8	3	111	7.1
9	3	127	7.1
10	3	107	8.3
<i>Average</i>	3	104.6	5.9

Similarly for FIR approximate filters, we have used simulations of approximate filtering algorithms to show that reduction in power consumption by an order of magnitude is achieved over fixed-order filter implementations. The context for these simulations is frequency-division demultiplexing of pairs of speech waveforms. We now describe a speech processing experiment. Each of the speech signals used in our simulations was sampled at 8000 Hz and normalized to have maximum amplitude of unity. Each speech signal corresponds to a complete sentence with negligible silence at its beginning and end. To begin, each digitized speech waveform was pre-filtered to have a maximum frequency of 1500 Hz. A guard band of 1000 Hz was used in multiplexing a reference speech signal (corresponding to the sentence, “that shirt seems much too long,”) with each of the other speech signals. The reference signal always occupied the 0 to 1500 Hz band, while the other signals always occupied the 2500 Hz to 4000 Hz band.

Demultiplexing involves lowpass filtering (cutoff frequency 2 KHz) to isolate the reference speech signal. An approximate filtering algorithm was used to perform this lowpass filtering for each of the 10 FDM signals. The parameter values were chosen to be $\text{OSNR}_{\text{tol}} = 100$ and $L = 100$. An FIR replacement filter structure with rectangularly-windowed ideal filter constituent elements was used.

In Table 5.1 we have listed various measures obtained for the performance of the approx-

imate filter as it was applied to each FDM signal. The first column contains the sentence number for the stopband component of the input signal. The second and third columns respectively list the minimum and maximum filter orders used by the approximate filter in each case. The final column shows the relative power consumption of the approximate filter with respect to a fixed-order filter which is guaranteed to satisfy the same output quality constraint as the approximate filter. We observe that approximate filtering reduces the average power consumption by a factor of 5.9.

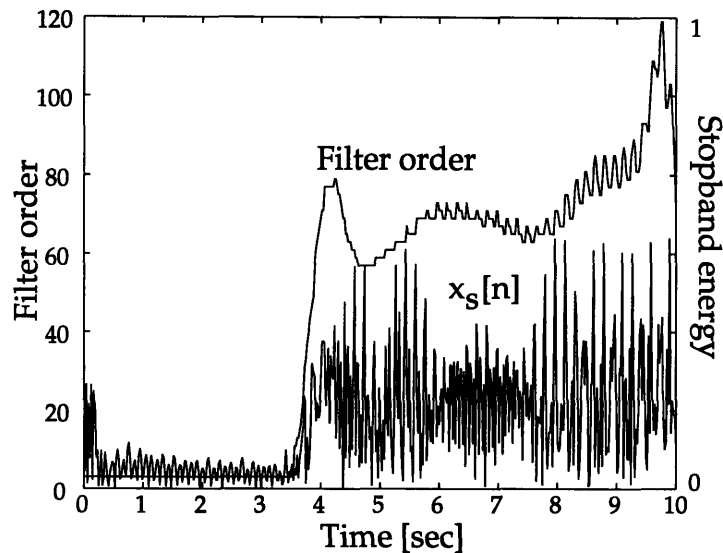


Figure 5-2: Evolution of filter order for the FDM speech signal processing example. Overlays of the approximate filter order and the stopband power in the input signal over time are shown. The approximate filter order clearly traces the envelope of the stopband power in the input signal.

To gain further insight into the source for this power reduction, in Fig. 5-2 we illustrate the nature of the adaptation performed by our technique in the case of one of the FDM signals. One of the curves shows the evolution of the filter order while the other curve shows the energy profile of the stopband power in the input signal. Note that since the passband speech has approximately constant power, the approximate filter order approximately traces the envelope of the stopband power in the input signal. The most power savings is achieved during the silence regions of the stopband signal.

Longer periods of speech communication generally include significantly larger fractions of silence periods than an individual sentence. To factor this into our analysis, we repeated our simulations while inserting additional silence at the end of each speech signal. The

