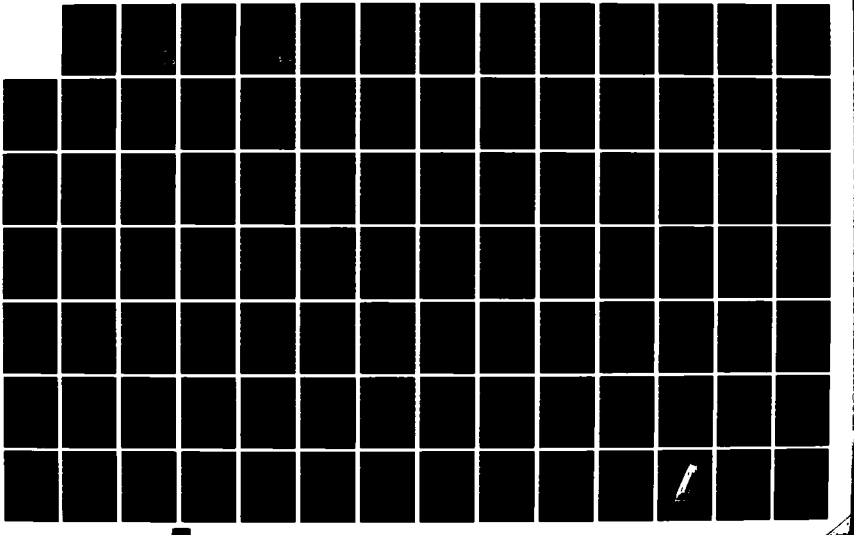
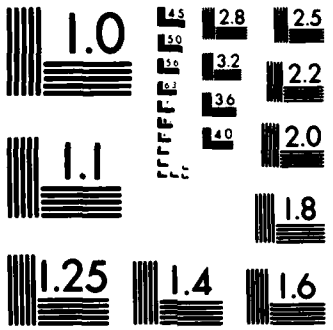
AD-A151 320 SPEECH ANALYSIS/SYNTHESIS BASED ON PERCEPTION(U) 1/3
MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB
J C ANDERSON 05 NOV 84 TR-707 ESD-TR-84-048
UNCLASSIFIED F19628-85-C-0002 F/G 17/2 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

REPRODUCED AT GOVERNMENT EXPENSE

ESD-TR-84-048

Technical Report
707

Speech Analysis/Synthesis Based on Perception

AD-A151 320

J.C. Anderson

5 November 1984

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

Prepared for the Department of the Air Force
Under Electronic Systems Division Contract F19628-85-C-0002.

Approved for public release; distribution unlimited.

DTIC FILE COPY

DTIC
ELECTE
MAR 15 1985
S D
A

85 03 04 005

REPRODUCED AT GOVERNMENT EXPENSE

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the Department of the Air Force under Contract F19628-85-C-0002.

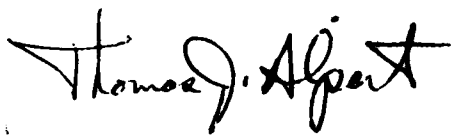
This report may be reproduced to satisfy needs of U.S. Government agencies.

The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Government.

The ESD Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER



Thomas J. Alpert, Major, USAF
Chief, ESD Lincoln Laboratory Project Office

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission is given to destroy this document when it is no longer needed.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

SPEECH ANALYSIS/SYNTHESIS BASED ON PERCEPTION

J.C. ANDERSON

Group 42

TECHNICAL REPORT 707

5 NOVEMBER 1984

Approved for public release; distribution unlimited.

DTIC
ELECTE
MAR 15 1985
A

LEXINGTON MASSACHUSETTS

ABSTRACT

A speech analysis system based on a combination of physiological and psychoacoustic results has been developed. The system contains a nonuniform Filter/Detector bank. A new relationship between Filter/Detectors and the Short-Time Fourier Transform magnitude is derived, and a generalized version of the Short-Time Fourier Transform magnitude is used to implement the analysis system. The new relationship is also applied to a discussion of channel vocoders, spectrograms, the sliding Discrete Fourier Transform, average power spectrum estimation, and nonuniform bandwidth analysis. Next, a new synthesis approach is used to reconstruct signals from the magnitude data produced by the nonuniform analysis. Apart from an overall sign factor, the analysis/synthesis system achieves exact reconstruction in the absence of data modification. The ability of the system to reconstruct signals from modified data is also demonstrated. Suggestions for further research, including data reduction and Automatic Speech Recognition applications, are given.

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
Distribution/
Availability Codes
Avail and/or
Special
A-1

DTIC COPY INSPECTED

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

Information exchange between human beings often takes place in the form of audio communication, or talking and listening. This form of communication is convenient and provides a rapid means of information transfer. Audio communication between humans and computers is also useful. Computer voice synthesis can replace warning lights and other displays, and Automatic Speech Recognition (ASR) devices can act as keyboard replacements.

Computer speech input/output has several advantages over other forms of man-to-machine communication. Since audio communication devices occupy minimal physical volume they can be used where large displays and keyboards are unacceptable. Speech allows "hands-off" communication of data as required for parcel sorting or wheelchair control. In addition, speech can provide convenient access to computer information via the telephone.

Humans have a speech recognition ability which is superior to that of existing ASR machines. Disregarding effects such as visual cues and contextual information, humans make speech recognition judgements based on information from the auditory system. Therefore, if results from perceptual and physiological studies of the auditory system are applied, it may be possible to design improved ASR machines.

When applying auditory system results to the design of ASR machines it is useful to understand what information, if any, is lost in the first analysis stage (or "front-end") of the system. Inappropriate front-end information loss can degrade overall ASR system performance. For example, if a poorly designed front-end produces the same output in response to two perceptually different input words, subsequent processing stages must rely on contextual information rather than the analyzed acoustic waveform to make a correct identification. It may therefore be possible to improve system performance if such information loss is avoided.

The problem of front-end information loss can be discovered when a synthesis technique is used to test the analyzed speech data for suitable information content. Furthermore, analyzed data may be subjected to a variety of transformations in order to reduce the data rate or investigate various auditory processes. Effects of such transformations may be examined by application of appropriate inverse transformations and signal synthesis from the processed data.

1.2 HISTORICAL DEVELOPMENT OF THE PROBLEM

The Sound Spectrograph is a widely used tool for creating speech spectrum displays, or spectrograms. A number of researchers have devised machines for reconstructing speech from spectrograms (Flanagan [1]), thereby creating a speech analysis/synthesis system. The intelligible monotone speech produced by such machines has been used in extensive perceptual studies. The Sound Spectrograph itself provides an audio analysis which is uniform with respect to frequency, and thus does not model human perception. Development of an auditory spectrogram-like representation is a current research goal (Carlson and Granstrom [2]).

In a related area, spectrogram-like representations can be generated from the Short-Time Fourier Transform (STFT) magnitude (Rabiner and Schafer [3]). Since signals can be reconstructed from the STFT magnitude (Altes [4]; Nawab, Quatieri, and Lim [5], [6]), a speech analysis/synthesis system can be developed using STFT techniques. As with spectrograms, however, this approach provides an analysis which is uniform with respect to frequency. The STFT can be modified for nonuniform analysis (Gambardella [7]; Youngberg and Boll [8]), but the corresponding synthesis techniques reported in the literature require both magnitude and phase of the modified STFT to perform signal reconstruction. Since available magnitude-only reconstruction techniques (Nawab, Quatieri, and Lim [9]) use autocorrelation functions rather than performing reconstruction directly from spectral values, such techniques cannot be modified for nonuniform analysis/synthesis. Furthermore, available approaches do not generally achieve exact signal reconstruction in the absence of data modification (Griffin and Lim [10]). Exact reconstruction is a desirable feature for algorithmic verification purposes.

1.3 THE SCOPE OF THIS REPORT

This report presents a speech analysis/synthesis system based on perception. Physiological and psychoacoustic results suggest that a nonuniform bank of Filter/Detector (F/D) subsystems can be used in the speech analysis system, as shown in Chapter 2. A new relationship between F/D subsystems and the STFT magnitude (or, equivalently, the STFT magnitude squared) is described, and a generalized version of the STFT magnitude is used to implement the desired F/D bank. A new synthesis approach capable of reconstructing signals from the generalized STFT magnitude is described in Chapter 3. Examples of results produced by the analysis/synthesis system are presented in Chapter 4. Apart from an overall sign factor, the system achieves exact reconstruction in the absence of data modification. The ability of the system to reconstruct signals from modified data is also demonstrated. A summary is given in Chapter 5, along with suggestions for further research. Appendix A presents standard definitions for reference purposes. Prerequisite F/D theory, which is used throughout the report, is presented in Appendix B. Several approaches to computation of the generalized STFT magnitude are described in Appendix C. Appendix D applies the new relationship between F/D subsystems and the STFT magnitude to a discussion of channel vocoders, spectrograms, the sliding Discrete Fourier Transform, average power spectrum estimation, and nonuniform bandwidth analysis.

CHAPTER 2

SPEECH ANALYSIS SYSTEM

2.1 INTRODUCTION

In this chapter, a simplified model of the (monaural) human peripheral auditory system is developed from a combination of physiological and psychoacoustic data. Binaural effects will not be discussed, although such effects may be important for Automatic Speech Recognition applications in noisy environments (Lyon [11]). A generalized version of the Short-Time Fourier Transform magnitude squared is used to obtain a digital representation of the model. Short-Time timing is also computed, and minimum sampling rate issues are discussed.

2.2 A SIMPLIFIED AUDITORY MODEL BASED ON PHYSIOLOGICAL RESULTS

Fig. 2.1 is a diagram of the human peripheral auditory system showing the outer, middle, and inner ear structures (Flanagan [1]). The drawing is not to scale, and some structures are enlarged for illustrative purposes. In the auditory system, sounds entering the outer ear travel through the middle ear and generate pressures in the inner ear fluids. The cochlea, a structure in the inner ear, contains the basilar membrane which functions as a filter bank. Basilar membrane motion causes hair cells in the organ of Corti to produce firings on the auditory nerve, which contains approximately 30,000 fibers. A number of researchers have studied firing patterns by inserting microelectrodes into the auditory nerve fibers of anesthetized animals (Kiang [12]; Frishkopf [13]; Katsuki, Suga, and Kanno [14]). Such studies indicate that the peripheral auditory system can be roughly modeled as a Filter/Detector (F/D) bank, and model parameters can be derived from physiological data.

Fig. 2.2 presents a F/D subsystem of the type often used in auditory models (eg., see Siebert [15]). The input is analogous to pressure at the eardrum, and the output simulates various firing pattern features which will be described later in this section. For simplicity, the effects of spontaneous nerve firing activity have been omitted from the model. The F/D subsystem of Fig. 2.2 consists of a bandpass filter, memoryless nonlinearity, and lowpass smoothing filter (see Section B.2 for a detailed description of the various F/D subsystem components). The bandpass filter impulse response is a lowpass window function h(t) modulated by a sinusoid of frequency Ω_c. The window function has

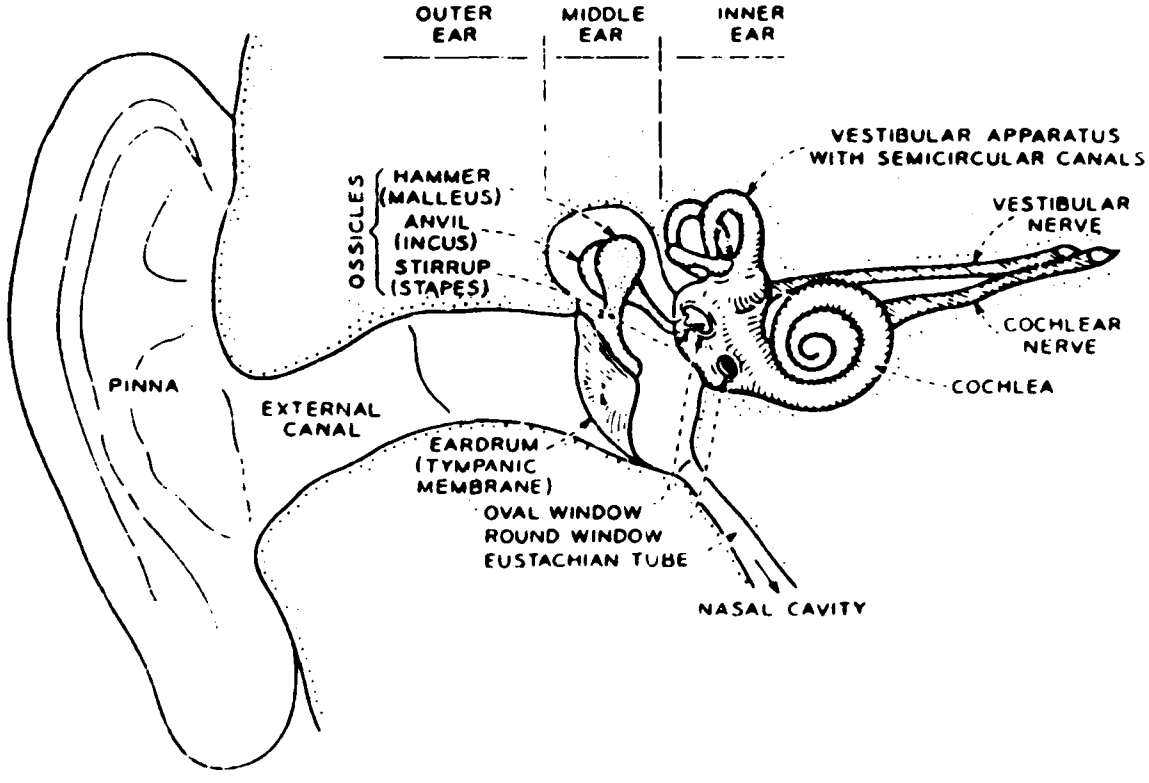


Figure 2.1: Diagram of the Human Peripheral Auditory System

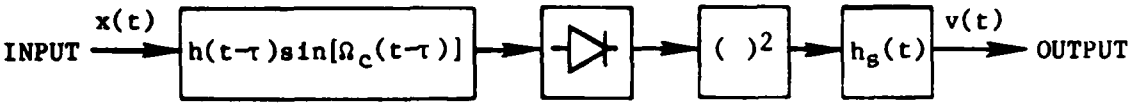


Figure 2.2: F/D Subsystem for the Simplified Auditory Model

---

bandwidth Ω_h, yielding a bandpass filter bandwidth of $2\Omega_h$. The window function is of the form

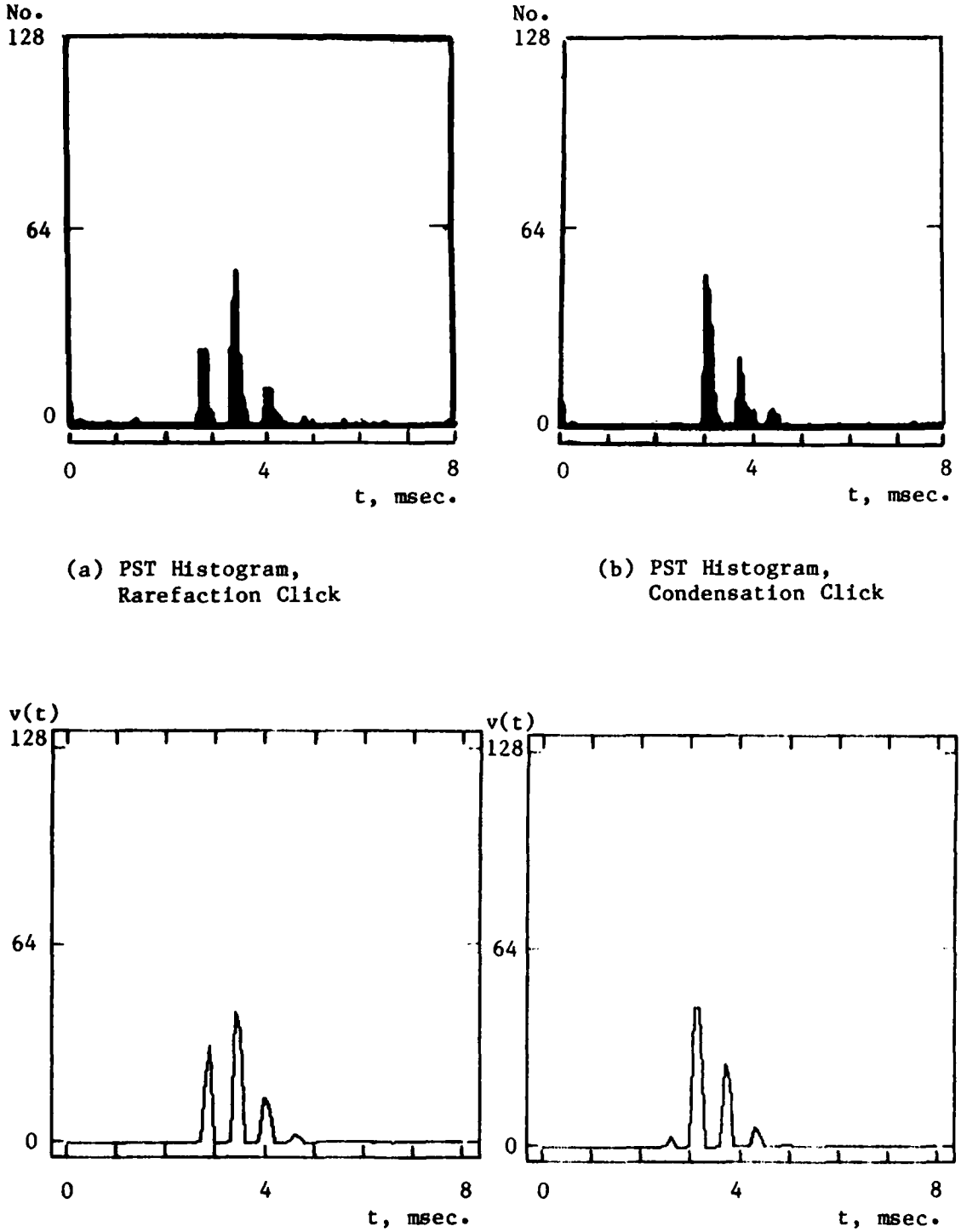
$$h(t) = \beta t^2 e^{-\alpha t}, \quad 0<t$$

$$= 0, \text{ otherwise,} \qquad (2.1)$$

where α and β are positive real constants. This window function, which is derived from basilar membrane models (Flanagan [1]), will be discussed further in Section 2.4.5. The window function of Equation 2.1 has been chosen for convenience, and the theory presented in this chapter is basically unchanged when other window functions are used instead. For example, Bessel filters (Chu [16]) can be used to obtain a better match between bandpass filter characteristics and the neural frequency response characteristics described by tuning curves (Kiang, Sachs, and Peake [17]). A fixed delay, τ, is included in the bandpass filter. The memoryless nonlinearity is approximately modeled as a half wave square law device (Siebert [18]), although some researchers have suggested use of a half wave piecewise linear device (Schroeder [19]). The smoothing filter, which has impulse response $h_s(t)$, acts as an envelope detector at high frequencies ($\Omega_c > 2\pi \times 4000$ radians per second) but follows details of the rectified waveform at low frequencies.

Appropriate model parameters can be obtained from an examination of physiological measurement techniques and the resulting physiological data. For example, the response of a nerve fiber to acoustic impulses, or clicks, is often described by a poststimulus-time (PST) histogram. A stimulus is repeated a large number of times, and the PST histogram depicts the density of firings as a function of time following the stimulus. Thus, a PST histogram indicates the likelihood that a particular nerve will fire at a given time following the stimulus. Firing patterns of individual nerves are not similar in appearance to a PST histogram. It is assumed, however, that firings from a large population of similar nerves could be combined to produce a deterministic pattern approximating a PST histogram.

A PST histogram is shown in Fig. 2.3a for rarefaction clicks, and in Fig. 2.3b for condensation clicks. The experimental animal was a cat, the click level was -70dB re 100 volt input to the condenser earphone, and the nerve fiber was maximally responsive at a frequency of 1.67 KHz (Kiang [11]). Fig. 2.4 presents eighteen further examples of PST histograms for various characteristic frequencies from a single cat. The click level for Fig. 2.4 was -50dB. Note the loss of timing details for high characteristic frequencies.



(a) PST Histogram, Rarefaction Click

(b) PST Histogram, Condensation Click

(c) Model Response to Negative Impulse Input

(d) Model Response to Positive Impulse Input

Figure 2.3: PST Histogram Data and Corresponding Model Results

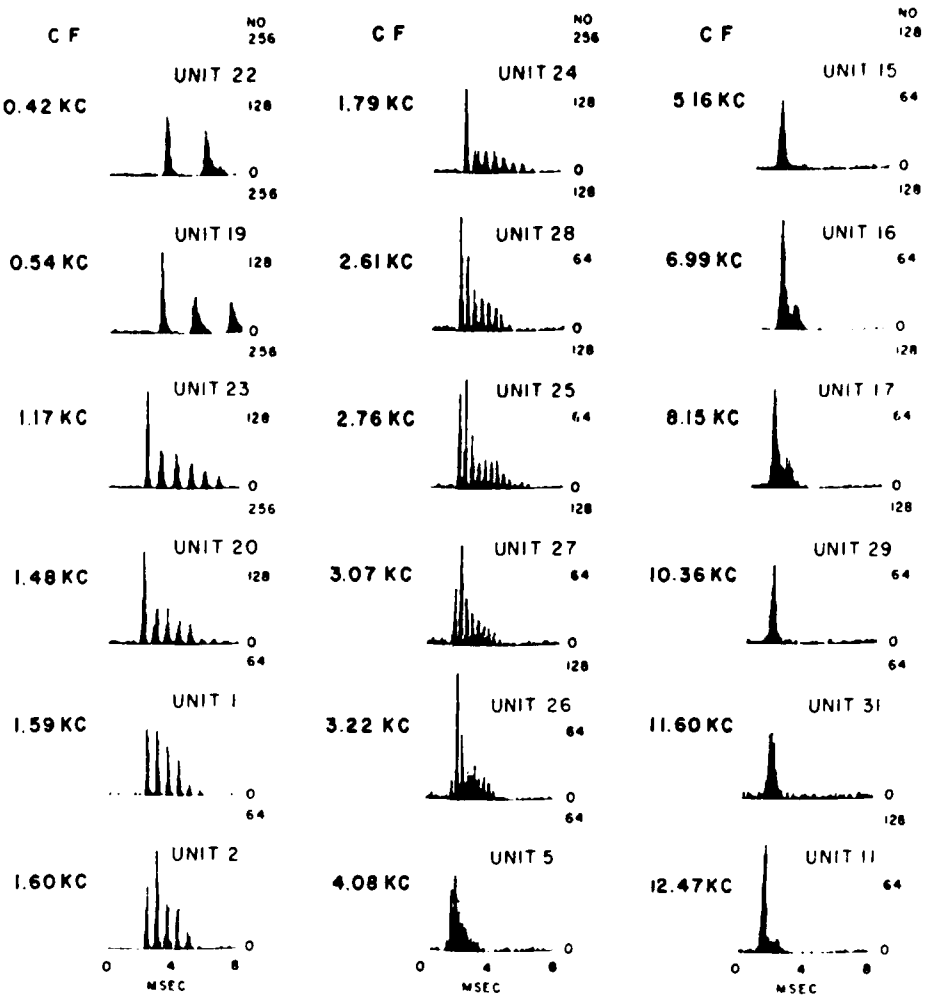
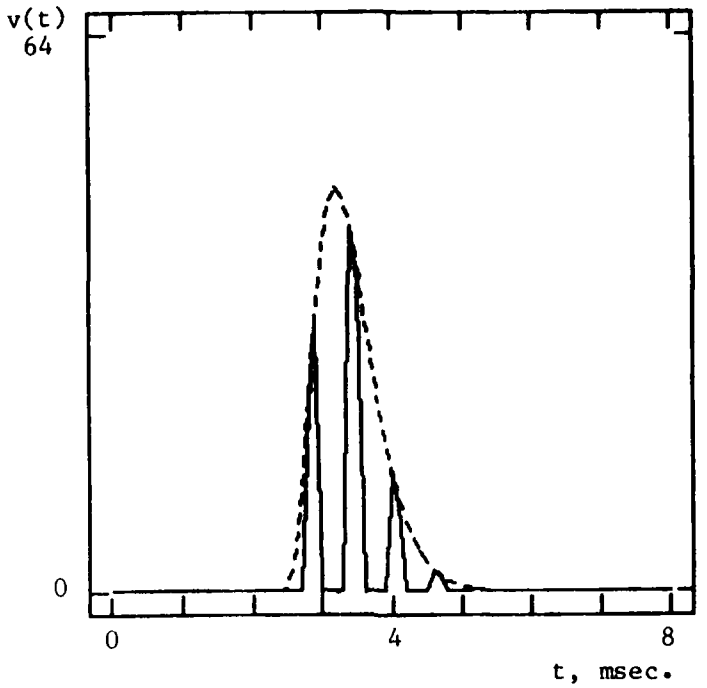


Figure 2.4: PST Histograms for Various Characteristic Frequencies

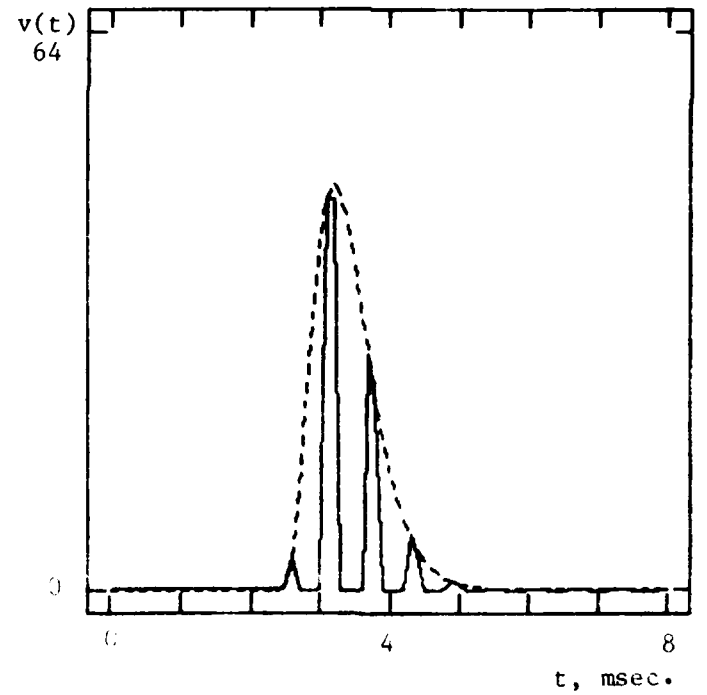
When the input to the F/D subsystem of Fig. 2.2 is an impulse, model parameters can be chosen so that the output mimics a PST histogram over a limited range of intensities. As a specific example, parameters are chosen so that the PST histograms of Figs. 2.3a and 2.3b are simulated. From Figs. 2.3a and 2.3b it can be seen that the delay is τ=.0024 second and the characteristic frequency is $\Omega_c=2\pi$x1670 radians per second. Since the characteristic frequency is low enough so that timing details are preserved, the smoothing filter has no effect and is eliminated by choosing $h_s(t)=\delta(t)$. Use of α=2500 sec^{-1} and $\beta=9x10^7$ sec^{-2} results in a reasonable match to the data. The F/D output, v(t), is shown in Fig. 2.3c for an input $x(t)=-\delta(t)$, and in Fig. 2.3d for an input $x(t)=\delta(t)$.

Model parameters can be chosen to mimic many features of auditory nerve patterns for clicks and steady sine waves over limited intensity ranges (Siebert [15]). Agreement over wider intensity ranges, and for stimuli such as tone bursts and noise, can be obtained by inserting an Automatic Gain Control (AGC) at the bandpass filter output. Recent research (Smith and Zwislocki [20]; Smith [21]; Harris and Dallos [22]) suggests use of a short-term adaptation function rather than an AGC. In any case, such improvements will not be considered here.

The PST histogram envelope, which represents a short-term average firing rate, is often a function of interest (Schroeder [19]). If the lowpass smoothing filter bandwidth Ω_s is chosen such that $2\Omega_h<\Omega_s<\Omega_c-2\Omega_h$, then the F/D subsystem impulse response mimics the PST histogram envelope rather than the detailed PST histogram (see Section B.3.6). Under these conditions the F/D output is proportional to $h^2(t-\tau)$, and Fig. 2.5 shows this function superimposed on the simulated PST histograms of Fig. 2.3.



(a) Negative Impulse Input



(b) Positive Impulse Input

Figure 2.7 Simulated PST Histograms and Envelopes

In general, it can be shown that the envelope of a half wave square law device output is proportional to the envelope of a square law device output (see section B.2.2). Thus, if the smoothing filter bandwidth is chosen so that the output follows the PST histogram envelope, then the half wave square law device can be replaced by a square law device. It will be shown in Section 2.4 that the resulting F/D subsystem can be implemented by a generalized form of the Short-Time Fourier Transform (STFT) magnitude squared. Therefore, a STFT magnitude squared approach can be used to roughly simulate PST histogram envelope functions at low frequencies ($\Omega_c < 2\pi \times 4000$ radians per second), and PST histograms at high frequencies. A simplified auditory model based on short-term average firing rates is thus implemented using STFT techniques.

The model described in this section does not attempt to account for all known aspects and limitations of the auditory system. The exact manner in which signals are encoded by the auditory system is a current research topic, and several theories have recently been developed (see for example Sachs and Young [23], [24]; Delgutte and Kiang [25]). Instead, the model demonstrates an approximate relationship between certain physiological results and F/D or STFT magnitude analysis techniques, and shows how standard analysis techniques must be modified for auditory modeling purposes. Although the auditory model presented in this section is crude, it will be shown in Chapter 3 that no important information is lost by such an approach since signals can be synthesized from the simplified auditory model outputs.

2.3 A SIMPLIFIED AUDITORY MODEL INCORPORATING PERCEPTUAL RESULTS

Although auditory model parameters can be derived from physiological data, there is no guarantee that the resulting model will simulate human perception. Recall that physiological data is generally obtained from experimental animals rather than humans. Furthermore, since available data mainly concerns the peripheral auditory system, effects of higher processing levels are not included in models based on such data alone. In order to develop a speech analysis system, it is desirable to account for at least some known aspects of human auditory perception. Supplementary information is therefore required for the determination of appropriate auditory model parameters.

The field of psychoacoustics provides an alternative means of investigating the auditory system. Listening experiments are performed on live human subjects, and the results indicate functional behavior of the complete auditory system. One useful psychoacoustic result is the concept of a critical band. A critical band has been defined (Scharf [26]) as the bandwidth at which subjective responses change abruptly. For example, assume that a listener is subjected to a bandlimited noise stimulus. The bandwidth of the stimulus is varied but a constant sound pressure level is maintained. As long as the bandwidth of the noise is less than a critical band, perceived loudness of the noise remains constant. When the bandwidth of the noise increases beyond a critical band, perceived loudness of the noise begins to increase. Since similar critical bands are encountered in a variety of different perceptual experiments, critical bands are often used to describe the filtering process assumed to take place within the auditory system.

A model of the human auditory system based on perception can be constructed by combining physiological and psychoacoustic results. The structure of the model is determined from physiology, as discussed in Section 2.2. Empirical critical bandwidth data from humans, rather than physiological tuning curve or PST histogram data from animals, is then used to determine the bandpass filter center frequencies and bandwidths.

Table 2.1 presents the necessary parameters for design of a critical bandwidth filter bank (Scharf [26]). Note that the critical bandwidth can be expressed as a continuous function of center frequency by interpolating the data of Table 2.1. Fifteen filters are chosen to adequately cover the 200-3675 Hz frequency range. The filters have nonuniform center frequency spacing, and bandwidth which increases with center frequency. The filters are roughly constant bandwidth (approximately 110 Hz) for frequencies below 700 Hz, and constant Q (center frequency to bandwidth ratio of approximately 6.4) above 700 Hz. Although recent estimates of auditory filter shape suggest use of different values below 500 Hz (Moore and Glasberg [27]), the data of Scharf will be used to design this speech analysis system.

TABLE 2.1

CRITICAL BANDWIDTH FILTER BANK PARAMETERS

Filter Number (k)	Center Frequency (Hz)	Critical Bandwidth (Hz)
1	250	100
2	350	100
3	450	110
4	570	120
5	700	140
6	840	150
7	1000	160
8	1170	190
9	1370	210
10	1600	240
11	1850	280
12	2150	320
13	2500	380
14	2900	450
15	3400	550

2.4 CRITICAL BANDWIDTH FILTER/DETECTOR BANK IMPLEMENTATION

This section describes implementation of a critical bandwidth Filter/Detector (F/D) bank, which will be employed as part of the speech analysis system. First, a new relationship between F/D subsystems and the continuous-time Short-Time Fourier Transform (STFT) magnitude squared is described. The new relationship demonstrates that a specific type of F/D subsystem can be implemented via the STFT magnitude squared. Next, the discrete-time case is described and then generalized to allow implementation of a critical bandwidth F/D bank. Finally, the specifications given in Sections 2.2 and 2.3 are used to design the desired F/D bank via STFT techniques.

2.4.1 CONTINUOUS-TIME SHORT-TIME FOURIER TRANSFORM DEFINITION

The STFT is a widely used approach to time-dependent frequency analysis. For the continuous-time case, the STFT evaluated at some fixed frequency Ω_c is defined as (Flanagan [1]):

$$X_t(j\Omega_c) = \int_{-\infty}^{+\infty} x(\tau)h(t-\tau)e^{-j\Omega_c\tau}\,d\tau. \qquad (2.2)$$

Note that if $h(t)=1$ for all t, the STFT becomes the continuous-time Fourier transform described in Appendix A. A block diagram for STFT computation, which expresses the STFT in terms of linear filtering operations, is shown in Fig. 2.6a. This interpretation indicates that the STFT, viewed as a function of time at the fixed frequency Ω_c, is a lowpass complex function bandlimited to the window function bandwidth.

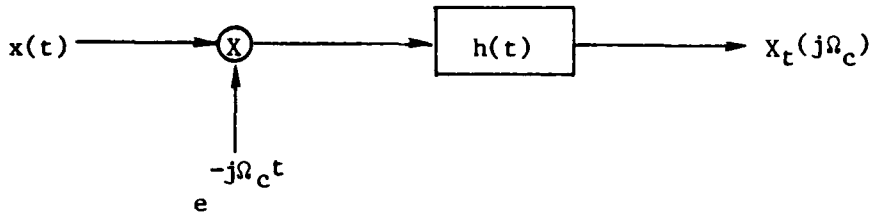
An equivalent STFT definition is:

$$X_t(j\Omega_c) = e^{-j\Omega_c t} \int_{-\infty}^{+\infty} x(t-\tau)h(\tau)e^{j\Omega_c \tau} d\tau \qquad (2.3)$$

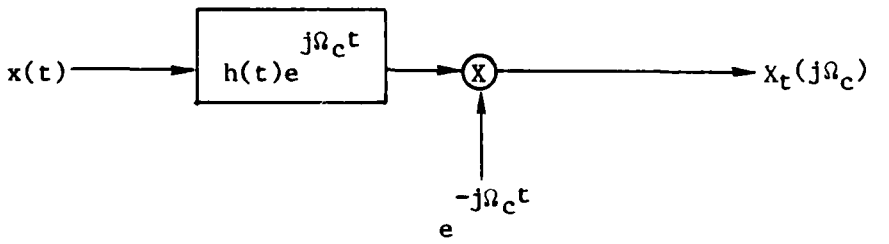
The corresponding block diagram is shown in Fig. 2.6b. In this approach, a complex modulation signal is used to downconvert the bandpass filter output into a lowpass function.

It follows from Equation 2.3 that the imaginary part of $e^{j\Omega_c t} X_t(j\Omega_c)$ is the output of a bandpass filter which has input x(t) and impulse response $h(t)\sin(\Omega_c t)$. Thus, the STFT could be used to implement the bandpass filter portion of the F/D subsystem shown in Fig. 2.2. This approach, however, will not be pursued.

The methods for STFT computation shown in Fig. 2.6 employ a local oscillator (Taub and Schilling [28]). Thus, the STFT is different from a detection process which uses a F/D. It will be shown in Section 2.4.2 that the STFT magnitude squared, rather than the STFT, corresponds to a detection process using a F/D.



(a) Modulator followed by Lowpass Filter



(b) Bandpass Filter followed by Modulator

Figure 2.6: Linear Filtering Interpretation of the STFT

2.4.2 CONTINUOUS-TIME F/D IMPLEMENTATION USING STFT MAGNITUDE SQUARED

F/D subsystems have long been used as a means of approximating the STFT magnitude squared. Early work by Fano [29] described a relationship between F/D subsystems and the STFT magnitude squared for special window functions. Schroeder and Atal [30] extended this work to include arbitrary window functions, and the results are discussed by Flanagan [1] and Gambardella [7]. However, these authors did not characterize basic F/D parameters such as lowpass smoothing filter bandwidth. Flanagan [1] discusses a relationship, valid only for certain signals under restrictive conditions, which links the STFT magnitude with speech spectrograms (see Section D.3). Flanagan also discusses a relationship between long-term average F/D outputs and an averaged version of the STFT magnitude squared.

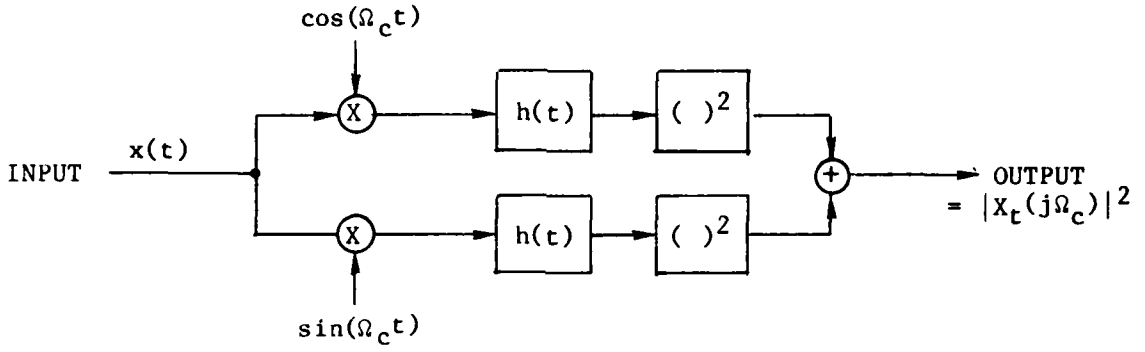
In this section, a new relationship between F/D subsystems and the STFT magnitude squared is described. The new relationship is more precise than those previously reported in the literature, and demonstrates the equivalence between the STFT magnitude squared and a specific type of F/D subsystem.

2.4.2.1 PLAUSIBILITY ARGUMENT

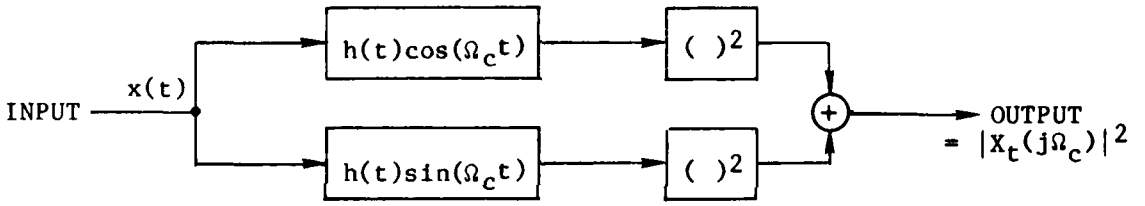
A system for computing the STFT magnitude squared is shown in Fig. 2.7a. From this system and the modulation property of Fourier transforms (described in Appendix A), it is easily seen that $|X_t(j\Omega_c)|^2$ is a lowpass real function of time which is bandlimited to twice the window function bandwidth.

An equivalent system for computing the STFT magnitude squared is shown in Fig. 2.7b. In this figure, the output of each square law device consists of a lowpass function and a high frequency bandpass function (see Section B.3.1). The high frequency bandpass functions cancel out in the adder, while the lowpass functions combine to form $|X_t(j\Omega_c)|^2$.

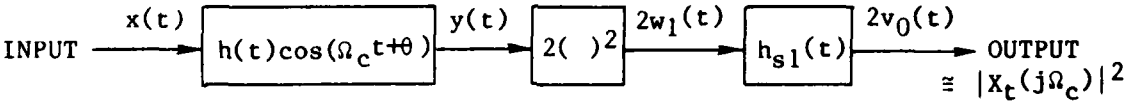
The fact that only lowpass functions are retained by the STFT magnitude squared suggests that a similar result could be produced by the F/D subsystem of Fig. 2.7c. Details of this F/D will be described in Section 2.4.2.2. In Fig. 2.7c, the high frequency components at the square law device output are eliminated by linear filtering rather than cancellation. Thus, the STFT magnitude squared and F/D outputs will generally be similar but not identical.



(a) Exact Computation using Lowpass Filters



(b) Exact Computation using Bandpass Filters



(c) Approximate Computation using a Bandpass Filter and Detector

Figure 2.7: Short-Time Fourier Transform Magnitude Squared Computation

2.4.2.2 PROOF OF F/D AND STFT MAGNITUDE SQUARED EQUIVALENCE

The F/D subsystem shown in Fig. 2.7c consists of a Linear Time-Invariant (LTI) bandpass filter, square law device with multiplicative constant, and LTI smoothing filter. The impulse response of the bandpass filter contains an arbitrary constant parameter θ. If $\theta = -\pi/2$ for example, the bandpass filter impulse response is $h(t)\sin(\Omega_c t)$. Nomenclature for the signals in Fig. 2.7c follows that of the general F/D theory presented in Appendix B.

Let the window function $h(t)$ be the impulse response of an ideal lowpass filter with bandwidth Ω_h:

$$h(t) = [\sin(\Omega_h t)]/\pi t. \tag{2.4}$$

The output of the bandpass filter in Fig. 2.7c is:

$$\begin{aligned} y(t) &= x(t)*[h(t)\cos(\Omega_c t+\theta)] \\ &= \int_{-\infty}^{+\infty} x(\tau)h(t-\tau)\cos(\Omega_c t-\Omega_c \tau+\theta)d\tau \\ &= f(t)\cos(\Omega_c t) + g(t)\sin(\Omega_c t), \end{aligned} \tag{2.5}$$

where

$$f(t) = [x(t)\cos(\Omega_c t-\theta)]*h(t) \tag{2.6}$$

and

$$g(t) = [x(t)\sin(\Omega_c t-\theta)]*h(t) \tag{2.7}$$

are lowpass functions. The square law device output is:

$$2w_1(t) = [f^2(t)+g^2(t)]$$

$$+ [\cos(2\Omega_c t)][f^2(t)-g^2(t)] + 2[\sin(2\Omega_c t)]f(t)g(t). \qquad (2.8)$$

Since $f(t)$ and $g(t)$ are lowpass bandlimited to the frequency domain region $|\Omega|<\Omega_h$, the function $f^2(t)+g^2(t)$ is lowpass bandlimited to $|\Omega|<2\Omega_h$. The remaining components of Equation 2.8 are high frequency bandpass signals limited to the region $2\Omega_c-2\Omega_h<|\Omega|<2\Omega_c+2\Omega_h$. Let the smoothing filter with impulse response $h_{s1}(t)$ be an ideal lowpass filter having bandwidth Ω_{s1}:

$$h_{s1}(t) = [\sin(\Omega_{s1}t)/\pi t]. \qquad (2.9)$$

Also, let $2\Omega_h<\Omega_{s1}<2\Omega_c-2\Omega_h$. It follows that $2\Omega_h<\Omega_c$. The F/D output is

$$2v_0(t) = f^2(t) + g^2(t), \qquad (2.10)$$

which is positive even though the impulse response of the smoothing filter is not positive for all t. Since the signals $x(t)$ and $h(t)$ are real, it follows from Equations 2.6, 2.7, and 2.10 that

$$2v_0(t) = \{[x(t)\cos(\theta-\Omega_c t)]*h(t)\}^2 + \{[x(t)\sin(\theta-\Omega_c t)]*h(t)\}^2$$

$$= |[x(t)e^{j(\theta-\Omega_c t)}]*h(t)|^2$$

$$= |X_t(j\Omega_c)|^2, \qquad (2.11)$$

where $X_t(j\Omega_c)$ is the STFT evaluated at a fixed frequency Ω_c. Thus, when ideal lowpass filter functions are used for $h(t)$ and $h_{sl}(t)$, the STFT magnitude squared is exactly the same as the F/D subsystem output of Fig. 2.7c. The F/D subsystem of Fig. 2.7c can therefore be used to measure the STFT magnitude squared, or the STFT magnitude squared can be used to implement this F/D subsystem. STFT magnitude squared (and therefore STFT magnitude) analysis of noise, impulse, sinusoid, and sinusoidal pair signals follows directly from the examples of Section B.3. Note that the parameter θ does not appear in the final result and has no effect on the F/D output.

When the window function $h(t)$ is the impulse response of a realizable non-ideal lowpass filter, the F/D subsystem of Fig. 2.7c is not necessarily equivalent to the STFT magnitude squared (although agreement is generally quite good). For non-ideal lowpass filter window functions, the lowpass and high frequency bandpass components of Equation 2.8 overlap in the frequency domain and cannot be separated by any LTI smoothing filter. Thus, although Equation 2.8 correctly describes the smoothing filter input, Equation 2.10 becomes an approximate description of the smoothing filter output. Under these conditions the F/D output is approximately, but not exactly, the same as the STFT magnitude squared.

It should be noted that many other window function and smoothing filter combinations exist which yield a F/D output identical to the STFT magnitude squared. As a simple example, let the window function be an impulse, $h(t)=\delta(t)$. If the smoothing filter has impulse response $h_{sl}(t)=\delta(t)/[2(\cos\theta)^2]$, $\cos\theta\neq 0$, then both the F/D subsystem and STFT magnitude squared produce the result $x^2(t)$.

2.4.2.3 DISCUSSION

In Section 2.4.2.2 it was shown that the STFT magnitude squared can be used to implement a F/D subsystem of the type shown in Fig. 2.7c. The fixed STFT analysis frequency, Ω_c, determines the center frequency of the bandpass filter in the F/D subsystem. Let Ω_h denote the one-sided main lobe bandwidth (see Section B.2.1) of any lowpass window function h(t). As long as the bandpass filter has a center frequency which is greater than its bandwidth, ie. $2\Omega_h \ll \Omega_c$, a lowpass smoothing filter operation is effectively implemented by the STFT magnitude squared computation. The effective smoothing filter can be considered to have the same bandwidth as the bandpass filter, ie. $2\Omega_h$. The window function thus determines the bandwidth of both the bandpass filter and the lowpass smoothing filter.

There are many advantages to implementing a F/D via the STFT magnitude squared. The STFT is widely used, so literature and computer programs are readily available. Since the magnitude squared computation automatically implements an effective smoothing filter, results may be obtained more efficiently than if a direct F/D implementation is used. Since there are no delay elements between the bandpass filter outputs and the adder output of Fig. 2.7b, the effective smoothing filter implemented by the STFT magnitude squared has zero delay regardless of the window function used. When the STFT magnitude squared is used to implement a F/D subsystem, difficulties normally associated with smoothing filter design (as discussed in Section B.2.3) are eliminated and the output is guaranteed to be positive at all times. This feature is desirable for auditory modeling purposes since nerve firing rates are always positive.

F/D implementation via the STFT magnitude squared has disadvantages as well. The STFT magnitude squared does not generally produce results identical to those produced by direct F/D subsystem implementations. Design flexibility is limited since the F/D bandpass filter must be of a specific type, the memoryless nonlinearity must be a square law device, and the lowpass smoothing filter must have the same bandwidth as the bandpass filter. Despite these limitations, however, F/D subsystems implemented via the STFT magnitude squared are appropriate for many applications.

2.4.3 DISCRETE-TIME F/D IMPLEMENTATION USING STFT MAGNITUDE SQUARED

For convenience, a discrete-time implementation of the speech analysis system is desired. The "analog" continuous-time theory presented in Section 2.4.2 must therefore be extended to the "digital" discrete-time case. One procedure for transforming an analog filter design to a digital filter design is known as the impulse invariant method (Oppenheim and Schafer [31]). In this procedure, the unit-sample response of the digital filter is equally spaced samples of the impulse response of the analog filter. For example, if h(t) is the impulse response of an analog lowpass filter, then the unit-sample response of the corresponding digital filter is:

$$h(n) = \{h(t)\}|_{t=nT}, \tag{2.12}$$

where T is the sampling period. The continuous-time F/D subsystem of Fig. 2.7c can be transformed, via the impulse invariant method, into the discrete-time F/D subsystem of Fig. 2.8. The bandpass filter center frequency is $\omega_c = \Omega_c T$, and θ is an arbitrary constant parameter.