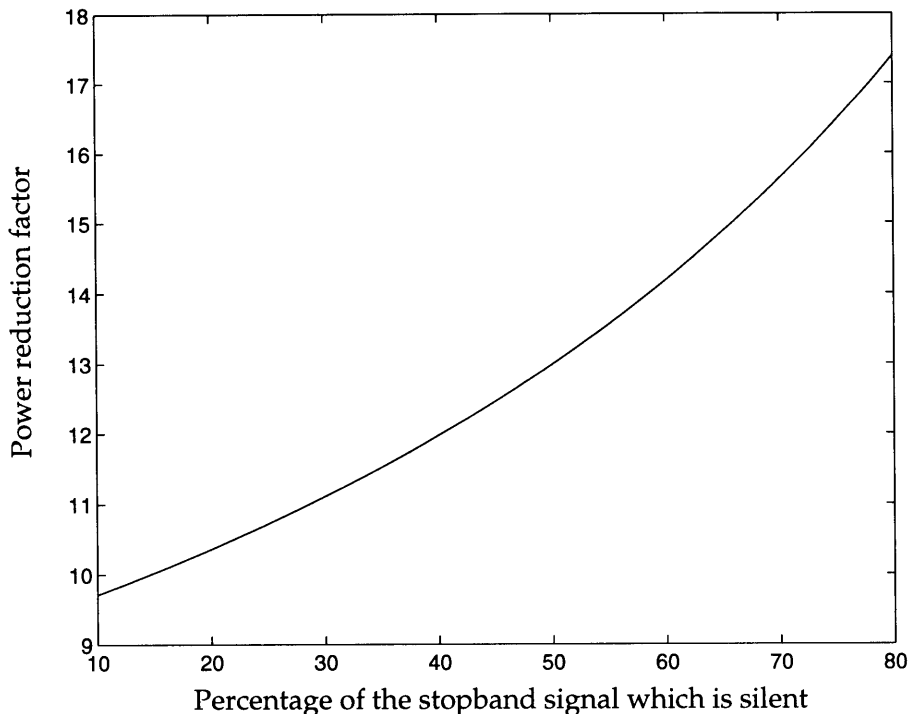


Figure 5-3: Filter performance vs. percentage silence in stopband signal.

average over all 10 cases of the relative power consumption is displayed in Fig. 5-3 as a function of the silence duration relative to the duration of the entire signal. As expected, the power reduction improves as the relative amount of silence in the speech signals is increased.

5.2 Interpolation and Decimation

A research group at Stanford University has designed and implemented an approximate filtering algorithm for the application of low power interpolation and decimation [46]. This application is important for analog-to-digital and digital-to-analog data converters. Included in the experimental circuit are a pair of decimation filters and two pairs of interpolation filters in a single-chip multimedia audio system. Excellent results have been obtained and were reported at The International Solid-State Circuits Conference in February, 1997. A conventional interpolation/decimation system was implemented and shown to consume an average of approximately 86.4 mW with a 5V power supply. Using an approximate filtering algorithm compiled onto a programmable processor, the power consumption was reduced

by 36% for the decimation system and 17% for the interpolation system.

5.3 Future Directions

Beyond what was presented in this chapter, approximate filtering algorithms provide potential for the incorporation of more general dynamic cost vs. quality tradeoffs for a host of applications in which power savings is needed. For example, one such application is spatial filtering or beamforming. In adaptive beamforming the number of sensors used in a linear array could be dynamically varied based on the direction-of-arrival or spatial dispersion characteristics of the received input signal. Varying the number of sensor array elements which are used is conceptually equivalent to varying the order of the underlying FIR filter which is accomplishing the spatial filtering. Thus, what results is the approximate spatial filtering dual to the approximate frequency-selective filtering we have discussed thus far. Significant power savings could be achieved when the order of a spatial filter or beamformer is dynamically varied to provide time-varying spatial stopband attenuation in proportion to the time-varying energy in the spatially-defined SNR the input signal, while maintaining a fixed level of output quality.

Another natural extension of approximate filtering algorithms which provides a future direction of potentially fruitful research is the incorporation of approximate filters into a binary-tree structured filter bank, as depicted in Fig. 5-4. The binary split of a the signal into its highpass and lowpass components is iterated using the same filters. Such filter banks have been used, for example, to compute the discrete wavelet transform, and may be also be modified to accommodate more general subband decompositions.

By replacing each one of the fixed-length FIR filters in Fig. 5-4 with an approximate filter, an *approximate filter bank* is formed. The approximate filter bank is shown at the top of Fig. 5-5. This provides a low-power implementation for the front end of source coding algorithms. The output quality of the approximate filter bank may be measured in a mean-square error sense compared to the outputs generated from a conventional filter bank with fixed-order filters. The power consumption of the approximate filter bank may be computed relative to the required power consumption of a conventional filter bank which is designed to maintain the same level of output quality as the approximate filter bank.

As depicted at the top of Fig. 5-5, each of the highpass and lowpass filters in a filter bank may be implemented using an approximate filtering algorithm. To illustrate the potential

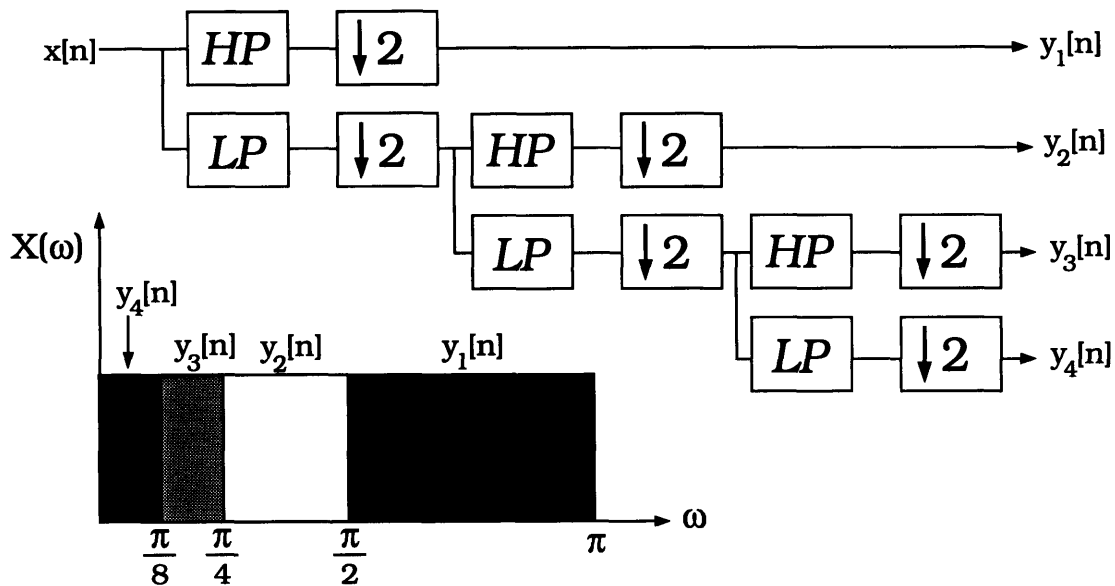


Figure 5-4: A conventional fixed-order FIR filter bank and associated spectral decomposition of the filter bank output signals.

for power savings in the first stage of the subband decomposition, an approximate FIR lowpass filter was applied to a speech signal, $x[n]$, corresponding to the sentence, “that shirt seems much too long.” The time-varying FIR filter order used by our technique is shown in the top plot of Fig. 5-5. The bottom plot in Fig. 5-5 shows the input’s stopband component, $x_s[n]$, to demonstrate that the filter order roughly tracks the stopband energy of the input signal.

In summary, this chapter has provided examples of how approximate filter algorithms may be used to adaptively reduce power consumption in practical applications. In addition, we discussed some domains in which the concepts of approximate filtering may be extended to provide interesting and potentially fruitful avenues for future research.

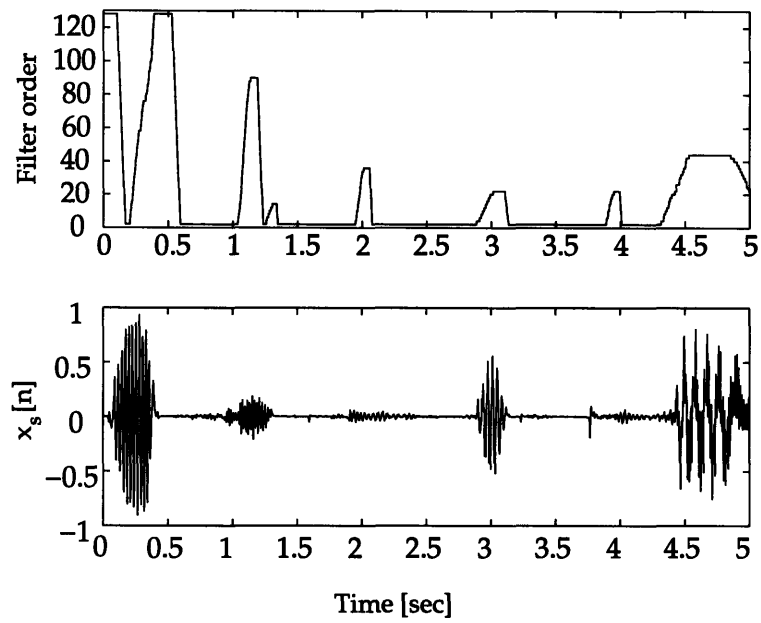
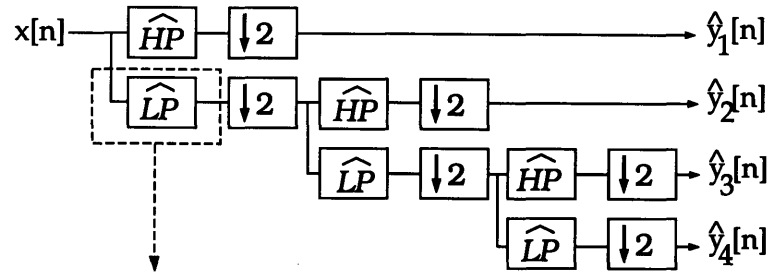


Figure 5-5: An FIR filter bank which has incorporated approximate filters, resulting in an *approximate filter bank*. The filter order evolution for the first lowpass filter has been enclosed in the dashed box. The filter order can be seen to follow the energy in the input's stopband component $x_s[n]$, which is shown in the bottom plot.