Let the window function h(n) be the unit-sample response of an ideal lowpass filter with bandwidth ω_h:

$$h(n) = [\sin(\omega_h n)]/\pi n. \tag{2.13}$$

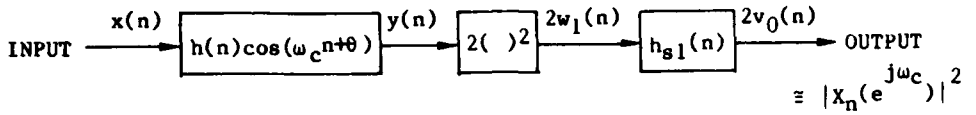
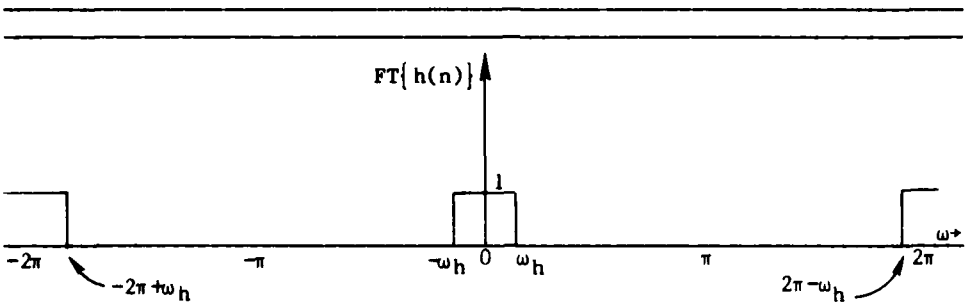
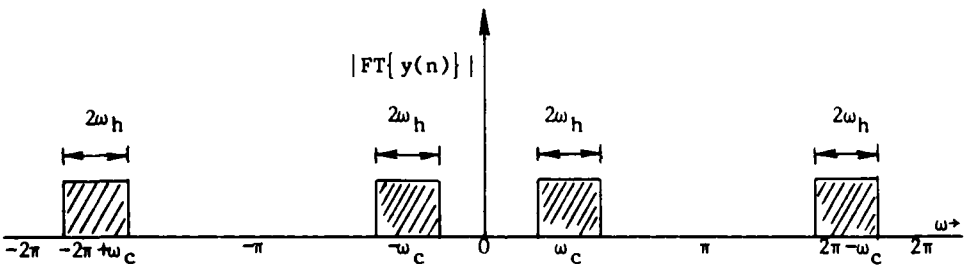


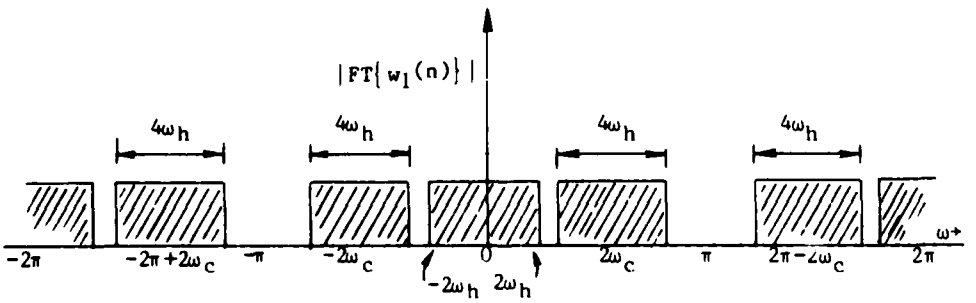
Figure 2.8: Discrete-Time F/D Subsystem



(a) Ideal Window Function Characteristic



(b) Spectral Regions Occupied by the Bandpass Filter Output



(c) Spectral Regions Occupied by the Square Law Device Output

Figure 2.9: Spectral Characteristics of Discrete-Time F/D Subsystem

The Fourier transform of the window function, $FT\{h(n)\}$, is shown in Fig. 2.9a. Spectral regions occupied by the bandpass filter output are shown in Fig. 2.9b. The bandpass filter output is:

$$y(n) = x(n)*[h(n)\cos(\omega_c n+\theta)]$$

$$= \sum_{m=-\infty}^{\infty} x(m)h(n-m)\cos(\omega_c n-\omega_c m+\theta)$$

$$= f(n)\cos(\omega_c n) + g(n)\sin(\omega_c n), \qquad (2.14)$$

where

$$f(n) = [x(n)\cos(\omega_c n-\theta)]*h(n) \qquad (2.15)$$

and

$$g(n) = [x(n)\sin(\omega_c n-\theta)]*h(n). \qquad (2.16)$$

On the interval $-\pi<\omega<\pi$, f(n) and g(n) are lowpass bandlimited to $|\omega|<\omega_h$. The square law device output is:

$$2w_1(n) = [f^2(n) + g^2(n)]$$

$$+ [\cos(2\omega_c n)][f^2(n)-g^2(n)] + 2[\sin(2\omega_c n)]f(n)g(n). \qquad (2.17)$$

By the modulation property, the function $f^2(n)+g^2(n)$ is lowpass bandlimited (on the interval $-\pi<\omega<\pi$) to the region $|\omega|<2\omega_h$, as shown in Fig. 2.9c. The remaining components of Equation 2.17 are high frequency bandpass signals which may be eliminated by the smoothing filter. Let

the smoothing filter with unit-sample response $h_{sl}(n)$ be an ideal lowpass filter having bandwidth ω_{sl}:

$$h_{sl}(n) = [\sin(\omega_{sl}n)]/\pi n. \tag{2.18}$$

Also, let $2\omega_h < \omega_{sl} < 2\omega_c - 2\omega_h$ and $2\omega_h < \omega_{sl} < 2\pi - 2\omega_c - 2\omega_h$. It follows that $2\omega_h < \omega_c < \pi - 2\omega_h$. The F/D output is:

$$2v_0(n) = f^2(n) + g^2(n). \tag{2.19}$$

Since the signals x(n) and h(n) are real, it follows that

$$2v_0(n) = |X_n(e^{j\omega_c})|^2, \tag{2.20}$$

where $X_n(e^{j\omega_c})$ is the discrete-time STFT evaluated at a fixed frequency ω_c, and is defined as (Rabiner and Schafer [3]):

$$X_n(e^{j\omega_c}) = \sum_{m=-\infty}^{\infty} x(n-m)h(m)e^{-j\omega_c(n-m)}. \tag{2.21}$$

The discrete-time STFT thus follows directly from application of the impulse invariant transformation to the continuous-time STFT. Note that the discrete-time STFT of Equation 2.21 corresponds to the continuous-time STFT of Equation 2.3, and the discrete-time F/D result of Equation 2.20 corresponds to the continuous-time result of Equation 2.11. The discussion of Section 2.4.2.3 therefore applies to both the discrete-time and continuous-time cases.

A difference between the discrete-time and continuous-time implementation occurs in the restriction on bandpass filter center frequency relative to bandwidth. For the continuous-time case, it is required that the bandpass filter have a center frequency which is greater than its bandwidth, ie., $2\Omega_h < \Omega_c$. This restriction also applies to the discrete-time case, ie., $2\omega_h < \omega_c$. However, an additional restriction must be applied in the discrete-time case because of the periodic spectral characteristics shown in Fig. 2.9c. An upper limit must be applied to the digital bandpass filter center frequency, resulting in the restriction $2\omega_h < \omega_c < \pi - 2\omega_h$. In other words, the digital bandpass filter must have a center frequency which is greater than its bandwidth but less than pi minus the bandwidth. The upper frequency limit is discussed further in Section 2.4.5.

It can easily be shown that if ω_c does not fall within the range $2\omega_h < \omega_c < \pi - 2\omega_h$, then the STFT magnitude squared does not implement a F/D subsystem. For example, if $\omega_c = 0$ it follows from Equation 2.21 that the STFT magnitude squared is equivalent to a lowpass filter with impulse response h(n) followed by a square law device. Similarly, if $\omega_c = \pi$ the STFT magnitude squared is equivalent to a highpass filter with impulse response $(-1)^n h(n)$ followed by a square law device. In all cases, however, the STFT magnitude squared is a lowpass function with bandwidth $2\omega_h$.

2.4.4 THE GENERALIZED SHORT-TIME FOURIER TRANSFORM (DISCRETE-TIME CASE)

It was shown in Section 2.4.2 that the STFT magnitude squared can be used to implement a F/D subsystem in which the bandpass filter bandwidth is fixed by choice of the window function. The simplified auditory system model described in Sections 2.2 and 2.3, however, uses a bank of F/D subsystems in which each bandpass filter has a different bandwidth. Therefore, a generalized version of the STFT which allows a different window function at each analysis frequency must be used to implement the auditory model. Only the discrete-time case will be discussed.

Let the STFT be evaluated at K discrete arbitrarily spaced frequencies ω_k, where k=1,2,...,K. A different window function $h_k(n)$ may be used at each frequency. It follows from Equation 2.21 that the Generalized Short-Time Fourier Transform (GSTFT) can be defined as (Rabiner and Schafer [3]):

$$X_n(e^{j\omega_k}) = \sum_{m=-\infty}^{\infty} x(n-m)h_k(m)e^{-j\omega_k(n-m)} . \qquad (2.22)$$

It is assumed that the signal x(n) and the set of window functions $h_k(n)$ are real. Several approaches to GSTFT computation are discussed in Appendix C.

The GSTFT magnitude squared can be used to implement a bank of F/D subsystems similar to the type shown in Fig. 2.8. The resulting bandpass filters have impulse response $h_k(n)\cos(\omega_k n+\theta_k)$, where θ_k is arbitrary. The bandpass filter with center frequency ω_k has a bandwidth determined by $h_k(n)$. As long as the bandpass filter has a center frequency which is greater than its bandwidth (and less than pi minus the bandwidth), a lowpass smoothing filter operation is effectively implemented by the GSTFT magnitude squared computation. The bandpass and smoothing filters can be considered to have the same bandwidth.

2.4.5 PERCEPTION-BASED SPEECH ANALYSIS SYSTEM IMPLEMENTATION

Since GSTFT magnitude squared results are proportional to the desired F/D bank outputs, the GSTFT magnitude squared can be used to implement the F/D bank specified in Sections 2.2 and 2.3. First, a bank of fifteen continuous-time bandpass filters must be designed using the critical bandwidth data of Table 2.1. Let each filter have impulse response $h_k(t)\sin(\Omega_k t)$, where Ω_k is the center frequency, and $h_k(t)$ is the set of window functions defined by:

$$h_k(t) = \beta_k t^2 e^{-\alpha_k t}, \quad 0<t$$

$$= 0, \text{ otherwise,} \tag{2.23}$$

for k=1,2,...,15. The delay τ, as shown in Fig. 2.2, will be neglected. The Laplace transform (see Appendix A) of each window function is:

$$LT\{h_k(t)\} = 2\beta_k(s-\alpha_k)^{-3}. \tag{2.24}$$

The bandpass filters can be designed to have unity gain at center frequency by choosing $\beta_k=(\alpha_k)^3$. A gain factor of two is lost in the process of converting the lowpass window into a bandpass filter. The Laplace transform of each normalized window function can thus be written as:

$$LT\{h_k(t)\} = 2[(s+\alpha_k)-\alpha_k]^{-3}. \tag{2.25}$$

Frequency domain characteristics of the windows are shown in Fig. 2.10. Each window has a 3dB bandwidth of $.509\alpha_k$ rad/sec, so each bandpass filter has a 3dB bandwidth of $1.018\alpha_k$ rad/sec. Specific values for Ω_k and α_k can be obtained from Table 2.1. For example, $\Omega_1=2\pi x250$ rad/sec, and $\alpha_1=(2\pi x100)/1.018$ sec^{-1}.

The impulse invariant method can now be applied to obtain a digital implementation. Let T represent the digital system sampling period in seconds. The window functions of Equation 2.23 are transformed as:

$$h_k(n) = (\alpha_k)^3 n^2 T^3 e^{-\alpha_k nT}, \quad 1 \leq n,$$

$$= 0, \text{ otherwise,} \tag{2.26}$$

where an additional factor of T has been included to compensate for the analog to digital transformation (*Oppenheim and Schafer [31]*). The window functions of Equation 2.26, which have rational z-transforms (see Appendix A and Section C.2), can be written in the form

$$h_k(n) = \sum_{\psi=1}^{\Psi_k} p_k(\psi) h_k(n-\psi) + \sum_{r=1}^{R_k} q_k(r)\delta(n-r) \tag{2.27}$$

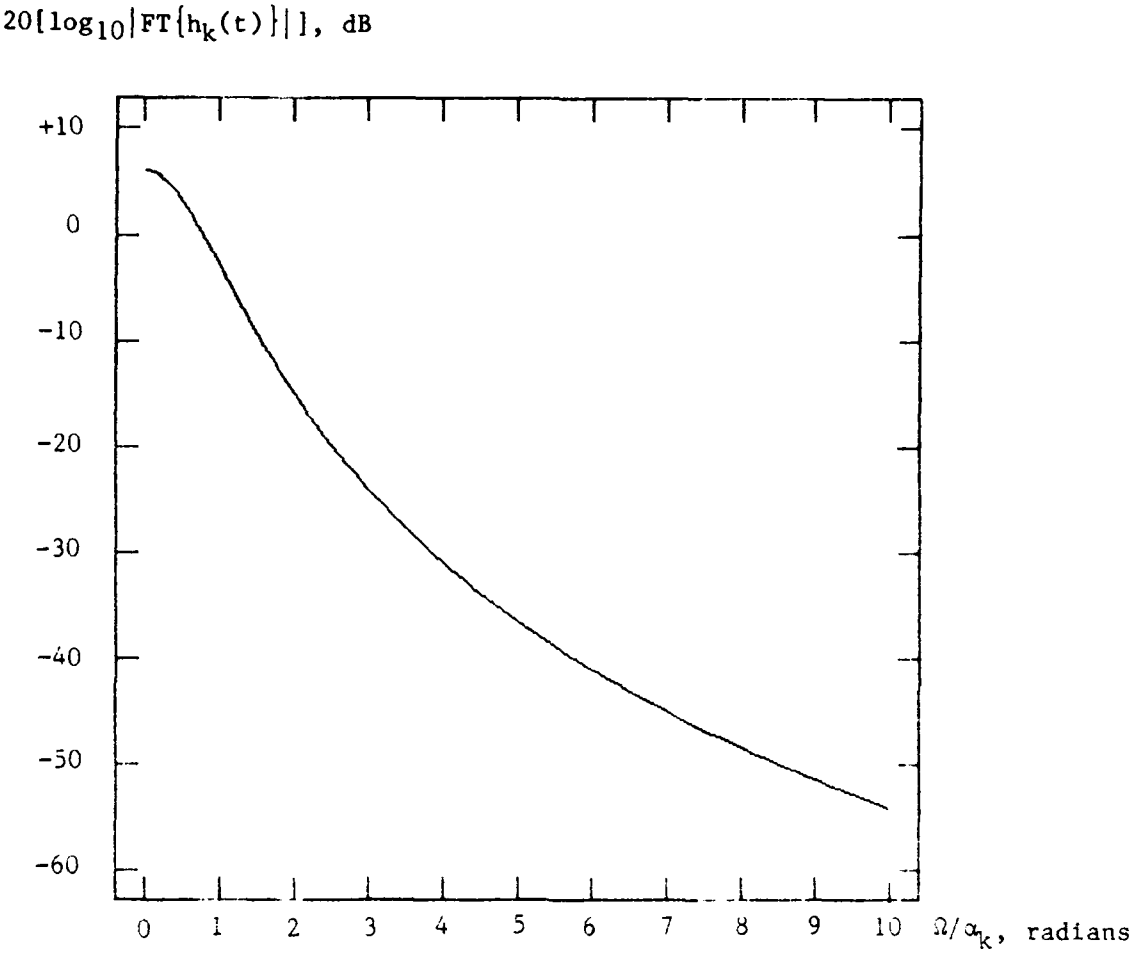


Figure 2.10: Window Function Frequency Characteristics

by choosing

$$i=1,\ \Psi_k=3,\ R_k=2,$$

$$q_k(1)=(\alpha_k T)^3 e^{-\alpha_k T},\ q_k(2)=(\alpha_k T)^3 e^{-2\alpha_k T},$$

$$p_k(1)=3e^{-\alpha_k T},\ p_k(2)=-3e^{-2\alpha_k T},\ \text{and}\ p_k(3)=e^{-3\alpha_k T}. \qquad (2.28)$$

Substitution of the Infinite-duration Impulse Response (IIR) window function defined by Equation 2.27 into Equation 2.22 yields a recursive formula for the GSTFT (Rabiner and Schafer [3]):

$$X_n(e^{j\omega_k}) = \sum_{\psi=1}^{\Psi_k} p_k(\psi) X_{n-\psi}(e^{j\omega_k}) + \sum_{r=i}^{R_k} q_k(r) x(n-r) e^{-j\omega_k(n-r)}. \qquad (2.29)$$

Note that the recursive GSTFT is computationally efficient for small values of Ψ_k and R_k. An implementation suitable for real-time applications is presented in Section C.2. Values for α_k and ω_k, $k=1,2,\ldots,15$, are obtained from Table 2.1 via the formulas $\alpha_k=(6.17)$(Critical Bandwidth in Hz) and $\omega_k=(2\pi T)$(Center Frequency in Hz).

Each bandpass filter of the digital F/D bank implemented via the GSTFT magnitude squared must meet two requirements. First, each filter must have a center frequency which is greater than its bandwidth. Second, each filter must have a center frequency less than pi minus the bandwidth. Since the analog filters of Table 2.1 meet the first requirement, so do the corresponding digital filters. The second requirement depends upon the sampling period T. In terms of analog filter parameters, the sum of center frequency and critical bandwidth

(both in Hz) must be less than 1/2T for each filter. The value T=.0001 second (ie., a 10 KHz sampling rate) is used to ensure that the second requirement is met.

Fig. 2.11 shows the F/D bank response to an impulse input applied at t=.0032 sec. The figure has linear amplitude and time scales. The graph of each F/D subsystem output, or "channel," has been normalized to the same peak value. Apart from a scale factor, the graphs of Fig. 2.11 are comparable in shape and duration to PST histogram envelopes (refer to Fig. 2.4). The impulse response of each F/D subsystem is proportional to $[h_k(t)]^2$.

When the F/D bank input is a sine wave, the output of each F/D subsystem is a constant. The graphs of Fig. 2.12 were obtained from average steady-state sinusoidal response measurements for each F/D subsystem, and these graphs correspond well with the critical bandwidth filter bank parameters given in Table 2.1. Since the F/D bank was not designed to match physiological tuning curves, the graphs of Fig. 2.11 do not possess the steep skirts exhibited by tuning curves. However, it is possible to obtain a better match to tuning curve data by using a different window function as discussed in Section 2.2.

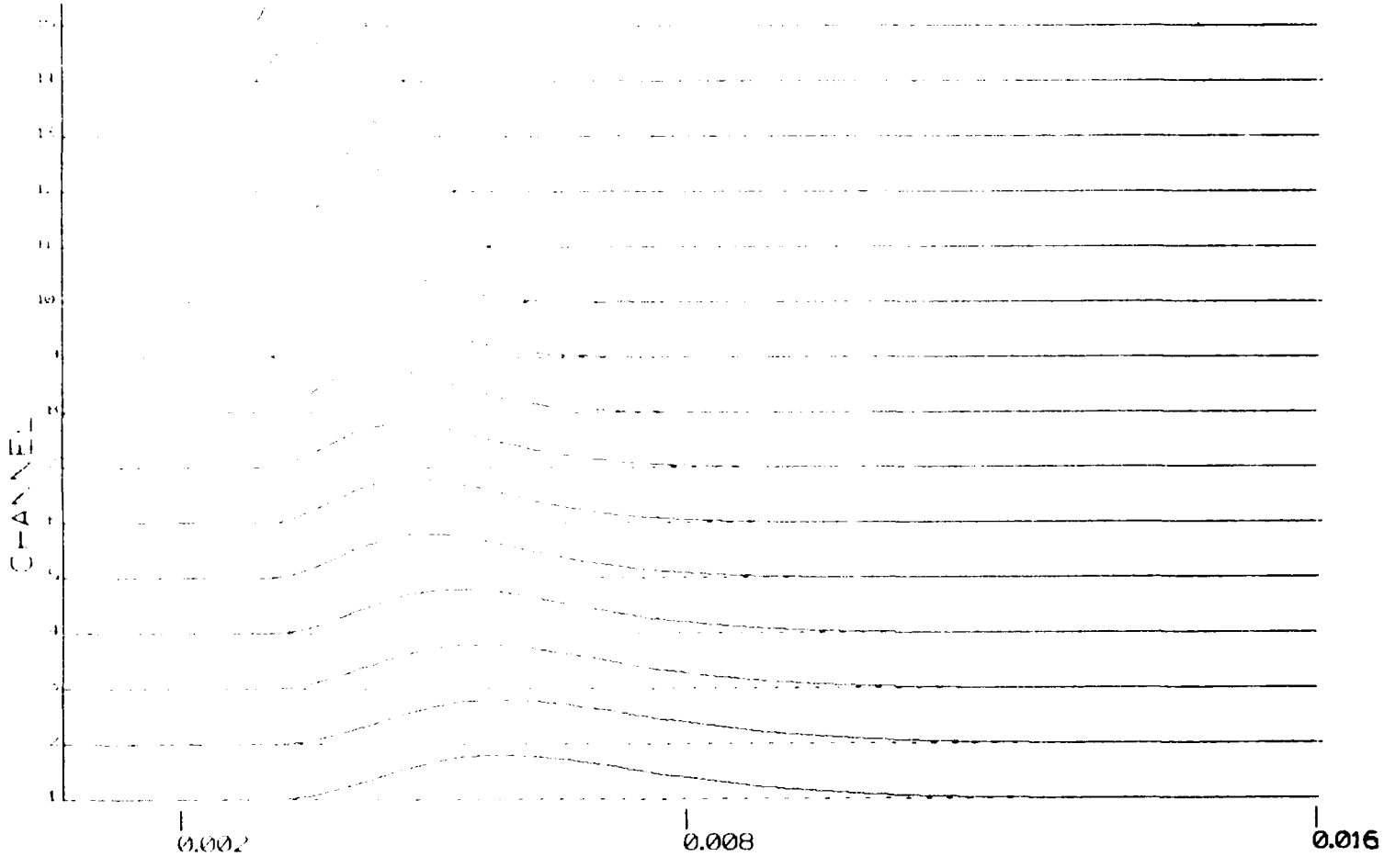


Figure 2.11: F/D Bank Impulse Response (Normalized)

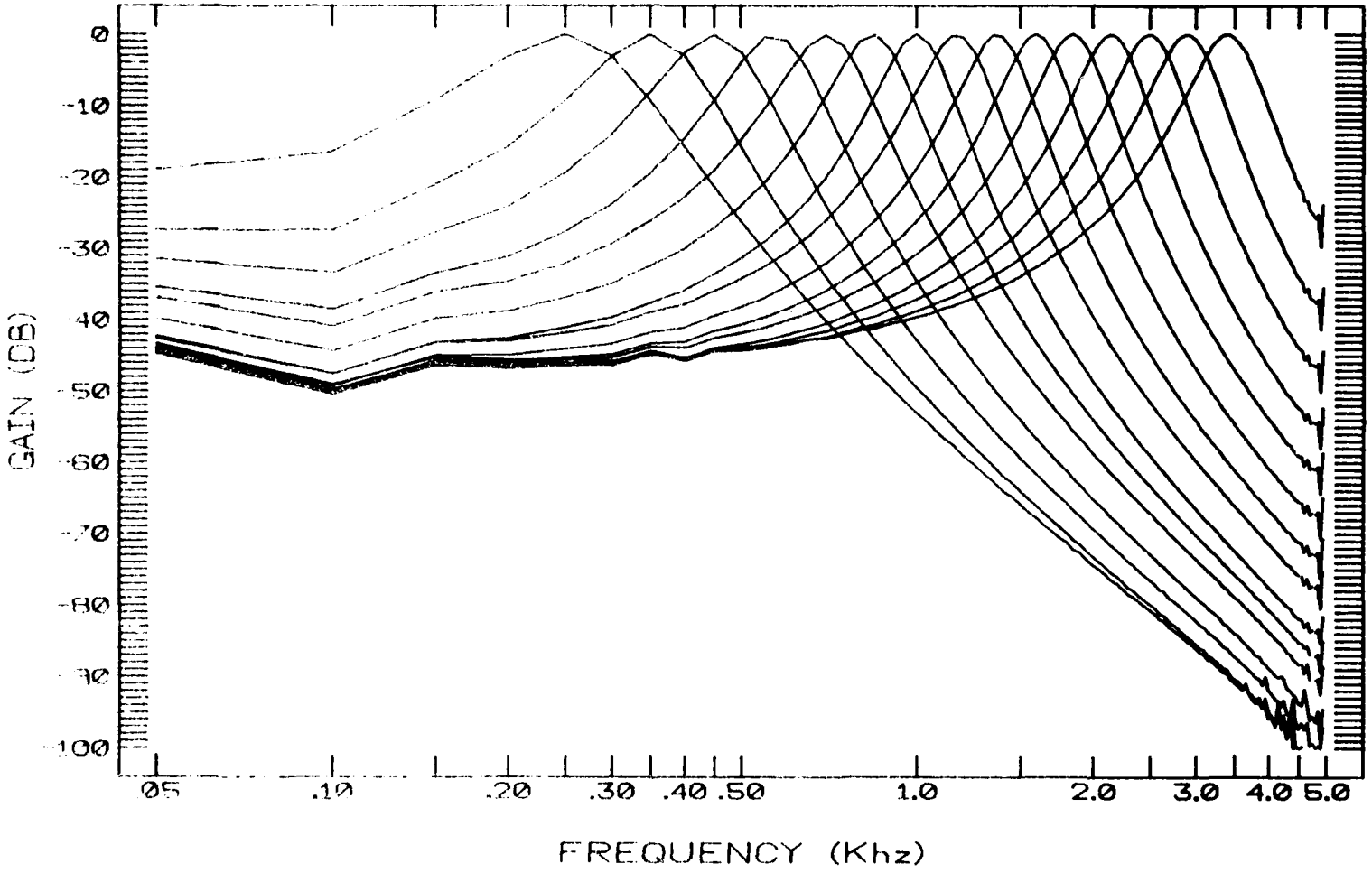


Figure 2.12: F/D Bank Sinusoidal Response

2.5 SHORT-TIME ENERGY

Short-Time Energy (STE) is a quantity which will prove useful in signal synthesis, as described in Chapter 3. For the discrete-time case, STE is defined by (Rabiner and Schafer [3]):

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h_0(n-m), \tag{2.30}$$

where $h_0(n)$ is the STE window function. Since the energy E_n must be non-negative for all real sequences $x(n)$, including $x(n)=\delta(n-n_0)$ for any integer n_0, the STE window function must be non-negative; ie., $h_0(n)\geq 0$ for all n. Note that the set of GSTFT window functions $h_k(n)$, $k=1,2,\ldots,K$, need not be non-negative in general.

A block diagram of the STE computation is shown in Fig. 2.13. A comparison of Figs. 2.8 and 2.13 reveals that the STFT (or GSTFT) magnitude squared essentially computes the STE within a given frequency band.

The STE can be computed recursively if $h_0(n)$ has a rational z-transform:

$$E_n = \sum_{\psi=1}^{\Psi_0} p_0(\psi)E_{n-\psi} + \sum_{r=i}^{R_0} q_0(r)x^2(n-r) \tag{2.31}$$

For example, let

$$h_0(n) = (\alpha_0 T)^3 n^2 e^{-\alpha_0 nT}, \quad 1 \leq n,$$

$$= 0, \text{ otherwise.} \qquad (2.32)$$

Choosing $\alpha_0 = 827$ results in a lowpass filter with a 3dB bandwidth of 67 Hz. The set of coefficients p_0 and q_0 are defined by Equations 2.28 with $k=0$.

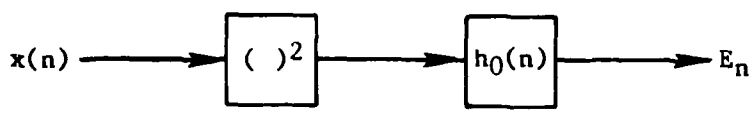


Figure 2.13: Short-Time Energy Computation

2.6 MINIMUM SAMPLING RATES

Although data reduction is not an essential part of an auditory model, it is often desirable to reduce the amount of data from speech analysis systems for practical purposes. A modest amount of data reduction can be achieved by sampling the STE and F/D bank outputs. If desired, the original outputs can be approximately recovered by passing the samples through an appropriate smoothing filter. A smoothing filter with positive impulse response (see Section B.2.3) can be used to ensure that the upsampled smoothed data is always positive. This downsampling/upsampling approach is also known as decimation and interpolation (Rabiner and Schafer [3]).

The output of each F/D subsystem is bandlimited to twice the window function bandwidth for that subsystem, so each output must be sampled at a rate which is greater than four times the corresponding window function bandwidth. The STE must be sampled at a rate which is greater than twice the STE window function bandwidth. Each output may, in general, be sampled at a different rate.

Since the F/D subsystems are implemented via the GSTFT magnitude squared, the sampling rates for F/D subsystem outputs also apply to GSTFT magnitude squared functions. Once the GSTFT magnitude squared has been sampled, any invertible operation such as square root or logarithm can be applied to the data. For non-negative numbers, knowledge of the square root or logarithm of a number is the same as knowledge of the number itself. It follows from the results of Sections B.3.6 and B.3.8 that GSTFT magnitude (as opposed to magnitude squared) functions are not bandlimited in general. The minimum sampling rate is therefore determined by the magnitude squared functions, but is equally applicable to magnitude or log magnitude functions, even though such functions may not be bandlimited.

Note that the minimum sampling rate requirements were derived from system theory considerations, and conditions for reconstruction of the original signal from the GSTFT magnitude data are not considered in this chapter (see Chapter 3). The minimum sampling rate arises when each channel is examined independently, and a sampling rate is determined which accurately preserves all available information in each channel. When the complete analysis system is considered, however, channels may overlap and contain redundant information. The overall sampling rate required for signal reconstruction may therefore be less than the product of the number of channels and the sampling rate per channel determined from system theory considerations.

2.7 CONCLUSION

In this chapter, it was shown that the peripheral auditory system can be roughly modeled as a F/D bank. To obtain a speech analysis system based on perception, a model structure determined by physiological data from animals was combined with model parameters determined by perceptual experiments performed on humans. It was shown, via a new relationship, that a F/D subsystem of the desired type can be implemented using the STFT magnitude squared. Further applications of this relationship are described in Appendix D. A generalized version of the STFT magnitude squared was used to implement the speech analysis system based on the simplified auditory model, and a STE function was also computed. Minimum sampling rates for the STE and F/D outputs have been specified.

CHAPTER 3

SPEECH SYNTHESIS SYSTEM

3.1 INTRODUCTION

This chapter describes a speech synthesis system which reconstructs a signal from spectral magnitude data, as provided by the analysis system of Chapter 2. Apart from an overall sign factor, the synthesis system can obtain exact signal reconstruction in the absence of data modification. It will be shown in Chapter 4 that the system also performs well given modified data. Only the discrete-time case will be discussed.

The overall speech analysis/synthesis system is depicted in Fig. 3.1. The signal x(n) is analyzed by a Filter/Detector (F/D) bank, which is implemented via the Generalized Short-Time Fourier Transform (GSTFT) magnitude squared as described in Chapter 2. An optional Short-Time Energy (STE) constraint may also be computed. The GSTFT magnitude squared and STE values are subjected to an analysis transformation A. The analysis transformation may consist of lowpass filtering, downsampling, logarithmic operations, or a variety of processes such as principal components analysis (Chu [16]). The analysis transformation may also include a delay in each channel which allows the impulse responses of Fig. 2.11 to attain their peak values simultaneously. Such delays are useful for data display purposes. The resulting data is sent through a transmission channel. At the channel output, received values are subjected to a synthesis transformation S. The synthesis transformation may consist of exponentiation, upsampling, lowpass filtering, or other operations. It will be assumed that the synthesis transformation attempts to reverse effects of the analysis transformation. Thus, the synthesis transformation produces modified data values which approximate the original values, ie., $|\tilde{X}_n(e^{j\omega_k})|^2 \cong |X_n(e^{j\omega_k})|^2$ and $\tilde{E}_n \cong E_n$. Finally, a sequence $\hat{x}(n)$ is reconstructed from the modified data. Let $\hat{X}_n(e^{j\omega_k})$ denote the GSTFT and $\hat{E}_n$ denote the STE of $\hat{x}(n)$. The sequence $\hat{x}(n)$ may be reconstructed by choosing values so that $|\hat{X}_n(e^{j\omega_k})|^2$ and $\hat{E}_n$ match the available data

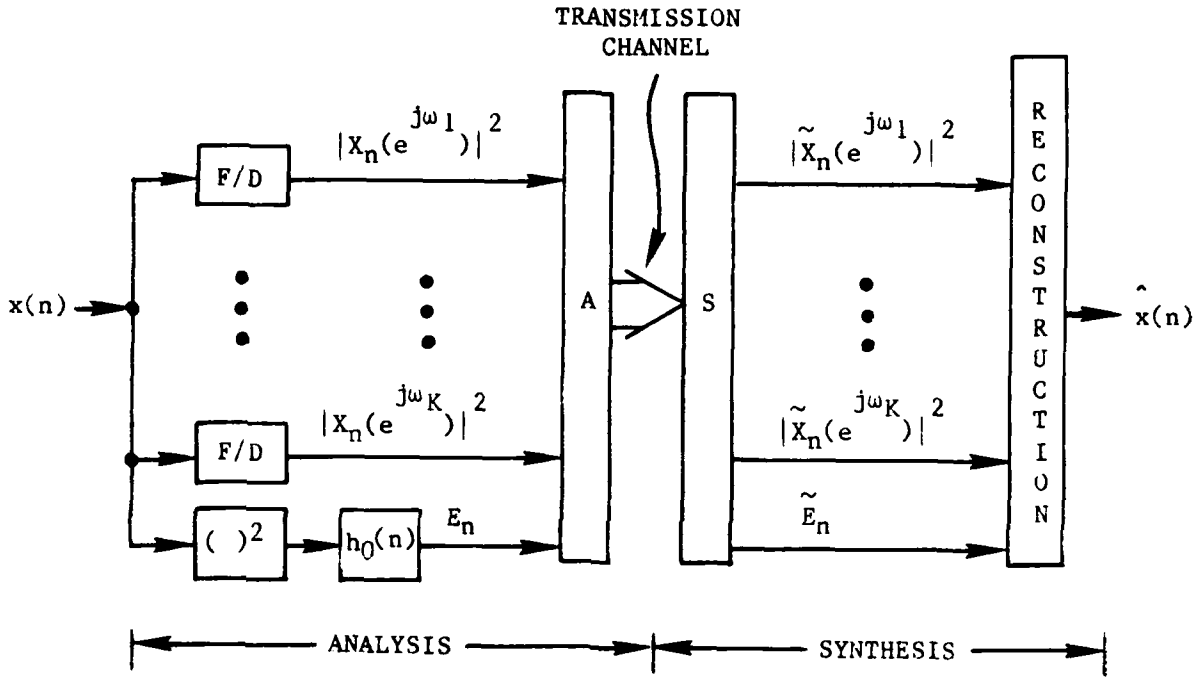


Figure 3.1: Overall Speech Analysis/Synthesis System

$|\tilde{X}_n(e^{j\omega_k})|^2$ and $\tilde{E}_n$. In this sense, the reconstructed signal $\hat{x}(n)$ approximates the original signal x(n). The reconstruction process is illustrated in Fig. 3.2. Note that the reconstruction process contains a model of the analysis system. Signal generation is accomplished by a set of equations which will be derived in Section 3.3.2.

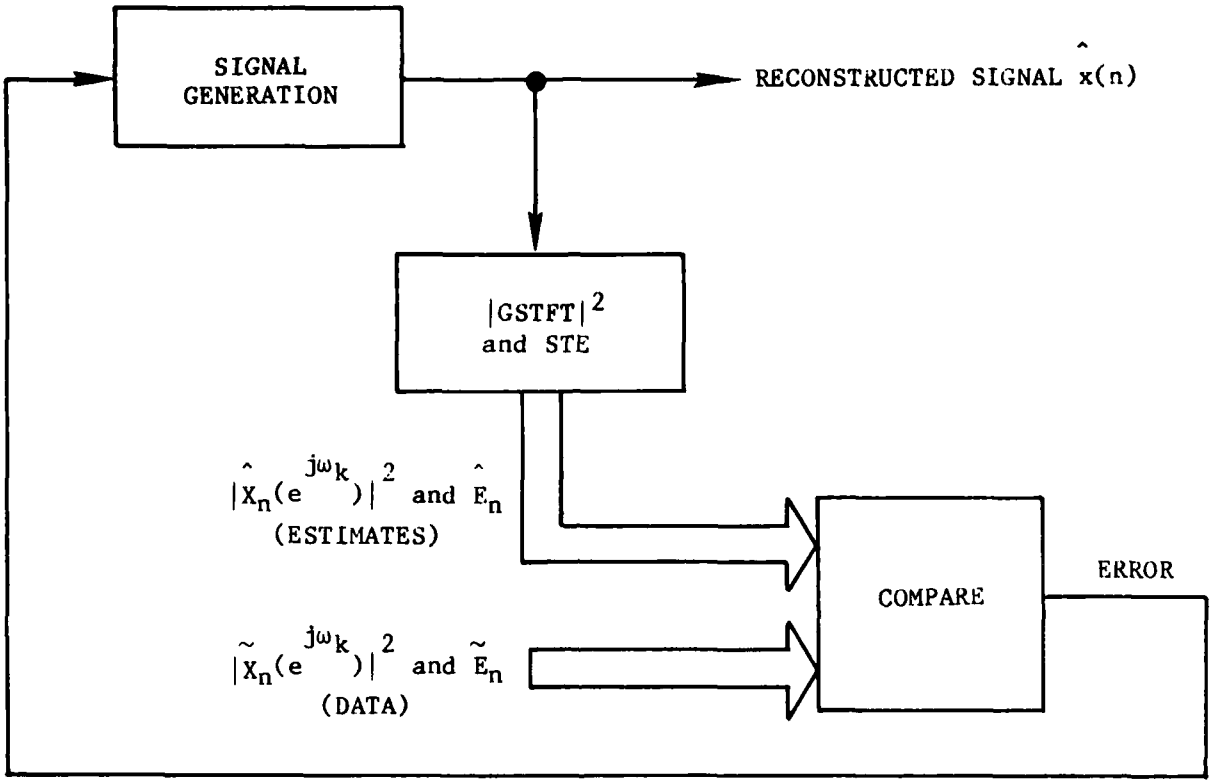


Figure 3.2: Reconstruction Process

3.2 ANALYSIS/SYNTHESIS SYSTEM DESIGN GUIDELINES

Although a signal can theoretically be recovered from the unmodified GSTFT magnitude squared (as will be shown in Section 3.3), several guidelines must be applied to design a practical analysis/synthesis system. These guidelines are a consequence of the F/D and GSTFT magnitude squared equivalence described in Chapter 2.

3.2.1 SHORT-TIME ENERGY

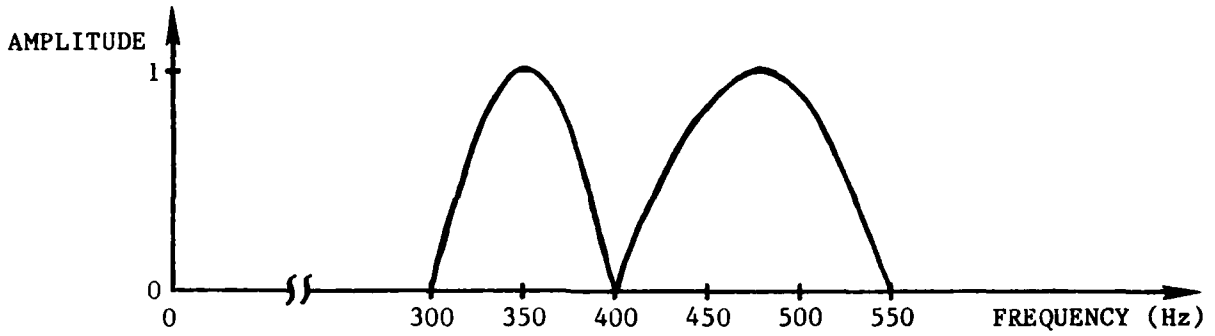
Under certain conditions, unwanted out-of-band components may be introduced by the reconstruction process if STE is not used. To illustrate, let the F/D bank analyzer of Fig. 3.1 examine the 200-3675 Hz frequency region. Information about other frequency components of x(n) is not transmitted through the channel. Assume $|\hat{X}_n(e^{j\omega_k})|^2$ exactly matches the data $|\tilde{X}_n(e^{j\omega_k})|^2$. For reconstruction based on magnitude information alone, nothing prevents the reconstructed signal $\hat{x}(n)$ from having large components at low frequencies (below 200 Hz) or high frequencies (above 3675 Hz). Such components could be eliminated by bandpass filtering $\hat{x}(n)$, but the reconstruction process must then employ a wide dynamic range to maintain a small signal with an arbitrarily large offset.

Although there are many ways to eliminate out-of-band components from the reconstructed signal, use of STE has proven most practical. As

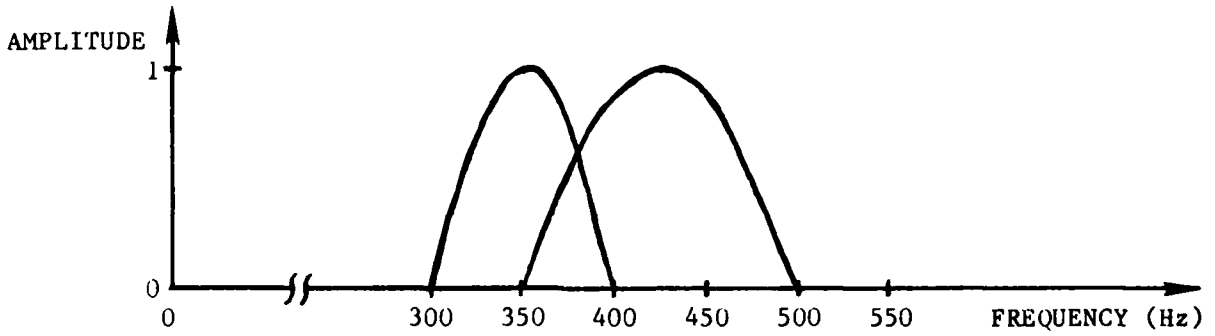
long as the original signal x(n) has been bandpass filtered to reject components outside the F/D bank analysis range, the STE constraint prevents out-of-band components from entering the reconstruction process. In cases where little data modification is ir olved, the STE constraint is unnecessary if some information about out-of-band components is allowed through the transmission channel. This occurs when non-ideal bandpass filters are used (refer to Fig. 2.12 for examples of appropriate filter characteristics). Thus, it is possible to achieve exact signal reconstruction without STE (see Anderson and Searle [32] for examples). For most practical applications, however, use of STE is recommended.

3.2.2 BANDPASS FILTER CHARACTERISTICS

If the bandpass filter frequency-domain characteristics do not meet certain overlap and shape requirements, then practical signal reconstruction is impossible. For example, consider a bank consisting of two F/D subsystems. Let the bandpass filters have non-overlapping frequency characteristics as shown in Fig. 3.3a. Assume that a sinusoidal tone burst in the 350-400 Hz frequency range is fed into the F/D bank, and the tone burst is of sufficient duration that the F/D outputs reach a steady-state value. In the steady-state condition, one F/D output is a constant positive value while the other is essentially zero (see Section B.3.7). The amplitude and frequency of the tone burst, which are two independent parameters of interest, cannot be determined from the single non-zero F/D output even if the filter characteristics



(a) Non-Overlapping Filters



(b) Overlapping Filters

Figure 3.3: Bandpass Filter Characteristics

are known. Steady-state F/D outputs will be identical for a variety of input amplitudes and frequencies. It can only be determined that the sinusoidal input frequency lies within a particular filter passband, and the amplitude is indeterminite. In theory, exact signal reconstruction can be achieved from unmodified F/D outputs if transient as well as steady-state values are examined. However, slight modifications (such as truncation error) are always present in actual systems, and practical reconstruction cannot be achieved from F/D banks using non-overlapped bandpass filters.

When overlapping bandpass filter characteristics are used, as shown in Fig. 3.3b, amplitude and frequency of an input sinusoid can easily be determined from the steady-state F/D outputs. For example, the frequency can be obtained from a ratio of the F/D outputs. The amplitude can then be determined from either bandpass filter characteristic. Thus, it is not necessary to rely on transient or low-level components to achieve reconstruction when overlapped filters are used. The need for overlapped filters in speech analysis systems has also been noted by Klatt [33].

It should be noted that bandpass filter overlap is necessary, but not sufficient, for a practical analysis/synthesis system. For example, if the bandpass filters are overlapped but possess both constant passband gain and steep skirts, then steady-state F/D outputs will be identical for a range of tone burst frequencies. The speech analysis system based on perception uses overlapped filters which do not have constant passband gain (see Fig. 2.12). It will be demonstrated in Chapter 4 that such a configuration performs well in a practical analysis/synthesis system.

3.2.3 TRANSMISSION CHANNEL DATA RATE

Since the transmission channel data rate (see Fig. 3.1) must be chosen in accordance with the bandpass filter characteristics, this rate is affected by the filter overlap requirement. For example, assume that the original signal is sampled at a 10 KHz rate, and a non-overlapped bank of bandpass filters is used to cover the full 0-5 KHz frequency range. Each F/D subsystem output must be sampled at a rate which is twice the associated bandpass filter bandwidth (see Section 2.6), leading to an overall data rate of 10 KHz (not including STE). When an overlapped filter bank is used, the required minimum transmission channel data rate is doubled (ie., 20 KHz). Of course, if the full range of possible frequencies is not covered by the F/D bank, then the required transmission channel data rate is correspondingly less.

It should be noted that the transmission channel data rate discussed in this section is based on system theory considerations, and does not consider the possibility of efficient waveform encoding to achieve data reduction. Data reduction is discussed in Section 5.3.

3.3 GENERAL SYNTHESIS EQUATIONS

In this section, it is shown that (apart from an overall sign factor) right-sided sequences can be exactly reconstructed from the GSTFT magnitude squared. Left-sided sequences can similarly be reconstructed when appropriate initial conditions are specified. The algorithms presented in this section are theoretically capable of performing signal reconstruction whether or not the practical guidelines of Section 3.2 are followed. Thus, in order to obtain a practical analysis/synthesis system, the guidelines of Section 3.2 are a prerequisite to application of the reconstruction algorithms. Note that synthesis equations of a general form are derived in this section. The procedure by which these equations are applied to a specific case is described in Section 3.4.

3.3.1 PLAUSIBILITY ARGUMENT

A simple approach described by Nawab, Quatieri, and Lim [34] can be used to recover a sequence from its GSTFT magnitude squared. Although this approach does not employ the reconstruction process depicted in Fig. 3.2, it serves to illustrate the issues involved in signal reconstruction from magnitude and to motivate the practical approach presented in Section 3.3.2. STE will not be used in this section.

Assume that two different F/D subsystems are implemented via the GSTFT and each Finite-duration Impulse Response (FIR) window function $h_k(n)$, k=1 or 2, is nonzero only for $0 \leq n \leq M_k-1$. It follows from Equation 2.22 that the GSTFT magnitude squared can be written as:

$$|X_n(e^{j\omega_k})|^2 = a_k x^2(n) + b_k(n)x(n) + c_k(n), \tag{3.1}$$

where

$$a_k = [h_k(0)]^2, \tag{3.2}$$

$$b_k(n) = 2h_k(0) \sum_{m=1}^{M_k-1} x(n-m)h_k(m)\cos(\omega_k m), \tag{3.3}$$

and

$$c_k(n) = \left| \sum_{m=1}^{M_k-1} x(n-m)h_k(m)e^{j\omega_k m} \right|^2. \tag{3.4}$$

Therefore,

$$x(n) = (1/2a_k)\{-b_k(n) \pm \sqrt{[b_k(n)]^2 - 4a_k[c_k(n)-|X_n(e^{j\omega_k})|^2]}\} \quad (3.5)$$

Note that care must be taken to ensure the quantity under the square root sign is always positive.

To illustrate the signal reconstruction process, assume $x(n)=0$ for $n<0$. It follows that $b_k(0)=c_k(0)=0$, and

$$x(0) = \pm |X_0(e^{j\omega_k})|/h_k(0). \quad (3.6)$$

Thus the output from either F/D subsystem may be used to determine the first reconstructed value within a sign factor. The positive value for $x(0)$ may be arbitrarily chosen, as choice of the negative value only changes the reconstructed sequence by an overall sign factor. Given the value of $x(0)$, values of $b_k(1)$ and $c_k(1)$ can be computed. Note that $b_k(n)$ and $c_k(n)$ are always computed using previously reconstructed signal values. Given $|X_1(e^{j\omega_k})|^2$ for two F/D subsystems appropriately spaced in frequency, the value of $x(1)$ can be determined using Equation 3.5. Each of the two F/D subsystems yields two possible values for $x(1)$, and the ambiguity is resolved by choosing the solution which is consistent with both F/D outputs. Given $x(0)$ and $x(1)$, the value for $x(2)$ can be determined, and so forth to reconstruct the entire sequence.

This simple reconstruction algorithm is subject to many practical difficulties. First of all, the reconstructed sequence may not be unique. Recall that the window functions $h_k(n)$, k=1 or 2, are nonzero for

$0 \leq n \leq M_k - 1$ (otherwise, uniqueness problems may be caused by "gaps" in the window function, as described by Nawab [35]). The reconstructed sequence is unique, to within an overall sign factor, unless a sequence of zero values having length $\{M_k - 1\}_{max}$ or more is encountered in the data. A sign ambiguity is introduced whenever such a sequence of zeros is encountered. For example, Fig. 3.4 shows four possible reconstructed sequences which can result when the two window functions are of length four or less. Studies suggest that such effects may not be important for speech if the analysis uses at least two F/D subsystems with impulse response duration of 10 milliseconds or more (Warren and Wrightson [36]; Flanagan and Guttman [37]). In any case, the multiple sign ambiguity problem can be alleviated by use of Infinite-duration Impulse Response (IIR) windows.

Another problem with the simple algorithm is its inability to perform reconstruction from modified data. Slight modifications such as truncation error can cause the two F/D outputs to produce contradictory results. For example, if data from one F/D indicates that x(1)=-2.000 or .919 while another F/D indicates x(1)=-2.002 or 2.832, then no consistent solution for the value of x(1) exists. It may be desirable, however, to use the value x(1)=-2.001 for future computations, although this value must be chosen by some algorithm which processes inconsistent results. Such inconsistent results can be treated in an organized manner by defining an error criterion, as will be shown in Section 3.3.2. An error criterion is used to determine the choice of an appropriate real value [illegible] information from many F/D outputs and ITE measurements.

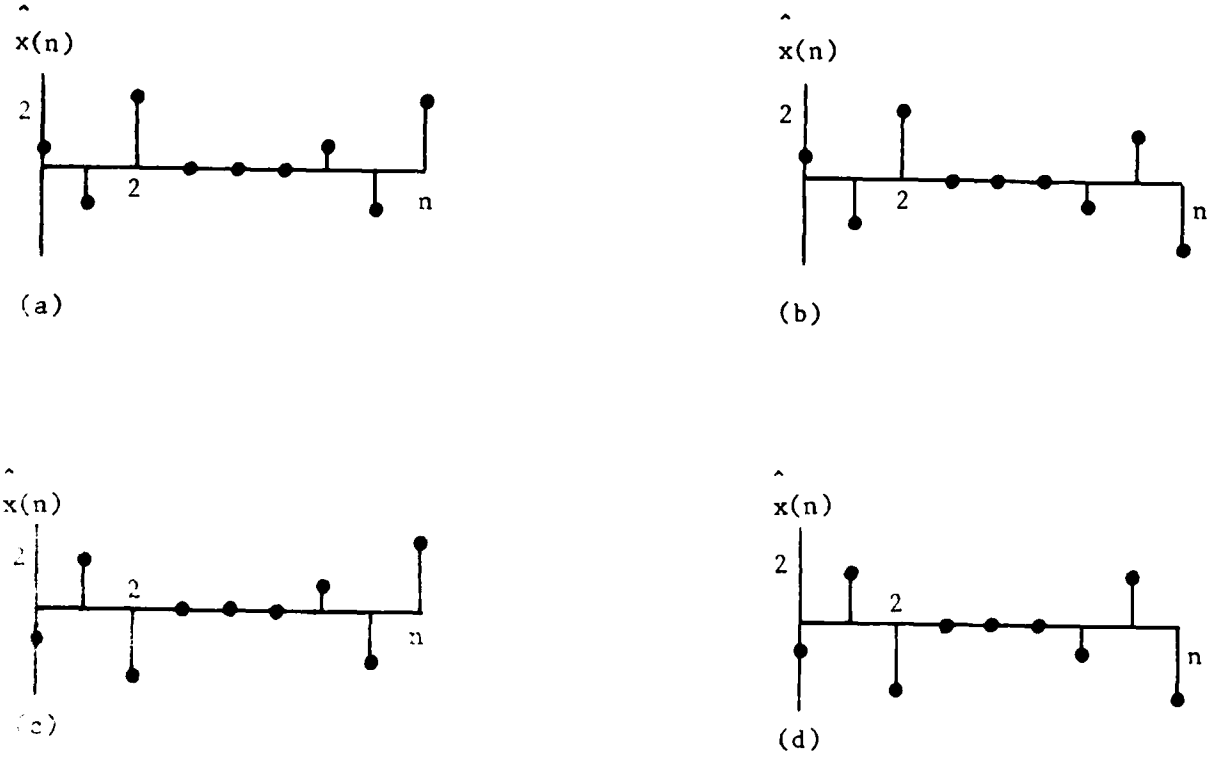


Figure 3.4: Four Reconstructed Sequence Possibilities

Next, note that the simple algorithm reconstructs x(n) given $|X_n(e^{j\omega_k})|^2$, ignoring information about x(n) contained in the future data $|X_m(e^{j\omega_k})|^2$ for $m=n+1,\ldots,n+M_k-1$. This observation suggests that an algorithm using non-causal processing, such as filtering with delay, may achieve superior results. For example, consider the reconstruction process of Fig. 3.2 which uses an error criterion. Assume that, once reconstructed, the value of a point is held constant. The feedback system of Fig. 3.2 reduces the error by changing only the value of $\hat{x}(n)$ at one specific time n. If previous points were not reconstructed exactly, the system attempts to compensate by changing the value of $\hat{x}(n)$ accordingly. Such a change may lead to further cumulative error, causing poor reconstruction. However, if previously reconstructed values can be modified on the basis of new information, the error can be distributed among a large number of points and reconstruction is improved. This non-causal approach is especially useful for reconstruction from modified spectra.

Note that since the simple reconstruction algorithm achieves exact reconstruction from only two F/D channel outputs, an overall data rate which is at least twice the sampling rate of x(n) can be used in the transmission channel of Fig. 3.1. This result is the same as that derived in Section 3.2.3. For data rates less than twice the sampling rate, exact reconstruction cannot be achieved in general (see Theorem 2.3 of Nawab [35]).

3.3.2 EQUATIONS FOR PRACTICAL SIGNAL RECONSTRUCTION

The simple reconstruction algorithm of Section 3.3.1 can be modified to obtain the practical algorithm shown in Fig. 3.2. The required modifications include use of an error criterion and non-causal processing.

To develop a practical algorithm, the GSTFT is rewritten in a more convenient form. For any integers ℓ and γ it follows from Equation 2.22 that:

$$\hat{X}_{n-\ell}(e^{j\omega_k}) = \hat{x}(n-\gamma)h_k(\gamma-\ell)e^{-j\omega_k(n-\gamma)} + \sum_{m\neq\gamma-\ell} \hat{x}(n-\ell-m)h_k(m)e^{-j\omega_k(n-\ell-m)}, \quad (3.7)$$

where the summation over $m\neq\gamma-\ell$ is defined as the summation from minus infinity to $\gamma-\ell-1$, plus the summation from $\gamma-\ell+1$ to infinity. Taking the magnitude squared of Equation 3.7 yields:

$$|\hat{X}_{n-\ell}(e^{j\omega_k})|^2 = \hat{a}_k\hat{x}^2(n-\gamma) + \hat{b}_k(n)\hat{x}(n-\gamma) + \hat{c}_k(n), \quad (3.8)$$

where

$$\hat{a}_k = [h_k(\gamma-\ell)]^2, \quad (3.9)$$

$$\hat{b}_k(n) = 2h_k(\gamma-\ell) \sum_{m\neq\gamma-\ell} \hat{x}(n-\ell-m)h_k(m)\cos[\omega_k(m-\gamma+\ell)], \quad (3.10)$$

and

$$\hat{c}_k(n) = \left| \sum_{m \neq \gamma - \ell} \hat{x}(n-\ell-m) h_k(m) e^{j\omega_k m} \right|^2. \qquad (3.11)$$

The GSTFT magnitude squared at any time $n-\ell$ can thus be expressed as a quadratic function of the sequence at any time $n-\gamma$. Note that $\hat{a}_k$, $\hat{b}_k(n)$, and $\hat{c}_k(n)$ are independent of $\hat{x}(n-\gamma)$. However, if a value of $\hat{x}(n-\gamma)$ and its corresponding value of $\hat{X}_{n-\ell}(e^{j\omega_k})$ are known, then it is easily verified from Equations 3.7 and 3.10 that:

$$\hat{b}_k(n) = 2h_k(\gamma-\ell)\left[\mathrm{Re}\left\{e^{j\omega_k(n-\gamma)} \hat{X}_{n-\ell}(e^{j\omega_k})\right\} - \hat{x}(n-\gamma)h_k(\gamma-\ell)\right]. \qquad (3.12)$$

Also, it follows from Equation 3.8 that:

$$\hat{c}_k(n) = \left|\hat{X}_{n-\ell}(e^{j\omega_k})\right|^2 - \hat{a}_k \hat{x}^2(n-\gamma) - \hat{b}_k(n)\hat{x}(n-\gamma). \qquad (3.13)$$

Equations 3.12 and 3.13 can often be computed more easily than Equations 3.10 and 3.11.

Using a similar approach for STE, it follows from Equation 2.30 that:

$$\hat{E}_{n-\ell} = \hat{a}_0 \hat{x}^2(n-\gamma) + \hat{c}_0(n), \qquad (3.14)$$

where

$$\hat{a}_0 = h_0(\gamma-\ell), \qquad (3.15)$$

and

$$\hat{c}_0(n) = \sum_{m \neq \gamma - \ell} h_0(m)\hat{x}^2(n-\ell-m). \tag{3.16}$$

Therefore,

$$\hat{c}_0(n) = \hat{E}_{n-\ell} - \hat{a}_0\hat{x}^2(n-\gamma) \tag{3.17}$$

For convenience, a weighted mean squared error criterion is chosen. The error is defined as:

$$\varepsilon(n) = \sum_{\ell=-\infty}^{\infty} \{ (\hat{E}_{n-\ell} - \tilde{E}_{n-\ell})^2 W_0(\ell)$$

$$+ \sum_{k=1}^{K} [|\hat{X}_{n-\ell}(e^{j\omega_k})|^2 - |\tilde{X}_{n-\ell}(e^{j\omega_k})|^2]^2 W_k(\ell) \}, \tag{3.18}$$

where $W_m(\ell)$, $m=0,1,\ldots,K$, and $-\infty<\ell<\infty$, is a weighting function. The weighting function specifies which data values contribute to the error at time n. Although weighting functions which vary with time or signal level can be used, such functions will not be considered here. When the weighting function is constant for all values of ℓ, the error is $\varepsilon(n)=\varepsilon_{total}$ where ε_{total} is a constant total error independent of n. Any reconstructed sequence $\hat{x}(n)$ which minimizes ε_{total} is an "optimum" solution in the mean squared error sense. In order to achieve reasonable results with less computation, a "local" sub-optimum error criterion may be preferable to the "global" optimum error criterion. A local error

criterion can be obtained by choosing a weighting function which is narrow in the ℓ dimension. Even when a sub-optimum error criterion is chosen, error minimization may require solution of an infinite number of simultaneous nonlinear equations if the window functions are infinitely long. It is not generally possible to solve such a set of equations with a finite amount of computation.

A practical sub-optimum approach to signal reconstruction, which avoids the problem of solving simultaneous nonlinear equations, can be obtained by holding all synthesized values constant with the exception of $\hat{x}(n-\gamma)$. An appropriate value of $\hat{x}(n-\gamma)$ can then be determined by substituting Equations 3.8 and 3.14 into 3.18 and setting $\partial\varepsilon(n)/\partial\hat{x}(n-\gamma)=0$. Under these conditions, $\varepsilon(n)$ is reduced by choosing $\hat{x}(n-\gamma)$ as a root of:

$$u_3\hat{x}^3(n-\gamma) + u_2\hat{x}^2(n-\gamma) + u_1\hat{x}(n-\gamma) + u_0 = 0, \tag{3.19}$$

where

$$u_3 = 2\sum_{\ell=-\infty}^{\infty}[(\hat{a}_0)^2W_0(\ell) + \sum_{k=1}^{K}(\hat{a}_k)^2W_k(\ell)], \tag{3.20}$$

$$u_2 = 3\sum_{\ell=-\infty}^{\infty}\sum_{k=1}^{K}\hat{a}_k\hat{b}_k(n)W_k(\ell), \tag{3.21}$$

$$u_1 = 2\sum_{\ell=-\infty}^{\infty} \hat{a}_0[\hat{c}_0(n) - \tilde{E}_{n-\ell}]W_0(\ell)$$

$$+ \sum_{\ell=-\infty}^{\infty} \sum_{k=1}^{K} \{[\hat{b}_k(n)]^2 + 2\hat{a}_k[\hat{c}_k(n) - |\tilde{X}_{n-\ell}(e^{j\omega_k})|^2]\}W_k(\ell), \quad (3.22)$$

and

$$u_0 = \sum_{\ell=-\infty}^{\infty} \sum_{k=1}^{K} \hat{b}_k(n)[\hat{c}_k(n) - |\tilde{X}_{n-\ell}(e^{j\omega_k})|^2]W_k(\ell). \quad (3.23)$$

Values for the reconstructed sequence can be generated by solving Equation 3.19 for the roots of a cubic expression. The real root which yields the smallest value of $\varepsilon(n)$ is chosen as the sequence value. Since cubics have one, two, or three distinct real roots, a real sequence value can always be found which satisfies the error criterion. Furthermore, a closed-form solution exists for computing the roots (CRC Standard Mathematical Tables [38]). In the actual implementation, double-precision computer arithmetic was used to obtain an accurate solution of the cubic expression. Such accuracy, however, is not required elsewhere in the reconstruction algorithm.

3.4 SPEECH SYNTHESIS PROCEDURE

Equations of Section 3.3.2 can be used to reconstruct speech signals from data produced by the analysis system of Chapter 2. Although the equations may be applied in many different ways, only one approach will be described in detail. This approach has been used to generate a number of examples, which are presented in Chapter 4.

From Equations 2.22 and 2.27 it follows that $X_n(e^{j\omega_k})$ contains information about $x(n-\gamma)$ for $\gamma \geq i$, where $i=1$ for the present application. For practical purposes, however, it is assumed that $X_n(e^{j\omega_k})$ contains significant information about $x(n-\gamma)$ only for the finite set of values $i \leq \gamma \leq \gamma_{max}$, where γ_{max} is some arbitrary integer. Therefore, if the values of $\hat{x}(n-\gamma)$ for $i \leq \gamma \leq \gamma_{max}$ are changed during the reconstruction process, then the values of $\hat{X}_{n-\ell}(e^{j\omega_k})$ for $0 \leq \ell \leq \gamma_{max}-i$ must also be changed accordingly. The value $\gamma_{max}=20$, which results in a 2 millisecond synthesis window, will be used throughout. Note that this value is not critical. Small values ($\gamma_{max}=3$) can be used to rapidly obtain exact reconstruction from unmodified spectral data, while large values commensurate with the maximum effective window length ($\gamma_{max}=100$) may improve the quality of reconstruction from modified data.

To completely specify the reconstruction error criterion, an appropriate weighting function $W_m(\ell)$, $m=0,1,\ldots,K$, and $-\infty \leq \ell \leq \infty$, must be chosen. The F/D outputs are bandlimited functions which do not generally change rapidly. Therefore, the weighting function can be chosen narrow

in the ℓ dimension. A weighting function which is wide in the ℓ dimension may be advantageous for reconstruction from highly modified data, but causes an increase in computation time and implementation complexity. Since the bandpass filters have normalized gains as described in Section 2.4.5, and are roughly of equal importance for speech intelligibility (Beranek [39]), the F/D weighting coefficients are equal. Let $W_k(\ell)=1$ for $\ell=0$ and $k=1,\ldots,K$, and $W_k(\ell)=0$ otherwise. With this choice of F/D weighting coefficients, an empirically determined energy weight $W_0(\ell)=.03$ for $\ell=0$, $W_0(\ell)=0$ otherwise, is appropriate. The energy weight is small because energy values are often large, and also because the energy function is intended as a constraint and not as an information-bearing element. The resulting error expression is:

$$\varepsilon(n) = (\hat{E}_n - \tilde{E}_n)^2 W_0 + \sum_{k=1}^{K} [|\hat{X}_n(e^{j\omega_k})|^2 - |\tilde{X}_n(e^{j\omega_k})|^2]^2, \qquad (3.24)$$

where $W_0=.03$ and $K=15$. The error given by Equation 3.24 is used for all reconstruction examples of Chapter 4 (see Anderson and Searle [32] for examples using a different weighting function). The total error can be computed as:

$$\varepsilon_{total} = \sum_{n=-\infty}^{\infty} \varepsilon(n). \qquad (3.25)$$

For comparison purposes, it is useful to define a total error which is normalized with respect to the original signal:

$$\varepsilon_{total,norm} = \varepsilon_{total} / \{ \sum_{n=-\infty}^{\infty} [(E_n)^2 W_0 + \sum_{k=1}^{K} |X_n(e^{j\omega_k})|^4] \}. \qquad (3.26)$$

The total normalized error does not change with input signal level, and provides a form of error-to-signal ratio (Griffin, Deadrick, and Lim [40]).

The synthesis procedure will now be described in detail. For convenience, assume $x(n)=0$ for $n<0$. The reconstructed sequence $\hat{x}(n)$, estimated GSTFT $\hat{X}_n(e^{j\omega_k})$, and estimated STE $\hat{E}_n$ are initially set to zero for all n. The synthesis procedure begins at any time $n<i$, where $i=1$ for the present application. The index n is incremented one point at a time, and each newly reconstructed point is used to update previously reconstructed values.

The first reconstruction step advances the time index, and updates estimated GSTFT and STE values based on available reconstructed sequence values. Previously calculated GSTFT and STE values which are unaffected by any changes in $\hat{x}(n-\gamma)$, $i<\gamma<\gamma_{max}$, are used as initial conditions for the update. Equations 2.29, 2.31, and the present values of $\hat{x}(n-\gamma)$ for $\gamma>i$ are used to generate GSTFT and STE estimates up to time n.

The present estimated value of $\hat{x}(n-i)$, which was set to zero during initialization, is likely to be in error. An improved estimate for $\hat{x}(n-i)$ is obtained by a procedure which will be described shortly. Improving the estimate for $\hat{x}(n-i)$ provides a reconstructed sequence value. Using this improved value, new values for previously reconstructed points can also be determined.

Due to the shape of the window functions, which have small initial values as shown in Fig. 2.11, many refinements are necessary in the estimates of $\hat{x}(n-\gamma)$ for small γ. To make refinements, all points other than one specified point are held constant, and the specified point is allowed to vary in a fashion which reduces the error. Thus, adjustments to $\hat{x}(n-\gamma)$ for large γ must not be made until the more recently reconstructed points are thoroughly corrected. Estimated values of the reconstructed points must therefore be refined in a certain order. To develop the examples shown in Chapter 4, the following order of refinement in values of $\hat{x}(n-\gamma)$ was used:

$$\begin{aligned}\gamma = {} & 1,2,1,2,1,2,1,2,1,2, \\ & 1,2,3,1,2,3,1,2,3,1,2,3,1,2,3, \\ & \text{etc.}, \\ & 1,\ldots,7,1,\ldots,7,1,\ldots,7,1,\ldots,7,1,\ldots,7, \\ & 1,\ldots,8,1,\ldots,9,1,\ldots,10,\text{etc.},1,\ldots,\gamma_{max},\end{aligned} \tag{3.27}$$

where $\gamma_{max}=20$. After this procedure has been performed to reconstruct one new point $\hat{x}(n-i)$ and adjust values of previously reconstructed points through $\hat{x}(n-\gamma_{max})$, the time index is incremented, GSTFT and STE estimates are updated based on the new sequence values, and the procedure is repeated to reconstruct the entire sequence.

To refine the estimate of any point $\hat{x}(n-\gamma)$, Equations 3.12, 3.13, and 3.17 are used with $\ell=0$ to obtain $\hat{b}_k(n)$, $\hat{c}_k(n)$, and $\hat{c}_0(n)$. Note that $\hat{a}_0$ and $\hat{a}_k$ are pre-computed constants which do not depend on the data. The contribution of the present sequence estimate $\hat{x}(n-\gamma)$ is now subtracted from the present GSTFT estimates $\hat{X}_n(e^{j\omega_k})$ and STE estimate $\hat{E}_n$ by using Equations 3.7 and 3.14. Next, u_0, u_1, and u_2 are computed from Equations 3.21, 3.22, and 3.23. Note that u_3 can be pre-computed, as shown in Equation 3.20. Equation 3.19 is solved, resulting in up to three new candidate estimates for $\hat{x}(n-\gamma)$. The first candidate is evaluated by adding its contribution to the GSTFT and STE estimates using Equations 3.7 and 3.14. The resulting error is evaluated using Equation 3.24. A similar procedure is applied to each remaining candidate, the one producing minimum error is chosen as the new estimated value for $\hat{x}(n-\gamma)$, and the corresponding GSTFT and STE estimates are retained. Note that, for a fixed time n, the error is reduced with each application of this procedure.

Note that the algorithm described in this section can be applied to reconstruction of right-sided sequences or other types of sequences for which appropriate initial conditions have been specified. If necessary, however, initial conditions may be generated by repeated application of the reconstruction equations for some fixed time n. Once the initial conditions have been established, n is incremented and the sequence is reconstructed.

Finally, it is worth noting that reconstruction can be performed directly from sampled data. For example, assume that only every other time-domain sample is available from the analysis. The synthesis can be advanced two time steps, rather than one step at a time, and smoothing accomplished by an order of refinement different than that described by Equation 3.27. Alternatively, a weighting function $W_m(\ell)$ which is nonzero only for $\ell=0$ and $\ell=2$ can be used in a modified version of the reconstruction algorithm. Although these approaches produce results comparable to those produced by simply smoothing the sampled data prior to reconstruction, they require considerably more computation time and are therefore less practical.

3.5 CONCLUSION

In this chapter, general guidelines for practical analysis/synthesis systems have been established. These guidelines indicate that STE (or some other constraint) must be used to prevent out-of-band components from dominating the reconstructed signal. The analysis must use an overlapped bandpass filter bank in which the filters do not possess both constant passband gain and steep skirts. The speech analysis/synthesis system based on perception meets these requirements.

In general, a transmission channel data rate which is twice the original sampling rate must be used to achieve exact signal reconstruction. However, if the F/D bank does not cover the full range of possible frequencies, then a lower rate can be used. Under these conditions, only signals within the range of the F/D bank can be reconstructed. Thus, unlike other systems which require an increase in transmission channel bandwidth when the sampling rate is increased, this system produces results which are independent of the original signal sampling rate.

The new signal reconstruction algorithm described in this chapter is presently the only one known which is capable of performing reconstruction from data produced by a critical bandwidth F/D bank. The algorithm is an extension of an algorithm described by Nawab, et al [34]. The extension introduces a weighted mean squared error criterion and non-causal processing to achieve practical results. The new algorithm is

applicable to systems using both IIR and FIR analysis filters, and exact reconstruction can be obtained in the absence of data modification. In the absence of substantial data modification, reconstruction can be accomplished in very little time by choosing a small value for γ_{max}. The algorithm can incorporate measurements of different types (such as Short-Time Energy), reconstruction can be accomplished from a limited range of frequencies, and contributions to error can be weighted according to frequency band if desired.

The new algorithm uses a sub-optimum reconstruction approach with a sub-optimum error criterion, and does not generally minimize the total error ε_{total}. However, it may not be possible to determine the optimum solution with a finite amount of computation when infinite-length window functions are involved. When the special case of an analysis using uniformly spaced constant-bandwidth FIR filters spanning the full frequency range is considered, other techniques are available which attempt to minimize total error (eg., Griffin and Lim [10]; Musicus [41]). The error criterion for the constant-bandwidth case, however, is not perception-based. Although the new algorithm presented in this chapter does not necessarily minimize ε_{total}, the error value $\varepsilon(n)$ is reduced with each refinement of the estimated sequence values. Note that the error criterion can be either local or global, but a local criterion is used to simplify the algorithm and reduce computation time.

CHAPTER 4

EXAMPLES

4.1 INTRODUCTION

In this chapter, operation of the speech analysis/synthesis system based on perception is demonstrated. Examples of tone bursts, tone pair bursts, synthetic vowels, and natural speech signals are analyzed, subjected to a short-time spectral modification, and synthesized. Although the analysis/synthesis system is actually implemented using a discrete-time approach, the examples are presented as continuous-time functions.

4.2 TONE BURST

Fig. 4.1 presents a 1 KHz tone burst of 32 millisecond duration. Since the tone burst is essentially a bandlimited signal, no pre-filtering was applied to suppress components outside the 200-3675 Hz frequency range. The tone burst was analyzed by the speech analysis system described in Sections 2.4.5 and 2.5. The resulting Filter/Detector (F/D) outputs, which are computed via the Generalized Short-Time Fourier Transform (GSTFT) magnitude squared, are shown in Fig. 4.2. The symbol "E" denotes Short-Time Energy (STE), and channel numbers correspond to the filter numbers of Table 2.1. The amplitude scale of Fig. 4.2 is logarithmic. A logarithmic scale is used in order to reveal features which might otherwise be obscured, and to approximate perceived loudness effects (Siebert [15]). After an initial transient, all F/D outputs reach a steady-state value, and a final transient occurs at the end of the tone burst. The highest value is attained in Channel 7 since this channel has a center frequency of 1 KHz. From Fig. 2.12 it follows that the steady-state level of Channel 1 is -55dB and Channel 15 is -43dB re Channel 7. Fig. 4.2 can be re-plotted to show log amplitude as a function of frequency with time as a parameter. Such a three-dimensional (3D) running spectrum plot is presented in Fig. 4.3.

The reconstruction algorithm described in Section 3.4 was applied to the data of Fig. 4.3, and the resulting signal is shown in Fig. 4.4. The reconstructed signal of Fig. 4.4 is indistinguishable from the original signal of Fig. 4.1, and is generally accurate to four significant digits. Thus, the GSTFT magnitude is a complete means of signal representation (apart from an arbitrary overall sign factor).

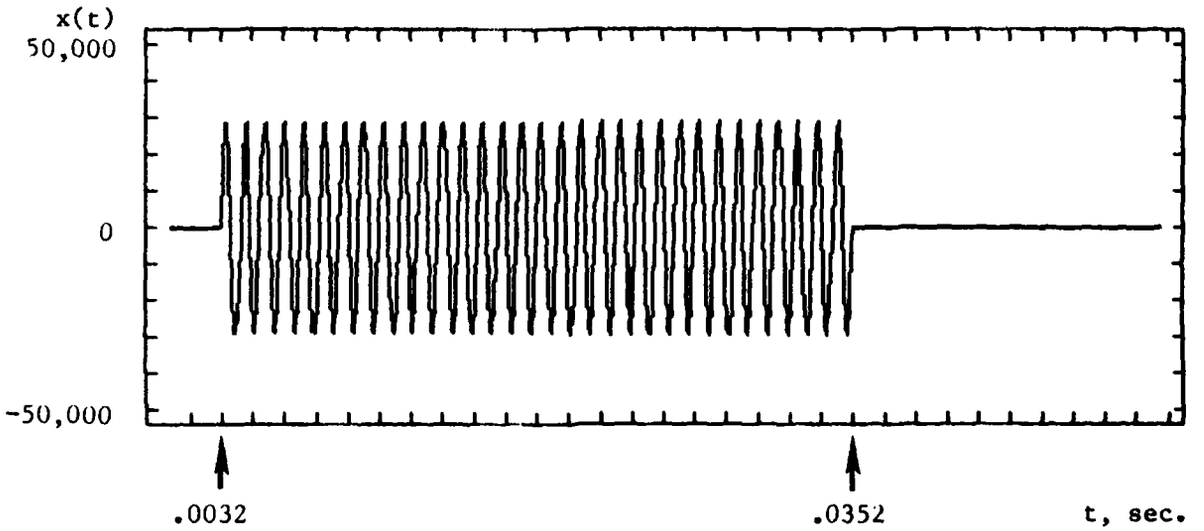


Figure 4.1: Original Signal (1000 Hz)

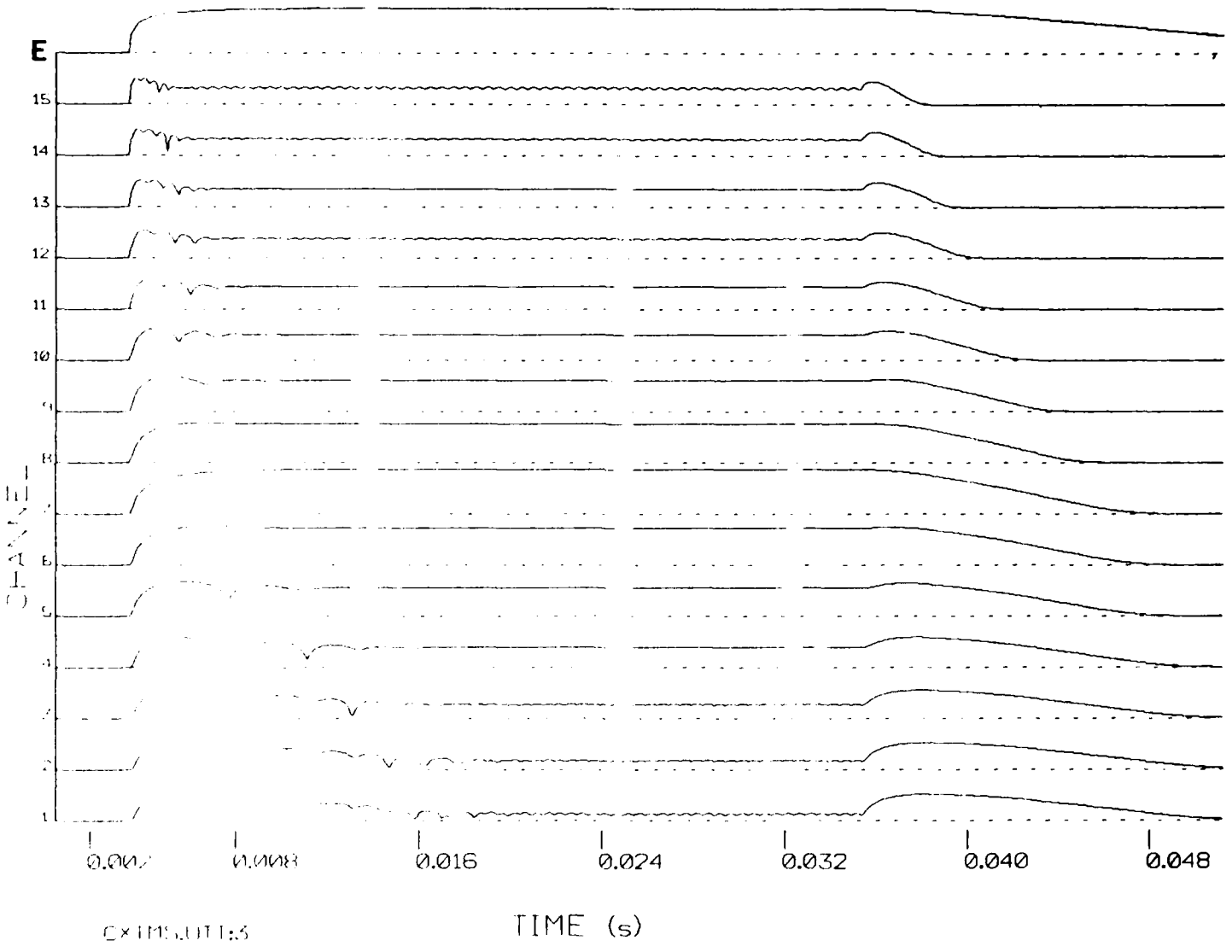


Figure 4.2: F/D Outputs (Log Amplitude)

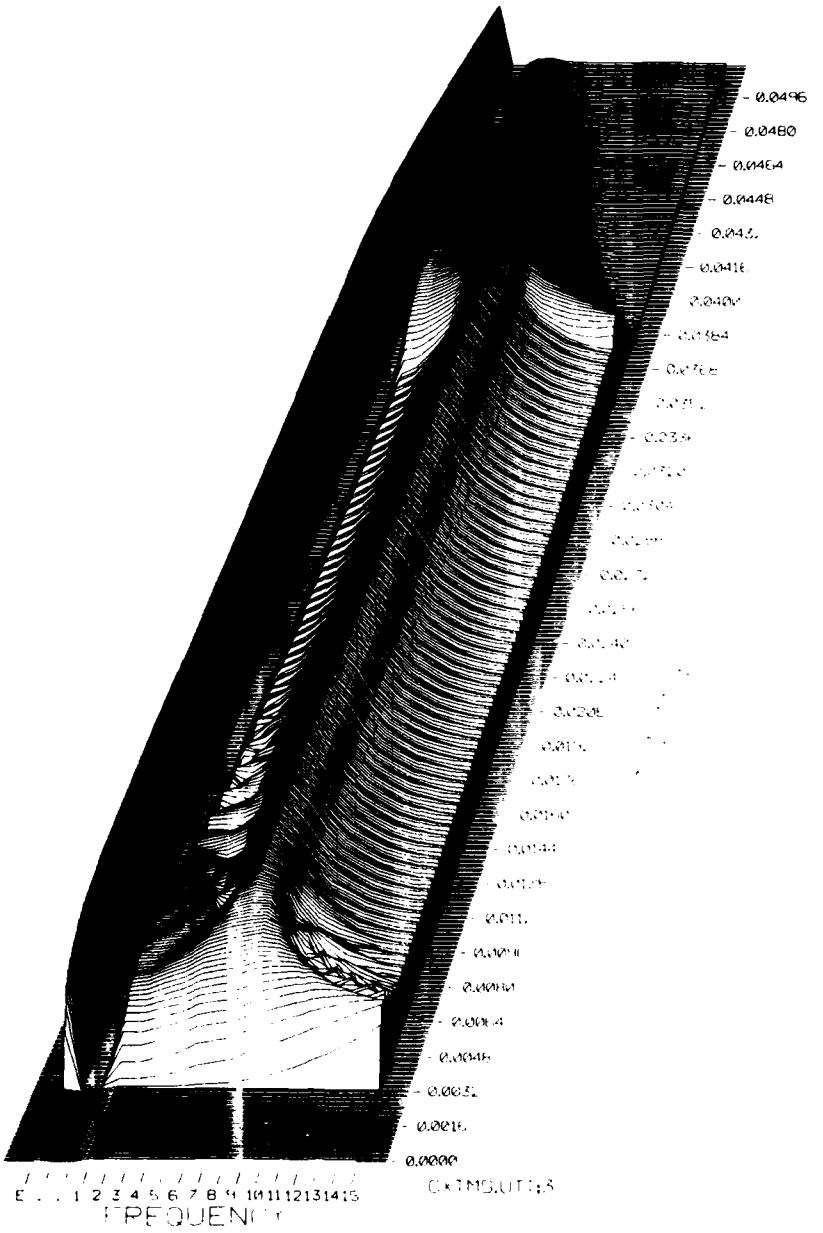


Figure 4.3: 3D Plot of Unmodified Data

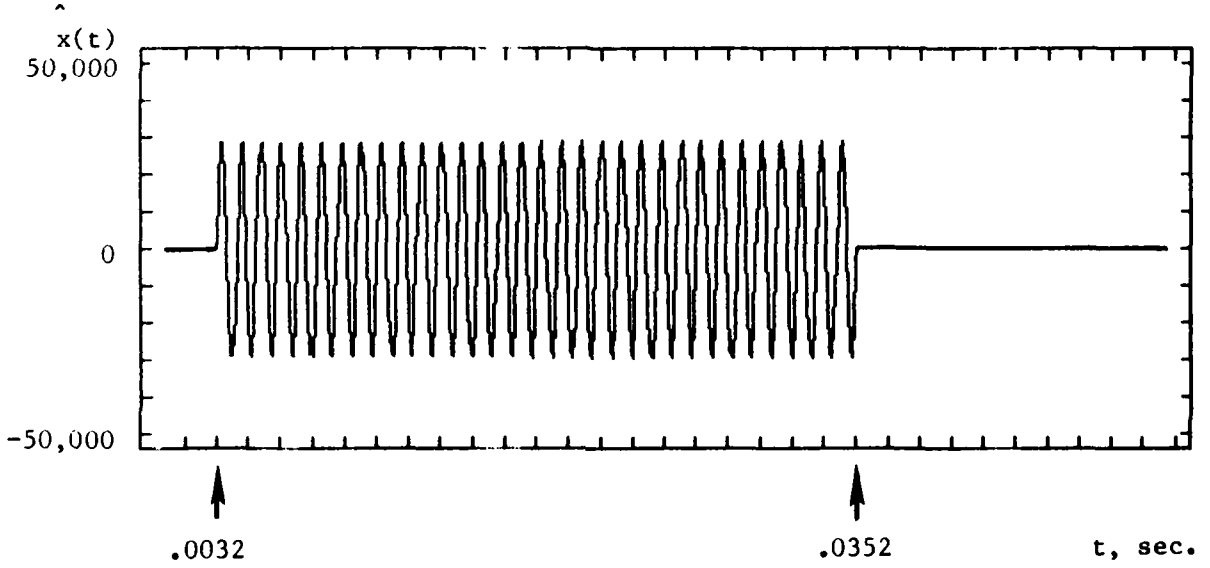
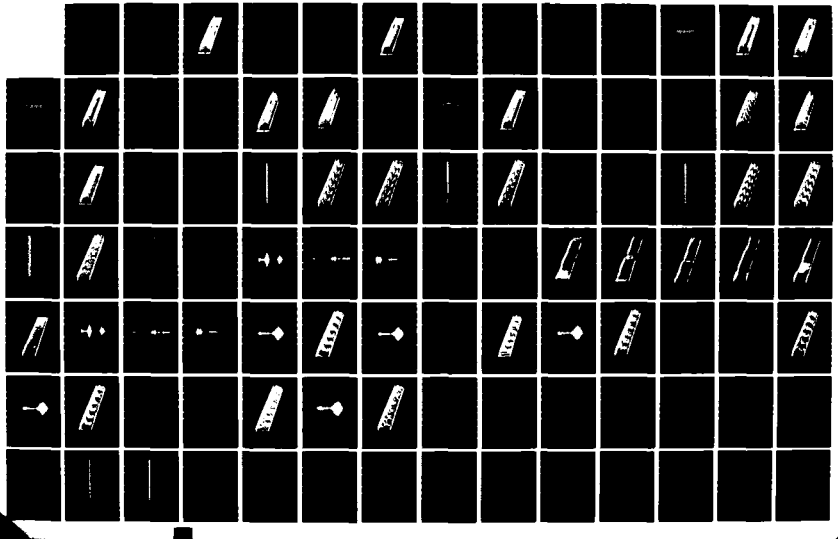


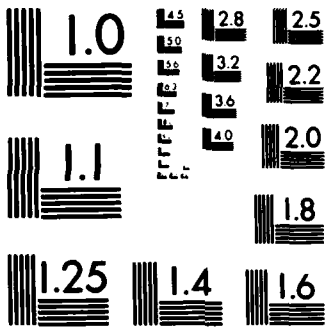
Figure 4.4: Reconstruction from Unmodified Data

The reconstructed signal was analyzed, and the resulting values were compared to the values of Fig. 4.3 in accordance with Equation 3.24. The resulting error is shown in Fig. 4.5. This plot is normalized to the peak error value of 4.9×10^{9}. The area under the graph of Fig. 4.5 corresponds to the total error, ε_{total}. Note that the error plot of Fig. 4.5 is a function of data values raised to the fourth power. Thus, reducing the tone burst amplitude by a factor of two reduces the error plot by a factor of sixteen. In order to obtain an error measure which is independent of signal level, the total normalized error is computed in accordance with Equation 3.26. For this reconstruction example, $\varepsilon_{total,norm} = 1.0 \times 10^{-9}$.

Next, a short-time spectral modification was employed in which sixteen time-domain samples in each channel were averaged, and each sample was replaced with the average value. The resulting modified data is shown in Fig. 4.6. This modification, which is employed for demonstration purposes, can be described in terms of Fig. 3.1. Since sixteen channels are used and the data rate of each channel has been reduced by a factor of sixteen, the transmission channel data rate of Fig. 3.1 is the same as the sampling rate of the original signal. Thus, in general, exact reconstruction from this modified data is impossible. The analysis transformation A uses a "boxcar" lowpass filter (ie., a filter with a constant amplitude, finite length unit-sample response) followed by downsampling in each channel. The corresponding synthesis transformation S uses upsampling followed by a boxcar lowpass filter in each channel. Thus, the data of Fig. 4.3 is the input to A, and the data of Fig. 4.6 is the output from S.

AD-A151 320 SPEECH ANALYSIS/SYNTHESIS BASED ON PERCEPTION(U) 2/3
MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB
J C ANDERSON 05 NOV 84 TR-707 ESD-TR-84-048
UNCLASSIFIED F19628-85-C-0002 F/G 17/2 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

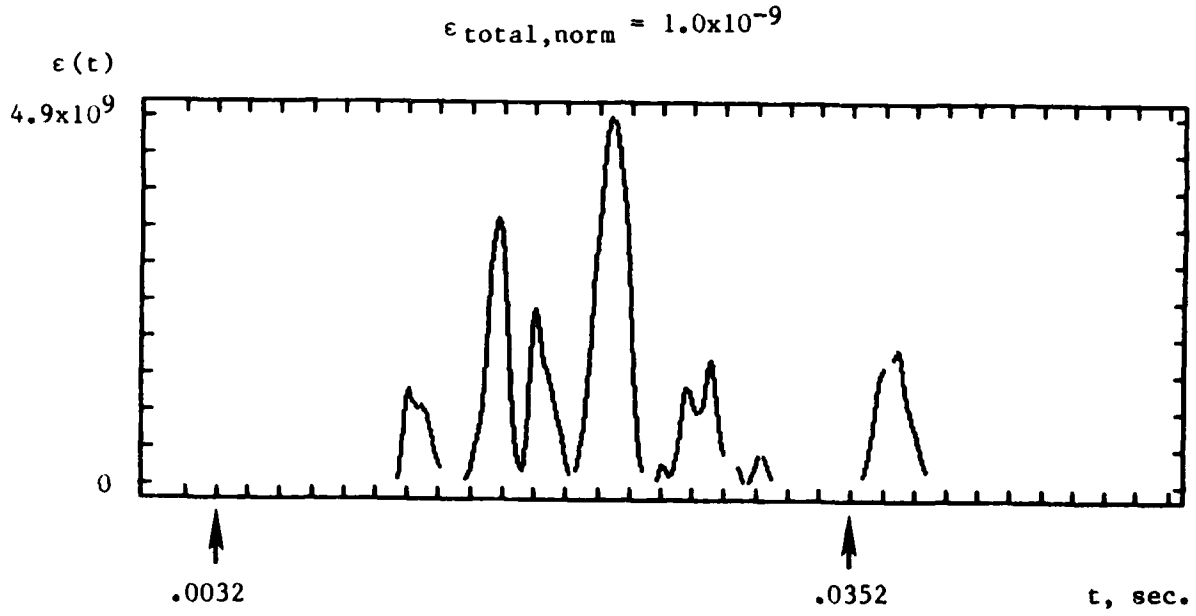


Figure 4.5: Error (1000 Hz, Unmodified Data)

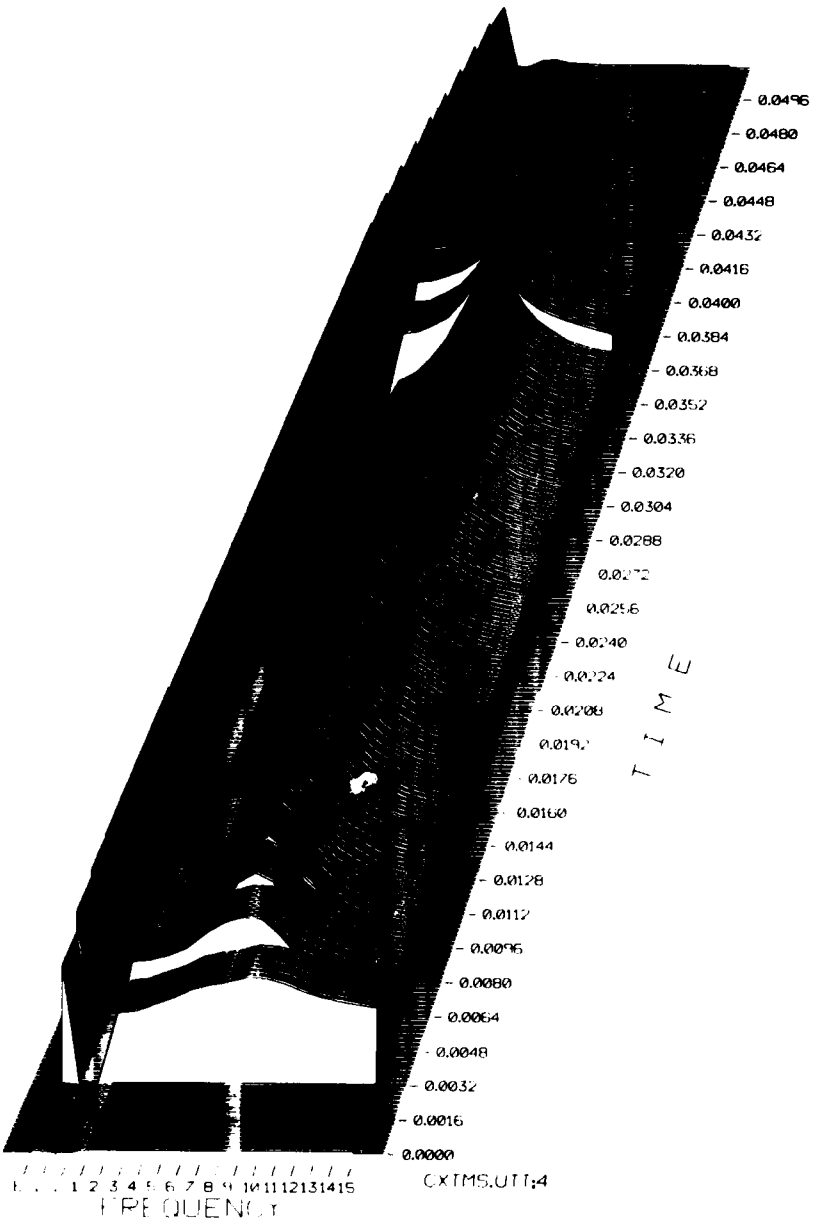


Figure 4.6: 3D Plot of Modified Data

The short-time spectral modification used to obtain Fig. 4.6 is of the type commonly employed in Automatic Speech Recognizer front-ends (Section 5.4), channel vocoders (Section D.2), and power spectrum estimation techniques (Section D.5) for data reduction purposes, although such applications typically average together a far greater number of samples. This simple data modification technique will be used to demonstrate many aspects of the reconstruction algorithm.

The reconstruction algorithm was applied to the data of Fig. 4.6, and the signal of Fig. 4.7 was obtained. The reconstruction is roughly a tone burst of correct amplitude, frequency, and duration. Since the modified data of Fig. 4.6 differs most from the unmodified data of Fig. 4.3 at the beginning and end of the burst, the reconstruction bears least resemblance to the original signal at the beginning and end of the burst.

In order to verify the algorithm operation, the reconstructed signal of Fig. 4.7 was analyzed, producing the 3D plot of Fig. 4.8. Comparison of Figs. 4.3, 4.6, and 4.8 demonstrates the ability of the algorithm to reconstruct a real-valued signal having short-time spectral characteristics which match the given data. Since the plots are on a logarithmic scale, low level differences may appear exaggerated.

Reconstruction error is plotted in Fig. 4.9. The peak error value of 9.7×10^{16} and the total normalized error value of 6.8×10^{-3} are many orders of magnitude greater than values for the previous example. Since significant error occurs only at the beginning and end of the tone burst, the total normalized error decreases with increasing tone burst duration for this example.

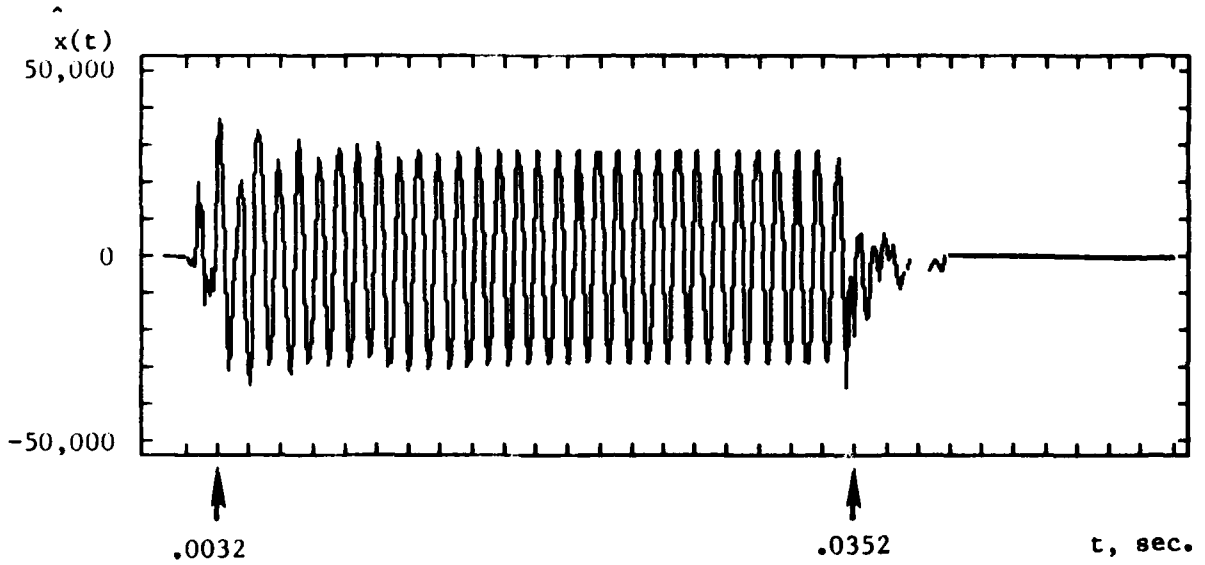


Figure 4.7: Reconstruction from Modified Data

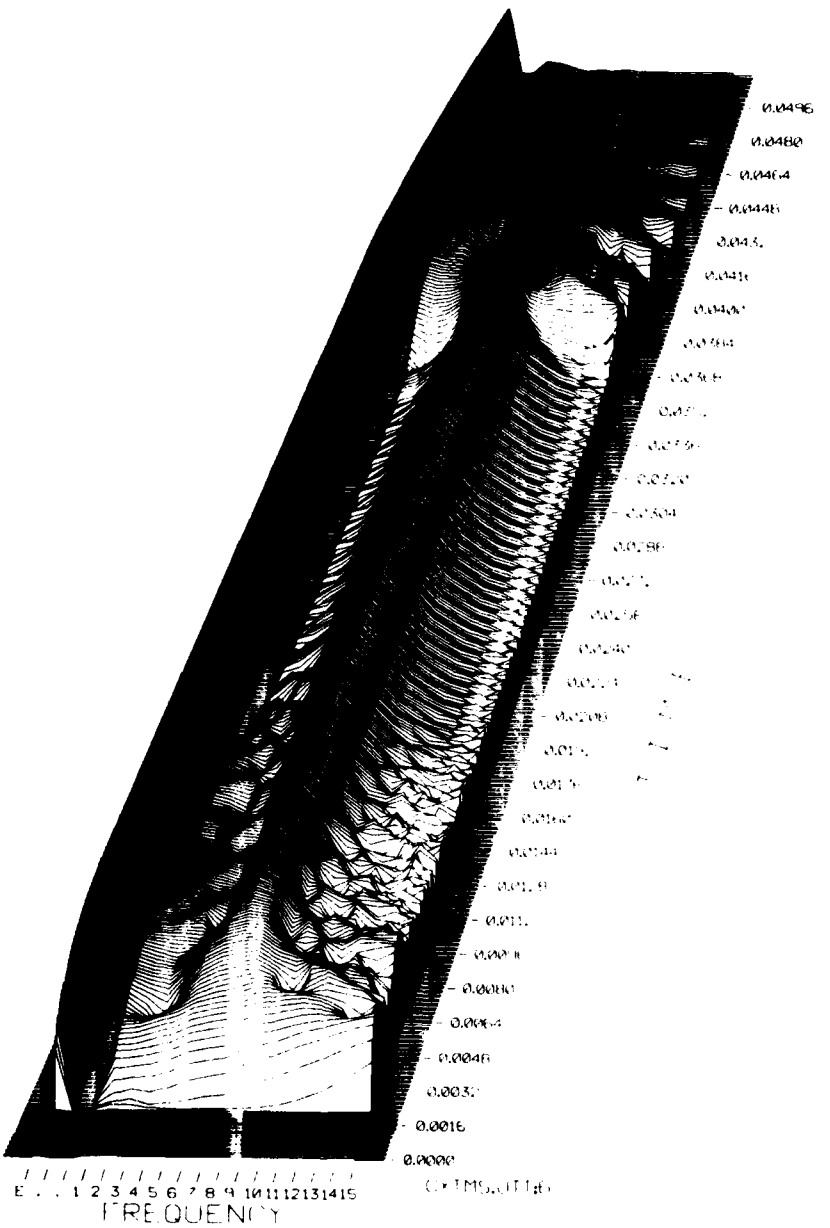


Figure 4.8: 3D Plot of Analyzed Reconstruction

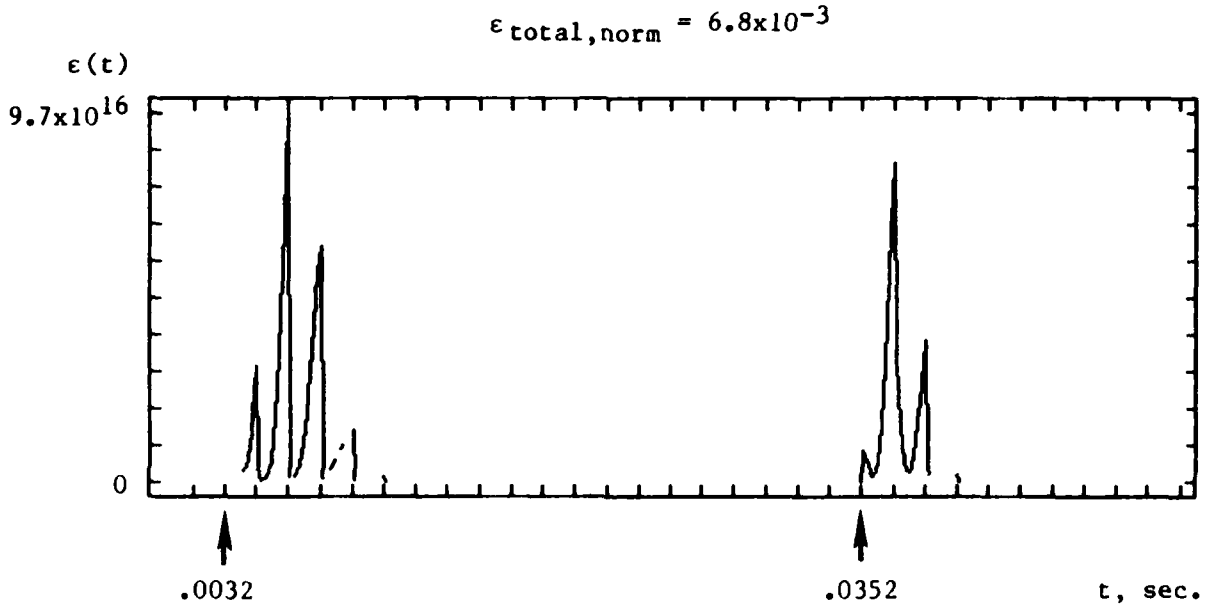


Figure 4.9: Error (1000 Hz, Modified Data)

It will be shown, via several examples, that the reconstruction algorithm of Section 3.4 performs well in the presence of short-time spectral modifications. Although no attempt will be made to optimize the algorithm for any particular modification, it is possible to reduce reconstruction error by doing so. For example, error may be reduced by choosing an error weighting function which extends over several periods of the modification. This approach, however, significantly increases computation time and implementation complexity, and will not be considered here. Note that the largest error peak of Fig. 4.9 can be reduced by simply setting the reconstructed sequence values to zero prior to t=.0032 sec. This can be done based on the fact that STE and all F/D channels are zero prior to this time. The reconstruction algorithm produced nonzero values in Fig. 4.7 because the modified data changed abruptly rather than in a bandlimited fashion. Since the model of the analysis system contained in the reconstruction process (see Fig. 3.2) produces only bandlimited functions, and the modified data does not agree with the model, a spike occurs in the error whenever a discontinuity occurs in the data. This effect can be seen by comparing Figs. 4.6, 4.8, and 4.9. In order to reduce error spike amplitudes, a smoother short-time spectral modification must be chosen. For example, considering each channel separately, when the value at each discontinuity in Fig. 4.6 is replaced by an average of the surrounding steady-state values, the maximum error value is reduced nearly 20%. This error reduction was accomplished without setting any values to zero prior to t=.0032 sec. Since the resulting surface is somewhat smoother, the reconstruction algorithm produces a signal having a short-time spectrum

which better matches the smoothed data. However, the fact that reconstruction error is reduced does not generally indicate that a signal reconstructed from smoothed modified data will bear closer resemblance to the original signal. Thus, such smoothing will not be employed as an aid to reconstruction from modified data. For demonstration purposes, the algorithm described in Section 3.4 will be applied directly in all examples.