Chapter 6

Conclusion

The signal processing demands of portable multimedia devices have increased dramatically in the last decade as products continue to shrink in size and require increasing computational speed. Due to the nature of portability, these increased processing demands are accompanied by definite constraints on power consumption. The important task of designing low power, computationally powerful processors has emerged and spurred great interest and activity in signal processing research. In this thesis we have directly addressed this task and formulated a new algorithmic approach to low power frequency-selective digital filtering. We have demonstrated that significant power savings may be achieved in digital filtering applications when the order of a digital filter is dynamically varied to provide time-varying stopband attenuation in proportion to the time-varying signal-to-noise ratio (SNR) of the input signal, while maintaining a fixed SNR at the filter output.

By considering the practical problem of conserving power via dynamically reducing the order of a frequency-selective digital filter, we abstracted the theoretical problem of determining an optimal filter order based on observations of the input data and a set of concrete statistical assumptions. Two approaches to solving this problem helped us interpret and understand the practical low power filtering problem and improve the performance of approximate filtering algorithms. One solution was guided by a low power approach and achieved suboptimal performance with an extremely low computational cost. A second solution was guided by a maximum likelihood objective and provided superior performance while requiring much more computation. While computationally impractical, the maximum likelihood approach provided valuable insight as well as a performance benchmark for comparison with the low power solution. To study these problems we developed a

theory for approximate filtering based on the concepts of approximate signal processing. We constructed a framework to explore the statistical properties of approximate filtering algorithms, and showed that under certain assumptions the performance of approximate filtering algorithms is asymptotically optimal.

We considered the transient effects of dynamically changing the filter order in approximate filtering. For this purpose, the output of an approximate filter was related to the output of a fixed digital filter by introducing the concept of *state transition error*. We statistically analyzed the corruptive effects of the state transition error on approximate filtering algorithms, and demonstrated analytically and empirically that the state transition error is essentially negligible.

We developed a framework for analyzing approximate filter structures. In so doing we demonstrated that approximate filter structures represent a critical element in the characterization of approximate filtering algorithms. Two classes of approximate filter structures, *truncation* and *replacement* filter structures, were studied extensively. We found that truncation filter structures may be described with fewer independent filter coefficients than replacement filter structures. Associated with this property we found that approximate filtering using a truncation filter structure requires less memory, chip area, and bus accesses than approximate filtering using a replacement filter structure. This property of truncation filter structures makes them especially attractive for low power applications. We also showed that replacement filter structures lead to approximate filtering algorithms with performance superior to that of approximate filtering algorithms using truncation filter structures. Thus, the decision to use a truncation or replacement filter structure depends on the application as well as the associated power constraints and performance specifications.

Finally, computer simulation experiments involving speech signals were used to demonstrate the practical viability of approximate filtering for low power signal processing. We demonstrated that an order of magnitude reduction in power consumption over fixed-order filters is possible using approximate filtering algorithms.

Thus, the main contribution of this thesis is the development of a framework for the design and implementation of approximate filters using signal-dependent algorithms which meet fixed performance specifications while dynamically minimizing power consumption.

Appendix A

All-Pole Filter Matrix

In this appendix we derive an expression for \mathbf{Q}_{N_2} , the $L \times N_2$ matrix whose elements are defined in terms of the polynomial coefficients a_{kN_2} . The roots of the polynomial coefficients a_{kN_2} are the pole locations of the filter $h_{N_2}[n]$. We assume the reader is familiar with the notation presented in Chapter 3 of the thesis.

We begin by writing out the relevant set of equations, beginning with

$$y_{\text{tr}}[0] = \mathbf{a}_0^T \mathbf{y}_0 \tag{A.1}$$

$$= \mathbf{A}_0 \mathbf{y}_0, \tag{A.2}$$

where $\mathbf{a}_0^T = [-a_{1N_2} - a_{2N_2} \cdots -a_{N_2N_2}]$ and $\mathbf{A}_0 = \mathbf{a}_0^T$. For the sake of notational simplicity, we define \mathbf{a}_k^T for $1 \leq k \leq (N_2 - 1)$ to be equal to \mathbf{a}_0^T with k zeros inserted in the beginning of the vector. The k zeros are inserted such that the rightmost k elements of \mathbf{a}_0^T are pushed out, such that the length of the vector \mathbf{a}_k^T is L for all k . For example,

$$\mathbf{a}_3^T = [0 \ 0 \ 0 \ -a_{1N_2} - a_{2N_2} \cdots -a_{(N_2-3)N_2}], \tag{A.3}$$

and so on. Now we proceed to write out

$$y_{\text{tr}}[1] = (\mathbf{a}_0^T a_{1N_2} + \mathbf{a}_1^T) \mathbf{y}_0 \quad (\text{A.4})$$

$$= \mathbf{A}_1 \mathbf{y}_0, \quad (\text{A.5})$$

and

$$y_{\text{tr}}[2] = [(\mathbf{a}_0^T a_{1N_2} + \mathbf{a}_1^T) a_{1N_2} + a_{2N_2} \mathbf{a}_0^T + \mathbf{a}_2^T] \mathbf{y}_0 \quad (\text{A.6})$$

$$= \mathbf{A}_2 \mathbf{y}_0. \quad (\text{A.7})$$

This leads to

$$\mathbf{y}_{\text{tr}} = \mathbf{Q}_{N_2} \mathbf{y}_0, \quad (\text{A.8})$$

where $\mathbf{Q}_{N_2}^k$ is the k th row of the matrix \mathbf{Q}_{N_2} defined recursively as

$$\mathbf{Q}_{N_2}^k = \left(\sum_{l=0}^{k-1} \mathbf{Q}_{N_2}^l \cdot a_{k-l, N_2} \right) + \mathbf{a}_k^T \quad 1 \leq k \leq (N_2 - 1), \quad (\text{A.9})$$

$$\mathbf{Q}_{N_2}^k = \left(\sum_{l=k-N_2}^{k-1} \mathbf{Q}_{N_2}^l \cdot a_{k-l, N_2} \right) + \mathbf{a}_k^T \quad N_2 \leq k \leq L, \quad (\text{A.10})$$

with initial row $\mathbf{Q}_{N_2}^0 = \mathbf{a}^T$. Thus, we have defined the matrix \mathbf{Q}_{N_2} by defining each of its L rows.

Appendix B

Bound on Vector Norm

In this appendix our goal is to prove that for an $N_2 \times 1$ vector $\mathbf{c} = [c_1 \ c_2 \ \cdots \ c_{N_2}]$ the following inequality is true

$$\sum_{k=1}^{N_2} |c_k| \leq N_2 \|\mathbf{c}\|, \quad (\text{B.1})$$

where

$$\|\mathbf{c}\| = \sqrt{\sum_{k=1}^{N_2} |c_k|^2} \quad (\text{B.2})$$

is the standard 2-norm of the vector \mathbf{c} . The inequality in Eq. (B.1) may be rewritten as

$$\frac{1}{N_2} \sum_{k=1}^{N_2} |c_k| \leq \|\mathbf{c}\|. \quad (\text{B.3})$$

Expanding the left-hand side of Eq. (B.3), we obtain

$$\frac{1}{N_2} \sum_{k=1}^{N_2} |c_k| = \frac{1}{N_2} (|c_1| + |c_2| + \cdots + |c_{N_2}|) \quad (\text{B.4})$$

$$\leq \frac{1}{N_2} \left(N_2 \max_{1 \leq k \leq N_2} c_k \right) \quad (\text{B.5})$$

$$\leq \max_{1 \leq k \leq N_2} c_k. \quad (\text{B.6})$$

Defining

$$c_{\max} = \max_{1 \leq k \leq N_2} c_k, \quad (\text{B.7})$$

we arrive at

$$\frac{1}{N_2} \sum_{k=1}^{N_2} |c_k| \leq c_{\max}. \quad (\text{B.8})$$

Clearly $c_{\max} < \|\mathbf{c}\|$, except for the special case in which $c_k = 0$ for all k except for one value of $k = k_0$. In this case $c_{\max} = \|\mathbf{c}\| = c_{k_0}$. Therefore

$$\frac{1}{N_2} \sum_{k=1}^{N_2} |c_k| \leq c_{\max} \quad (\text{B.9})$$

$$\leq \|\mathbf{c}\|, \quad (\text{B.10})$$

and the proof of Eq. (B.1) is complete.

Appendix C

Autoregressive Parameter Values

In the numerical example of Chapter 2 we use an autoregressive random process with order 30, whose power spectral density was plotted in Fig. 2-6. The system function of this power spectral density is

$$H(z) = \frac{G}{D(z)} \quad (\text{C.1})$$

The numerator gain was chosen to be $G = 65.3497$. The denominator polynomial of the autoregressive power spectral density used in the numerical example is given by

$$\begin{aligned} z^{-30} + 0.4433z^{-29} + -0.8229z^{-28} + -0.9432z^{-27} + 0.1484z^{-26} + 0.4942z^{-25} + \\ 0.0588z^{-24} + 0.1376z^{-23} + 0.3110z^{-22} + -0.0910z^{-21} + -0.2907z^{-20} + \\ -0.0341z^{-19} + -0.0397z^{-18} + -0.1610z^{-17} + 0.0351z^{-16} + 0.1436z^{-15} + \\ 0.0152z^{-14} + 0.0456z^{-13} + 0.1494z^{-12} + -0.0092z^{-11} + -0.1112z^{-10} + -0.0282z^{-9} + \\ -0.0638z^{-8} + -0.0951z^{-7} + 0.0375z^{-6} + 0.0866z^{-5} + 0.0125z^{-4} + \\ 0.0472z^{-3} + 0.0760z^{-2} + -0.0265z^{-1} + -0.0722 \quad (\text{C.2}) \end{aligned}$$

The pole locations are the roots of this polynomial, and have been tabulated in Table C.1.