4.3 TONE PAIR BURSTS

A 450 and 2500 Hz tone pair burst is shown in Fig. 4.10. From Fig. 2.12 it can be seen that the filters having center frequencies at 450 and 2500 Hz overlap at the -40dB level. Thus, since the signals are widely separated in frequency, the analysis of Fig. 4.11 does not reveal any interaction between the two component sinusoids. Each F/D output reaches a steady-state value (see Section B.3.7). A short-time spectral modification was applied by averaging sixteen samples in each channel, and each sample was replaced with the average value as shown in Fig. 4.12. The reconstruction algorithm was applied to the modified data, and the result is shown in Fig. 4.13. The reconstructed signal was then analyzed, and the results are shown in Fig. 4.14. Corresponding error is plotted in Fig. 4.15, and is comparable to the single tone burst case shown in Fig. 4.9, although the peak error value of 1.2×10^{16} is considerably less due to a reduction in average input signal level. The total normalized error value of 8.4×10^{-3} is comparable to that of the single tone burst case since both signals are of the same duration. As in the single tone burst data modification example, total normalized error is a function of tone pair burst duration for this example.

A 1000 and 1600 Hz tone pair burst is shown in Fig. 4.16. Since the filters at the corresponding center frequencies overlap at the -18dB level, some interaction between the spectral components is visible in the analysis of Fig. 4.17. Each F/D output consists of a constant and a beat frequency component (see Section B.3.8). When the short-time spectral modification is applied, the beat frequency component is eliminated as shown in Fig. 4.18. Thus, the surface of Fig. 4.18 represents

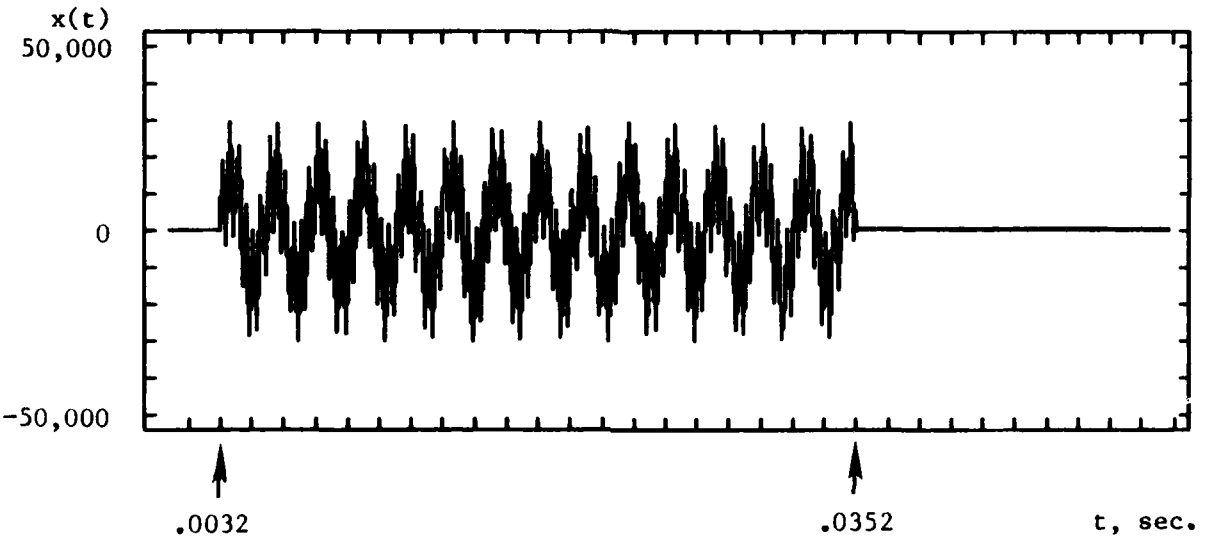


Figure 4.10: Original Signal (450 & 2500 Hz)

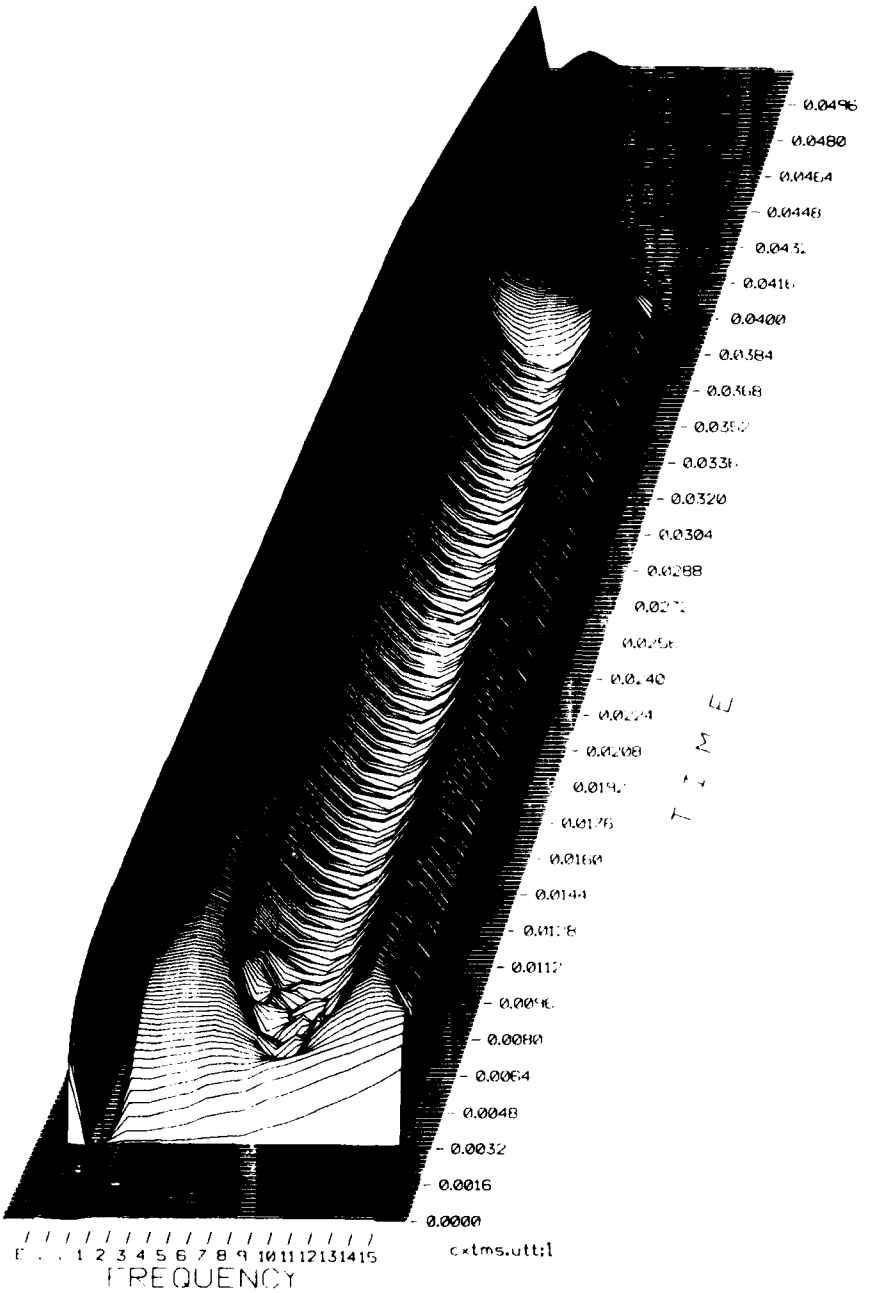


Figure 4.11: 3D Plot of Unmodified Data

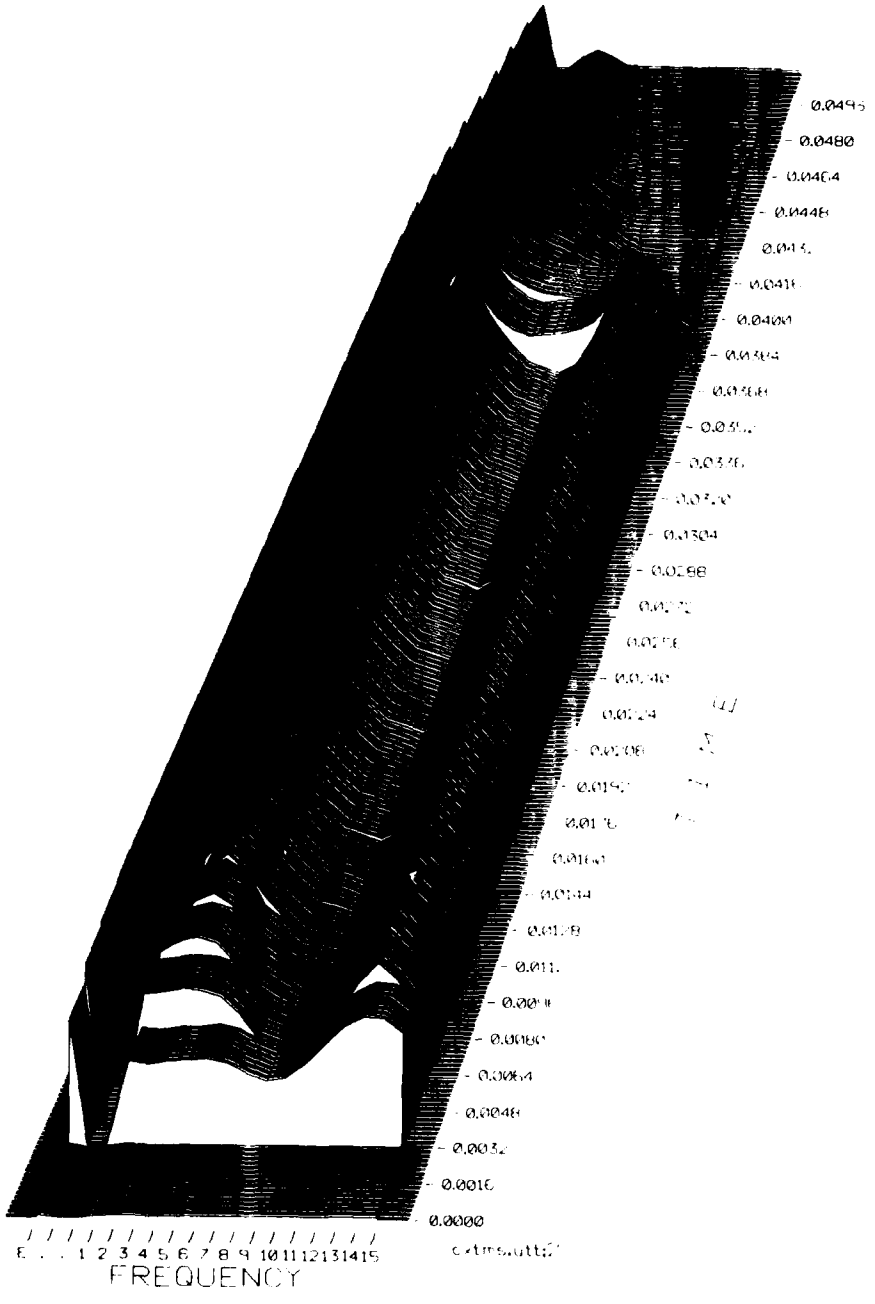


Figure 4.12: 3D Plot of Modified Data

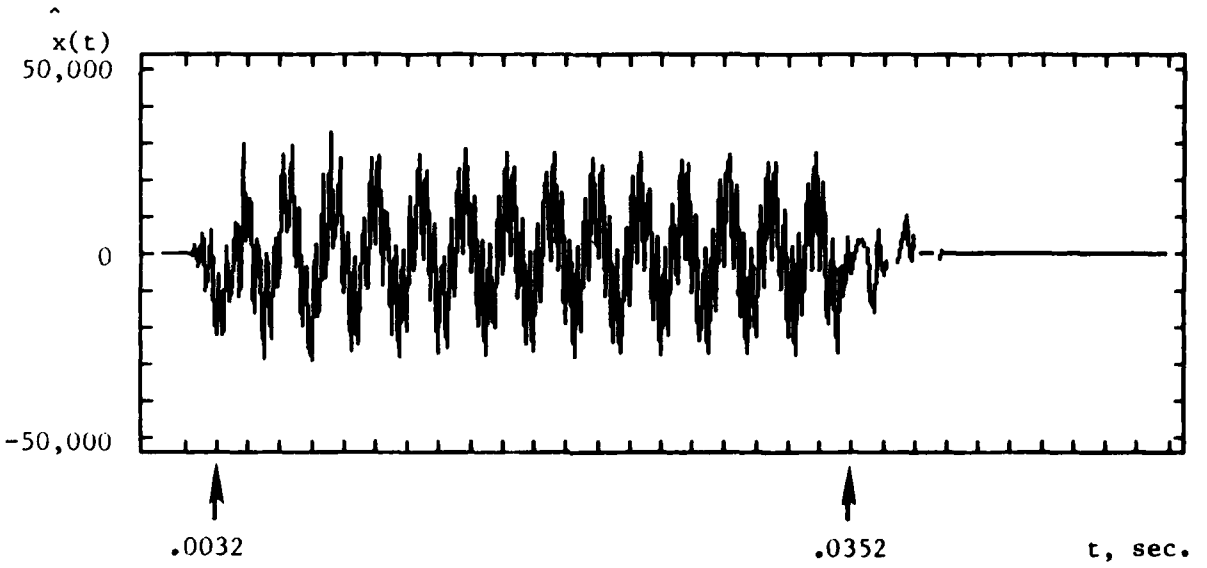


Figure 4.13: Reconstruction from Modified Data

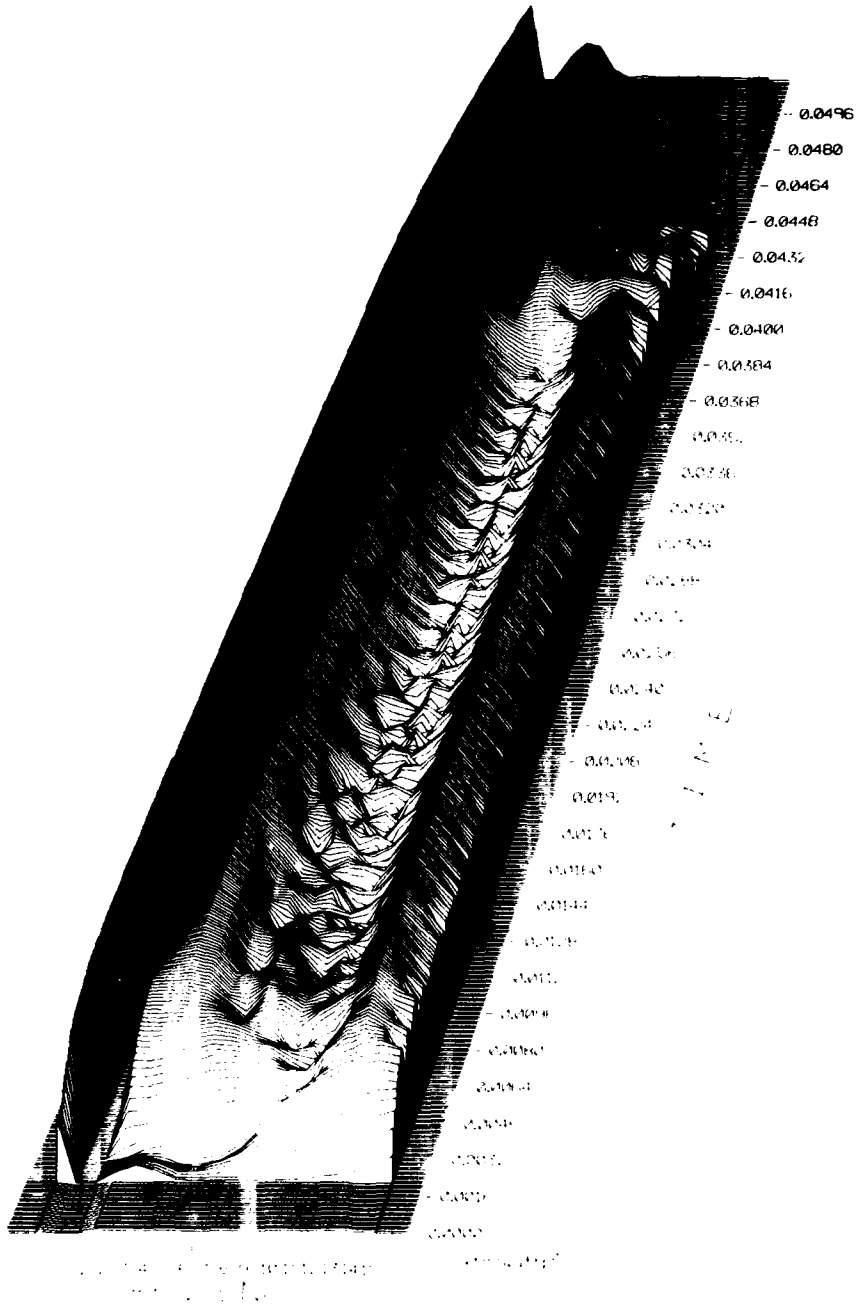


Figure 4.14: 3D Plot of Analyzed Reconstruction

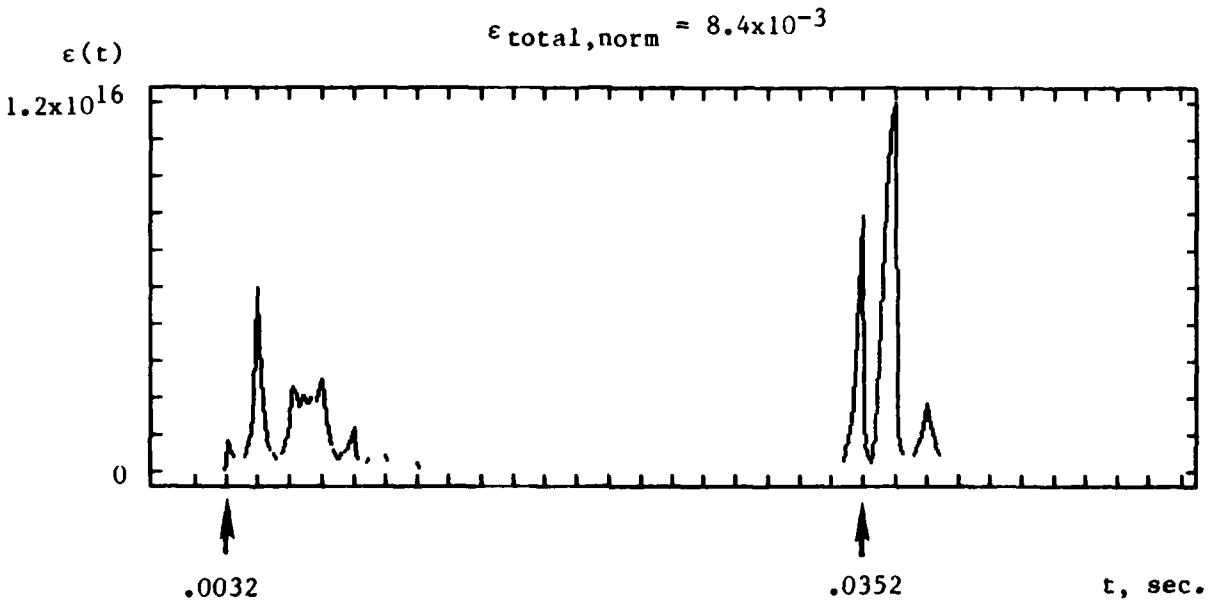


Figure 4.15: Error (450 & 2500 Hz)

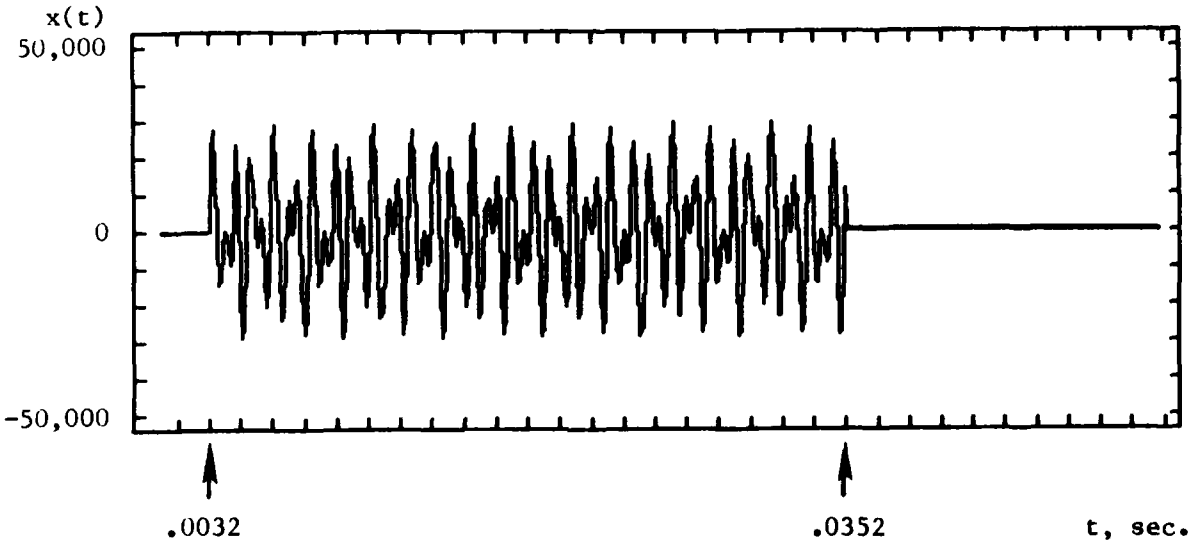


Figure 4.16: Original Signal (1000 & 1600 Hz)

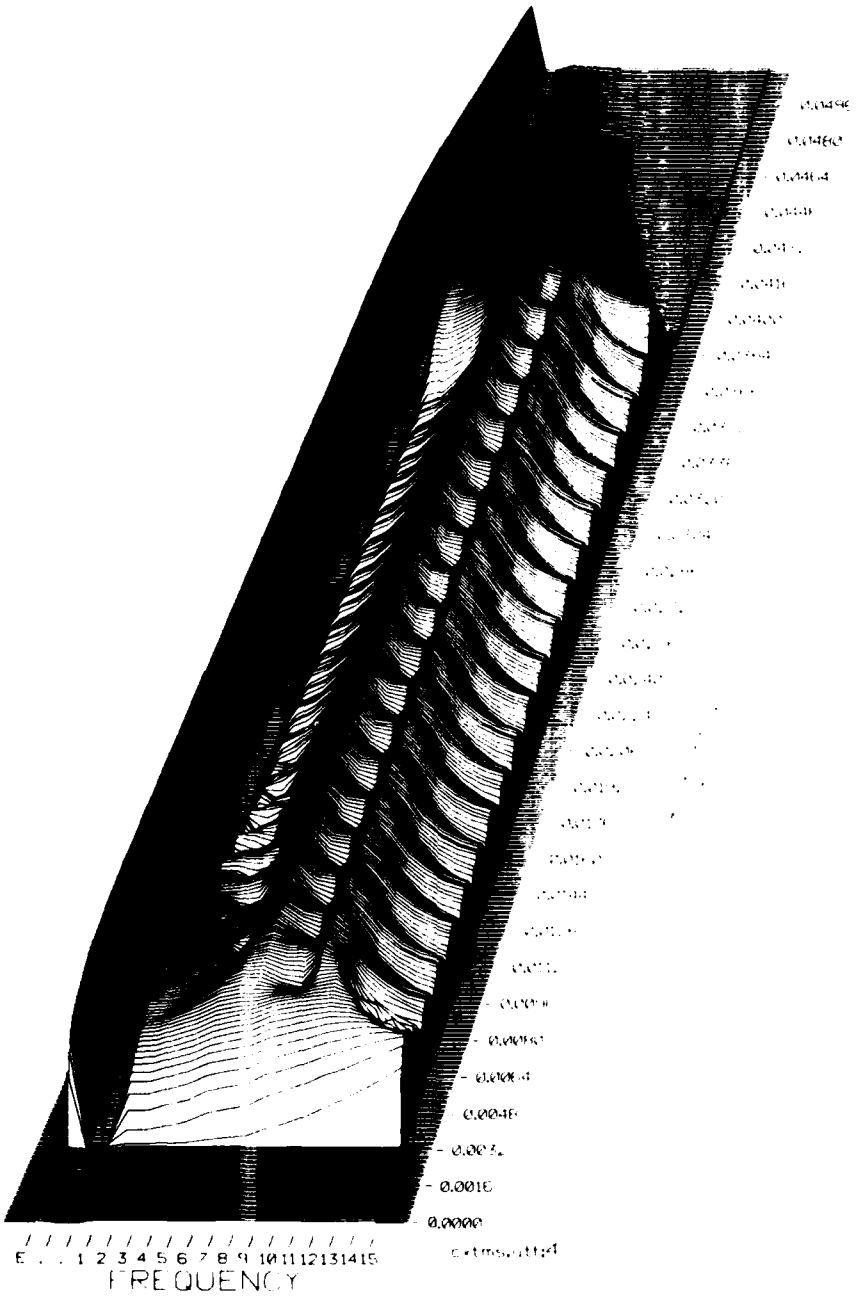


Figure 4.17: 3D Plot of Unmodified Data

Figure 4.18: 3D Plot of Modified Data

inconsistent information. On one hand, the data indicates presence of two sine waves because two spectral peaks are visible. On the other hand, if two sine waves are present then beat frequencies should occur, but none are visible in the data of Fig. 4.18. Thus, given the inconsistent data of Fig. 4.18, a reasonable signal reconstruction approach might be to first choose two sinusoidal components as indicated by the spectral peaks. A low-level periodic waveform having amplitude and frequency determined in accordance with an error criterion could then be added to the two sinusoids, thereby reducing the beat frequencies in order to approximate the data of Fig. 4.18. This was exactly the result obtained upon application of the reconstruction algorithm to the data of Fig. 4.18, as shown in the reconstruction of Fig. 4.19. Analysis of the reconstructed signal is shown in Fig. 4.20, and it can be seen that the reconstruction algorithm inserts a third sinusoidal component to compensate for the inconsistent data of Fig. 4.18. The plot of Fig. 4.21 reveals a 530 Hz oscillation in the error. Since there are few discontinuities in the data of Fig. 4.18, there are few spikes in the error plot of Fig. 4.21. Since the area under this error plot during the steady-state portion of the tone pair burst is greater than the area under the error transients at the beginning and end of the burst, the total normalized error does not depend strongly on tone pair burst duration. The total normalized error value of 3.1×10^{-2} is nearly four times that of the previous tone pair burst example, and the peak error value of 2.4×10^{16} is twice that of the previous example.

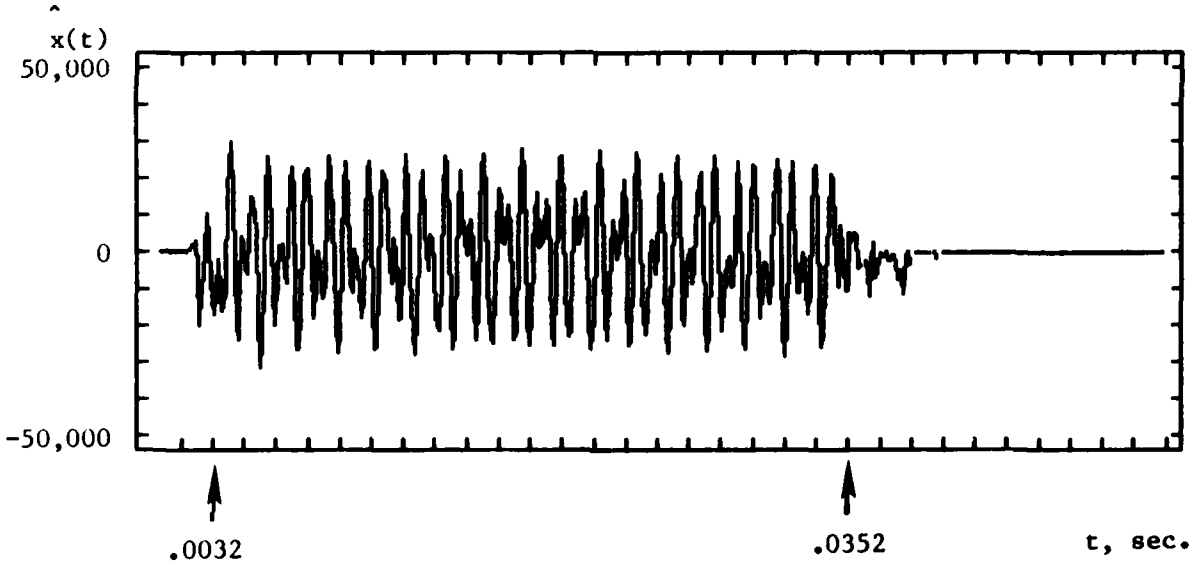


Figure 4.19: Reconstruction from Modified Data

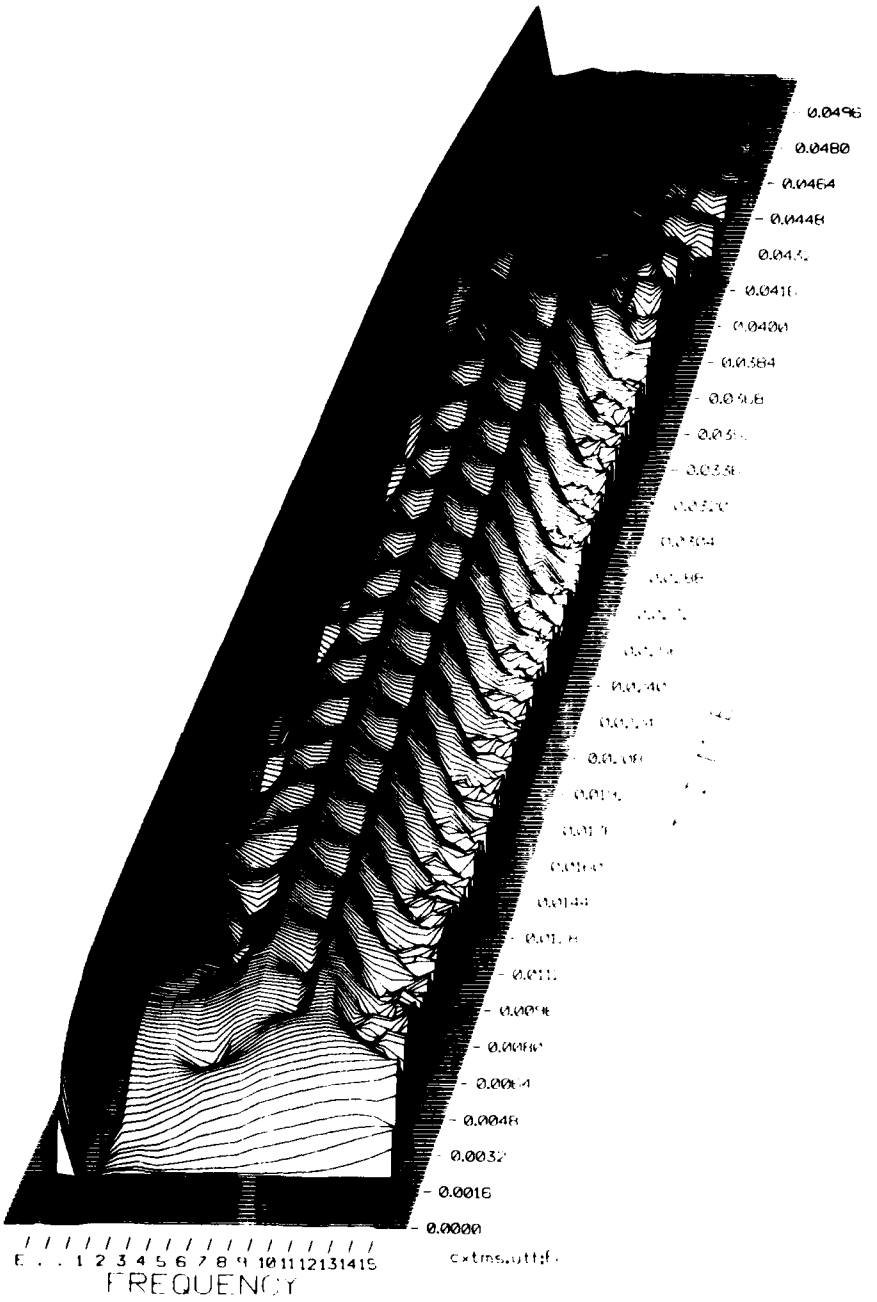


Figure 4.20: 3D Plot of Analyzed Reconstruction

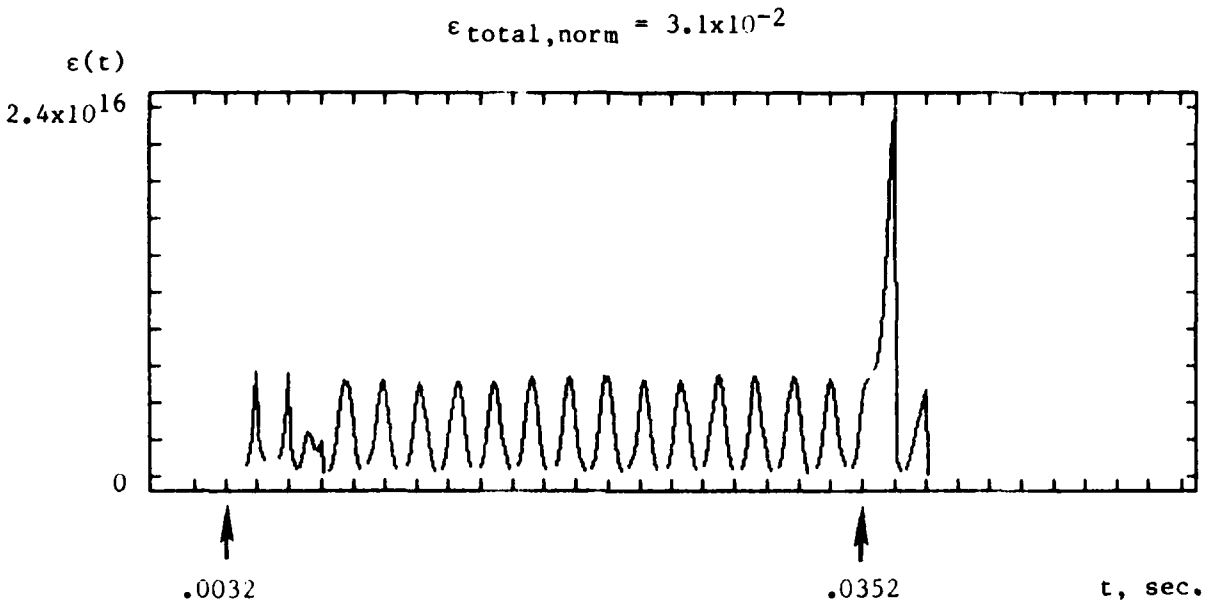


Figure 4.21: Error (1000 & 1600 Hz)

Finally, a 1000 and 1170 Hz tone pair burst is shown in Fig. 4.22. Since filters at the corresponding center frequencies overlap at the -3dB level, spectral components are not resolved in the frequency domain characteristics of Fig. 4.23, although beat frequencies are apparent in the time dimension. Short-time spectral modification severely distorts these beat frequencies, and inserts large discontinuities as shown in Fig. 4.24. A poor quality reconstruction is obtained, as shown in Fig. 4.25. The analyzed reconstruction is shown in Fig. 4.26, and the corresponding error in Fig. 4.27. The peak error value of 4.6×10^{16} and total normalized error value of 5.6×10^{-2} are the largest of any examples thus far. Again, since the area under the error plot during the steady-state portion of the tone pair burst is greater than the area under the error transients at the beginning and end of the burst, the total normalized error does not depend strongly on tone pair burst duration.

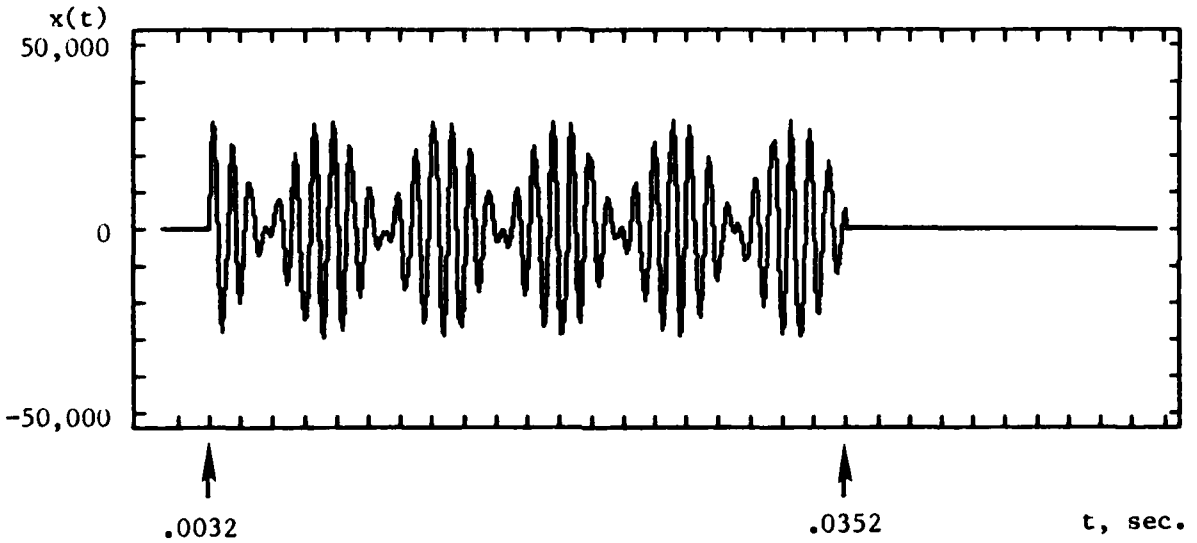


Figure 4.22: Original Signal (1000 & 1170 Hz)

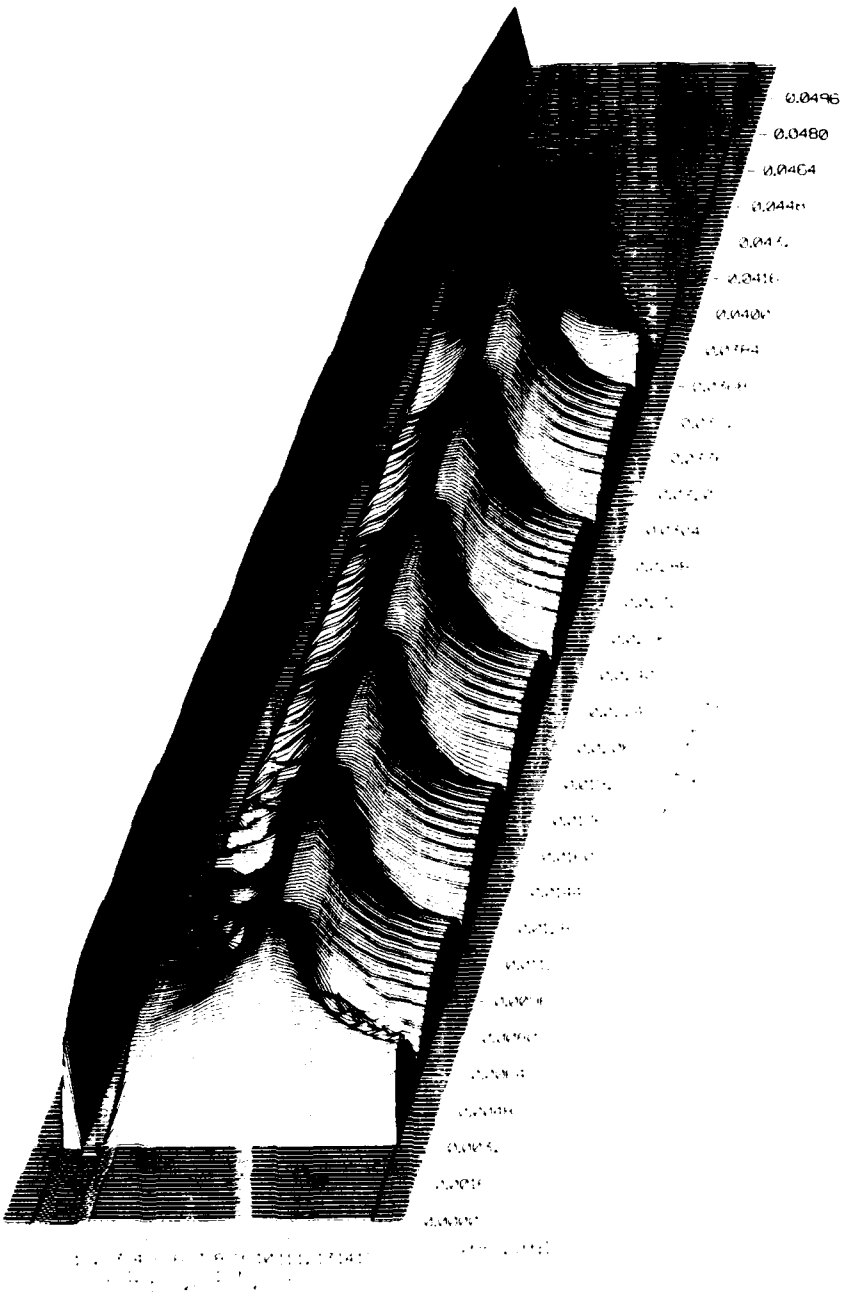


Figure 4.23: 3D Plot of Unmodified Data

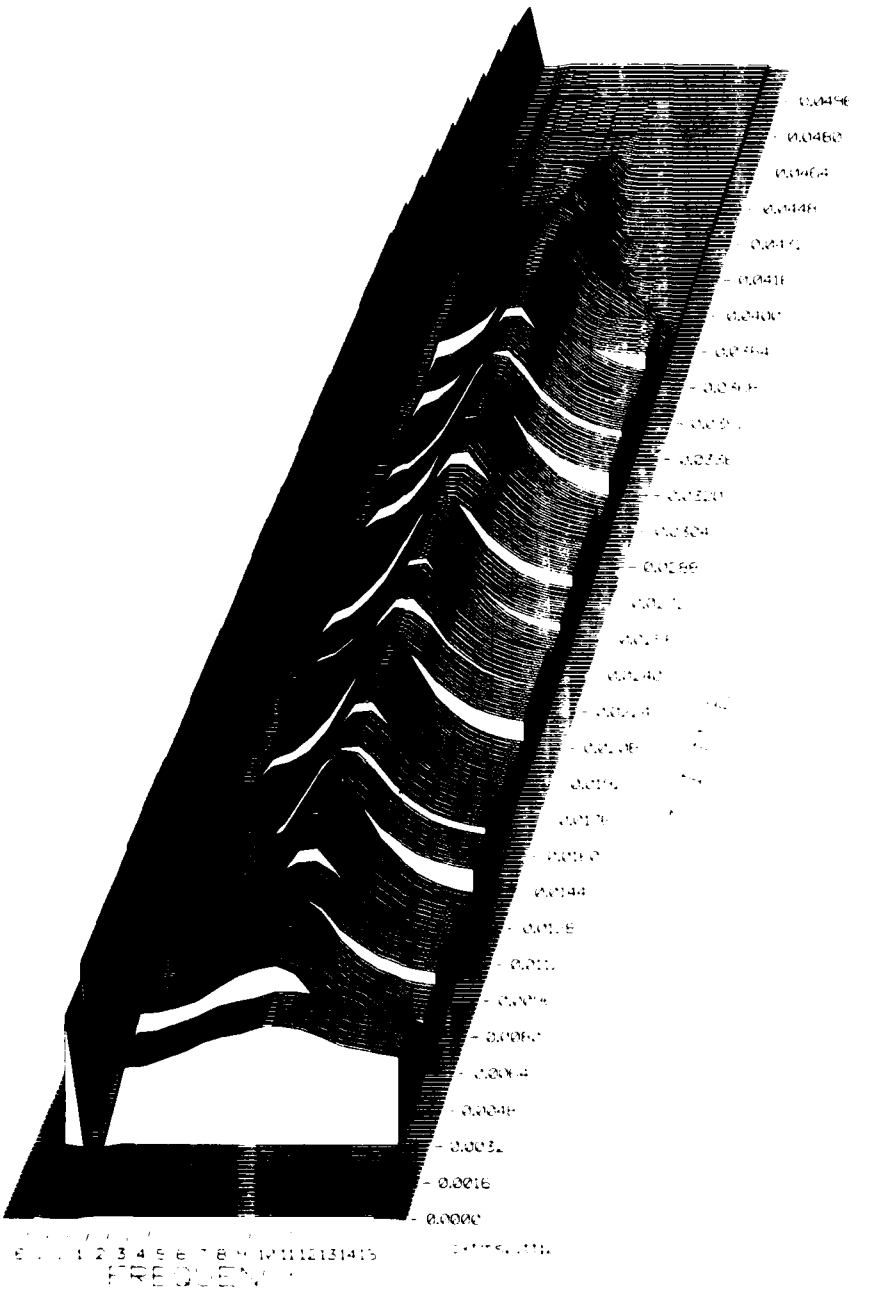


Figure 4.24: 3D Plot of Modified Data

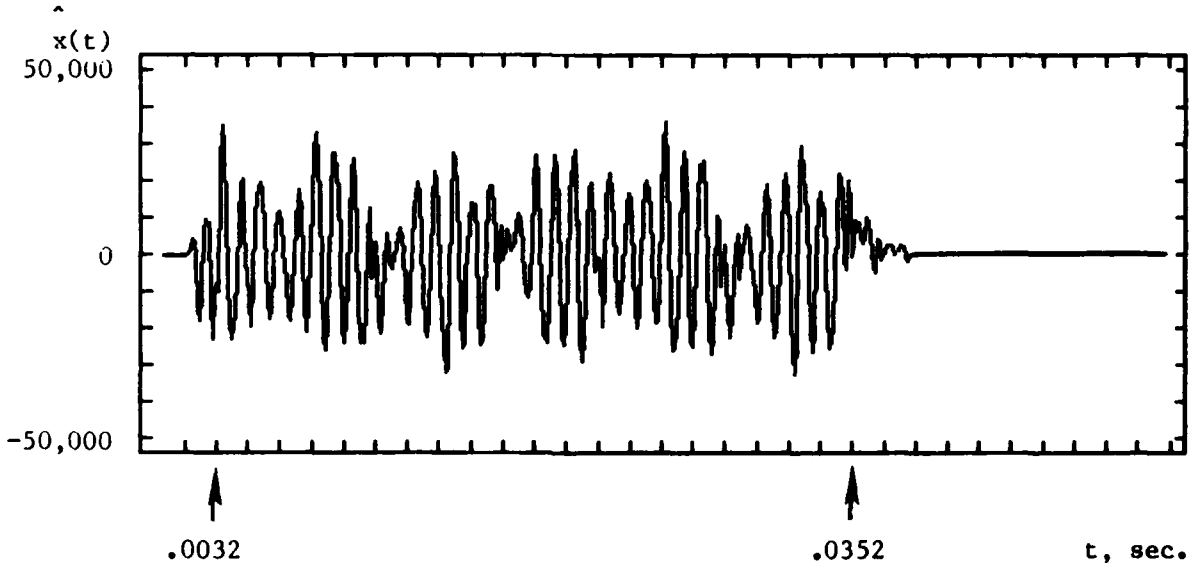


Figure 4.25: Reconstruction from Modified Data

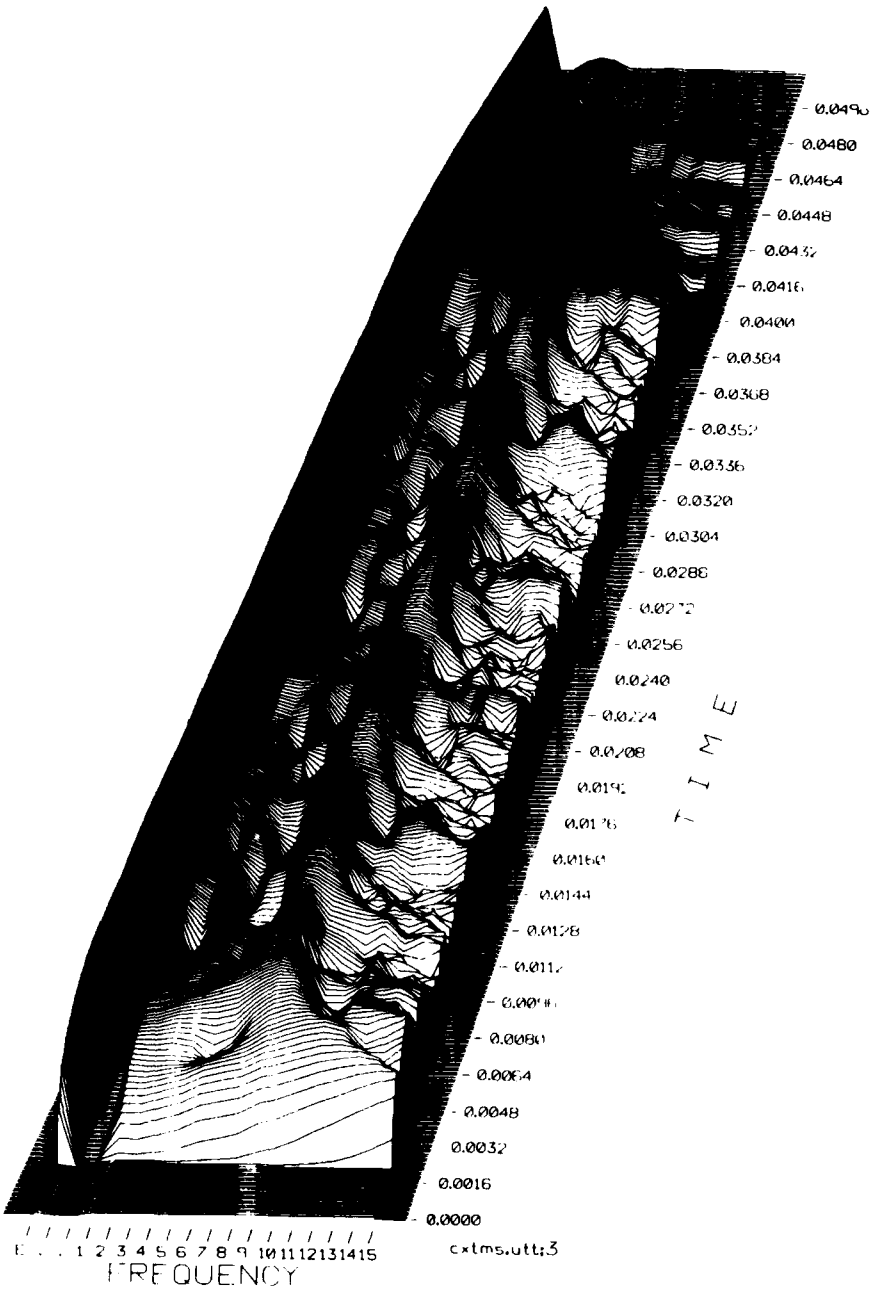


Figure 4.26: 3D Plot of Analyzed Reconstruction

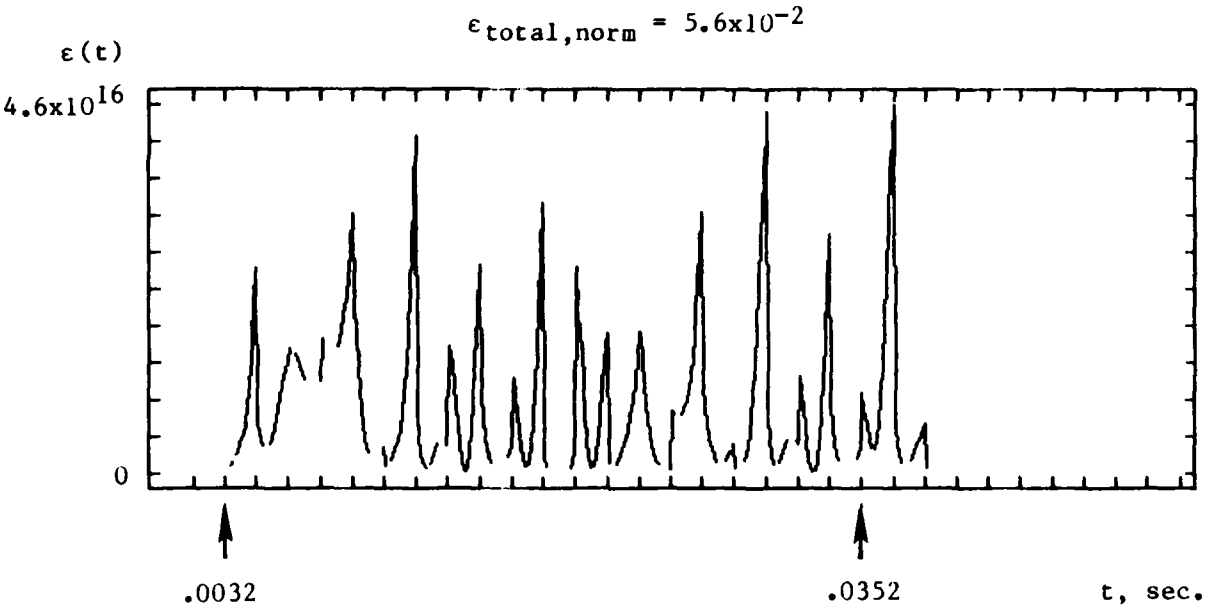


Figure 4.27: Error (1000 & 1170 Hz)

4.4 SYNTHETIC VOWELS

Synthetic vowels provide a controlled speech-like signal for testing and demonstration purposes. Such vowels can be conveniently generated via an acoustic tube vocal tract model (Rabiner and Schafer [3]). An example, the synthetic vowel /E/ as in "bet," is shown in Fig. 4.28. Vowel sounds are often characterized in terms of their spectral peaks, or formants (Peterson and Barney [42]). This vowel has a first formant frequency F1 of 530 Hz. The second and third formants are F2=1840 and F3=2480 Hz. Formant bandwidths are 40 Hz for F1, 60 Hz for F2, and 100 Hz for F3. The pitch frequency is F0=125 Hz, so a male speaker is simulated. The pitch is visible in the time dimension and formant peaks are visible in the frequency dimension of Fig. 4.29. Note that F1 has a far higher level than F2 or F3, and thus is the most important feature for reconstruction spectral matching purposes.