Table C.1: The pole locations of the 30th-order autoregressive random process used in the numerical example of Chapter 2.

<i>Pole Number</i>	<i>Location</i>
1,2	$-0.4242 \pm 0.8733i$
3,4	$-0.5454 \pm 0.7673i$
5,6	$-0.6524 \pm 0.6478i$
7,8	$-0.7695 \pm 0.5061i$
9,10	$-0.8365 \pm 0.3485i$
11,12	$-0.8975 \pm 0.1812i$
13,14	$-0.1137 \pm 0.8891i$
15,16	$0.0833 \pm 0.8900i$
17,18	$0.2746 \pm 0.8488i$
19,20	$0.4330 \pm 0.7703i$
21,22	$0.6143 \pm 0.6689i$
23,24	$0.8548 \pm 0.4606i$
25,26	$0.8864 \pm 0.3059i$
27,28	$0.9034 \pm 0.1236i$
29	-0.9119
30	0.8473

Bibliography

- [1] J. G. Ackenhusen. *Signal Processing Technology and Applications*. Institute of Electrical and Electronics Engineers, Inc., 1995.
- [2] M. Alidina, J. Monteiro, S. Devadas, A. Ghosh, and M. Papaefthymiou. Precomputation-based sequential logic optimization for low power. In *Proceedings of the 1994 International Workshop on Low-power Design*, pages 57–62, April 1994.
- [3] K. Asai, K. Ramchandran, and M. Vetterli. Image representation using time-varying wavelet packets, spatial segmentation and quantization. In *Proceedings of Conference on Information Science and Systems*, March 1993.
- [4] T. E. Bell. The incredible shrinking computer. *IEEE Spectrum*, pages 37–41, May 1991.
- [5] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.
- [6] J. B. Burr. Cryogenic ultra low power CMOS. *1995 IEEE Symposium on Low Power Electronics*, pages 82–83, October 1995.
- [7] T. Callaway and E. Swartzlander. Optimizing arithmetic elements for signal processing. *VLSI Signal Processing, IEEE Special Publications*, pages 91–100, 1992.
- [8] A. P. Chandrakasan. A low power chipset for multimedia applications. In *IEEE International Solid-state Circuits Conference*, February 1994.
- [9] A. P. Chandrakasan and R. W. Brodersen. *Low Power Digital CMOS Design*. Kluwer Academic Publishers, Norwell, MA, 1995.
- [10] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen. Low-power digital CMOS design. *IEEE Journal of Solid-State Circuits*, pages 473–484, April 1992.

- [11] P. A. Chou, T. Lookabaugh, and R. M. Gray. Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Transactions on Information Theory*, IT-35:299–315, March 1986.
- [12] R. Coifman, Y. Meyer, S. Quake, and V. Wickerhauser. Signal processing and compression with wave packets. Technical report, Yale University, 1990.
- [13] R. Coifman and M. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38:713–718, March 1992.
- [14] A. Crosier, D. Esteban, and C. Galand. Perfect channel splitting by use of interpolation/decimation tree decomposition techniques. In *Proceedings of the International Symposium on Circuits and Systems*, 1976.
- [15] E. Dorkan. *Approximate Processing and Knowledge-Based Reprocessing of Non-Stationary Signals*. PhD thesis, Boston University, September 1993.
- [16] A. P. Chandrakasan et al. Design of portable systems. In *Proceedings of the Custom Integrated Circuits Conference*, Santa Clara, CA, May 1994.
- [17] D. Byrne et al. An international comparison of long-term average speech spectra. *Journal of the Acoustical Society of America*, 96(4):2108–2120, October 1994.
- [18] G. A. Ghazal. Moments of the ratio two dependent quadratic forms. *Statistics and Probability Letters*, 20(4):313–315, 1994.
- [19] B. M. Gordon and T. Meng. A low power subband video decoder architecture. In *Proceedings of the International Conference on Speech, Acoustics, and Signal Processing*, Sydney, Australia, April 1994.
- [20] S. Haykin. *Adaptive Filtering Theory*. Prentice Hall, Englewood Cliffs, NJ, 1994.
- [21] C. Herley, J. Kovacevic, K. Ramachandran, and M. Vetterli. Tilings of the time-frequency plane: Construction of arbitrary orthogonal bases and fast tiling algorithms. *IEEE Transactions on Signal Processing*, 41(12):3341–3359, December 1993.
- [22] C. Herley, J. Kovacevic, K. Ramachandran, and M. Vetterli. Time-varying orthonormal tilings of the time-frequency plane. In *Proceedings of the International Conference on*

- Speech, Acoustics, and Signal Processing*, volume III, pages 205–208, Minneapolis, MN, May 1993.
- [23] C. Herley and M. Vetterli. Orthogonal time-varying filter banks and wavelet packets. *IEEE Transactions on Signal Processing*, 42(10):2650–2663, October 1994.
- [24] N. S. Jayant and P. Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, Englewood Cliffs, NJ, 1984.
- [25] S. M. Kay. *Modern Spectrum Estimation: Theory and Application*. Prentice Hall, Englewood Cliffs, NJ, 1988.
- [26] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, Upper Saddle River, NJ, 1993.
- [27] S. M. Kay and S. L. Marple Jr. Spectrum analysis—a modern perspective. *Proceedings of the IEEE*, 69(3):1380–1419, November 1981.
- [28] J. Lazzaro, J. Wawrzynek, and R. Lippmann. Anawake: Signal-based power management for digital signal processing systems. Technical report, MIT Lincoln Laboratory, 1997.
- [29] J. Lazzaro, J. Wawrzynek, and R. Lippmann. A micropower analog VLSI HMM state decoder for wordspotting. Technical report, MIT Lincoln Laboratory, 1997.
- [30] E. Lee and D. G. Messerschmitt. *Digital Communication*. Kluwer Academic Publishers, Boston, MA, 1994.
- [31] V. R. Lesser, J. Pavlin, and E. Durfee. Approximate processing in real-time problem solving. *AI Magazine*, pages 49–61, Spring 1988.
- [32] O. Lieberman. Saddlepoint approximation for the distribution of a ratio of quadratic forms in normal variables. *Journal of the American Statistical Association*, 89(427):924–928, 1994.
- [33] J. W. S. Liu, W. Shih, K. Lin, R. Bettati, and J. Chung. Imprecise computations. *Proceedings of the IEEE*, 82(1):83–94, January 1994.

- [34] J. T. Ludwig, S. H. Nawab, and A. P. Chandrakasan. Low-power filtering using approximate processing for DSP applications. In *Proceedings of the Custom Integrated Circuits Conference (CICC)*, pages 185–188, Santa Clara, CA, May 1995.
- [35] J. T. Ludwig, S. H. Nawab, and A. P. Chandrakasan. Convergence results on adaptive approximate filtering. In *Advanced Signal Processing Algorithms (F. T. Luk, ed.), Proceedings of SPIE*, Denver, CO, August 1996.
- [36] J. T. Ludwig, S. H. Nawab, and A. P. Chandrakasan. Low-power digital filtering using approximate processing. *IEEE Journal on Solid State Circuits*, 31(3):395–400, March 1996.
- [37] H. S. Malvar. *Signal Processing with Lapped Transforms*. Artech House, Norwood, MA, 1992.
- [38] J. D. Meindl. *Micropower Circuits*. Wiley, New York, NY, 1969.
- [39] F. Mintzer. Filters for distortion-free two-band multirate filter banks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33:626–630, June 1985.
- [40] S. H. Nawab and E. Dorken. A framework for quality versus efficiency tradeoffs in STFT analysis. *IEEE Transactions on Signal Processing*, pages 998–1001, April 1995.
- [41] S. H. Nawab, A. V. Oppenheim, A. P. Chandrakasan, , J. T. Ludwig, and J. M. Winograd. Approximate signal processing. *Journal on VLSI Signal Processing*, 15(1–2), January–February 1997.
- [42] S. H. Nawab and J. M. Winograd. Approximate signal processing using incremental refinement and deadline-based algorithms. In *Proceedings of the International Conference on Speech, Acoustics, and Signal Processing*, pages 2857–2860, Detroit, MI, April 1995.
- [43] K. Nayebi, T. P. Barnwell III, and M. J. T. Smith. Time domain filter bank analysis: a new design theory. *IEEE Transactions on Signal Processing*, 40(6):1412–1429, June 1992.
- [44] A. V. Oppenheim and R. Schaffer. *Discrete-Time Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1989.

- [45] J. Pachares. Note on the distribution of a definite quadratic form. *Annals of Mathematical Statistics*, 26:128–131, 1955.
- [46] C. J. Pan. A low power digital filter for decimation and interpolation using approximate processing. *International Solid State Circuits Conference*, pages 102–103, February 1997.
- [47] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, Inc., 1984.
- [48] J. K. Patel and C. B. Read. *Handbook of the Normal Distribution*. Marcel Dekker, Inc., New York, NY, 1996.
- [49] Boaz Porat. *Digital Processing of Random Signals: Theory and Methods*. Prentice Hall, Englewood Cliffs, NJ, 1994.
- [50] J. M. Rabaey and M. Pedram. *Low Power Design Methodologies*. Kluwer Academic Publishers, Norwell, MA, 1996.
- [51] K. Ramachandran. *Joint Optimization Techniques for Image and Video Coding and Applications to Digital Broadcasting*. PhD thesis, Columbia University, June 1993.
- [52] K. Ramchandran and M. Vetterli. Best wavelet packet bases in a rate-distortion sense. *IEEE Transactions on Signal Processing*, 2(2):160–175, April 1993.
- [53] B. D. Rao. Adaptive IIR filtering using cascade forms. In *Proceedings of the 27th Asilomar conference on signals, systems, and computers*, pages 194–198, Pacific Grove, CA, November 1993.
- [54] L. L. Scharf. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Wesley Publishing Company, Reading, MA, 1991.
- [55] L. L. Scharf and B. D. Van Veen. Low rank detectors for Gaussian random vectors. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35(11):1579–1582, November 1987.
- [56] B. K. Shah and C. G. Khatri. Distribution of a definite quadratic form for non-central normal variates. *Annals of Mathematical Statistics*, 32:883–887, 1961.