As in previous examples, the data was modified by averaging sixteen time-domain samples in each channel, and replacing each sample with the average value as shown in Fig. 4.30. The reconstruction algorithm was applied to the modified data, and the results are shown in Fig. 4.31. The reconstructed signal was then analyzed, and the result is shown in Fig. 4.32. A comparison of Figs. 4.32 and 4.30 reveals that F1 of the analyzed reconstructed signal provides a good match to the modified spectrum. This observation is supported by the corresponding error shown in Fig. 4.33. The peak error value of 2.0×10^{16} is comparable to that of the tone pair burst examples since similar average signal levels are used. The total normalized error value of 4.3×10^{-2} is also comparable to previous examples, and does not depend strongly on signal duration.

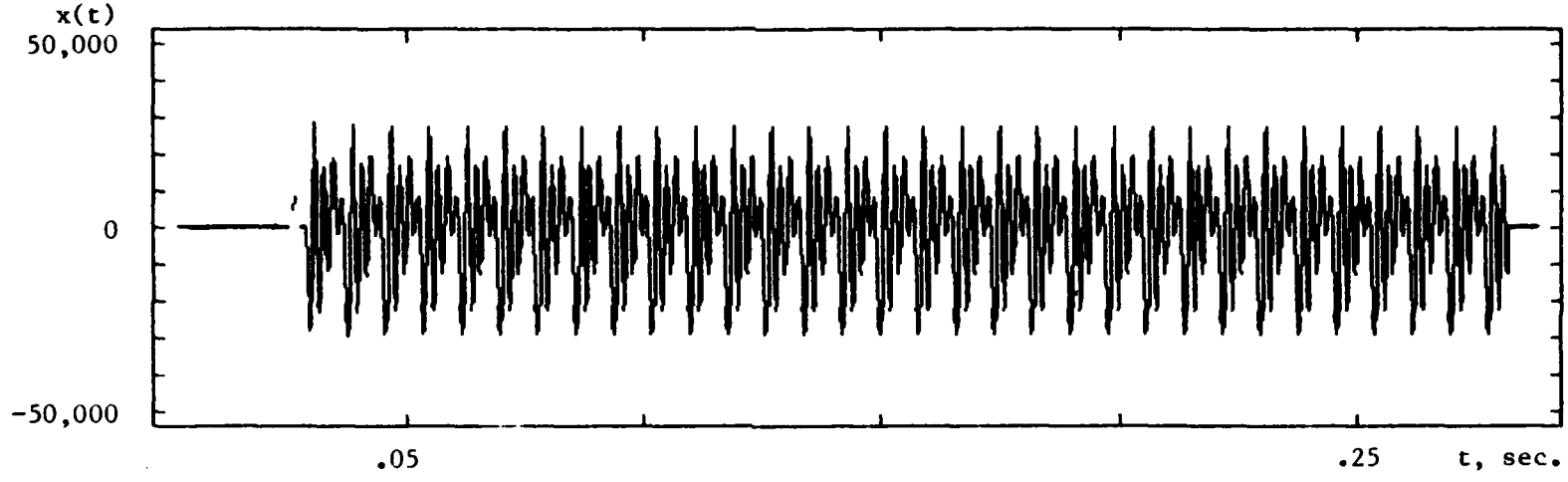


Figure 4.28: Original Signal (Vowel /E/)

Figure 4.29: 3D Plot of Unmodified Data

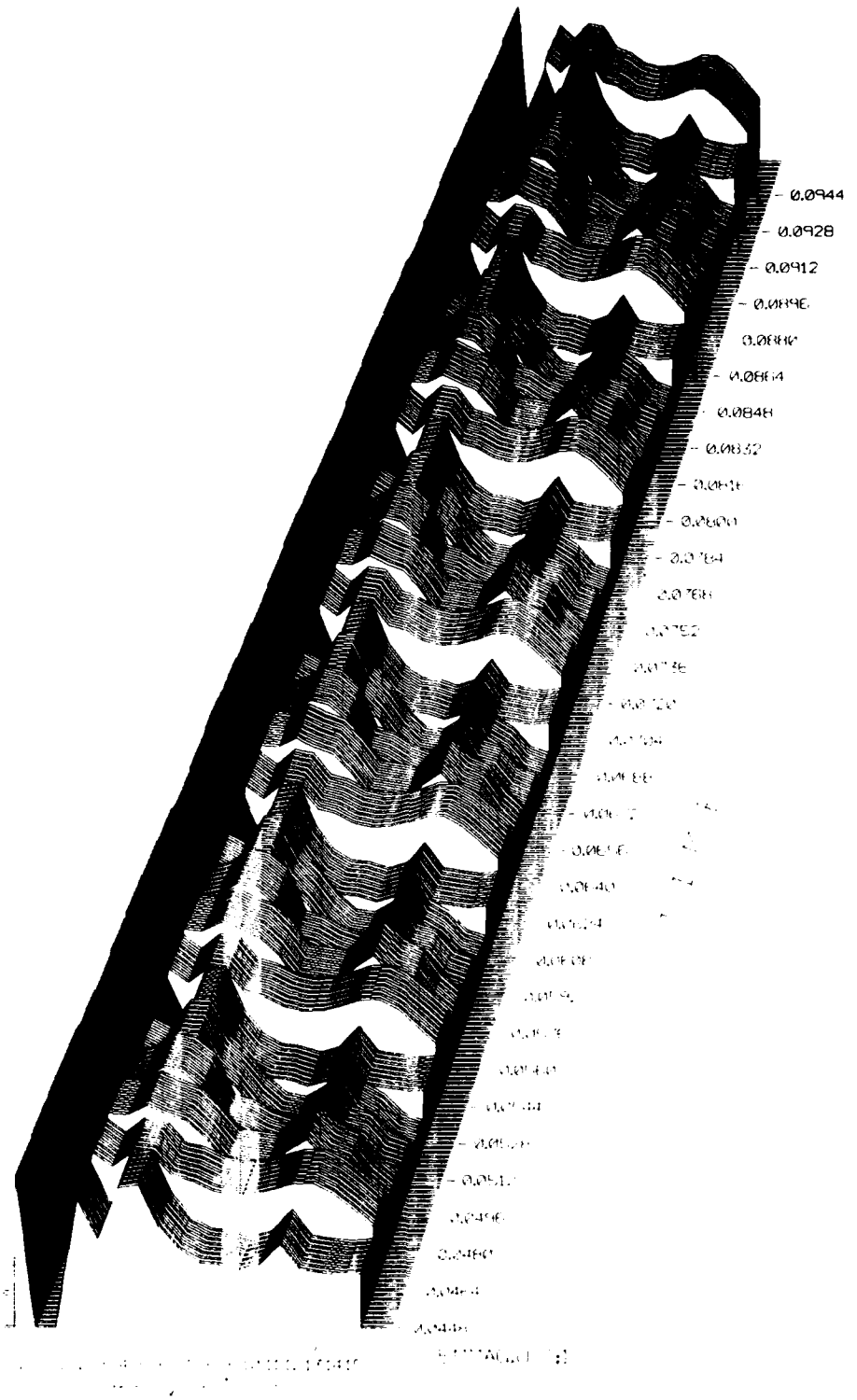


Figure 4.30: 3D Plot of Modified Data

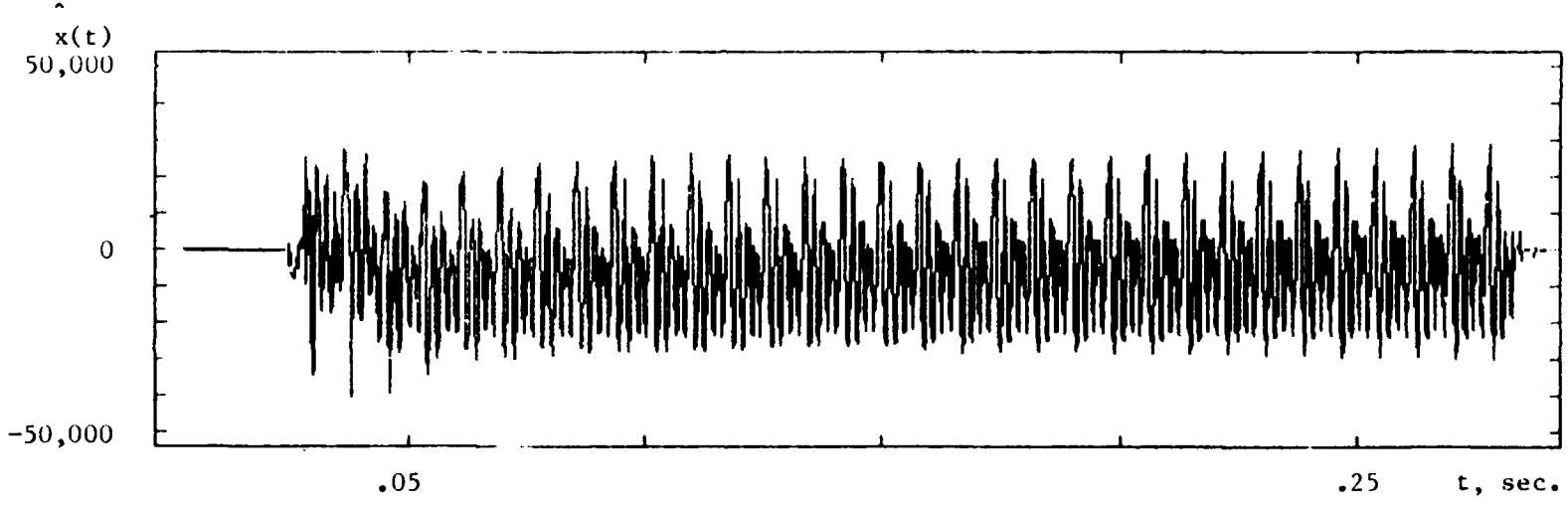


Figure 4.31: Reconstruction from Modified Data (Vowel /E/)

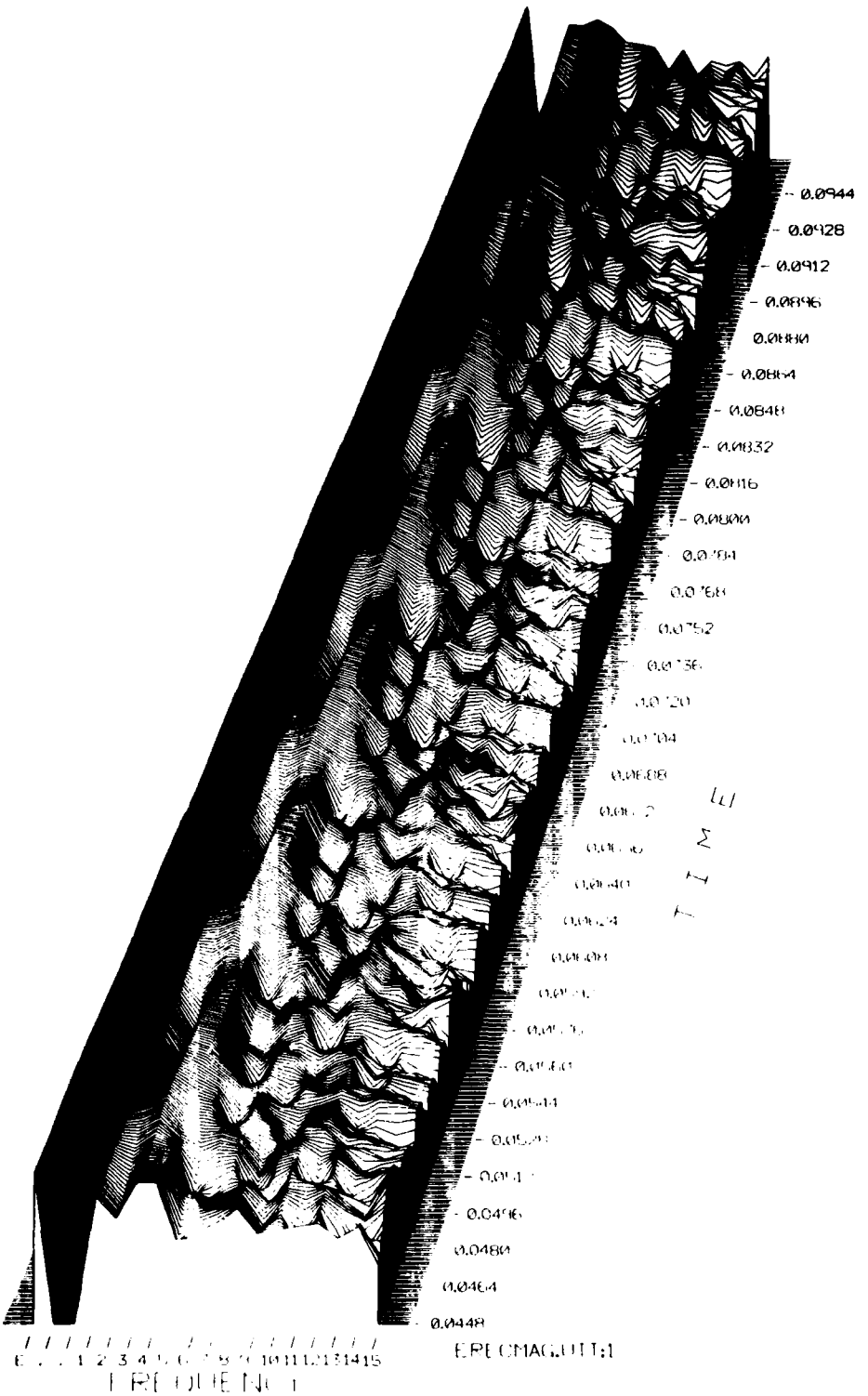


Figure 4.32: 3D Plot of Analyzed Reconstruction

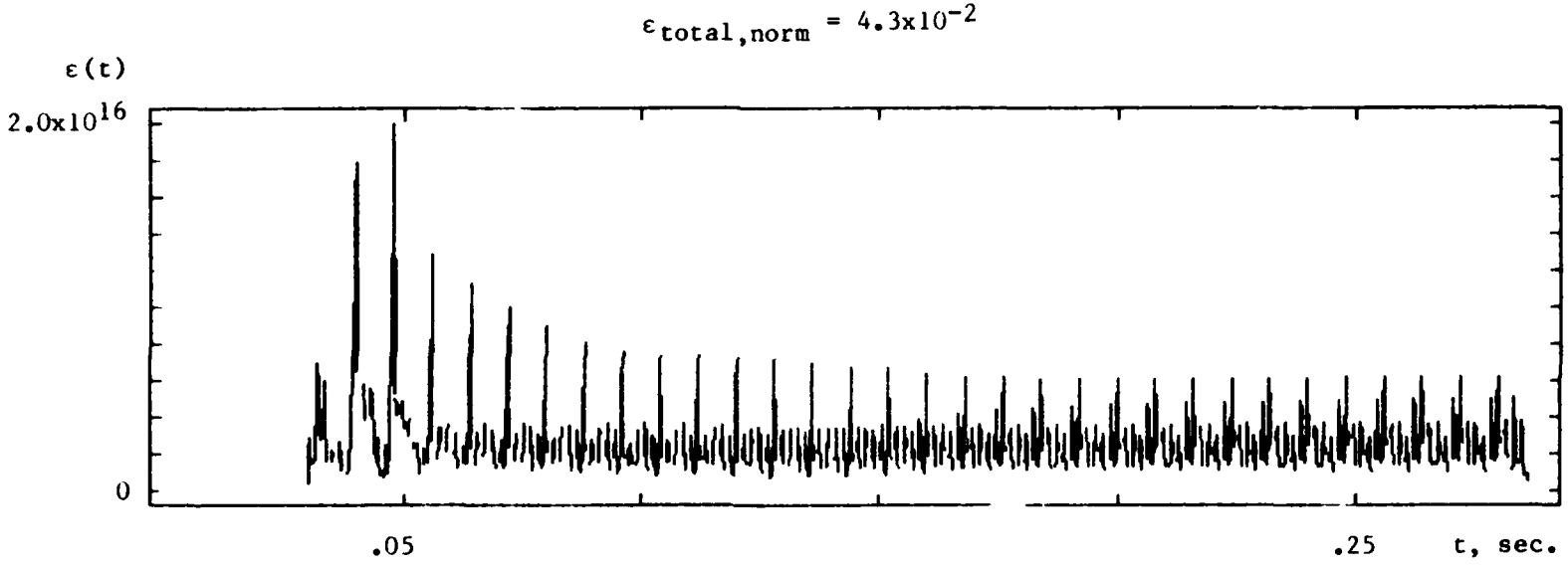


Figure 4.33: Error (Vowel /E/)

A similar synthetic vowel, /AE/ as in "bat," is shown in Fig. 4.34. This vowel has formant frequencies F1=660, F2=1720, and F3=2410 Hz. Formant bandwidths and pitch are the same as for the previous example. The analyzed signal is shown in Fig. 4.35. The data was modified as shown in Fig. 4.36. In this example, the short-time spectral modification produces large discontinuities in F1 as compared to the previous example. The reconstruction algorithm was applied to the modified data, and results are shown in Fig. 4.37. The reconstructed signal was analyzed as shown in Fig. 4.38. A comparison of Figs. 4.38 and 4.36 reveals a relatively poor match between F1 of the analyzed reconstructed signal and the modified spectrum, due to the reconstruction algorithm's inability to model discontinuities. The discontinuities also cause large spikes in the corresponding error of Fig. 4.39. Although the peak error value of 2.0×10^{16} is the same as the previous example, the total normalized error value of 8.4×10^{-2} is twice that of the previous example.

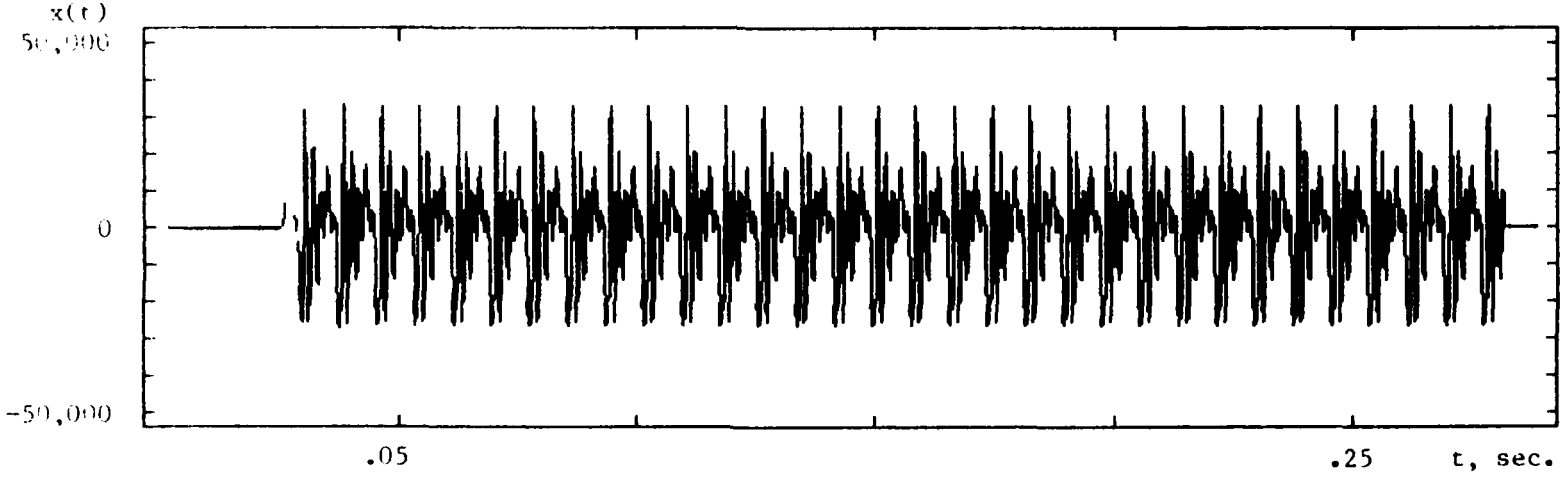


Figure 4.34: Original Signal (Vowel /AE/)

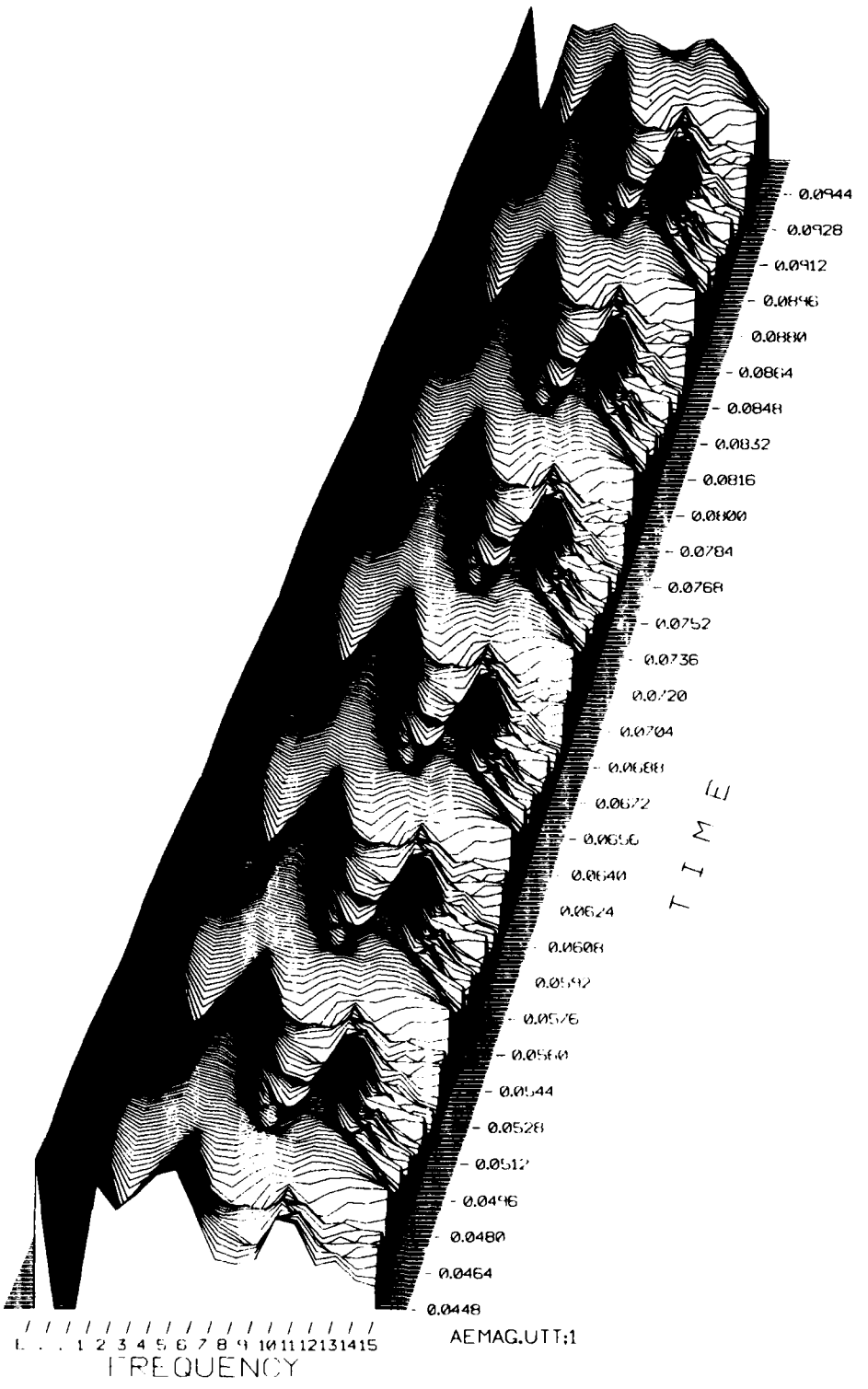


Figure 4.35: 3D Plot of Unmodified Data

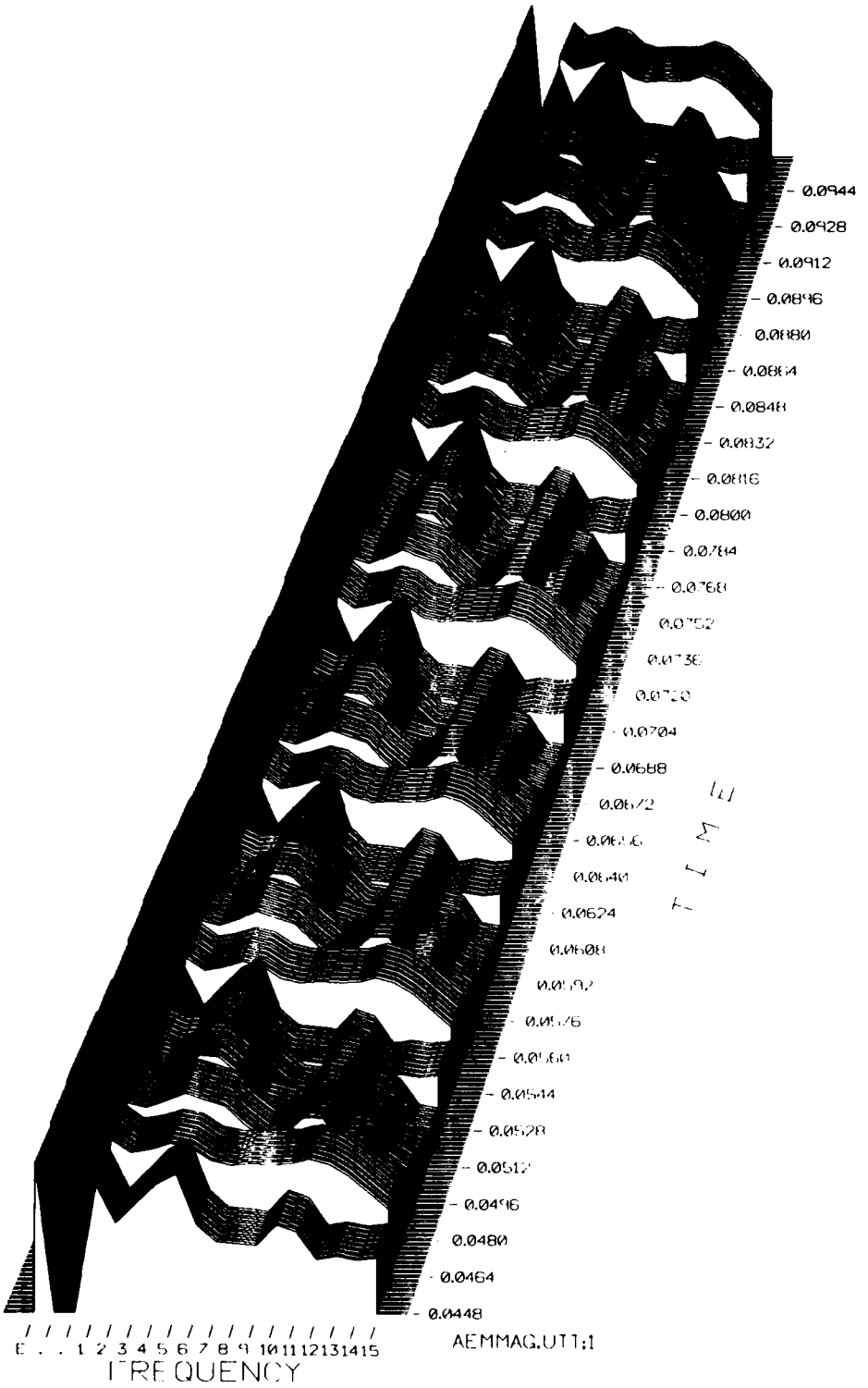


Figure 4.36: 3D Plot of Modified Data

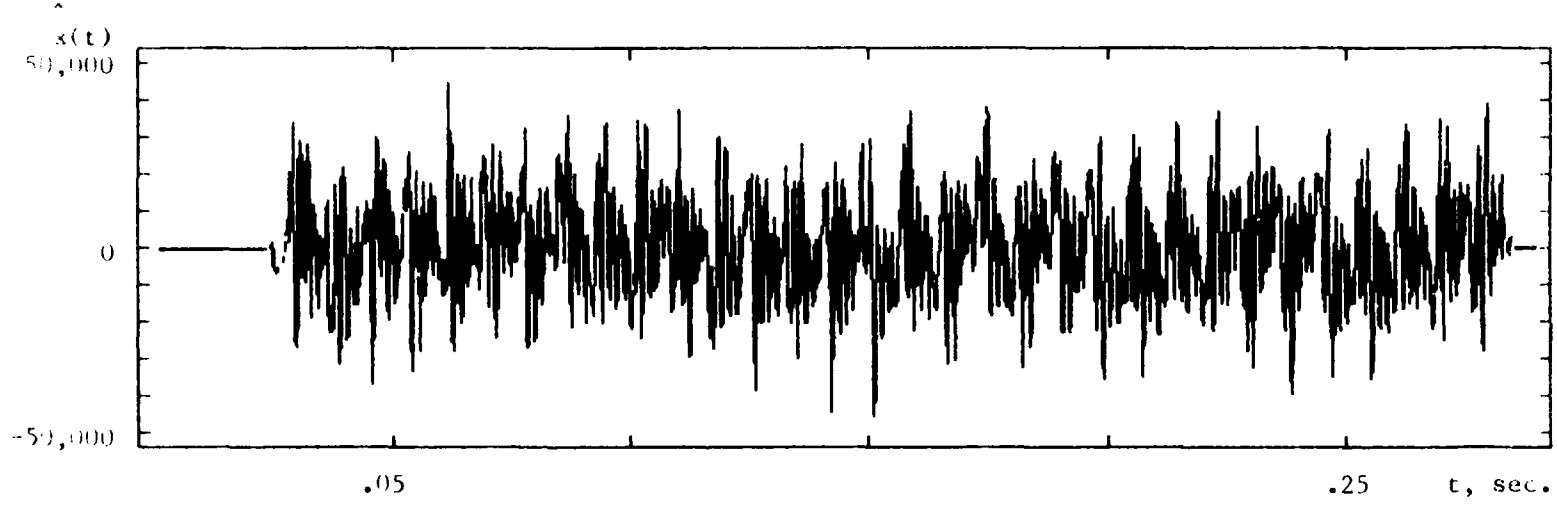


Figure 4.37: Reconstruction from Modified Data (Vowel /AE/)

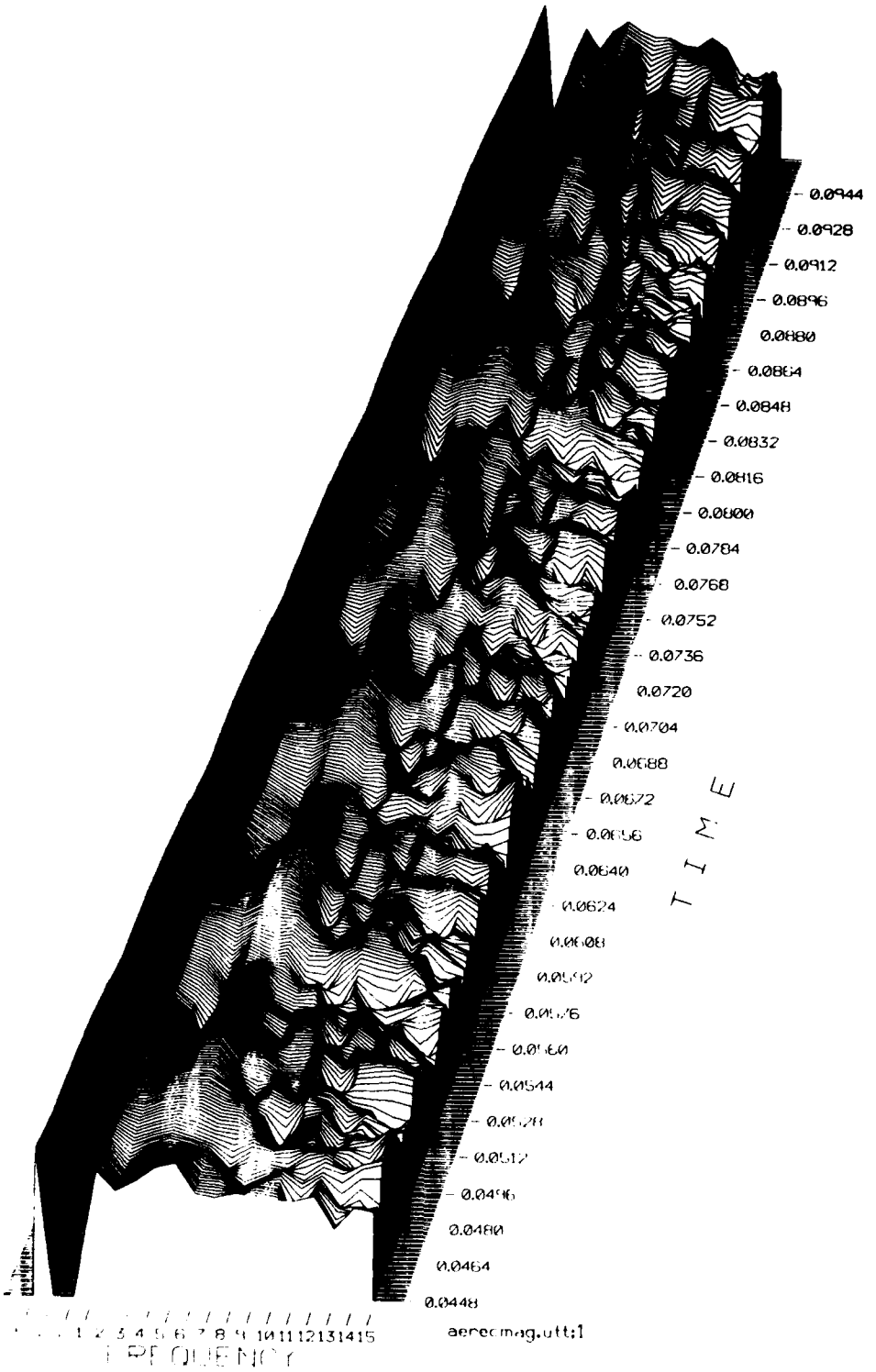


Figure 4.38: 3D Plot of Analyzed Reconstruction

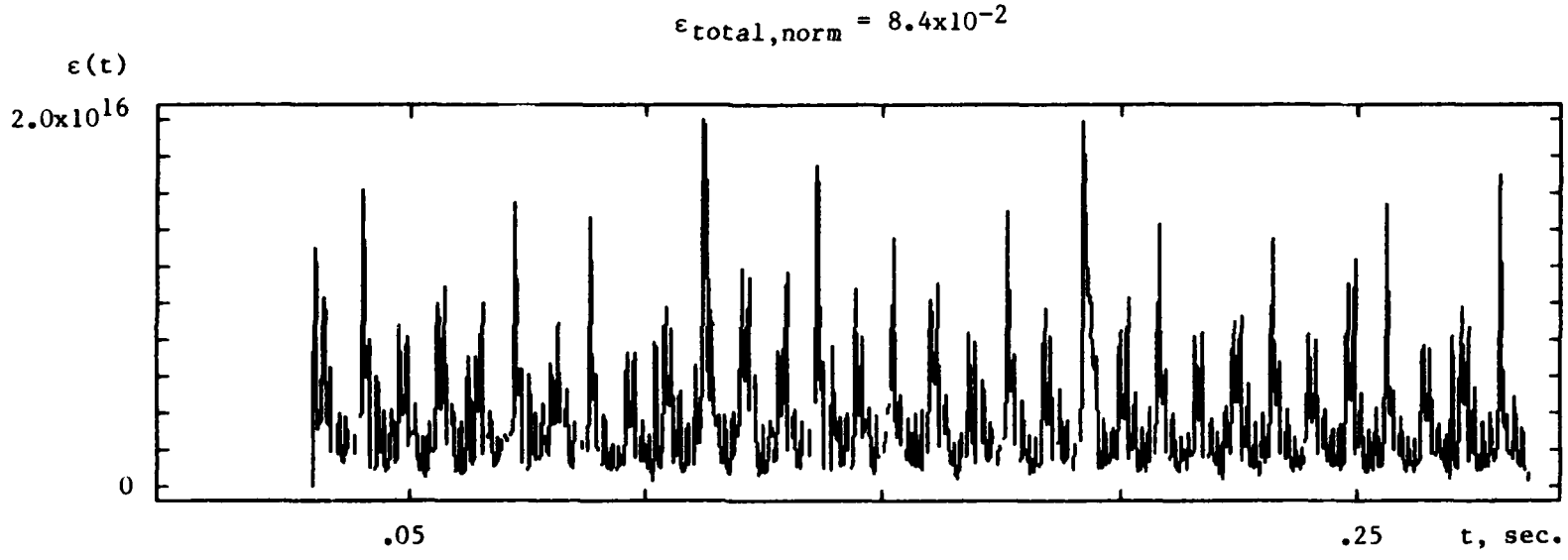


Figure 4.39: Error (Vowel /AE/)

4.5 NATURAL SPEECH SIGNALS

Fig. 4.40 presents a time-domain plot of the sentence "Their hot protein can pace on our breakdowns" as spoken by a male subject. This signal has been pre-filtered to suppress components outside the 200-3675 Hz frequency range. The signal of Fig. 4.40 was analyzed, and Fig. 4.41 is a plot of the resulting F/D outputs. In order to reduce the figure size, one of every eight F/D output samples (in the time domain) was used to create the 3D plot of Fig. 4.42. Many features of the speech signal, such as vowel structures, can be seen in the analysis of Fig. 4.42. Interpretation of this type of speech display is discussed by Searle [43], [44]. The reconstruction algorithm was applied to the non-downsampled data of Fig. 4.41, and the result is shown in Fig. 4.43. Except for an overall sign factor, the reconstruction of Fig. 4.43 is indistinguishable from the original signal of Fig. 4.40.

For demonstration purposes, a short phrase "their hot," shown in Fig. 4.44, was obtained from the sentence of Fig. 4.40. The short phrase was analyzed, and a portion of the results are shown in the 3D plot of Fig. 4.45. The reconstruction algorithm was applied to the unmodified F/D outputs, and the signal of Fig. 4.46 was obtained. The reconstructed signal of Fig. 4.46 is indistinguishable from the original signal of Fig. 4.44, and has the same overall polarity.

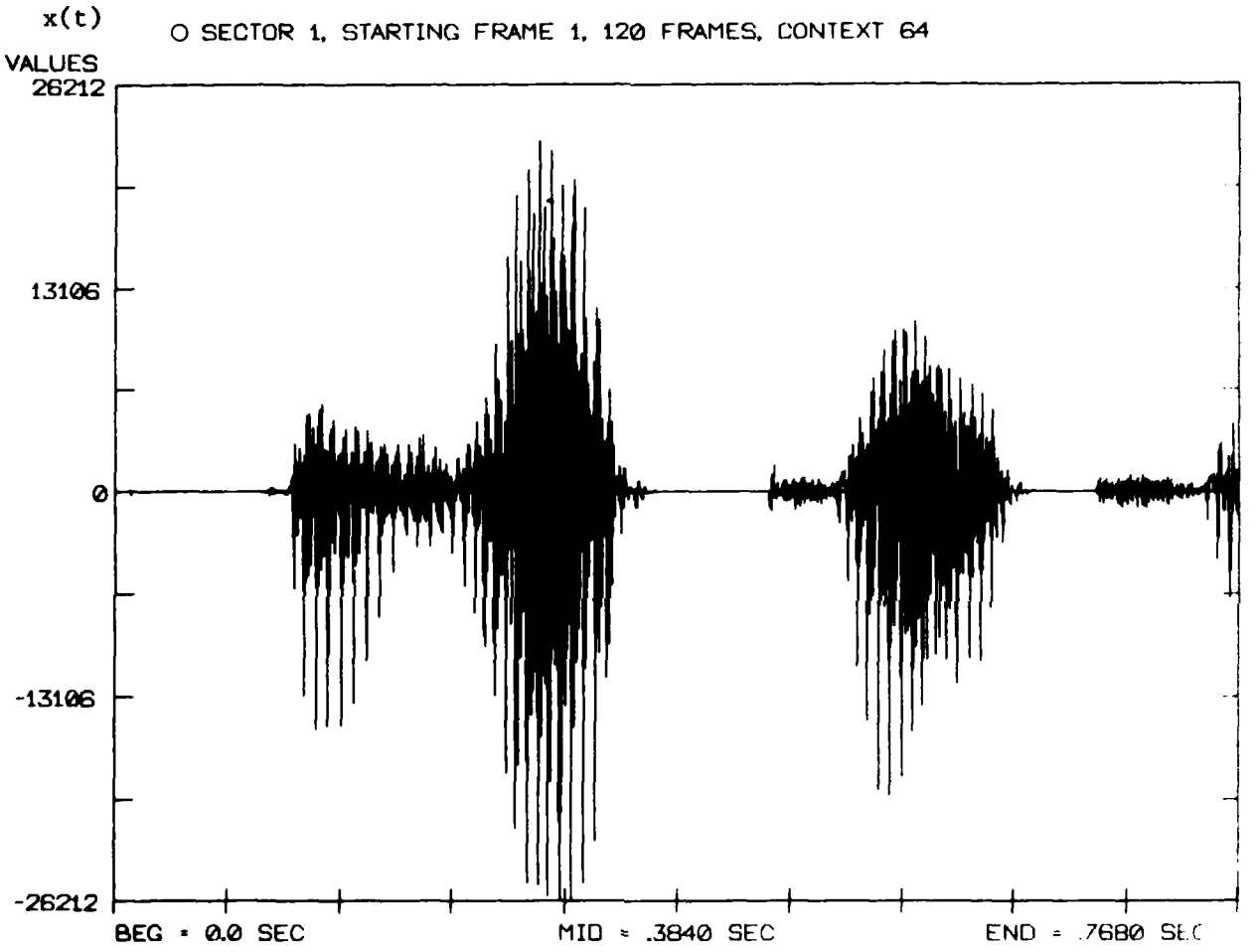


Figure 4.40: Original Signal

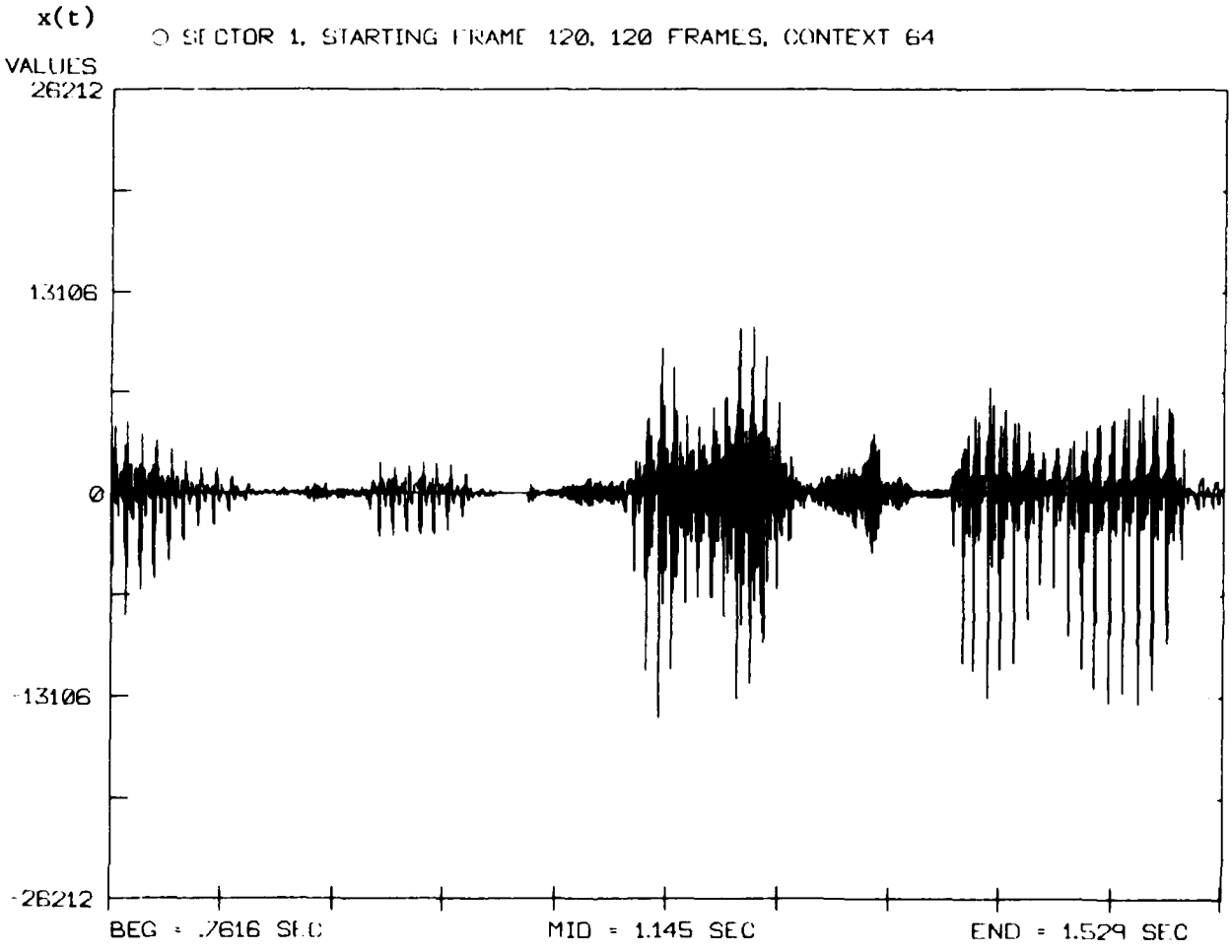


Figure 4.40 Continued

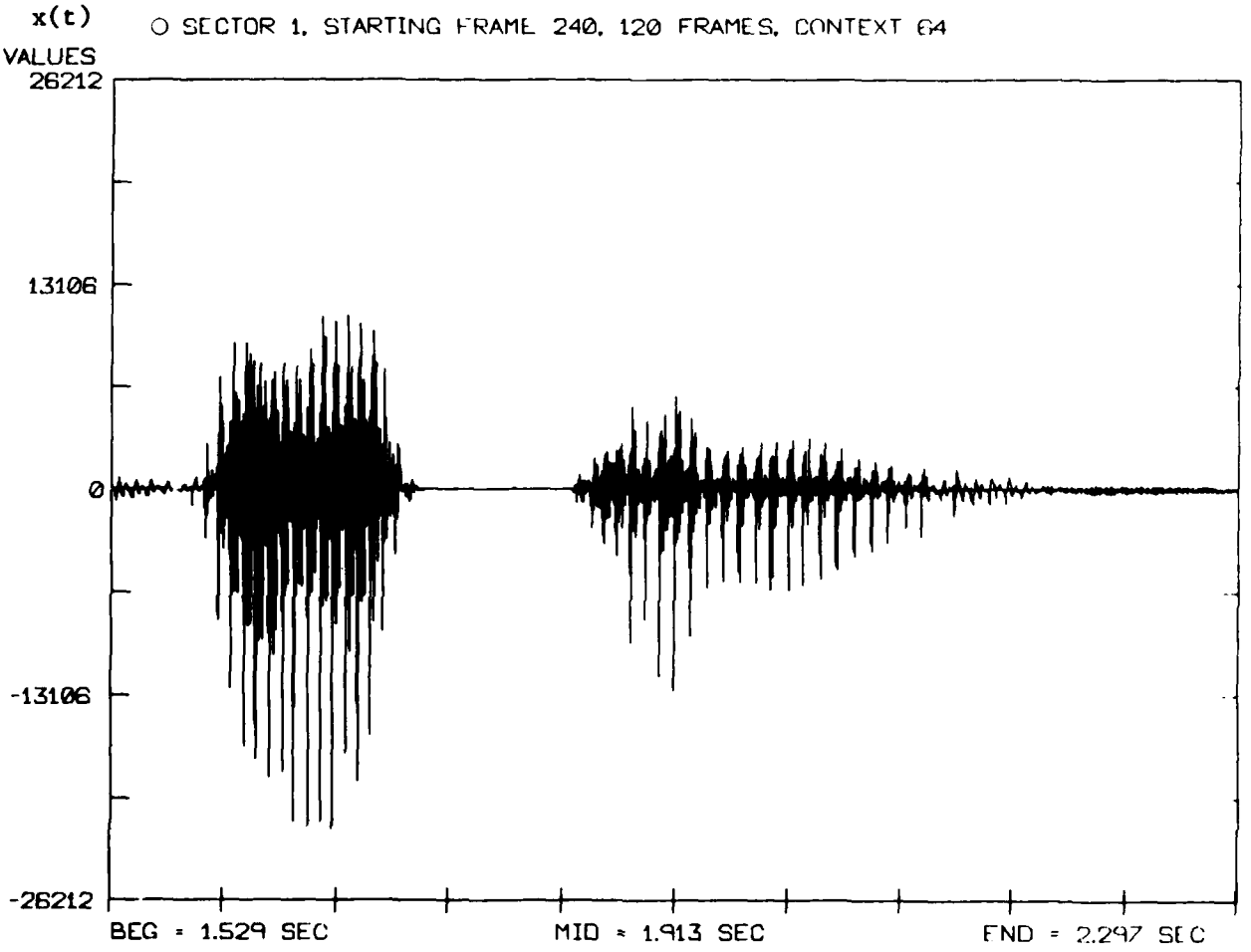


Figure 4.40 Continued

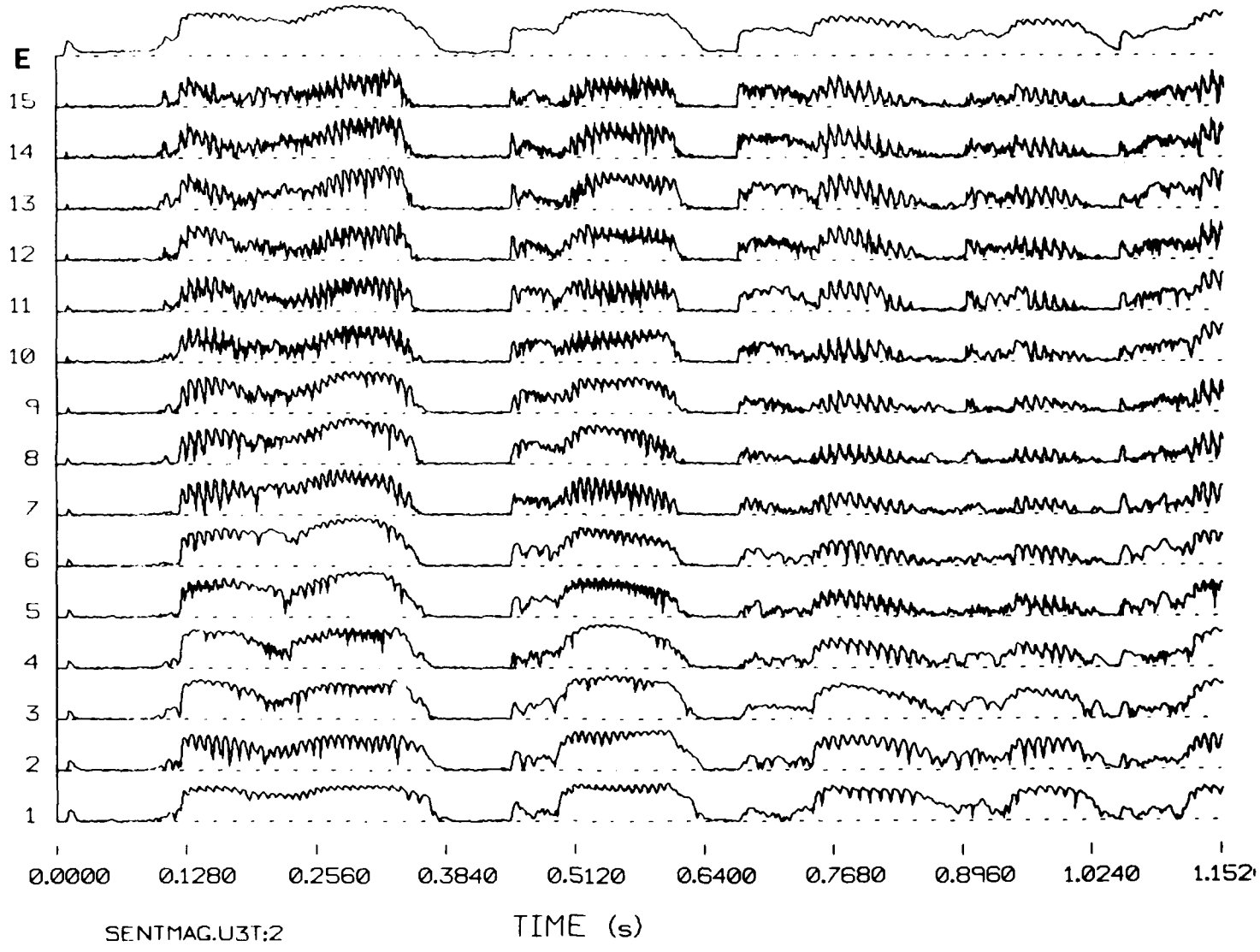


Figure 4.41: F/D Outputs (Log Amplitude)