- [57] N. R. Shanbhag and M. Goel. Low-power adaptive filter architectures and their application to 51.84 Mb/s ATM-LAN. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 45(5):1276–1290, May 1997.
- [58] J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, December 1993.
- [59] Y. Shoham and A. Gersho. Efficient bit allocation for an arbitrary set of quantizers. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(9):1445–1453, September 1988.
- [60] J. J. Shynk. Adaptive IIR filters. *IEEE ASSP Magazine*, 6:4–21, April 1989.
- [61] M. D. Smith. Expectations of ratios of quadratic forms. *Journal of Multivariate Analysis*, 31:244–257, 1989.
- [62] M. D. Smith. Comparing approximations to the expectation of a ratio of quadratic forms in normal variables. *Econometric Reviews*, 15(1):81–95, 1996.
- [63] M. J. T. Smith and S. L. Eddins. Analysis/synthesis techniques for subband image coding. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(8):1446–1456, August 1990.
- [64] M. J. T. Smith and T. P. Barnwell III. Exact reconstruction for tree-structured subband coders. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34:434–441, June 1986.
- [65] I. Sodagar, K. Nayebi, and T. P. Barnwell III. Time-varying filter banks and wavelets. *IEEE Transactions on Signal Processing*, 42(11):2983–2996, November 1994.
- [66] A. K. Soman and P. P. Vaidyanathan. Paraunitary filter banks and wavelet packets. In *Proceedings of the International Conference on Speech, Acoustics, and Signal Processing*, volume IV, pages 397–400, San Francisco, CA, May 1992.
- [67] G. Strang. *Linear Algebra*. Harcourt Brace, Jovanovich, Publishers, San Diego, CA, 1976.
- [68] A. Stuart, J. K. Ord, and M. G. Kendall. *Kendall's Advanced Theory of Statistics, Volume I, Distribution Theory*. Edward Arnold; Halsted Press, 1994.

- [69] R. M. Swanson and J. D. Meindl. Ion-implanted complementary MOS transistors in low-voltage circuits. *IEEE Journal of Solid-State Circuits*, SC-7(No. 2):146–152, April 1972.
- [70] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, Englewood Cliffs, NJ, 1993.
- [71] P. P. Vaidyanathan and T. Q. Nguyen. Eigenfilters: a new approach to least-squares FIR filter design and applications including Nyquist filters. *IEEE Transactions on Circuits and Systems*, CAS-34:11–23, January 1987.
- [72] M. J. Walsh. *Choosing and Using CMOS*. McGraw-Hill Book Company, New York, NY, 1985.
- [73] N. H. E. Weste and K. Eshraghian. *Principles of CMOS VLSI Design*. Addison-Wesley Publishing Company, Reading, MA, 1992.
- [74] G. A. Williamson. Structural issues in cascade-form adaptive IIR filters. In *Proceedings of the International Conference on Speech, Acoustics, and Signal Processing*, volume 5, pages 1436–1439, Detroit, MI, May 1995.
- [75] J. M. Winograd. *Incremental Refinement Structures for Approximate Signal Processing*. PhD thesis, Boston University, February 1997.
- [76] J. M. Winograd, J. T. Ludwig, S. H. Nawab, A. P. Chandrakasan, and A. V. Oppenheim. Approximate processing and incremental refinement concepts. In *Proceedings 2nd ARPA Rapid Prototyping of Application-Specific Signal Processors (RASSP) Conference*, pages 257–261, Washington, D.C., July 1995.
- [77] J. M. Winograd, J. T. Ludwig, S. H. Nawab, A. P. Chandrakasan, and A. V. Oppenheim. Flexible systems for digital signal processing. In *AAAI Fall Symposium on Flexible Computation in Intelligent Systems: Results, Issues, and Opportunities*, Cambridge, MA, November 1996.
- [78] Z. Xiong, K. Ramchandran, C. Herley, and M. T. Orchard. Flexible tree-structured signal expansions using time-varying wavelet packets. *IEEE Transactions on Signal Processing*, 45(2):333–345, February 1997.