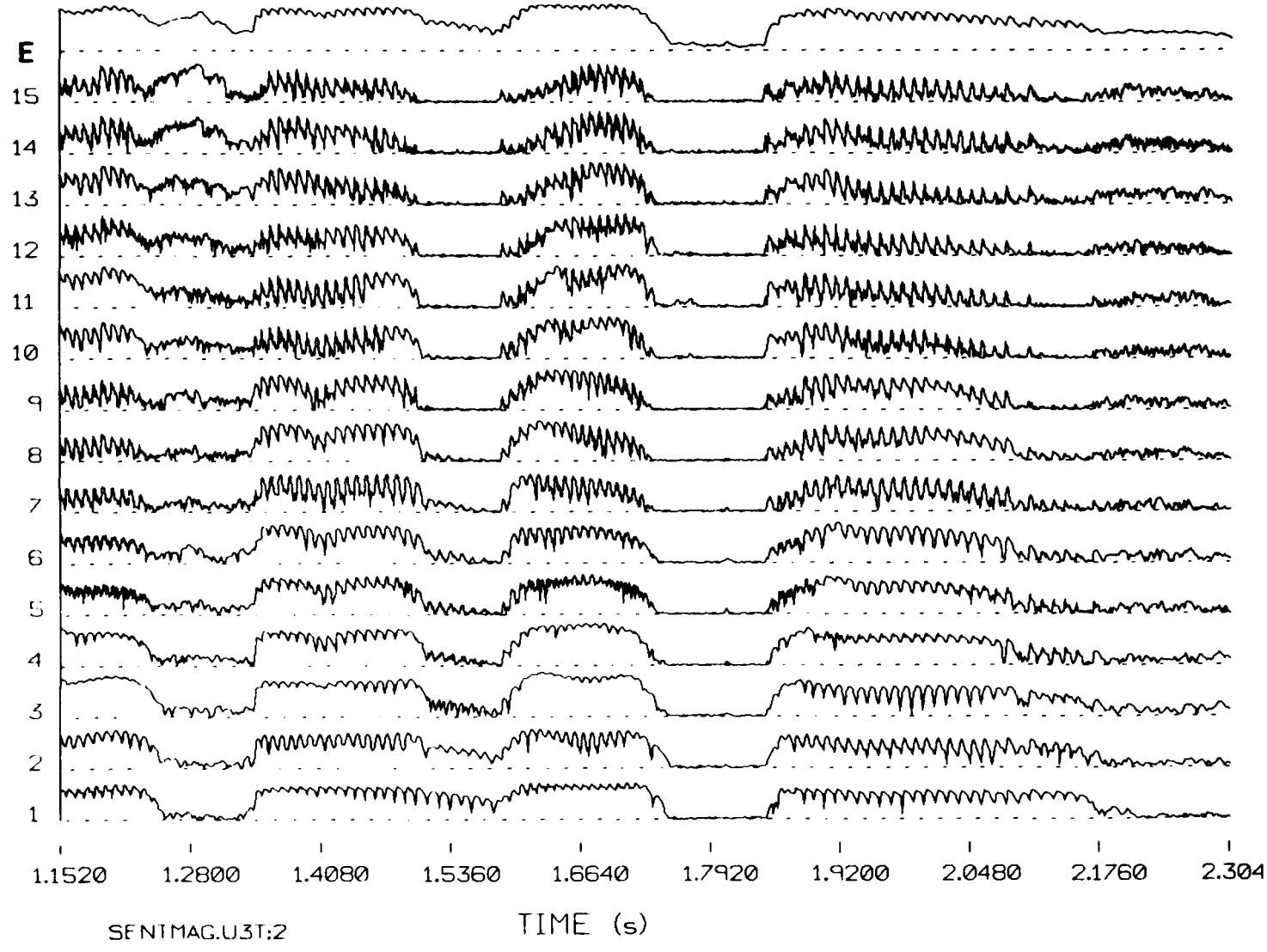


Figure 4.41 Continued

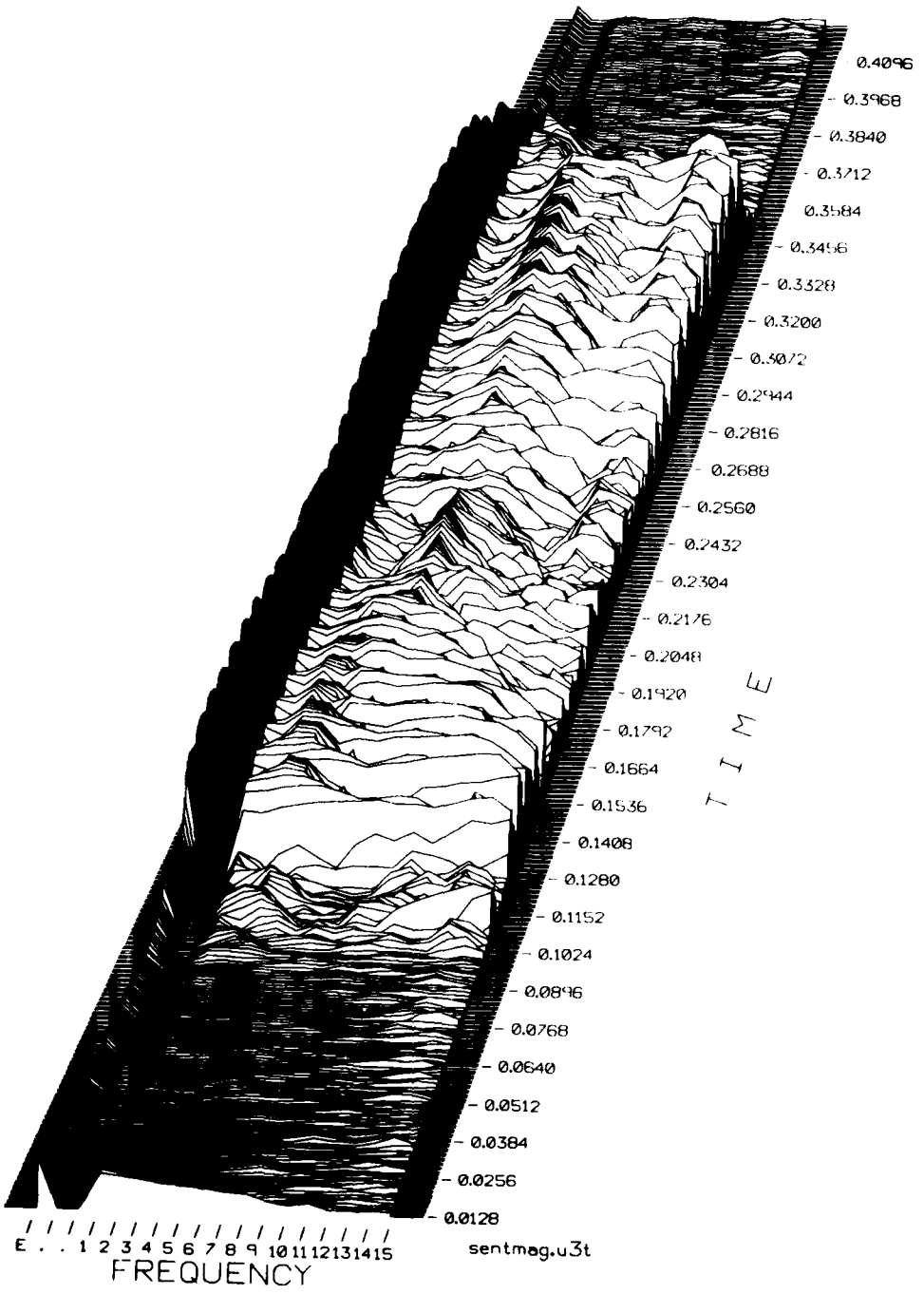


Figure 4.42: 3D Plot of Downsampled Data

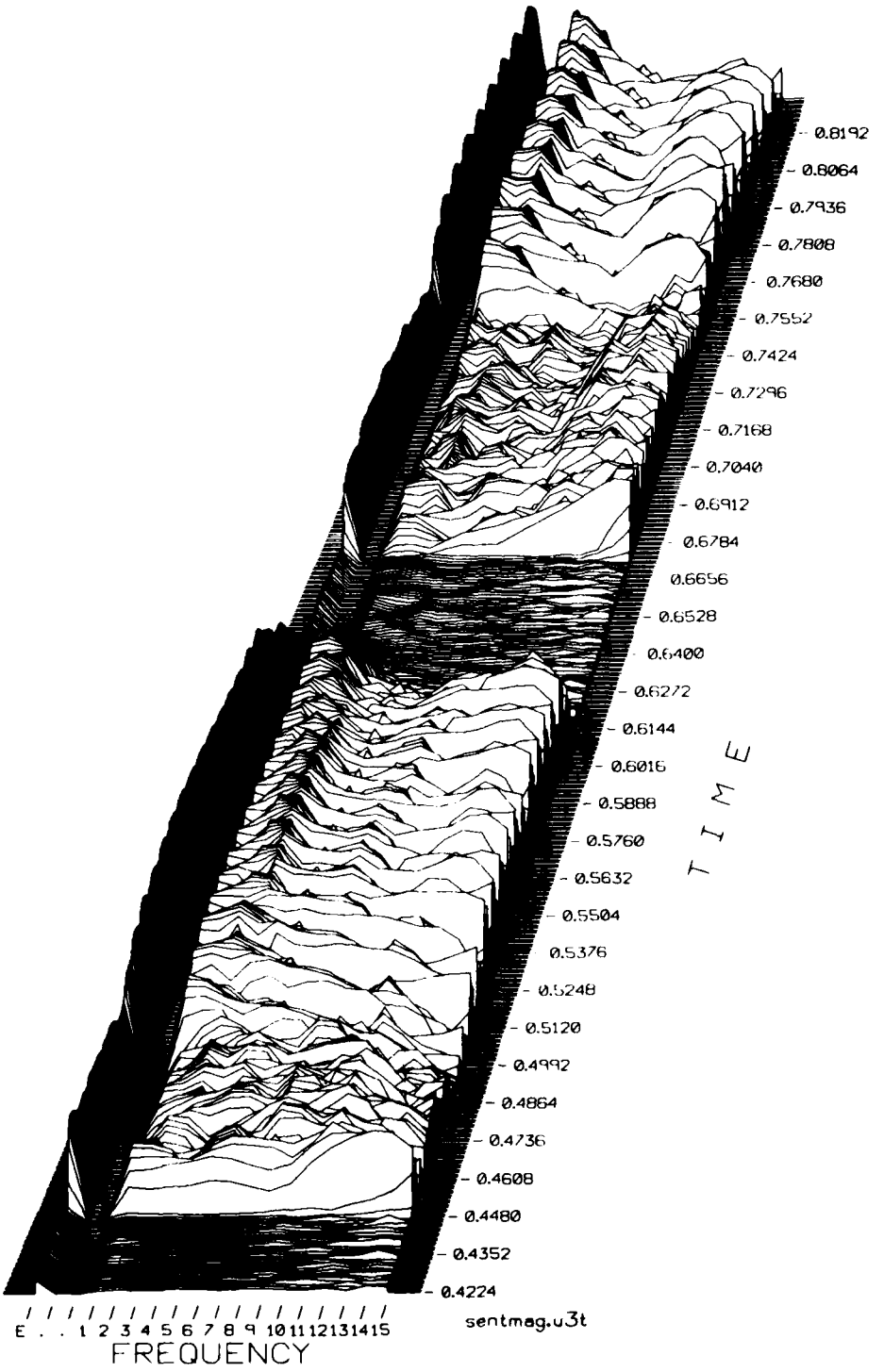


Figure 4.42 Continued

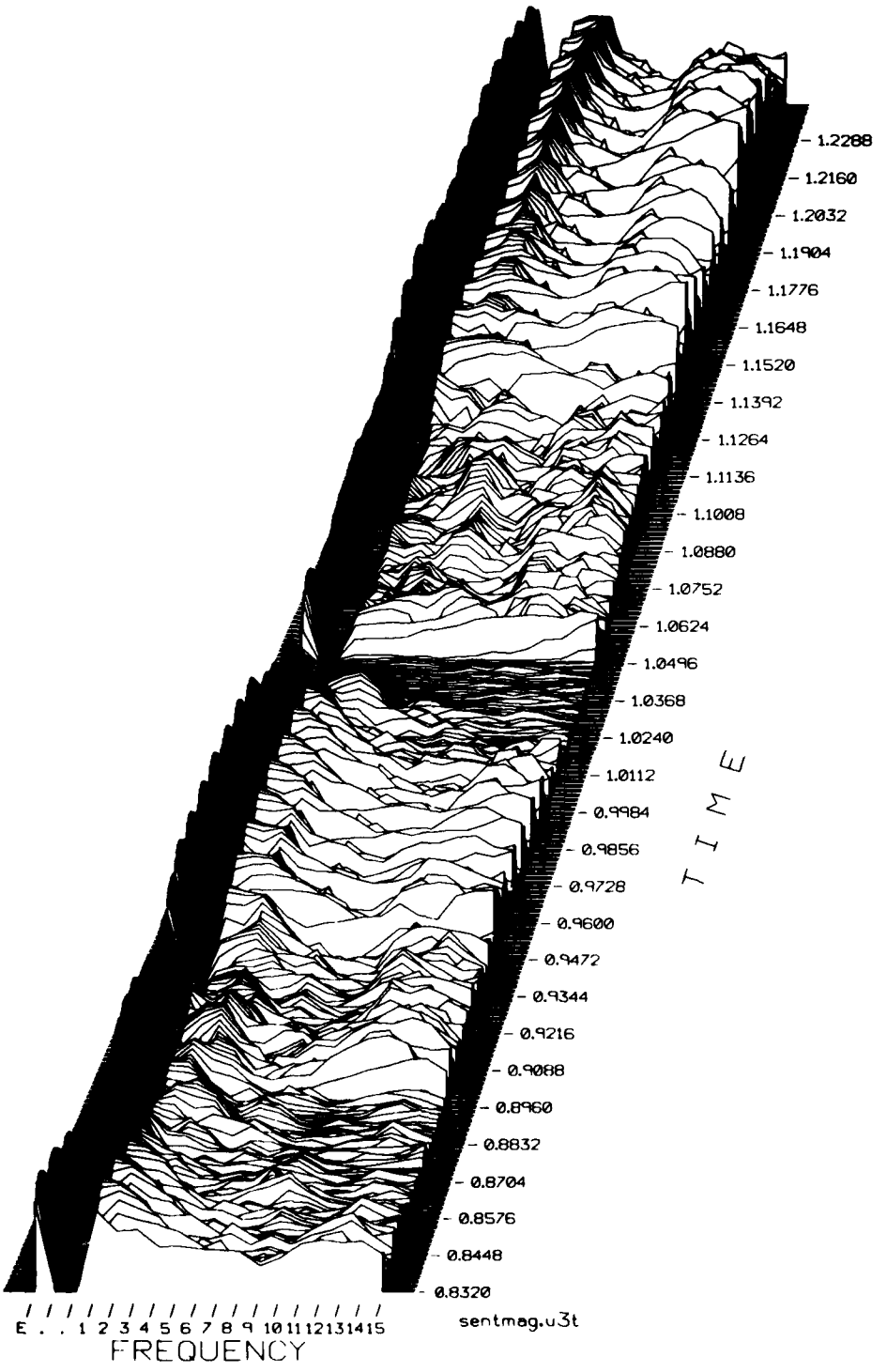


Figure 4.42 Continued

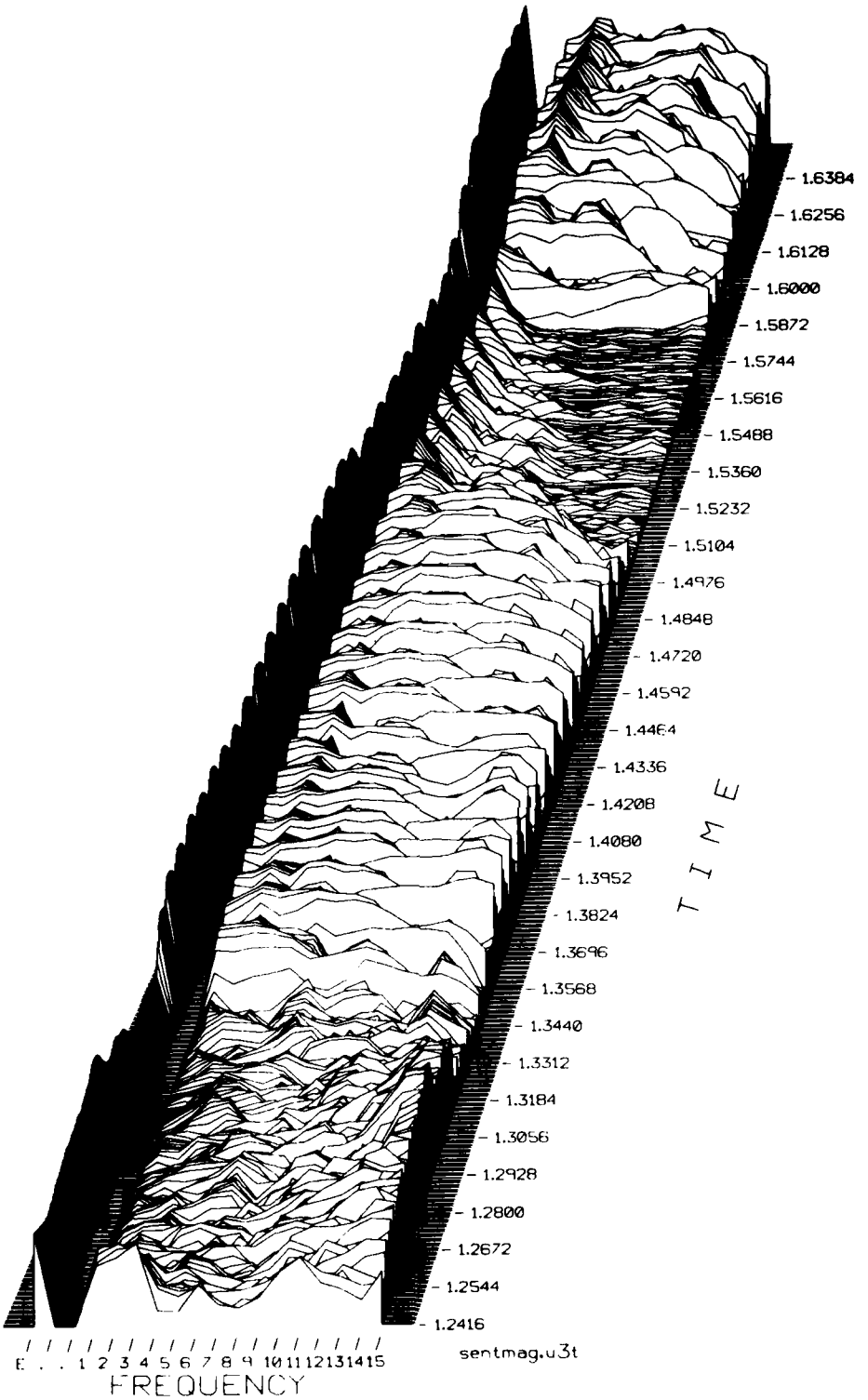


Figure 4.42 Continued

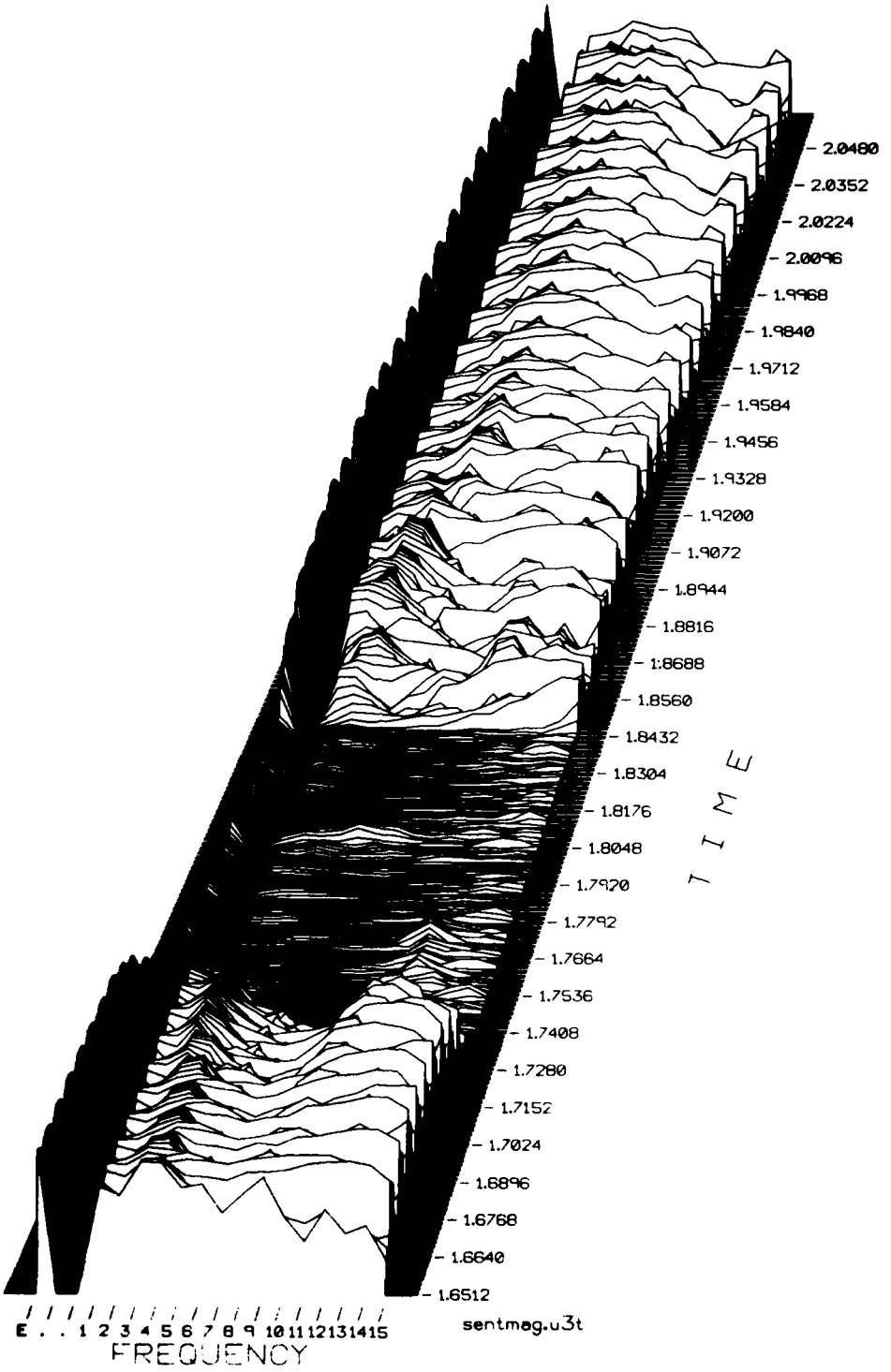


Figure 4.41 Continued

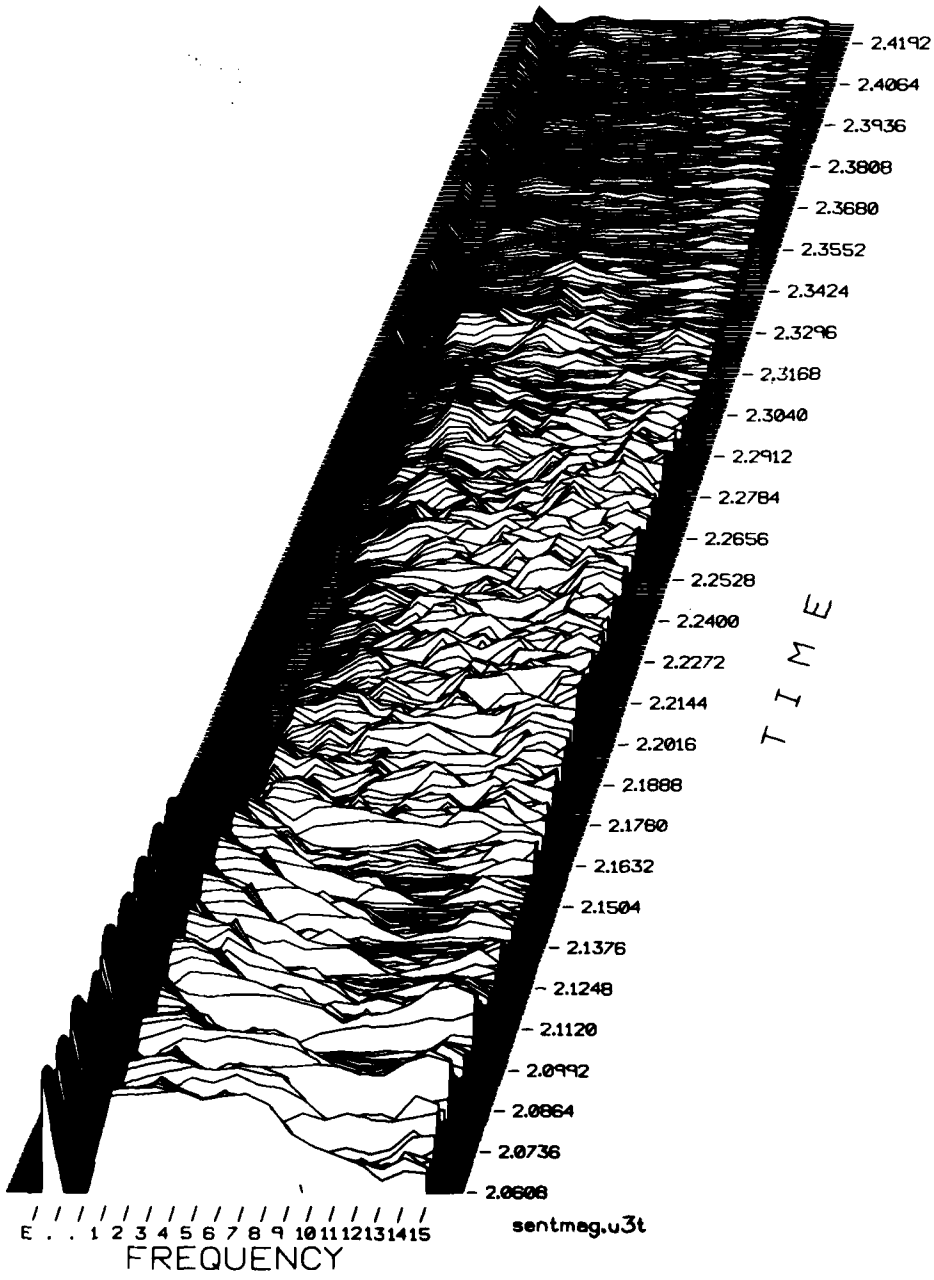


Figure 4.42 Continued

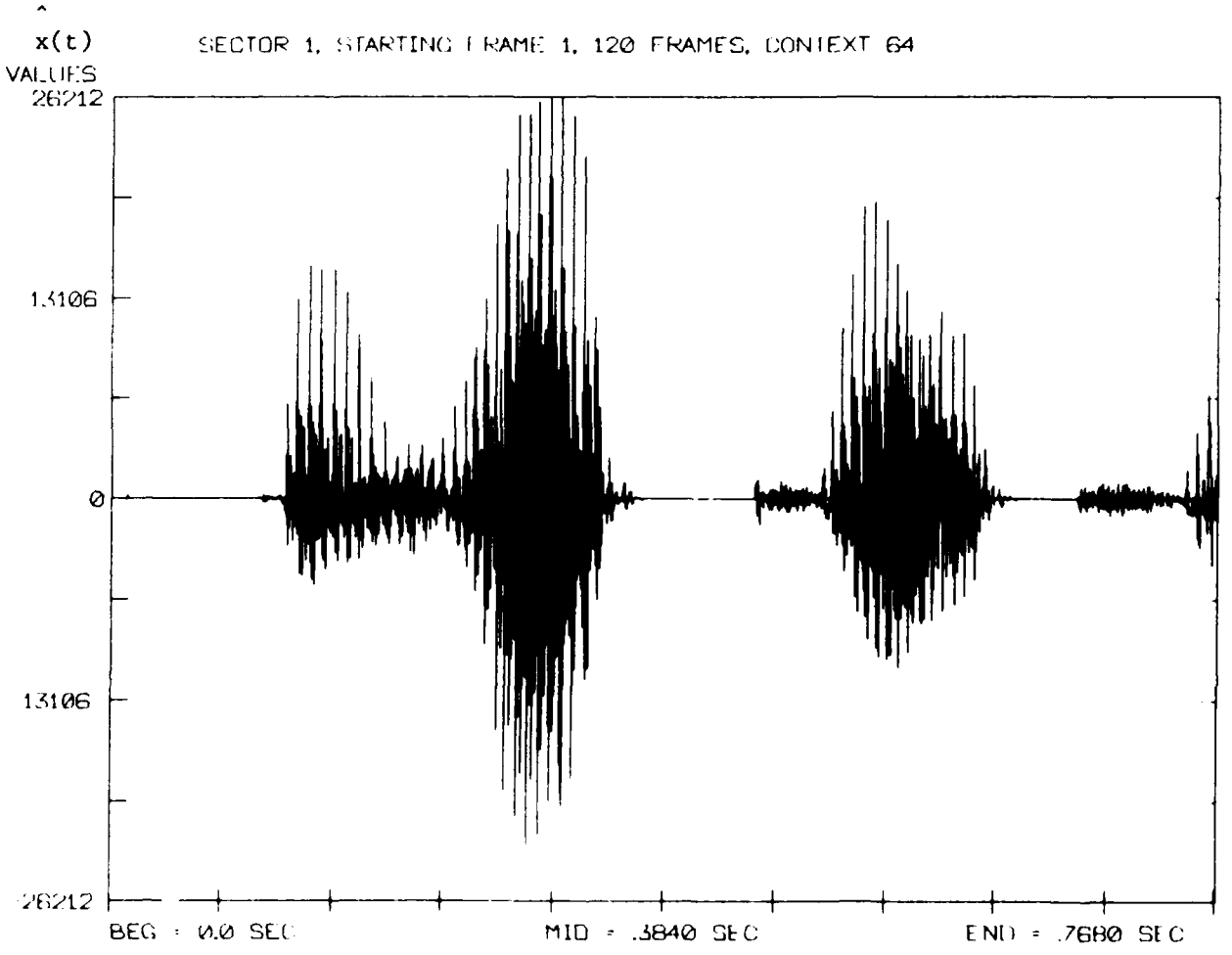


Figure 4.43: Reconstruction from Unmodified Data

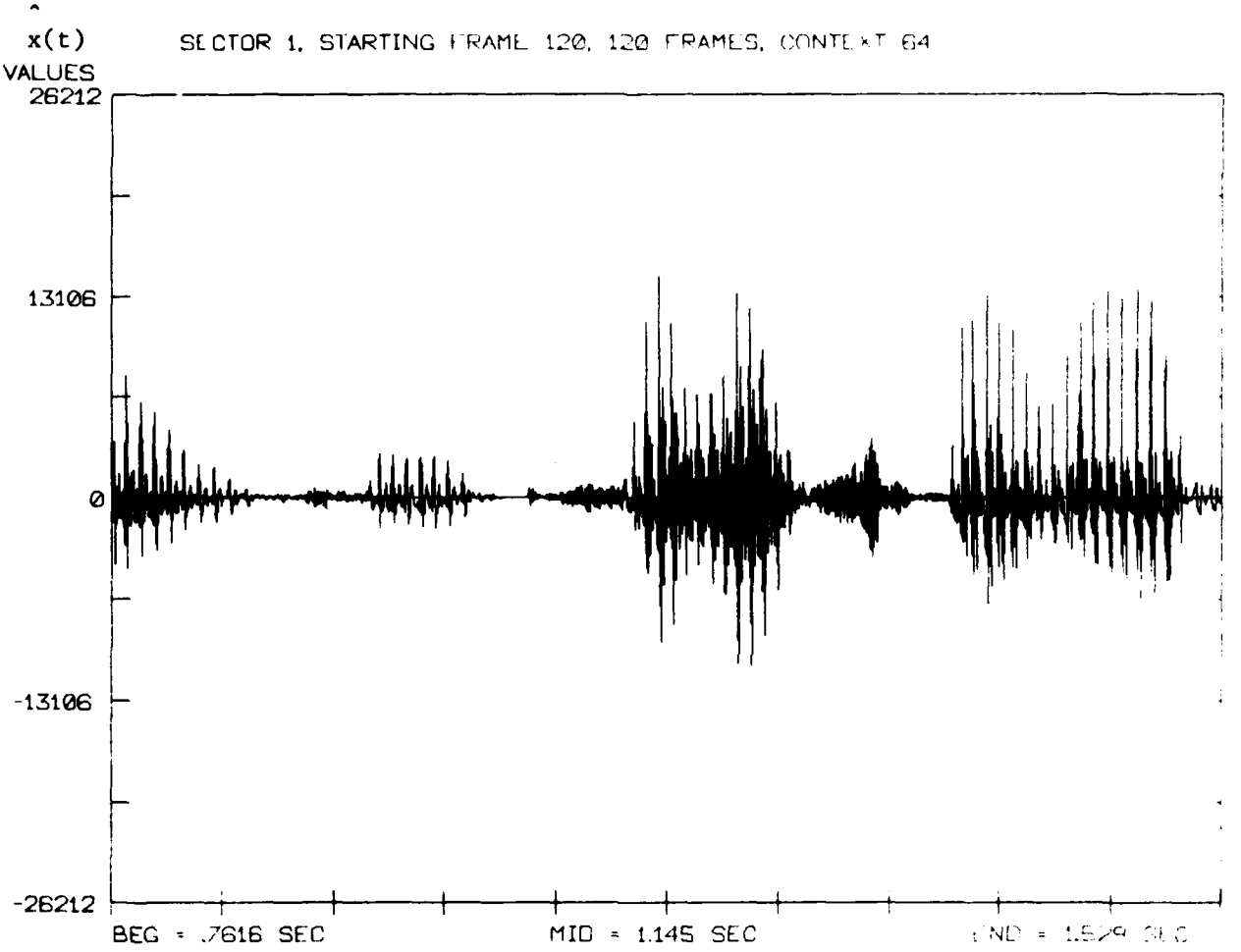


Figure 4.43 Continued

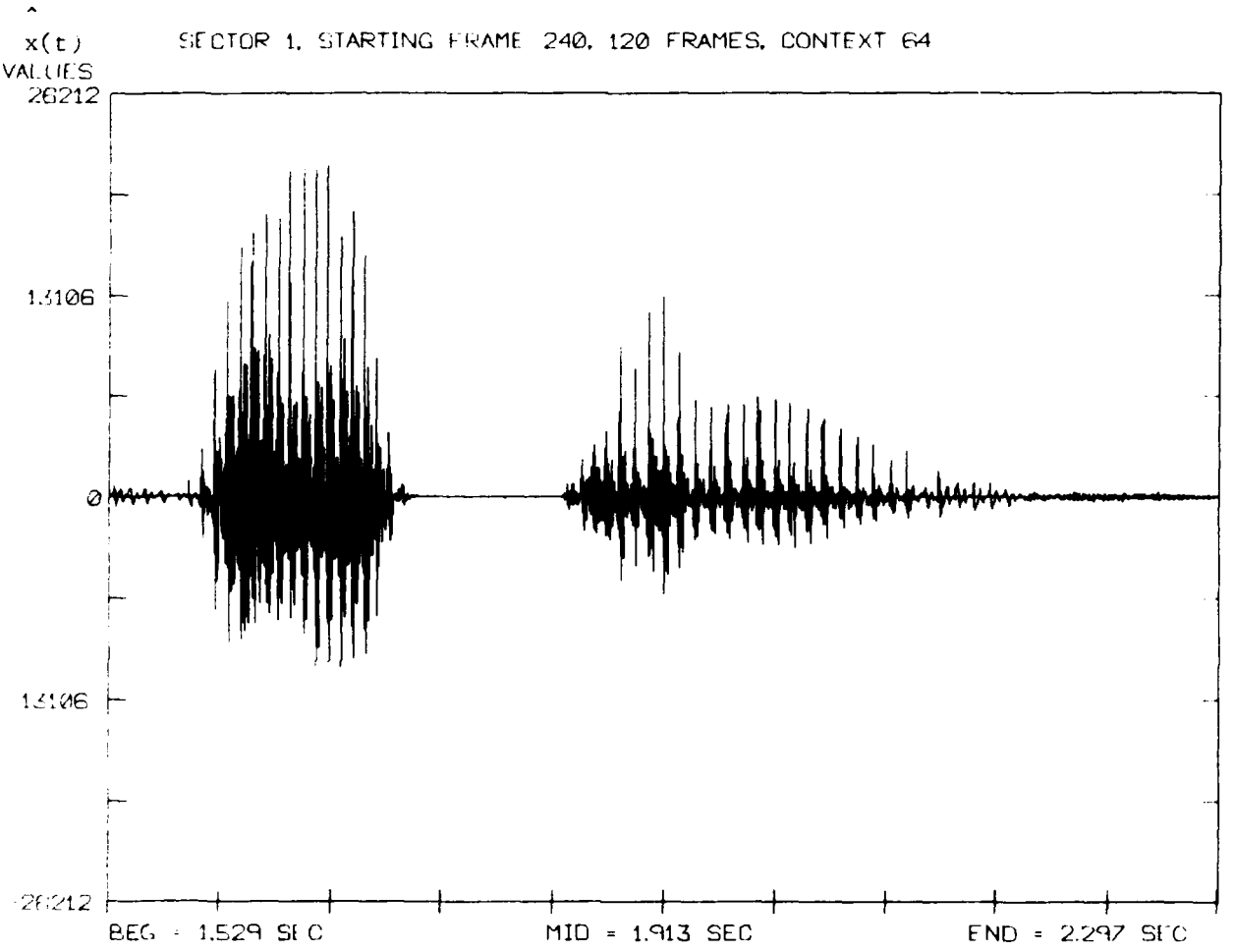


Figure 4.43 Continued

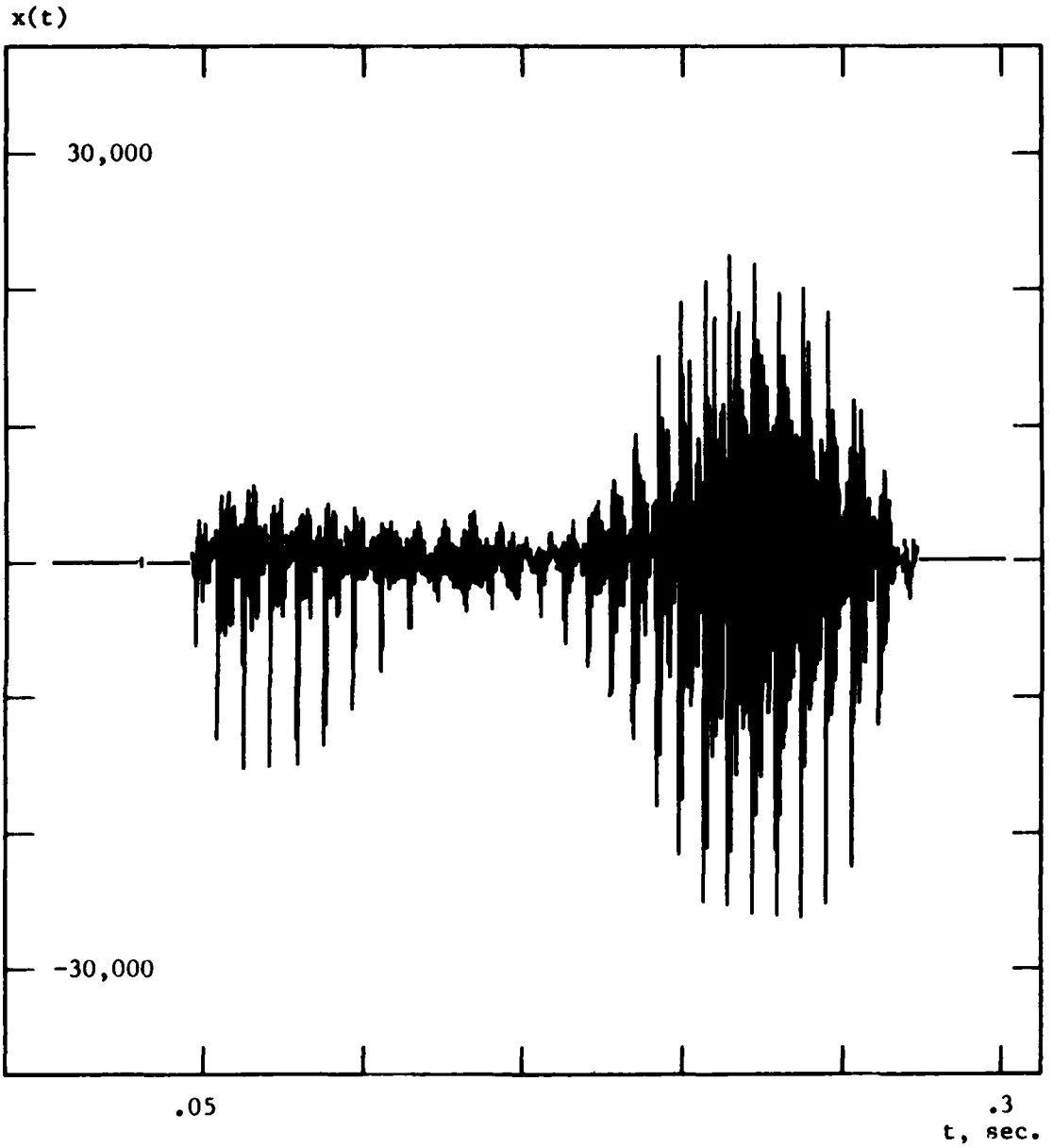


Figure 4.44: Original Signal

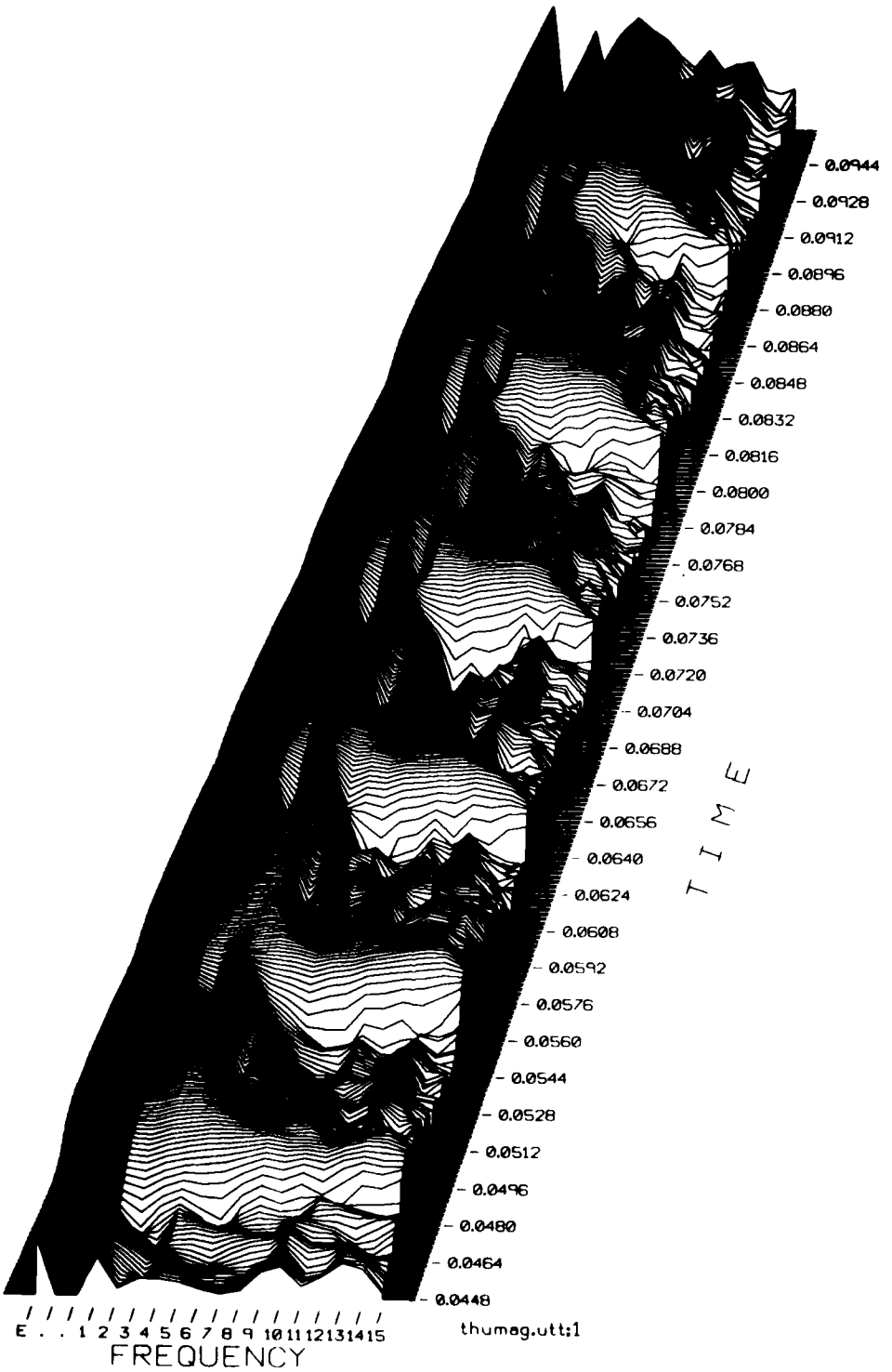


Figure 4.45: 3D Plot of Unmodified Data

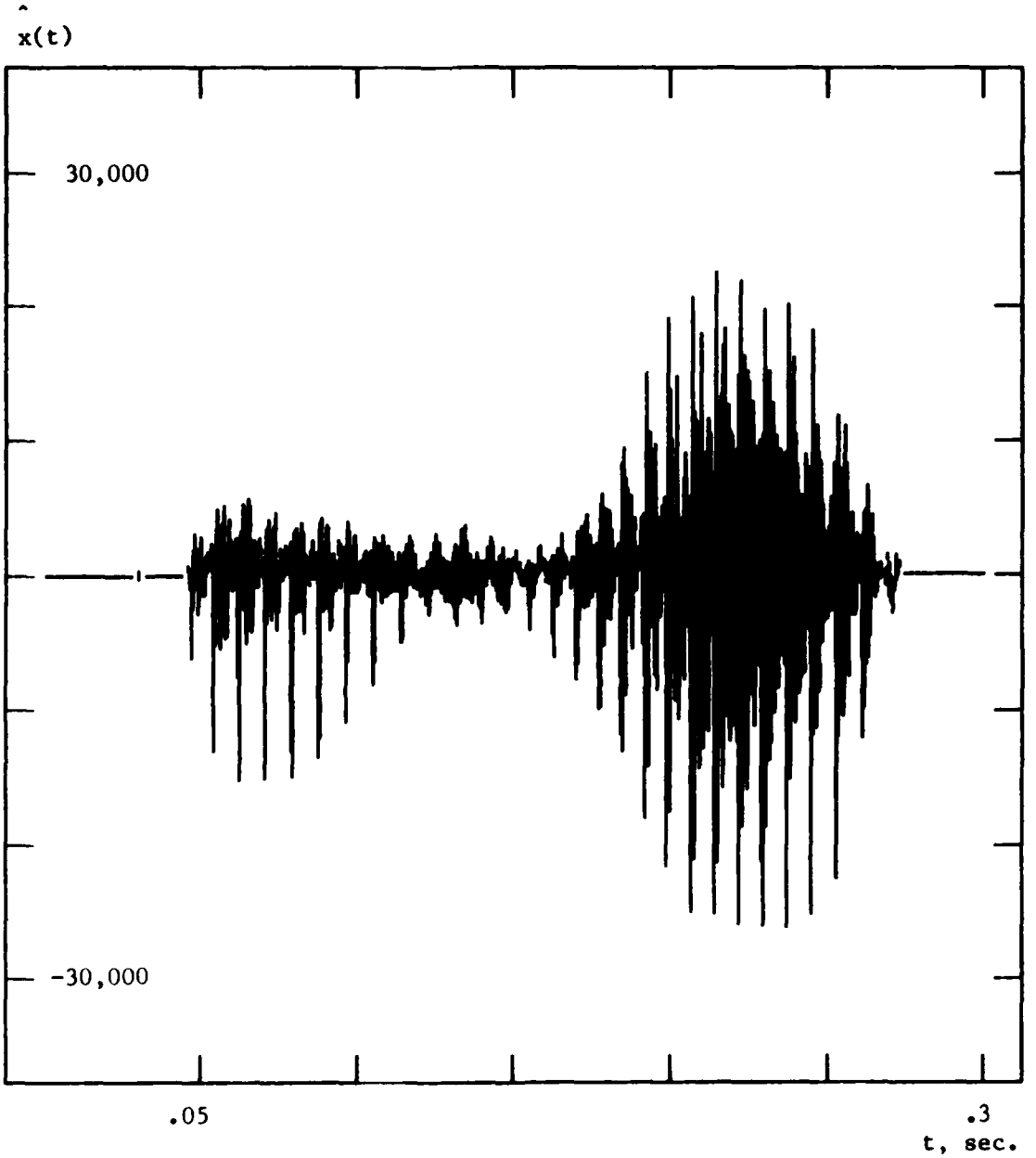


Figure 4.46: Reconstruction from Unmodified Data

The F/D outputs were then modified by averaging sixteen time-domain samples together in each channel, and replacing the samples with average values. A portion of the modified data is shown in the 3D plot of Fig. 4.47, and the resulting reconstruction is shown in Fig. 4.48. Note that the correct pitch has been retained, and the reconstructed signal appears somewhat noisy. The analyzed reconstruction is shown in Fig. 4.49, and the corresponding error in Fig. 4.50. Since the average signal level is less than in previous examples, the peak error value of 7.4×10^{15} is also less. The total normalized error value of 7.4×10^{-2}, however, is comparable to that of previous examples. The total normalized error does not depend strongly on signal duration for the examples given in this section.

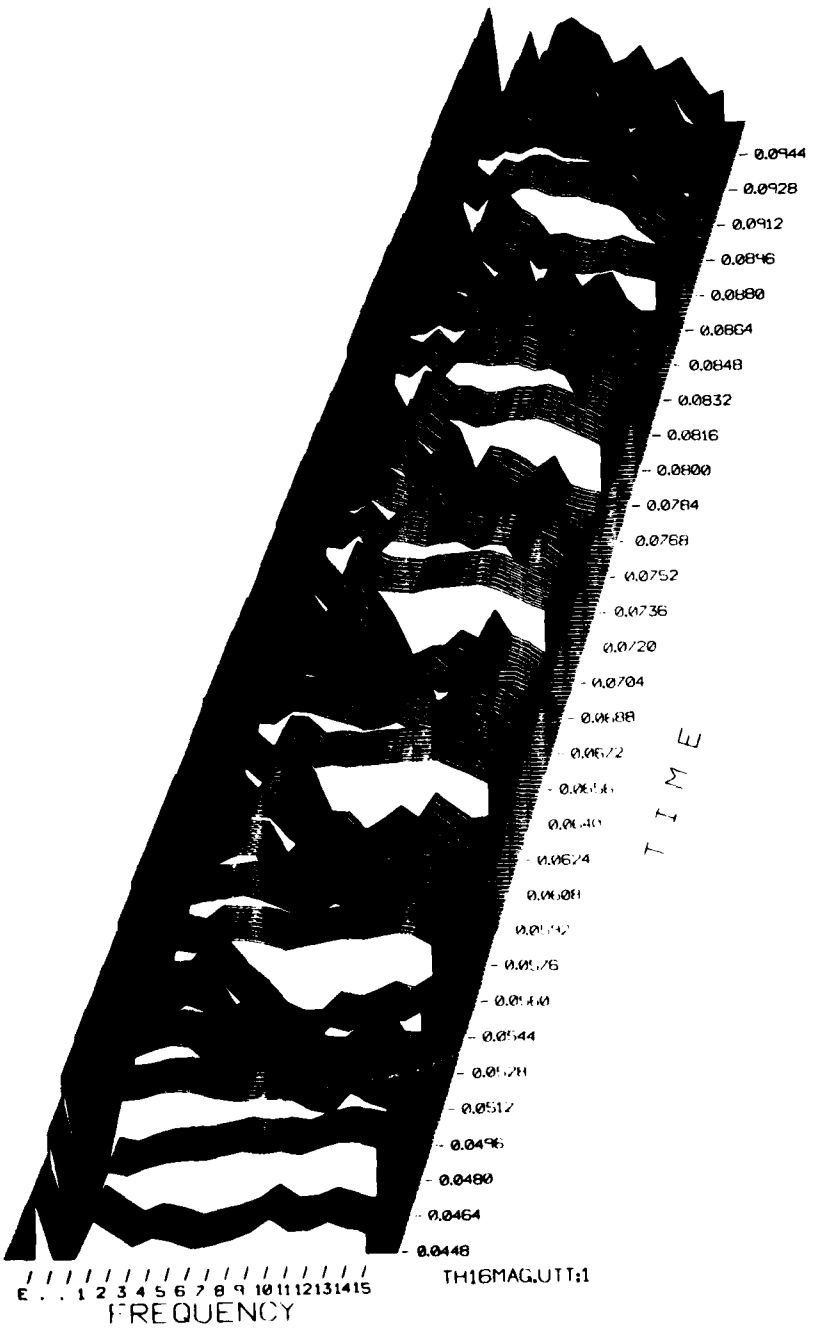


Figure 4.47: 3D Plot of Modified Data

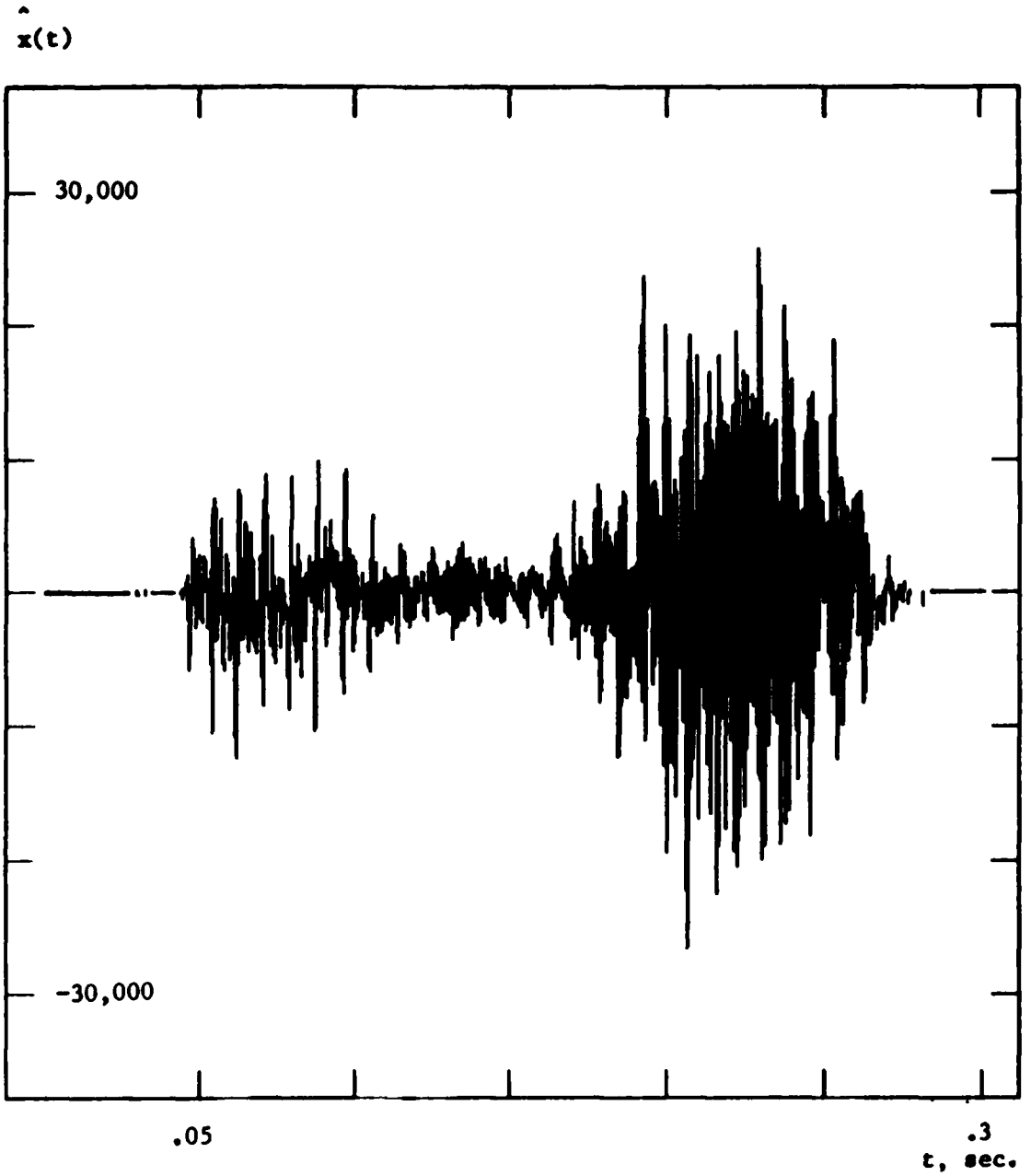


Figure 4.48: Reconstruction from Modified Data

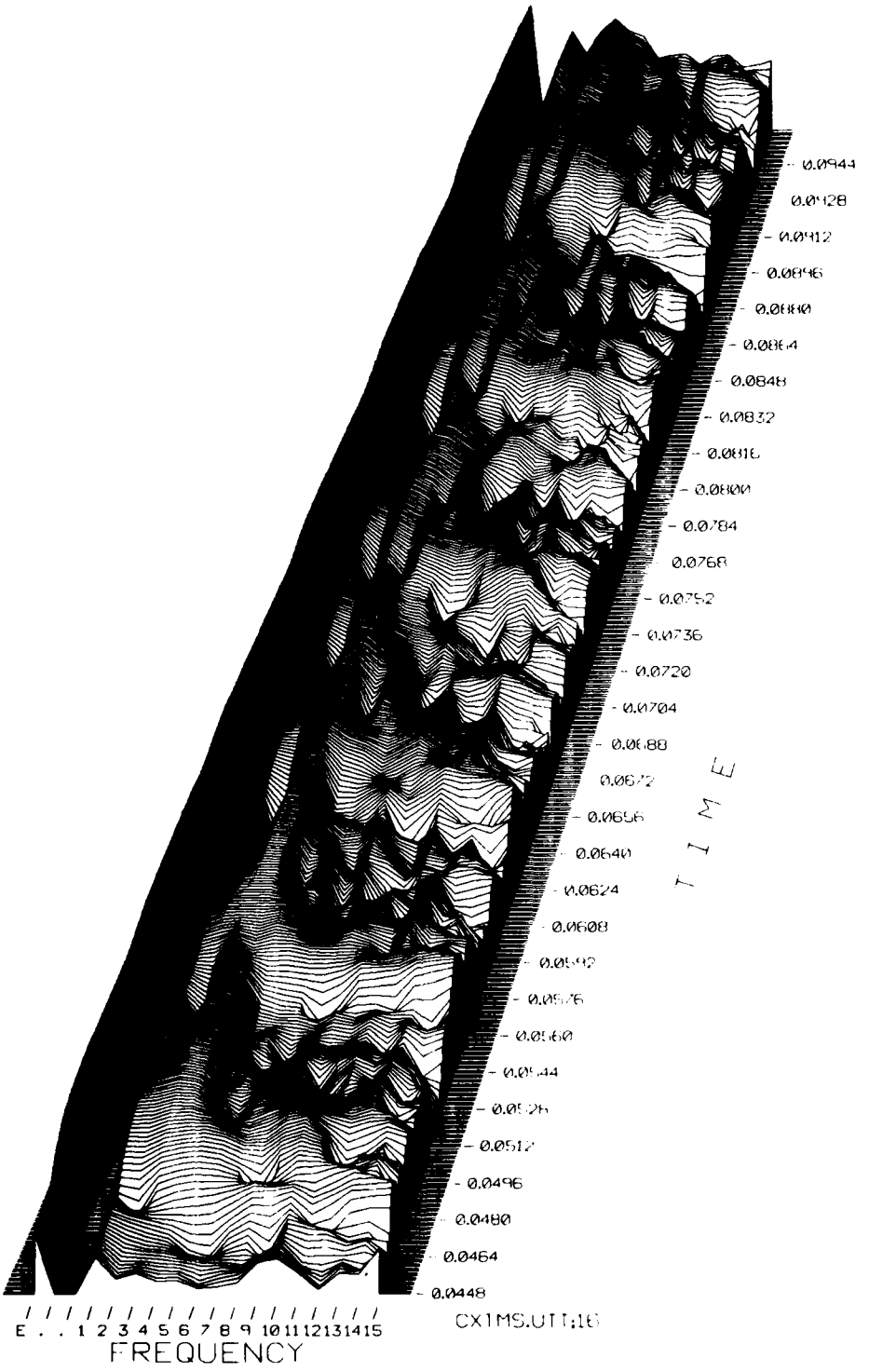


Figure 4.49: 3D Plot of Analyzed Reconstruction (Modified Data)

$\varepsilon_{total,norm} = 7.4 \times 10^{-2}$

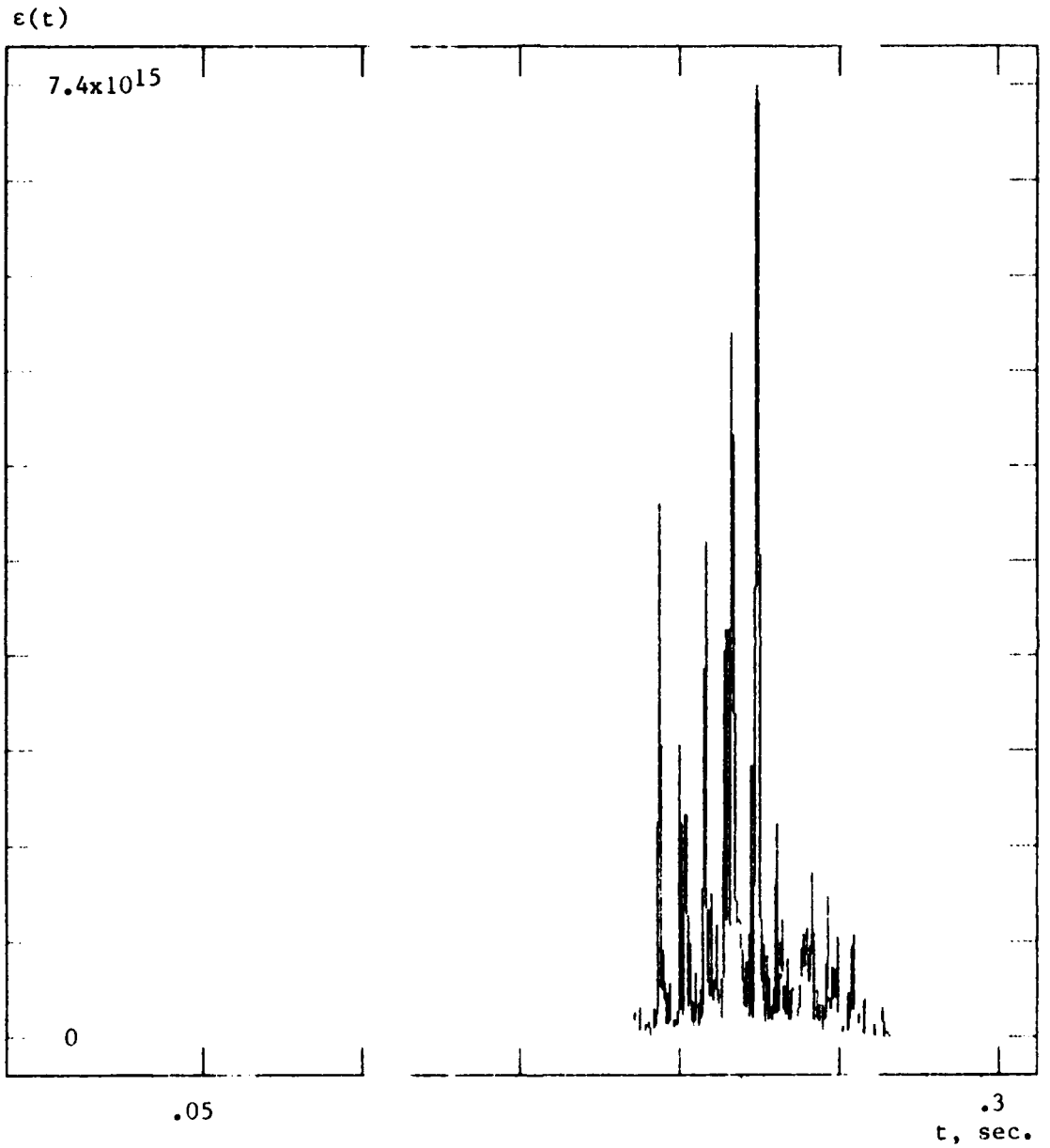


Figure 4.50: Error (Modified Data)

Next, a short-time spectral modification was considered in which the data was not so severely distorted. This modification averaged time-domain samples together in each channel and replaced the samples with average values, but fewer samples were averaged in the high frequency channels. To ensure that all channels were modified to some extent, two samples were averaged together in each of the high frequency channels. More samples were averaged in lower frequency channels according to their bandwidth. Specifically, 7 samples were averaged in the energy channel, 6 samples in Channels #1-2, 5 in #3-4, 4 in #5-7, 3 in #8-10, and 2 in #11-15. Note that this modification is different from the modification used in all previous examples, where the same number of samples was averaged regardless of filter bandwidth. This short-time spectral modification corresponds more closely to a simple decimation and interpolation of the F/D and STE outputs, as discussed in Section 2.6. A portion of the slightly modified data is shown in Fig. 4.51, and the resulting reconstruction is shown in Fig. 4.52. The reconstructed signal is similar to the original, although differences are clearly visible. The reconstructed speech sounds quite similar to the original signal, but the two signals are audibly distinguishable. The analyzed reconstruction is shown in Fig. 4.53, and the corresponding error in Fig. 4.54. The peak error value of 1.7×10^{15} and the total normalized error value of 9.3×10^{-3} are far less than values obtained for the previous example.

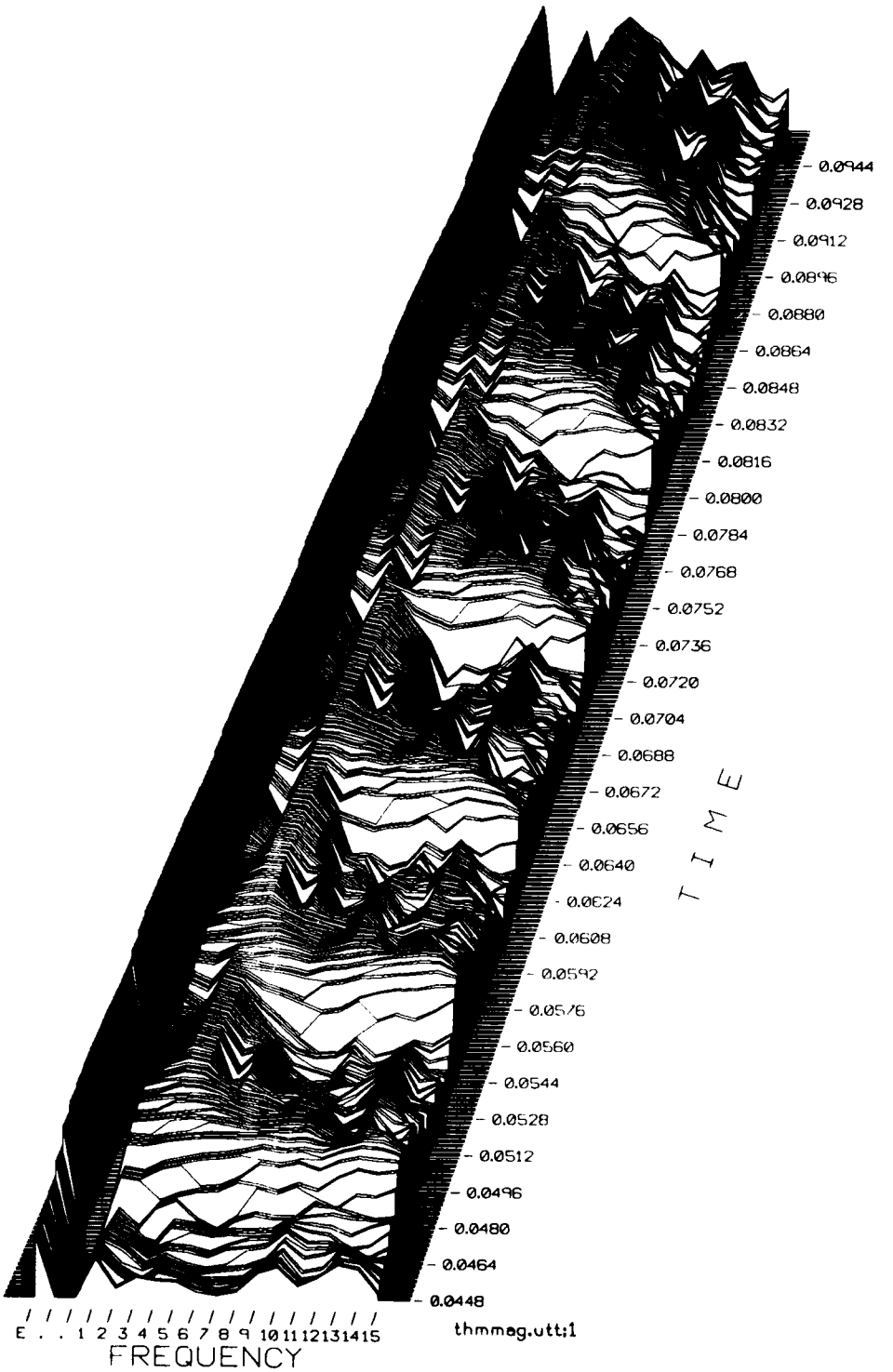


Figure 4.51: 3D Plot of Slightly Modified Data

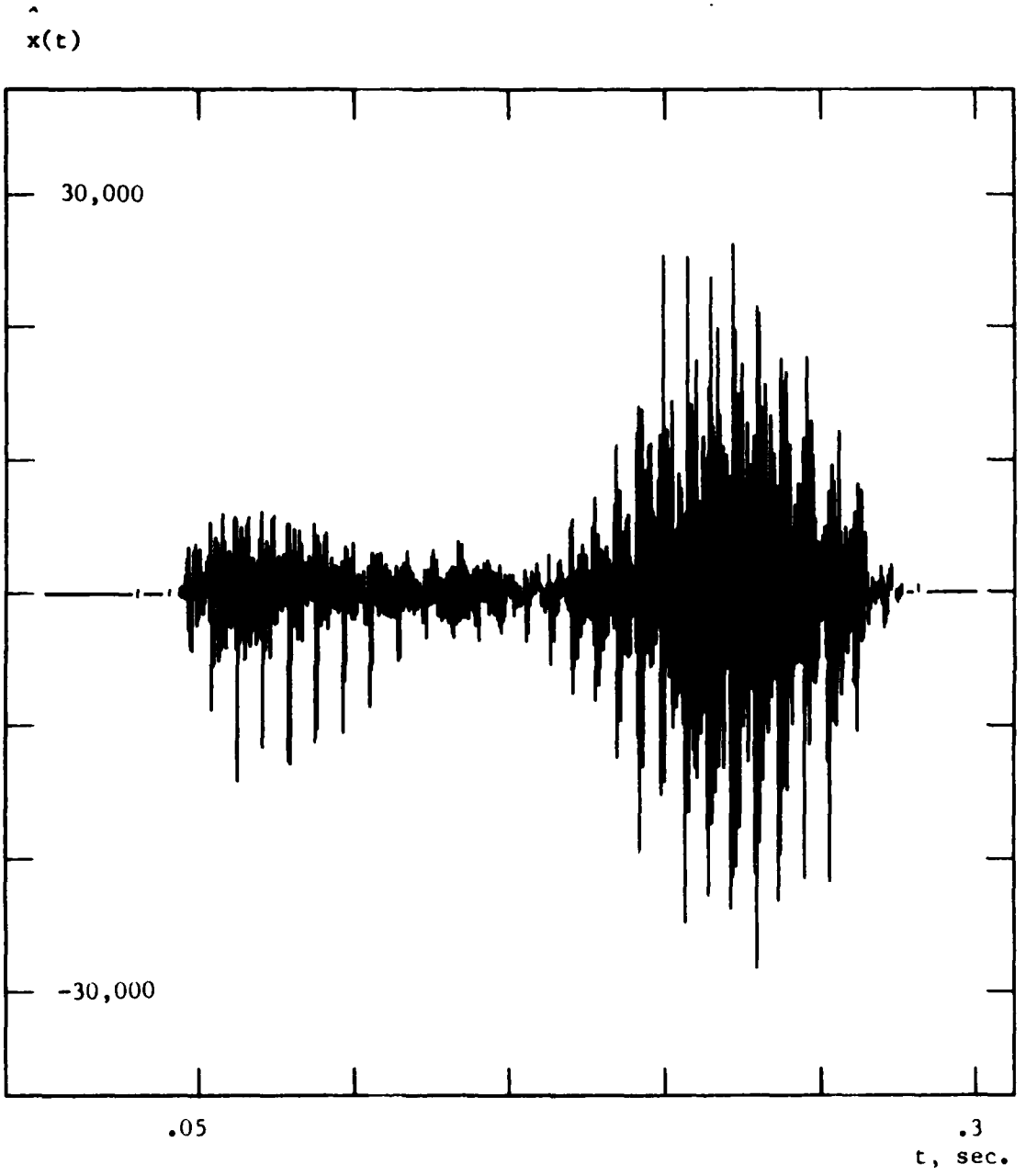


Figure 4.52: Reconstruction from Slightly Modified Data

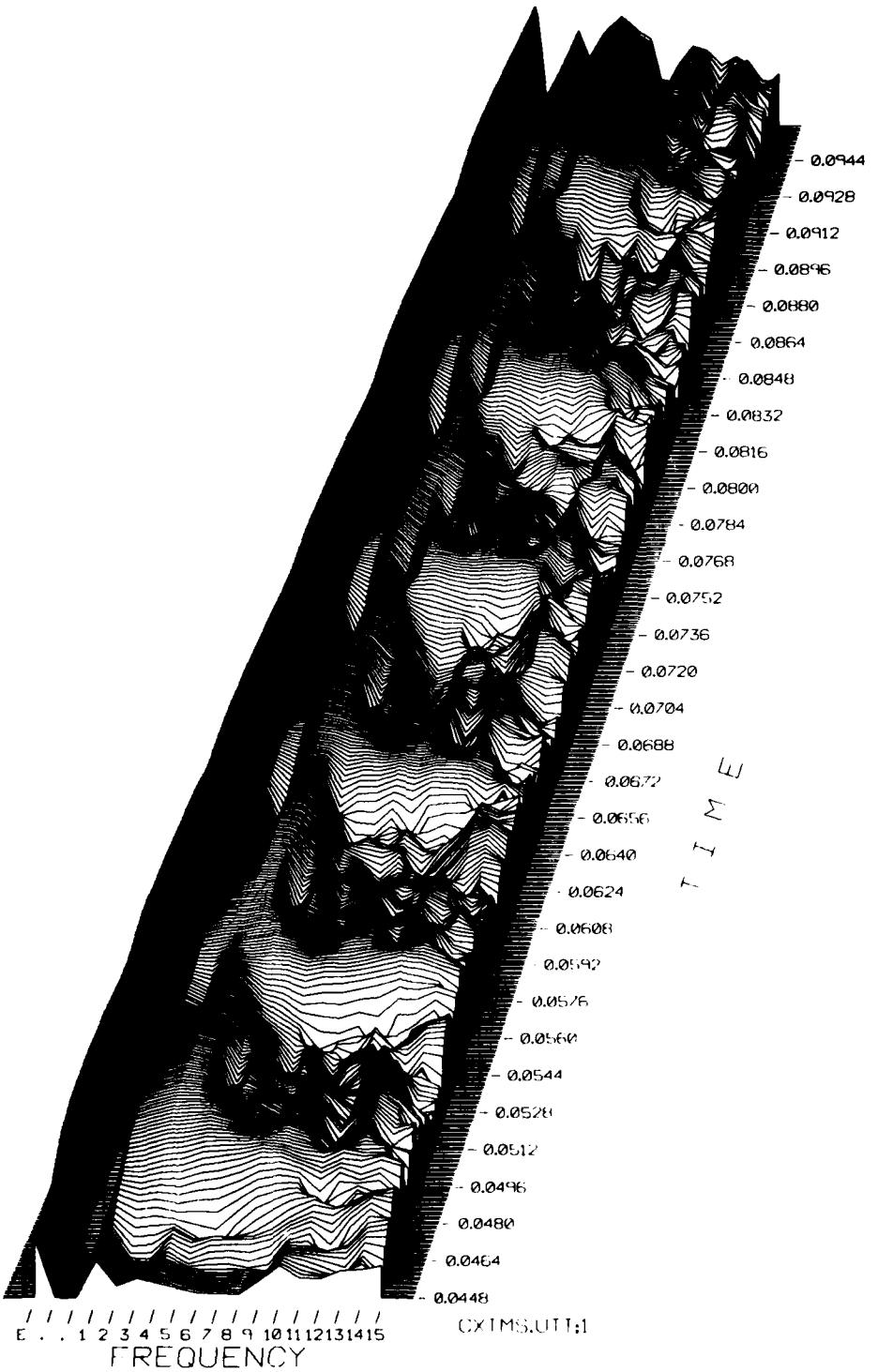


Figure 4.53: 3D Plot of Analyzed Reconstruction (Slightly Modified Data)

$\epsilon_{total,norm} = 9.3 \times 10^{-3}$

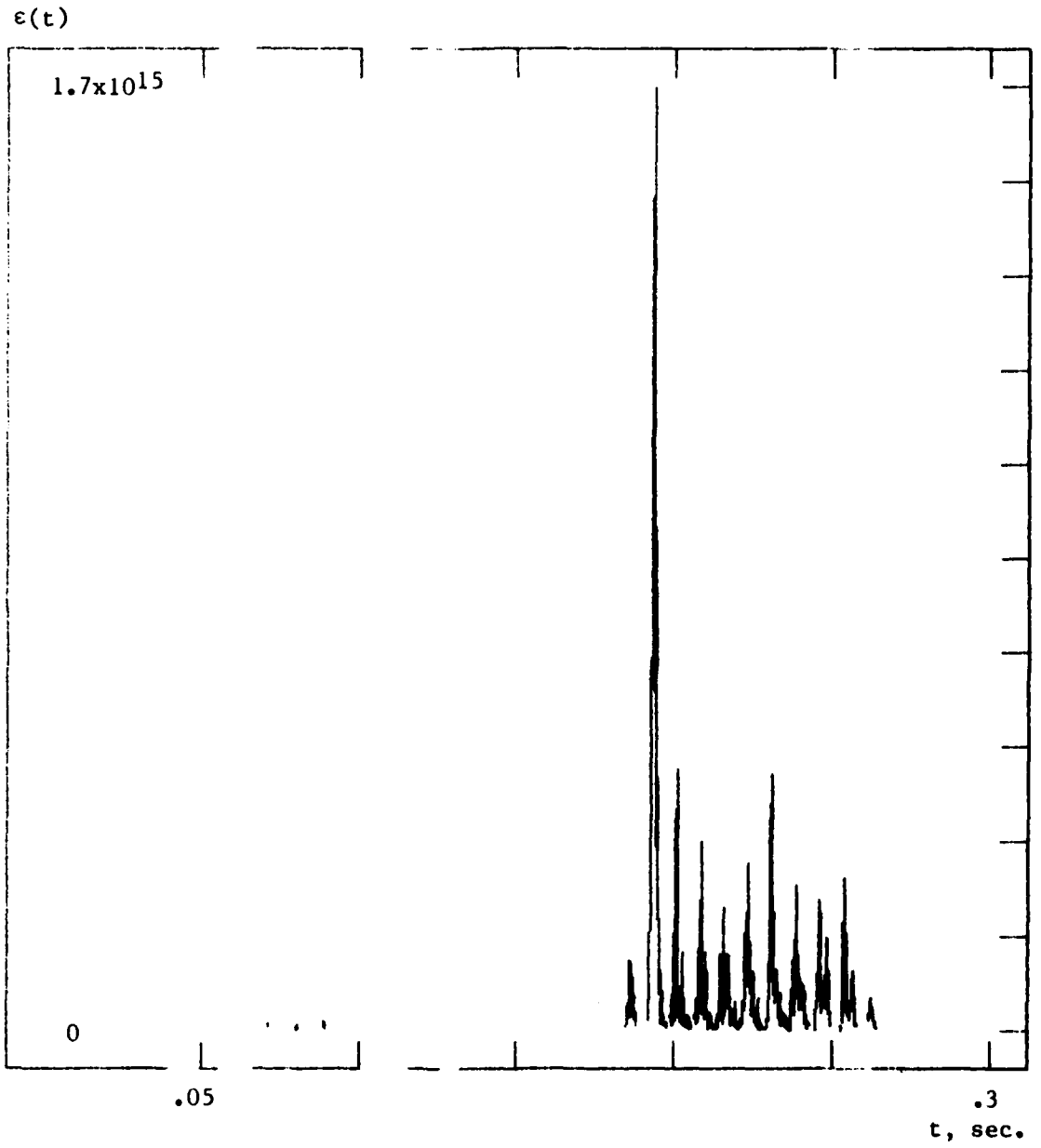


Figure 4.54: Error (Slightly Modified Data)

Finally, a short-time spectral modification was considered in which the data was highly distorted. The F/D outputs were modified by averaging many samples together in each channel, and replacing the samples with average values. Specifically, 74 samples were averaged in the energy channel, 50 samples in Channels #1-2, 45 in #3, 41 in #4, 35 in #5, 33 in #6, 31 in #7, 26 in #8, 23 in #9, 20 in #10, 17 in #11, 15 in #12, 13 in #13, 11 in #14, and 9 in #15. The number of samples averaged in each channel corresponds to the minimum sampling rate for the channel based on 3dB bandwidths, ie., 5000 divided by the critical bandwidth (see Section 2.6). Since values are averaged, however, the resulting data is highly modified and unsuitable for signal recovery purposes. The resulting transmission channel data rate of 7276 samples per second is less than the original signal sampling rate of 10,000 samples per second. Therefore, this spectral distortion is more severe than any considered in previous examples. A portion of the highly modified data is shown in Fig. 4.55, and the resulting reconstruction is shown in Fig. 4.56. The averaging process destroys periodic pitch information, and the reconstructed signal appears quite noisy. The overall envelope of the reconstructed waveform, however, is similar to the original waveform envelope. The waveform reconstructed from highly modified data sounds like very noisy speech. Analysis of the reconstructed signal is shown in Fig. 4.57, and the corresponding error in Fig. 4.58. Although the peak error value of 6.9×10^{15} is comparable with the value obtained in the modified data example of Fig. 4.50, the total normalized error value of 1.2×10^{-1} is greater since the area under the plot of Fig. 4.58 is greater than the area under the plot of Fig. 4.50.

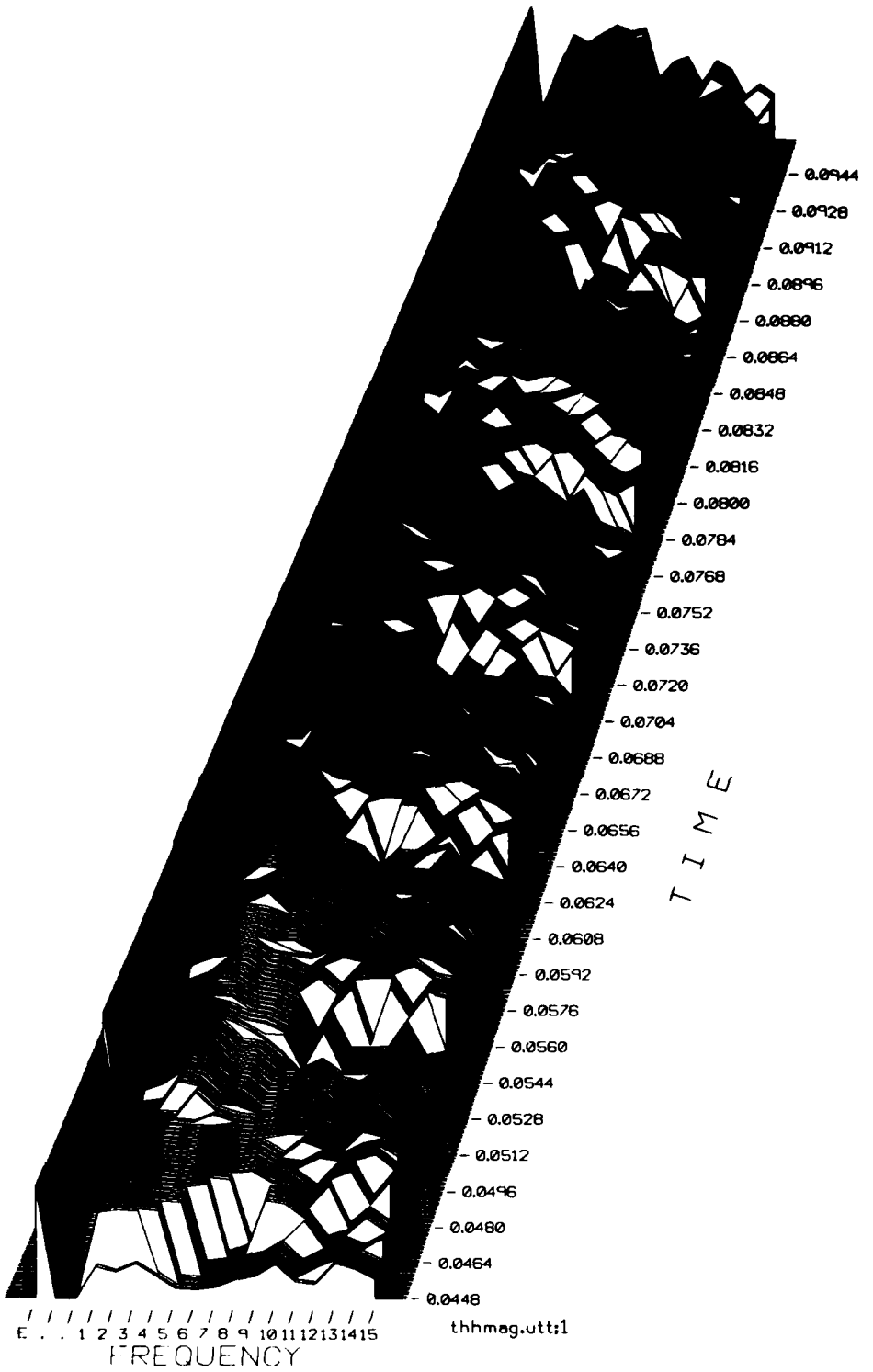


Figure 4.55: 3D Plot of Highly Modified Data

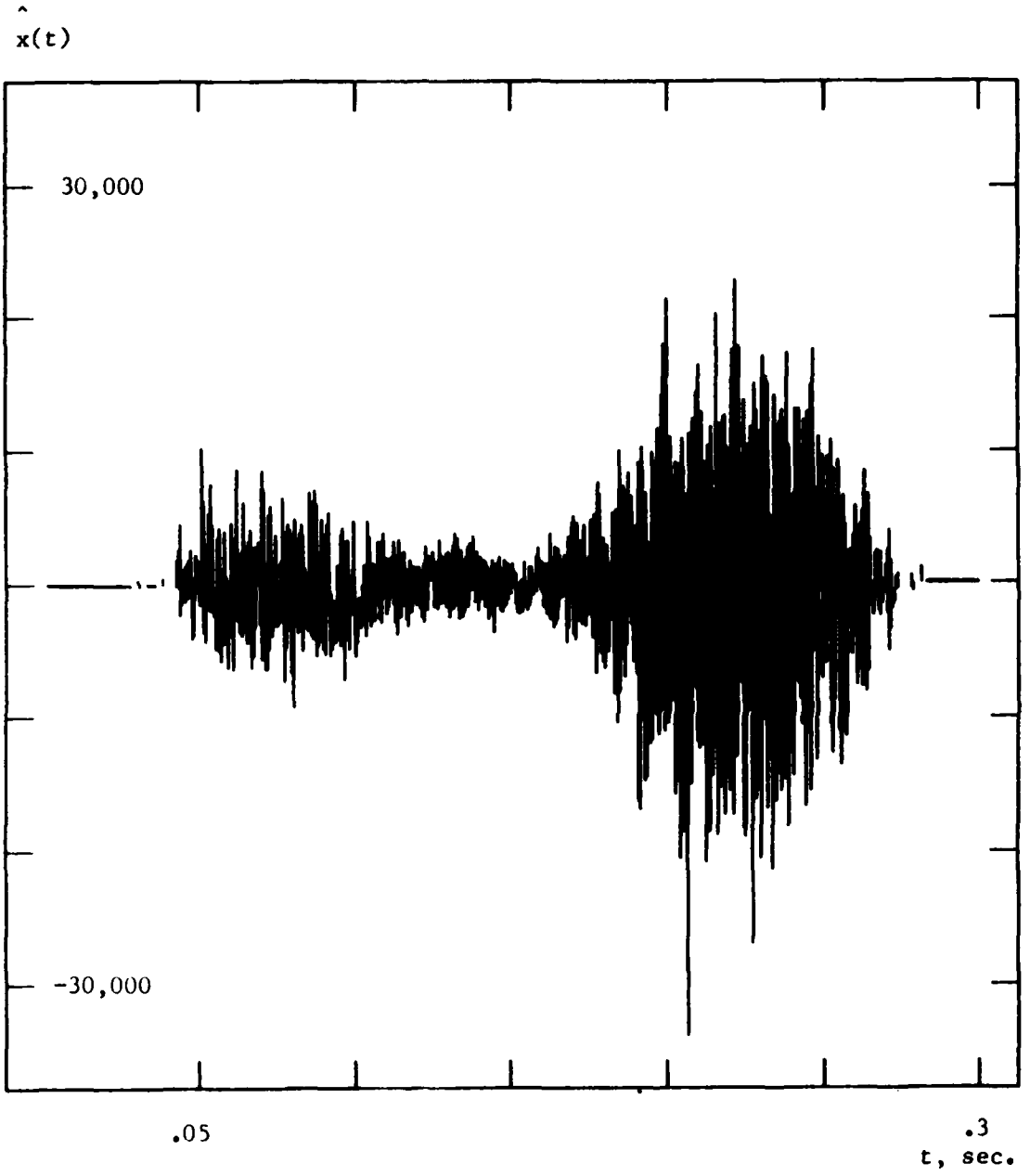


Figure 4.56: Reconstruction from Highly Modified Data

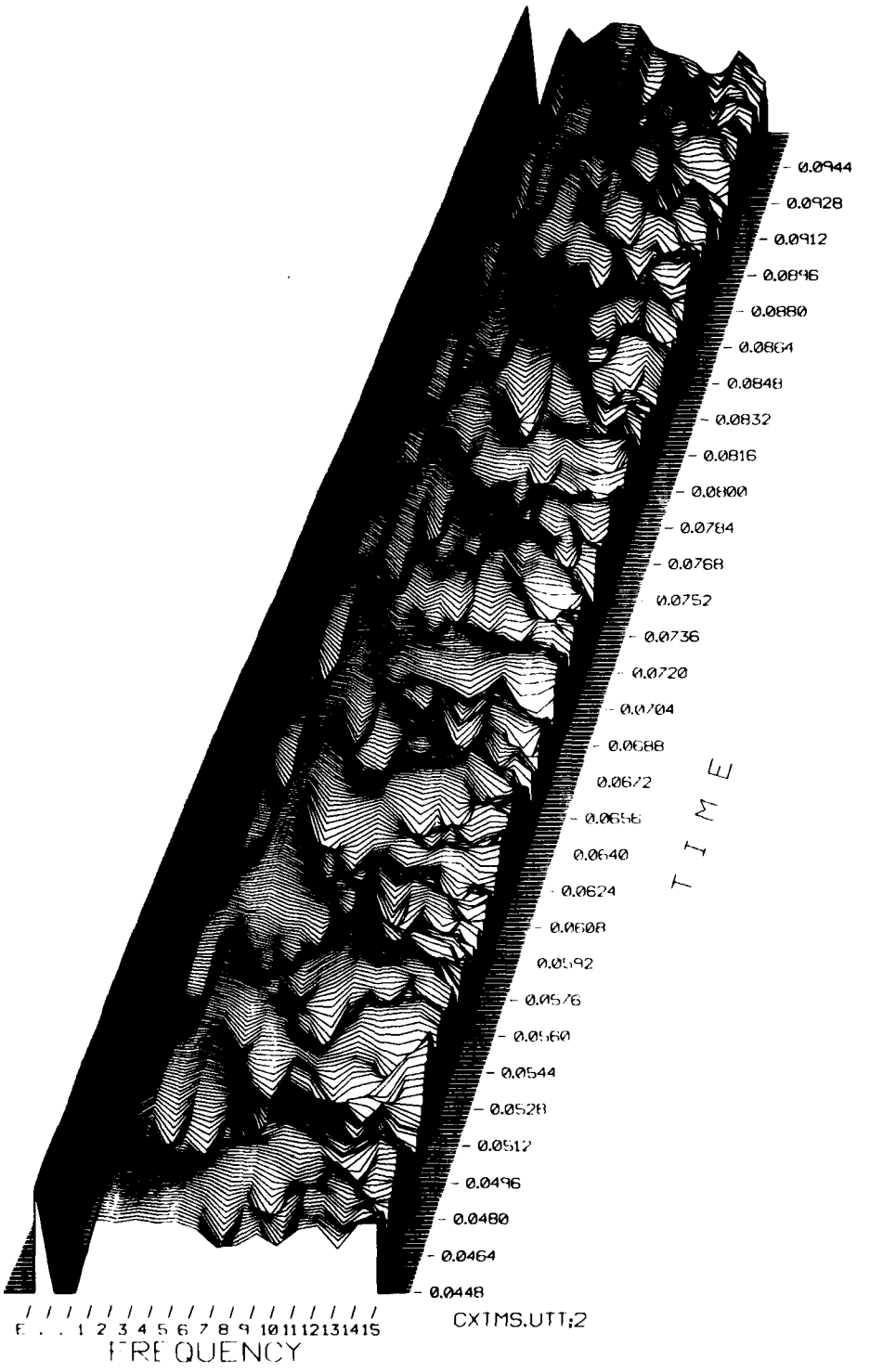


Figure 4.57: 3D Plot of Analyzed Reconstruction (Highly Modified Data)

$\epsilon_{total,norm} = 1.2 \times 10^{-1}$

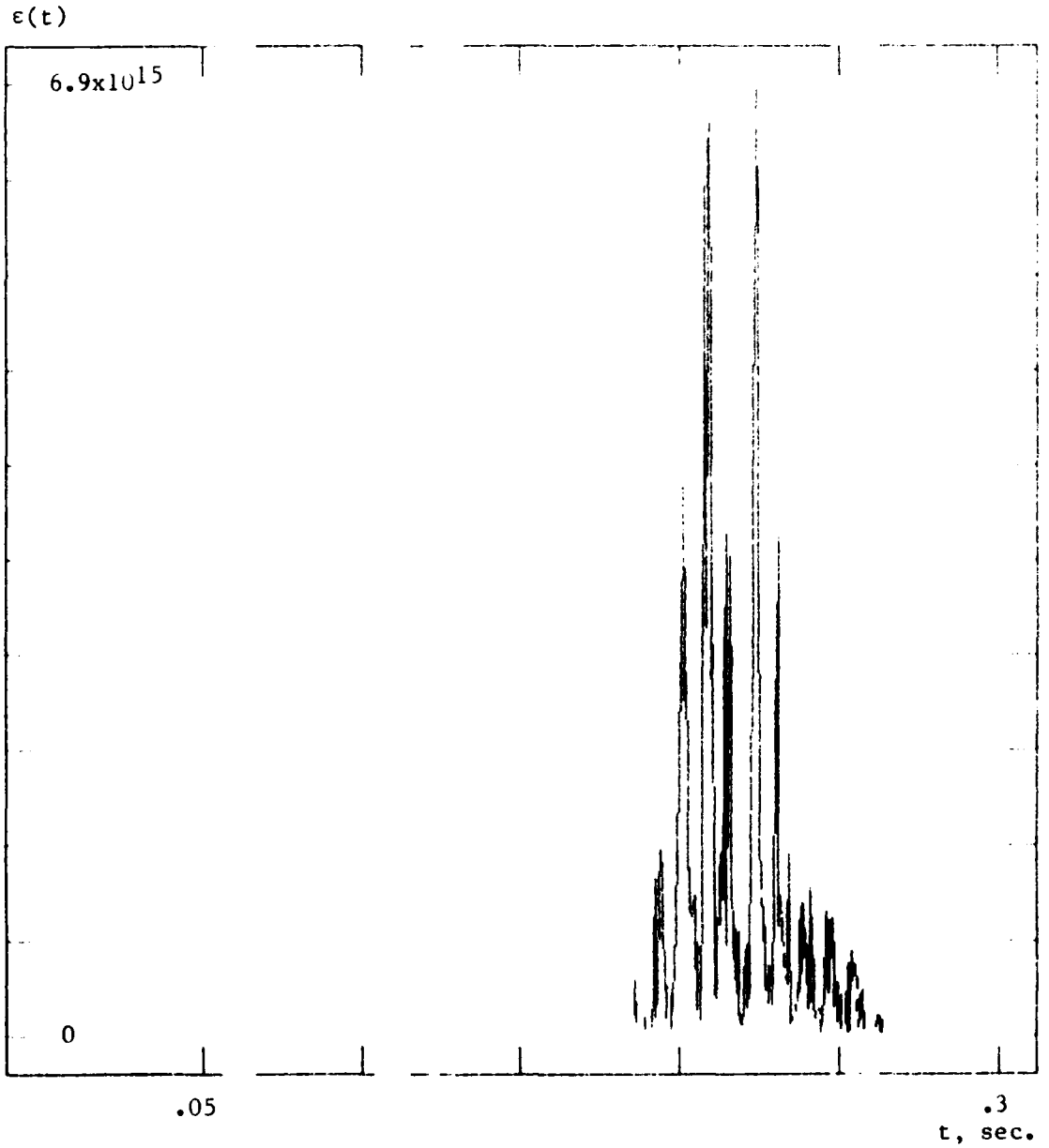


Figure 4.58: Error (Highly Modified Data)

4.6 CONCLUSION

Operation of a speech analysis/synthesis system based on perception has been demonstrated via several examples. The system achieves exact reconstruction (to within an overall sign factor) in the absence of data modificaton, and the ability of the system to reconstruct speech from modified data has also been demonstrated. Note that the data modification technique of this chapter, ie. averaging, was used solely for demonstration purposes and is not recommended for data reduction. Recommended data reduction techniques will be discussed in Section 5.3.

CHAPTER 5

SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

5.1 SUMMARY

This report has presented a speech analysis/synthesis system based on perception. A nonuniform Filter/Detector (F/D) bank and optional Short-Time Energy constraint formed the analysis system. F/D bank characteristics were determined from a combination of physiological and psychoacoustic results. A new relationship demonstrated that the F/D bank could be implemented by the Generalized Short-Time Fourier Transform (GSTFT) magnitude, and a digital implementation suitable for real-time analysis was given. For speech synthesis, a new approach capable of

reconstructing signals from the GSTFT magnitude was used. The speech analysis/synthesis system achieved exact reconstruction in the absence of data modification. The ability of the synthesis system to reconstruct speech from modified data was also demonstrated.

5.2 REAL-TIME SYNTHESIS

Although the analysis system described in Chapter 2 and Appendix C is suitable for real-time operation using existing technology, the synthesis system of Chapter 3 generally is not. Further improvements in the synthesis algorithm, however, may produce a real-time analysis/synthesis system based on perception. For example, the triangle and Schwartz inequalities (Churchill, Brown, and Verhey [45]) can be applied to the recursive GSTFT of Equation 2.29, resulting in an expression which directly relates reconstructed sequence values with the GSTFT magnitude. It may be possible to perform crude real-time synthesis from such results. Alternatively, it may be possible to use a synthesis approach similar to that employed by channel vocoders (Section D.2).

5.3 DATA REDUCTION

Although data reduction is not an essential part of an auditory system model, it may be useful in many applications. When little data reduction is required, the standard downsampling/upsampling approach of Section 2.6 is applicable. When a high degree of data reduction is required, more sophisticated approaches may be used. For example, it is clear from Figs. 4.41 and 4.42 that an efficient encoding can be accomplished by matching the STE and F/D output time-domain waveforms with a few well-chosen prototype wave shapes. Such an encoding can be performed automatically by a principal components approach (Chu [16]). Effectively, the principal components analysis applied to the temporal domain performs a type of pitch extraction. A principal components synthesis, followed by signal reconstruction from the resulting modified data, produces a signal which sounds quite similar to channel vocoded speech (see Section D.2). Speech can be obtained via this approach using transmission channel data rates on the order of 10,000 bits per second (not samples per second). Further research in this area may prove beneficial to the design of channel vocoders based on properties of the human auditory system (Gold and Tierney [46]).

5.4 AUTOMATIC SPEECH RECOGNITION MACHINE DESIGN

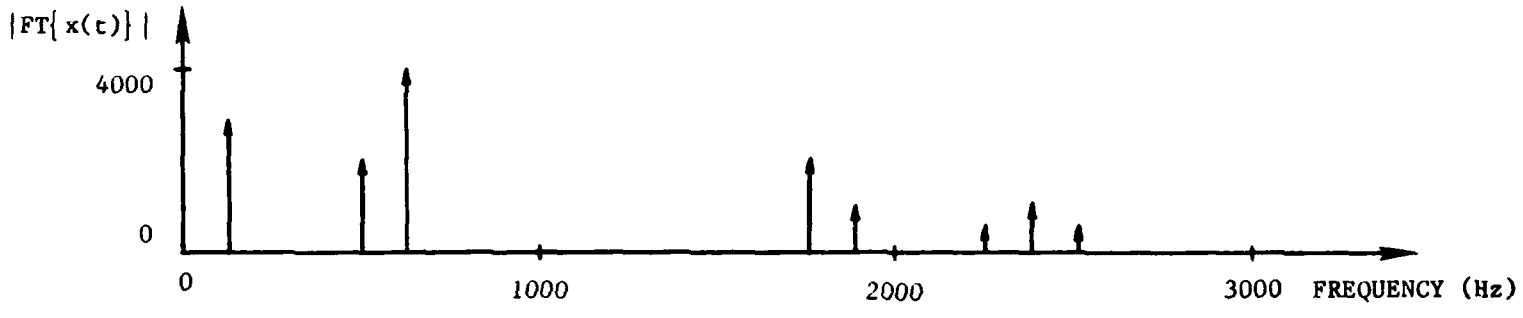
When an Automatic Speech Recognition (ASR) machine fails to correctly identify a spoken input word, the failure may be due to inadequacies in the first processing stage, or "front-end." Note that front-end inadequacies can cause unavoidable errors in subsequent stages. Since the new algorithm described in Chapter 3 is the only known means of reconstructing speech from critical bandwidth F/D outputs, it provides a new tool for ASR machine front-end design. Front-end inadequacies can now be discovered when a synthesis technique is used to test the analyzed speech data for suitable information content.

The need for a synthesis system in ASR machine front-end testing can be illustrated by a few simple examples. A bank of bandpass filters having constant passband gain and minimum passband overlap is often used in ASR front-ends (Schafer, Rabiner, and Herrmann [47]; Rubinstein and Silverman [48]; Dautrich, Rabiner, and Martin [49], [50]). If the filters are carefully designed, it is possible to reconstruct the input signal by simply adding the non-detected filter outputs together. When the filters are followed by detectors, however, practical reconstruction of signals from the resulting F/D bank outputs is impossible. For instance, tones of widely different frequencies produce identical steady-state F/D outputs (see Section 3.2.2), and reconstruction of such signals is impossible. Since humans have excellent frequency resolution ability, it is clear that this type of F/D bank cannot be used to perform many waveform discrimination tasks easily performed by humans. For

examples relevant to the task of speech recognition, consider the synthetic vowels /E/ and /AE/ depicted in Figs. 5.1, 5.2, and 5.3. In order to achieve control over each individual spectral component, these vowels were created by adding sine waves rather than using an acoustic tube vocal tract model as in Section 4.4. Assume that a critical bandwidth F/D bank having constant passband gain and minimum passband overlap is designed by interpolating the data of Table 2.1. When the F/D bank is used to analyze the synthetic vowels of Fig. 5.1, it follows from Sections 3.2.2, B.3.7, and B.3.8 that the two vowels yield identical steady-state outputs. Thus, it is impossible for an ASR machine equipped with such a front-end to distinguish between these steady-state sounds. A similar result holds for several other vowel pairs including the /OW/ sound in "bought" and the /U/ sound in "foot," as well as the /UH/ sound in "but" and the /ER/ sound in "bird." The importance of this effect with regard to specific speech recognition vocabularies is a topic for further research, and a speech synthesis system similar to that described in Chapter 3 can be applied to test the results. Of course, such problems are avoided altogether when the speech analysis system of Chapter 2 is used.

In addition to testing front-ends, the synthesis approach can be used to test effects of subsequent processing stages. Such tests can reveal loss of information relevant to recognition of a given vocabulary. Note that loss of irrelevant information may be useful for data reduction purposes. Such loss is acceptable so long as the nature of the loss is

(a) Synthetic Vowel /E/



(b) Synthetic Vowel /AE/

Figure 5.1: Two Vowels Yielding Identical F/D Outputs

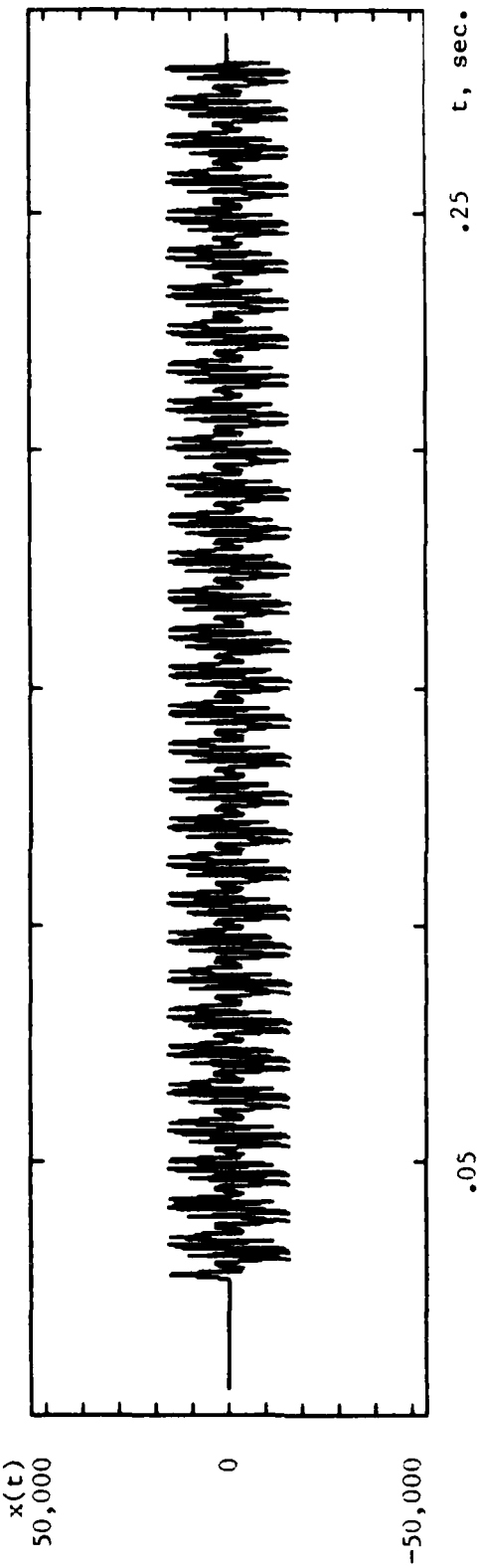


Figure 5.2: Synthetic Vowel /E/

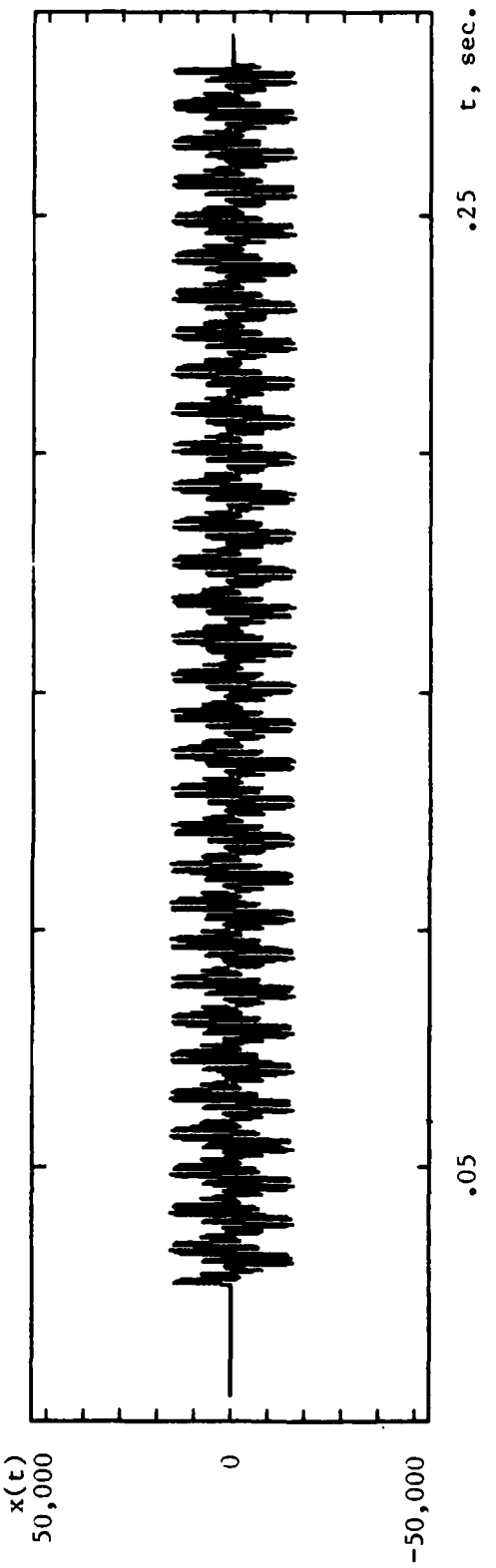


Figure 5.3: Synthetic Vowel /AE/

understood and the results can be tested. Testing is accomplished via synthesis from the modified data, as demonstrated in Chapter 4. For example, to achieve data reduction an additional narrow lowpass filter is often placed at each F/D output or, equivalently, a narrow lowpass smoothing filter is used in the detector. The short-time spectral modification examples of Sections 4.4 and 4.5 indicate that a great deal of information is lost when speech is processed by such a system. The information loss, however, may or may not be important for a specific speech recognition vocabulary. Again, this is a topic for further research. Note that approaches to data reduction other than narrow lowpass filtering can be used which do not sacrifice intelligibility of the reconstructed speech (Section 5.3). Such approaches are therefore suitable for a wider variety of vocabularies.

The preceding observations are consistent with experimental results reported in the literature. For example, a recent study (Dautrich, et al [49], [50]) has shown that a word recognizer based on Linear Predictive Coding (LPC) techniques performed better than a particular 13-channel critical band F/D bank design. In this study the lowpass smoothing filter cutoff frequencies were chosen so that each F/D output could be sampled at a 67 Hz rate regardless of the bandpass filter bandwidth, and the digital bandpass filters had constant passband gain and minimum passband overlap. In an earlier study (White and Neely [51]), LPC was compared with a 20-channel overlapped F/D bank (1/3 octave analog filters were used to cover the 100-10,000 Hz range) using a 100 Hz sampling rate

on each channel, and similar scores were produced by both the F/D and LPC approaches. Finally, a study using mel-frequency cepstrum coefficients, which are similar to processed critical band F/D bank outputs, achieved superior performance compared to LPC (Davis and Mermelstein [52]).

The comparison of speech recognizers using different front-ends is a difficult task. On one hand, if a high quality speech signal can be reconstructed from F/D bank front-end outputs, then any speech recognizer errors must be attributed to the recognition algorithms rather than front-end inadequacies. Since a high quality signal cannot generally be reconstructed from data produced by LPC front-ends (the signal may not fit the model assumed by LPC analysis/synthesis), the F/D bank approach can potentially outperform the LPC approach. On the other hand, the LPC approach may be more convenient since it achieves a high degree of data reduction. Therefore, an important topic for future research is a comparison of speech recognizers using LPC with those using F/D bank front-ends followed by data reduction approachs which do not sacrifice speech intelligibility.

REFERENCES

[1] J.L. Flanagan, Speech Analysis, Synthesis, and Perception (Academic Press, New York, 1965).

[2] R. Carlson and B. Granstrom, "Towards an Auditory Spectrograph," in The Representation of Speech in the Peripheral Auditory System, R. Carlson and B. Granstrom, Ed. (Elsevier Biomedical Press, New York, 1982).

[3] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals (Prentice-Hall, Englewood Cliffs, NJ, 1978).

[4] R.A. Altes, "Detection, Estimation, and Classification with Spectrograms," J. Acoust. Soc. Am. 67, 1232, (1980).

[5] H. Nawab, T.F. Quatieri, and J.S. Lim, "Signal Reconstruction from the Short-Time Fourier Transform Magnitude," ICASSP '82 Proceedings, Paris, France, 3-5 May 1982, pp. 1046-1048.

[6] S.H. Nawab, T.F. Quatieri, and J.S. Lim, "Signal Reconstruction from Short-Time Fourier Transform Magnitude," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-31, 986, 1983, DTIC AD-A137590.

[7] G. Gambardella, "A Contribution to the Theory of Short-Time Spectral Analysis with Nonuniform Bandwidth Filters," IEEE Trans. on Circuit Theory, CT-18, 455 (1971).

[8] J.E. Youngberg and S.F. Boll, "Constant-Q Signal Analysis and Synthesis," ICASSP '78 Proceedings, Tulsa, 10-12 April 1978, pp. 375-378.

[9] S.H. Nawab, T.F. Quatieri, and J.S. Lim, "Algorithms for Signal Reconstruction from Short-Time Fourier Transform Magnitude," ICASSP '83 Proceedings, Boston, 14-16 April 1983, pp. 800-803.

[10] D.W. Griffin and J.S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," ICASSP '83 Proceedings, Boston, 14-16 April 1983, pp. 804-807.

[11] R.F. Lyon, "A Computational Model of Binaural Localization and Separation," ICASSP '83 Proceedings, Boston, 14-16 April 1983, pp. 1148-1151.

[12] N.Y.S. Kiang (with the assistance of T. Watanabe, E.C. Thomas, and L.F. Clark), Discharge Patterns of Single Fibers in the Cat's Auditory Nerve, Research Monograph No. 35 (M.I.T. Press, Cambridge, MA, 1965).

[13] L.S. Frishkopf, "Excitation and Inhibition of Primary Auditory Neurons in the Little Brown Bat," J. Acoust. Soc. Amer. 36, 1016 (1964).

[14] Y. Katsuki, N. Suga, and Y. Kanno, "Neural Mechanism of the Peripheral and Central Auditory System in Monkeys," J. Acoust. Soc. Amer. 34 1396 (1962).

[15] W.M. Siebert, "Hearing and the Ear," in Engineering Principles in Physiology, Vol. 1, Chapter 7 (Academic Press, New York, 1973).

[16] T. Chu, "Principal Components Analysis of the Temporal Dimension of Speech Spectra," MS Thesis, M.I.T. Dept. of EE & CS, Cambridge, MA, 1984.

[17] N.Y.S. Kiang, M.B. Sachs, and W.T. Peake, "Shapes of Tuning Curves for Single Auditory-Nerve Fibers," J. Acoust. Soc. Amer. 42 1341 (1967).

[18] W.M. Siebert, "What Limits Auditory Performance?", Symposial Paper presented at the International Biophysics Congress, International Union for Pure and Applied Biophysics, Academy of Sciences of the USSR, Pushchino, USSR, 1973.

[19] M.R. Schroeder, "Models of Hearing," Proc. of the IEEE, Vol. 63, No. 9, pp. 1332-1350,Sept., 1975.

[20] R.L. Smith and J.J. Zwislocki, "Short-Term Adaptation and Incremental Responses of Single Auditory-Nerve Fibers," in Biol. Cybernetics 17 (Springer-Verlag, 1975), p. 169.

[21] R.L. Smith, "Adaptation, Saturation, and Physiological Masking in Single Auditory-Nerve Fibers," J. Acoust. Soc. Am. 65 166 (1979).

[22] D.M. Harris and P. Dallos, "Forward Masking of Auditory Nerve Fiber Responses," Journal of Neurophysiology, 42, 1083 (1979).

[23] E.D. Young and M.B. Sachs, "Representation of Steady-State Vowels in the Temporal Aspects of the Discharge Patterns of Populations of Auditory-Nerve Fibers," J. Acoust. Soc. Am. 66 1381 (1979).

[24] M.B. Sachs and E.D. Young, "Effects of Nonlinearities on Speech Encoding in the Auditory Nerve," J. Acoust. Soc. Am. 68 858 (1980).

[25] B. Delgutte and N.Y.S. Kiang, "Speech Coding in the Auditory Nerve," J. Acoust. Soc. Am. 75 866 (1984).

[26] B. Scharf, "Critical Bands," in Foundations of Modern Auditory Theory, Vol. 1, J.V. Tobias, Ed. (Academic Press, New York, 1970).

[27] B.C.J. Moore and B.R. Glasberg, "Suggested Formulae for Calculating Auditory-Filter Bandwidths and Excitation Patterns," J. Acoust. Soc. Am. 74 750 (1983).

[28] H. Taub and D.L. Schilling, Principles of Communication Systems, (McGraw-Hill, New York, 1971).

[29] R.M. Fano, "Short-Time Autocorrelation Functions and Power Spectra," J. Acoust. Soc. Am. 22 546 (1950).

[30] M.R. Schroeder and B.S. Atal, "Generalized Short-Time Power Spectra and Autocorrelation Functions," J. Acoust. Soc. Am. 34 1679 (1962).

[31] A.V. Oppenheim and R.W. Schafer, Digital Signal Processing (Prentice-Hall, Englewood Cliffs, NJ, 1975).

[32] J.C. Anderson and C.L. Searle, "Speech Analysis/Synthesis Based on Perception," ICASSP '83 Proceedings, Boston, 14-16 April 1983, pp. 1364-1367.

[33] D.H. Klatt, "A Digital Filter Bank for Spectral Matching," ICASSP '76 Proceedings, Philadelphia, 12-14 April 1976, pp. 573-576.

[34] T.F. Quatieri, S.H. Nawab, and J.S. Lim, "Frequency Sampling of the Short-Time Fourier Transform Magnitude," paper presented at Topical Meeting on Signal Recovery and Synthesis with Incomplete Information and Partial Constraints, Hyatt Lake Tahoe, Incline Village, Nevada, January 12-14, 1983.

[35] S.H. Nawab, "Signal Estimation From Short-time Spectral Magnitude," PhD Thesis, MIT Dept. of EE & CS, Cambridge, MA, 1982.

[36] R.M. Warren and J.M. Wrightson, "Stimuli Producing Conflicting Temporal and Spectral Cues to Frequency," J. Acoust. Soc. Am. 70 1020 (1981).

[37] J.L. Flanagan and N. Guttman, "On the Pitch of Periodic Pulses," J. Acoust. Soc. Am. 32 1308 (1960).

[38] CRC Standard Mathematical Tables, W.H. Beyer, Ed. (CRC Press, Inc., Florida, 1982).

[39] L.L. Beranek, Acoustics (McGraw-Hill, New York, 1954).

[40] D.W. Griffin, D.S. Deadrick and J.S. Lim, "Speech Synthesis from Short-Time Fourier Transform Magnitude and Its Application to Speech Processing," ICASSP '84 Proceedings, San Diego, 19-21 March 1984, pp. 2.4.1-2.4.4.

[41] B.R. Musicus, "Iterative Algorithms for Optimal Signal Reconstruction and Parameter Identification Given Noisy and Incomplete Data," RLE Technical Report 496, PhD Thesis, M.I.T. Dept. of EE & CS, Cambridge, MA, 1982.

[42] G.E. Peterson and H.L. Barney, "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am. 24, 175 (1952).

[43] C.L. Searle, J.Z. Jacobson, and S.G. Rayment, "Stop Consonant Discrimination Based on Human Audition," J. Acoust. Soc. Am. 65, 799 (1979).

[44] C.L. Searle, J.Z. Jacobson, and B.P. Kimberley, "Speech as Patterns in the 3-Space of Time and Frequency," in Perception and Production of Fluent Speech, R.A. Cole, Ed. (Lawrence Erlbaum Associates, Hillsdale, NJ, 1980).

[45] R.V. Churchill, J.W. Brown, and R.F. Verhey, Complex Variables and Applications, (McGraw-Hill, New York, 1974).

[46] B. Gold and J. Tierney, "Vocoder Analysis Based on Properties of the Human Auditory System," Technical Report 670, Lincoln Laboratory, M.I.T. (22 December 1983), NTIC AD-A138660.

[47] R.W. Schafer, L.R. Rabiner, and O. Herrmann, "FIR Digital Filter Banks for Speech Analysis," The Bell System Technical Journal, Vol. 54, No. 3, March, 1975.

[48] J.T. Rubinstein and H.F. Silverman, "Some Comments on the Design and Implementation of FIR Filterbanks for Speech Recognition," ICASSP '83 Proceedings, Boston, 14-16 April 1983, pp. 812-815.

[49] B.A. Dautrich, L.R. Rabiner, and T.B. Martin, "On the Use of Filter Bank Features for Isolated Word Recognition," ICASSP '83 Proceedings, Boston, 14-16 April 1983, pp. 1061-1064.

[50] B.A. Dautrich, L.R. Rabiner, and T.B. Martin, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-31, 793 (1983).

[51] G.M. White and R.B. Neely, "Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-24, 183 (1976).

[52] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-28, 357 (1980).

[53] A.V. Oppenheim and A.S. Willsky (with I.T. Young), Signals and Systems, (Prentice-Hall, Englewood Cliffs, NJ, 1983).

[54] M.M. Sondhi, C.E. Schmidt, and L.R. Rabiner, "Improving the Quality of a Noisy Speech Signal," The Bell System Technical Journal, Vol. 60, No. 8, pp. 1847-1859, Oct., 1981.

[55] W.M. Siebert, Signals and Systems, Lecture Notes for Course 6.003, M.I.T. Dept. of EE & CS, Cambridge, MA, Feb., 1979 (rev. Aug. 1980).

[56] L.R. Rabiner and B. Gold, Theory and Application of Digital Signal Processing, (Prentice-Hall, Englewood Cliffs, NJ, 1975).

[57] W.B. Davenport and W.L. Root, An Introduction to the Theory of Random Signals and Noise (McGraw-Hill, New York, 1958).

[58] A. Papoulis, Probability, Random Variables, and Stochastic Processes (McGraw-Hill, New York, 1965).

[59] M.R. Spiegel, Laplace Transforms, Schaum's Outline Series (McGraw-Hill, New York, 1965).

[60] R. Steele, Delta Modulation Systems, (Halsted Press, Wiley & Sons, New York, 1975).

[61] J.C. Anderson, "A *Low-Cost* Digital Voice Recognition and Response System," Telecommunications, Horizon House-Microwave, Inc., Dedham, MA, pp. 47-50, June, 1981.

[62] B. Gold and C.M. Rader (with A.V. Oppenheim and T.G. Stockham), Digital Processing of Signals (Robert E. Krieger Publishing Co., Malabar, FL, 1983).

[63] Manual for the Model 4691A Sound Spectrograph, (Voiceprint Laboratories, Sommerville, NJ, 1971).

[64] A.V. Oppenheim, "Speech Spectrograms using the Fast Fourier Transform," IEEE Spectrum, 7, 57 (1970).

[65] M.L. Wood and A.V. Oppenheim, "Speech Communication," Quarterly Progress Report No. 102, M.I.T. Research Laboratory of Electronics, Cambridge, MA, July, 1971.

[66] H.K. Dunn and S.D. White, "Statistical Measurements on Conversational Speech," J. Acoust. Soc. Am. 11, 278 (1940).

[67] P.D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," IEEE Trans. Audio and Electroacoust, AU-15, 70 (1967).

APPENDIX A

DEFINITIONS

This appendix presents standard definitions for reference purposes (for further information, see Oppenheim and Willsky [53]). In the continuous-time case, the time variable is "t" and the frequency variable is "Ω." In the discrete-time case, the time variable is "n" and the frequency variable is "ω."

The continuous-time Fourier transform of a signal x(t) is defined as:

$$FT\{x(t)\} = X(j\Omega)$$

$$= \int_{-\infty}^{+\infty} x(t)e^{-j\Omega t}dt. \qquad (A.1)$$

The continuous-time inverse Fourier transform is:

$$x(t) = (1/2\pi)\int_{-\infty}^{+\infty} X(j\Omega)e^{j\Omega t}d\Omega. \qquad (A.2)$$

The modulation property of continuous-time Fourier transforms is given by:

$$FT\{x(t)y(t)\} = (1/2\pi)[X(j\Omega)*Y(j\Omega)]$$

$$= (1/2\pi)\int_{-\infty}^{+\infty} X(j\lambda)Y(j\Omega-j\lambda)d\lambda. \qquad (A.3)$$

The discrete-time Fourier transform of a signal x(n) is defined as:

$$FT\{x(n)\} = X(e^{j\omega})$$

$$= \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n}. \qquad (A.4)$$

The discrete-time inverse Fourier transform is:

$$x(n) = (1/2\pi)\int_{-\pi}^{+\pi} X(e^{j\omega})e^{j\omega n}d\omega. \qquad (A.5)$$

The modulation property of discrete-time Fourier transforms is given by:

$$FT\{x(n)y(n)\} = (1/2\pi)[X(e^{j\omega})*Y(e^{j\omega})]$$

$$= (1/2\pi)\int_{-\pi}^{+\pi} X(e^{j\lambda})Y(e^{j\omega-j\lambda})d\lambda . \qquad (A.6)$$

The z-transform of a discrete-time signal x(n) is defined as:

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}, \qquad (A.7)$$

where z is a complex variable.

The Laplace transform of a continuous-time signal x(t), specified for t>0, is:

$$LT\{x(t)\} = \int_0^{\infty} x(t)e^{-st}dt, \qquad (A.8)$$

where s is a complex variable.

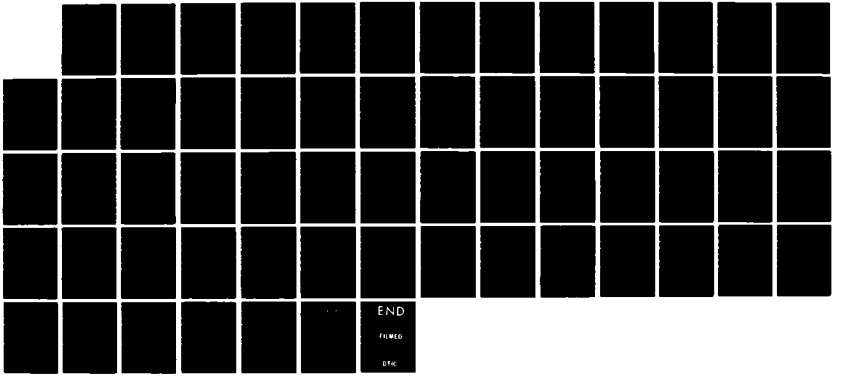
APPENDIX B

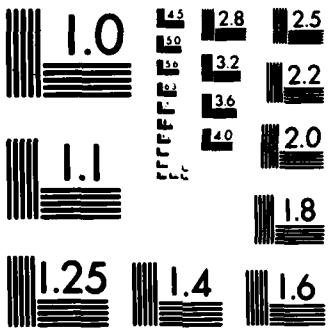
FILTER/DETECTOR THEORY

B.1 INTRODUCTION

This appendix presents details of Filter/Detector (F/D) theory, which is used throughout the main body of the report. First, each component of a continuous-time F/D subsystem is defined. Responses of several commonly used continuous-time F/D subsystems are then examined. Derivations are performed in the continuous-time domain so that results may be conveniently compared with the given references. Similar results can be derived for corresponding discrete-time cases if the sampling rate is adequate to prevent significant aliasing error.

AD-A151 320 SPEECH ANALYSIS/SYNTHESIS BASED ON PERCEPTION(U) 3/3
MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB
J C ANDERSON 05 NOV 84 TR-707 ESD-TR-84-048
UNCLASSIFIED F19628-85-C-0002 F/G 17/2 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

B.2 CONTINUOUS-TIME FILTER/DETECTOR COMPONENT DESCRIPTION

A F/D subsystem consists of a bandpass filter followed by a detector, as shown in Fig. B.1. The detector is comprised of a memoryless nonlinearity and a lowpass smoothing filter.

B.2.1 BANDPASS FILTER DESIGN

A simple design procedure for Linear Time-Invariant (LTI) bandpass filters involves modulating the impulse response of a prototype lowpass filter. Let the prototype lowpass filter impulse response be denoted by h(t). The function h(t) is also known as a window function because it sometimes serves as a time domain "window" through which signals are viewed. As a specific example of the design procedure, let h(t) be the impulse response of an ideal LTI lowpass filter,

$$h(t) = [\sin(\Omega_h t)]/\pi t. \qquad \text{(B.1)}$$

The window function's Fourier transform $FT\{h(t)\}$ is shown in Fig. B.2a. In the frequency domain, the window function has bandwidth Ω_h and unity gain. From the modulation property of Fourier transforms (see Appendix A), the function $h(t)\sin(\Omega_c t)$ is the impulse response of a bandpass filter. Frequency domain magnitude characteristics of the bandpass filter are shown in Fig. B.2b. When $0<\Omega_h<\Omega_c$ the bandpass filter designed in this manner has center frequency Ω_c, bandwidth $2\Omega_h$, and a gain of one-half.

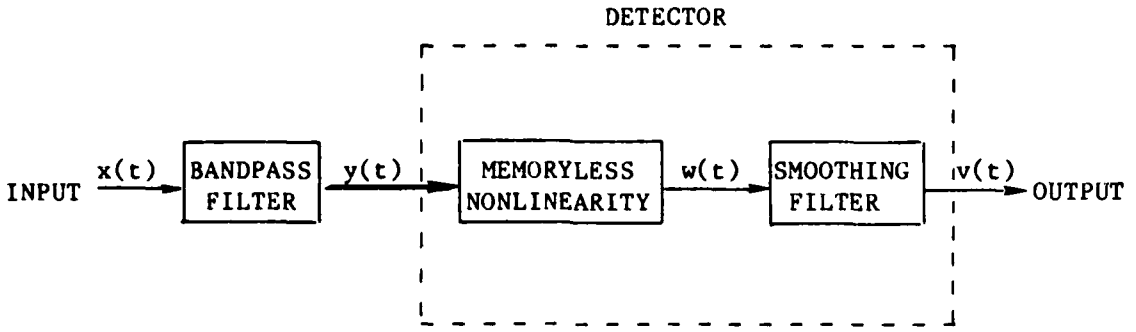
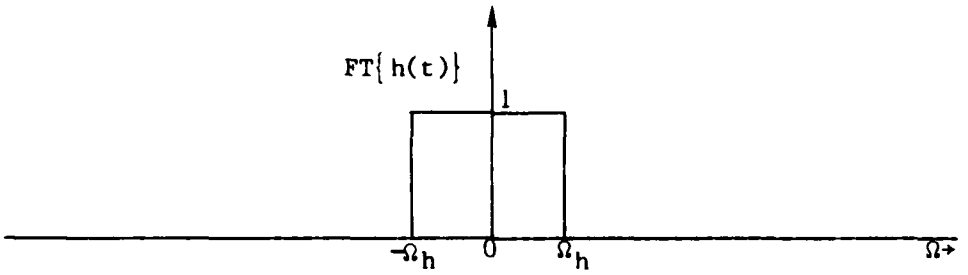
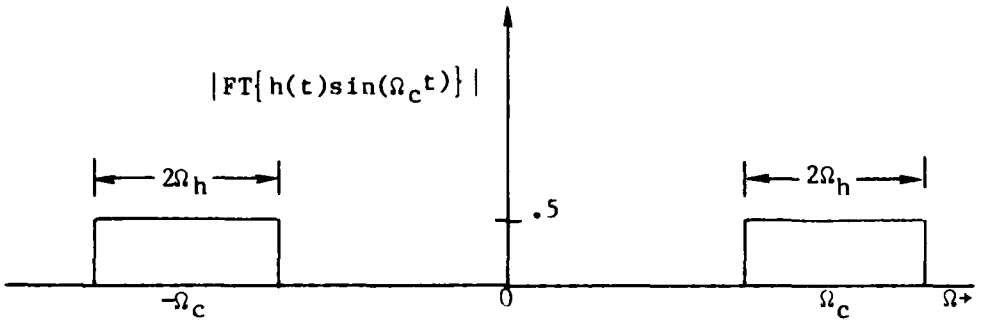


Figure B.1: General Filter/Detector Subsystem



(a) Ideal Window Function Characteristic



(b) Corresponding Bandpass Filter Characteristic

Figure B.2: Bandpass Filter Design Example

Although many bandpass filter design procedures exist, only the approach which modulates a prototype lowpass filter will be discussed. It is shown in Section 2.4 that this particular design technique is used in Short-Time Fourier Transform analysis. The technique is also useful in auditory system modeling, as shown in Section 2.2.

In practical applications a window function other than the impulse response of an ideal lowpass filter is used. When h(t) is the impulse response of a non-ideal lowpass filter, Ω_h is chosen such that frequency components in the region $|\Omega| > \Omega_h$ are negligible. In Sections 2.4.2.3 and D.4, this bandwidth is referred to as the one-sided main lobe bandwidth.

B.2.2 MEMORYLESS NONLINEARITIES

A device is memoryless if its output at any given time depends only upon the input at that time. For example, let the input to a device be y(t) and the output be w(t). The device is memoryless if w(t) at some time t_0 depends only upon $y(t_0)$.

Let the waveform $\alpha(t)$ be the output of a device in response to any input waveform a(t), and $\beta(t)$ be the response to b(t). The device is nonlinear in the system theory sense if the input $c_1a(t)+c_2b(t)$ does not yield an output $c_1\alpha(t)+c_2\beta(t)$, where c_1 and c_2 are constants.

An example of a device which is both memoryless and nonlinear is the square law device described by the input-output relationship:

$$w_1(t) = y^2(t). \qquad \text{(B.2)}$$

Another memoryless nonlinearity is the full wave piecewise linear device described by the input-output relationship:

$$w_2(t) = |y(t)|. \tag{B.3}$$

The half wave piecewise linear device is a memoryless nonlinearity described by the input-output relationship:

$$w_3(t) = [y(t)/2] + [|y(t)|/2]. \tag{B.4}$$

The half wave piecewise linear device can be followed by a square law device to implement a half wave square law device with input-output relationship:

$$w_4(t) = [y^2(t) + y(t)|y(t)|]/2. \tag{B.5}$$

In addition to those described above, other devices such as exponential and square root are often useful.

The F/D of Fig. B.1 will accomplish demodulation so long as the memoryless nonlinearity does not possess an input-output relationship with odd function symmetry (Taub and Schilling [28]). Devices with odd function symmetry produce signals with equal positive and negative excursions which may lead to a smoothing filter output of zero. Note that the half wave square law device of Equation B.5 consists of an even function $y^2(t)/2$ and an odd function $y(t)|y(t)|/2$. Since $y(t)$ is a narrowband signal, contributions from the odd function can be eliminated by the smoothing filter. Thus, a smoothed version of the square law device output $w_1(t)$ differs only by a factor of two from a smoothed version of the half wave square law device output $w_4(t)$.

B.2.3 SMOOTHING FILTERS

The smoothing filter can be implemented as a LTI lowpass filter with bandwidth Ω_s. The smoothing filter impulse response $h_s(t)$ is not necessarily the same as the window function h(t).

In many applications it is desirable to use a F/D whose output is always positive. For example, a F/D using a square law device may be used to measure average power spectra (Flanagan [1]), and a F/D with a half wave square law device can be used to model auditory nerve firing patterns (Siebert [18]). Since negative power spectra and negative firing rates are meaningless, a positive F/D output is required. Also, the F/D is often followed by a square root device (Sondhi, Schmidt, and Rabiner [54]) or a logarithmic amplifier (Searle [43]). A positive F/D output is clearly required in such cases. Unless otherwise stated, a positive F/D output will be assumed.

The requirement for positive F/D output may place a restriction on the smoothing filter design. Assume the memoryless nonlinearity output is always positive. From the F/D subsystem shown in Fig. B.1, it follows that the smoothing filter must produce a positive output v(t) in response to a positive input w(t). Since any LTI filter with positive impulse response will produce a positive output given a positive input, the restrictions $0 < w(t)$ and $0 < h_s(t)$ are sufficient to ensure that $0 < v(t)$ for all t. Although these restrictions are not always necessary (a counter-example is given in Section 2.4.2.2) they are practical design guidelines which conveniently guarantee a positive F/D output.

In certain cases it is easily shown that a smoothing filter with positive impulse response is necessary, as well as sufficient, to guarantee a positive F/D output. For example, assume the bandpass filter has no spectral zeros and the memoryless nonlinearity is a full wave piecewise linear device. Choosing x(t) so that the product of its Fourier transform and the bandpass filter transfer function are unity leads to an impulse at the bandpass filter output, $y(t)=\delta(t)$. An impulse also appears at the smoothing filter input, $w(t)=\delta(t)$. Since the resulting subsystem output v(t) must be positive, the smoothing filter must have a positive impulse response.

An ideal smoothing filter is a LTI filter having positive impulse response and constant magnitude across its lowpass bandwidth. Although the ideal smoothing filter is a useful concept, it can be shown that such a filter does not exist (Siebert [55]). When $h_s(t)\geq 0$, $|FT\{h_s(t)\}|$ evaluated at the frequency $\Omega=0$ is strictly greater than $|FT\{h_s(t)\}|$ evaluated at any other frequency $\Omega\neq 0$.

Despite the absence of an ideal smoothing filter, a variety of practical smoothing filter designs are possible. For example,

$$h_s(t) = [\sin^2(\Omega_s t/2)]/(\pi t)^2 \qquad (B.6)$$

has a Fourier transform which is zero for $|\Omega|>\Omega_s$. Another design is the causal filter

$$h_s(t) = \beta t^2 e^{-\alpha t},\ 0<t$$

$$= 0,\ \text{otherwise}, \qquad (B.7)$$

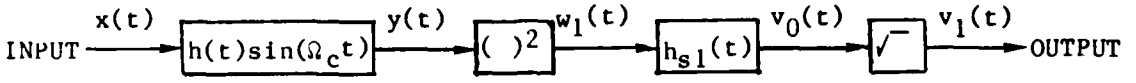
where α and β are positive real constants. This filter is discussed further Chapter 2. Channel vocoders sometimes use Bessel filters which have a small negative overshoot in the impulse response (Sondhi, et al [54]). To maintain an overall positive impulse response, a small positive offset must be added to the Bessel filter impulse response. When a finite duration impulse response is required, a function such as the Hamming window may be used to truncate the impulse responses of Equations B.6 or B.7 (Rabiner and Gold [56]). Alternatively, since a Hamming window is the impulse response of a lowpass filter and is always positive, it may be directly used as a smoothing filter.

B.3 CONTINUOUS-TIME FILTER/DETECTOR RESPONSES

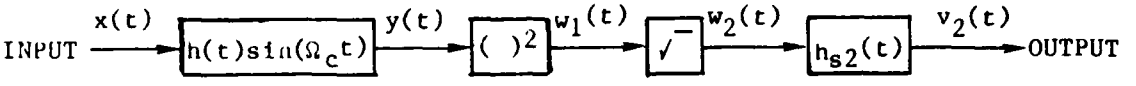
In this section, responses of several F/D subsystems to a variety of signals are examined in detail. Three commonly used continuous-time F/D subsystems which differ in memoryless nonlinearity type and smoothing filter bandwidth are shown in Fig. B.3. Smoothing filter bandwidths for the square law, full wave piecewise linear, and half wave piecewise linear detectors are Ω_{s1}, Ω_{s2}, and Ω_{s3}, respectively. For convenience, all three LTI smoothing filters are assumed to have the ideal characteristics of unity gain, zero delay, and positive output given a positive input.

Fig. B.3a depicts a F/D subsystem using a square law device in the detector. A square root device is present so that output levels are the same order of magnitude as those given by detectors using full wave or half wave piecewise linear devices. If the F/D outputs are followed by a logarithmic amplifier, as is often the case in practice, then power law devices at the output have little effect on the final result.

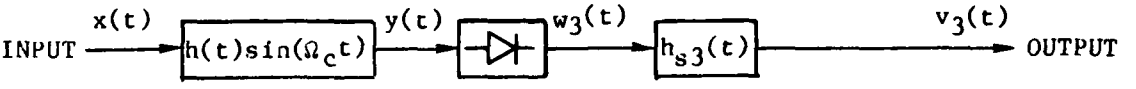
Fig. B.3b depicts a detector using a full wave piecewise linear device, which is drawn as a square law device followed by a square root device. The half wave piecewise linear device of Fig. B.3c is represented by a diode symbol.



(a) Square Law Device



(b) Full Wave Piecewise Linear Device



(c) Half Wave Piecewise Linear Device

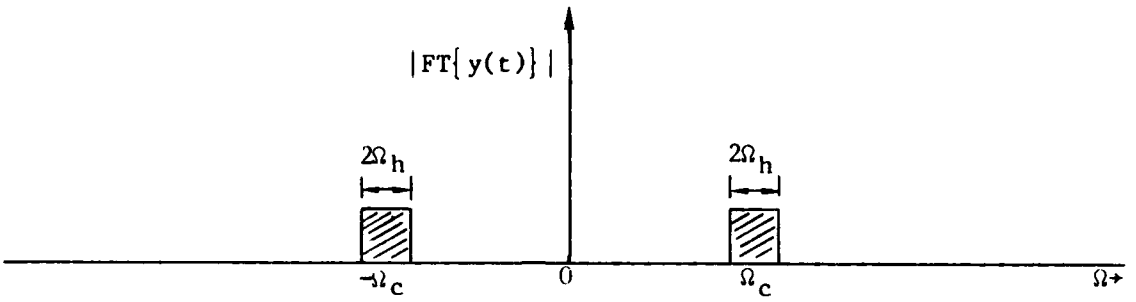
Figure B.3: Commonly Used Filter/Detector Subsystems

B.3.1 SQUARE LAW DETECTOR RESPONSE TO ARBITRARY INPUTS

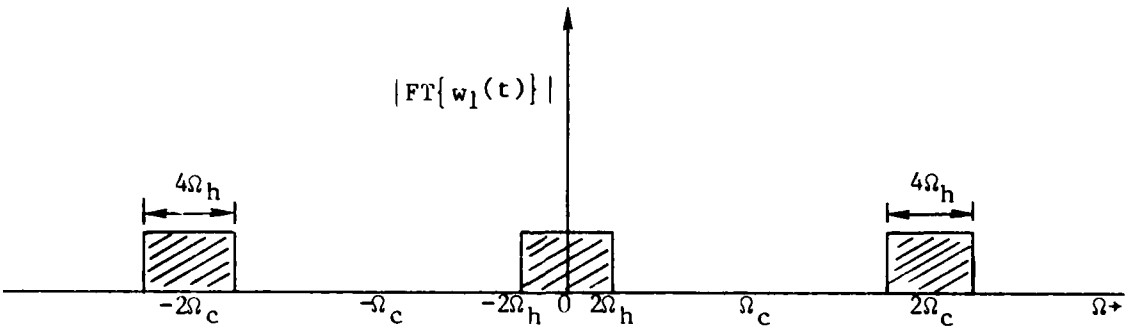
For any arbitrary input signal x(t), the spectrum of y(t) is bandlimited to the region $\Omega_c-\Omega_h<|\Omega|<\Omega_c+\Omega_h$ as shown in Fig. B.4a. Note that the graphs of Fig. B.4 do not represent the exact Fourier transform of any particular signal, but indicate regions where non-negligible spectral components may exist. From the modulation property of Fourier transforms (see Appendix A), it follows that the spectrum of $w_1(t)$ consists of low and high frequency regions as shown in Fig. B.4b. If the smoothing filter bandwidth is chosen so that $2\Omega_h<\Omega_{s1}<2\Omega_c-2\Omega_h$, then no low frequency information is lost but all high frequency components are eliminated from $v_0(t)$.

B.3.2 FULL AND HALF WAVE PIECEWISE LINEAR DETECTOR RESPONSES TO ARBITRARY INPUTS

In this section, it is shown that the full wave piecewise linear detector output $w_2(t)=|y(t)|$ can be expanded in terms of even powers of y(t). The spectrum of |y(t)| can therefore be determined from the spectrum of y(t) by repeated application of the modulation property. The result is a new Fourier transform operation which, given the spectrum of a signal, determines the spectrum of the absolute value of the signal. It follows from Equation B.4 that a similar result may be applied to the half wave piecewise linear detector.



(a) Spectral Regions Occupied by the Bandpass Filter Output



(b) Spectral Regions Occupied by the Square Law Device Output

Figure B.4: Square Law Detector Response to Arbitrary Inputs

The input-output characteristic of a full wave piecewise linear device is given by:

$$w_2(t) = |y(t)|. \tag{B.8}$$

Since the device is memoryless, time dependence of the signals is unimportant and the time parameter may be suppressed. Equation B.8 can thus be written as:

$$w_2 = |y|. \tag{B.9}$$

Assume the device input amplitude is limited to some arbitrary finite range $-R \leq y \leq R$. Using the Fourier series expansion for a triangle wave, which is identical to the input-output characteristic over the specified range,

$$w_2 = (R/2) - (4R/\pi^2) \sum_{n=1}^{\infty} (2n-1)^{-2} \cos[(2n-1)\pi y/R]. \tag{B.10}$$

The cosine function can be expanded via a power series for any y:

$$\cos[(2n-1)\pi y/R] = \sum_{m=0}^{\infty} (-1)^m [(2n-1)\pi y/R]^{2m}/(2m)!. \tag{B.11}$$

Substitution of Equation B.11 into B.10 yields:

$$w_2 = \sum_{m=1}^{\infty} a_m y^{2m} \tag{B.12}$$

where

$$a_m = -(4R/\pi^2)[(-1)^m(\pi/R)^{2m}/(2m)!][\sum_{n=1}^{\infty} (2n-1)^{2m-2}]. \tag{B.13}$$

The full wave piecewise linear device output $w_2(t)$ can thus be expressed in terms of even powers of its input $y(t)$. Note, however, that the coefficients a_m have infinite values.

If the number of terms in the series expansion is limited, $m=1,2,\ldots,M$, an appropriate set of finite values for a_m can be obtained from truncated versions of the Fourier series of Equation B.10 and the power series of Equation B.11. The Fourier series converges rapidly, and relatively few terms are required to obtain results within a specified accuracy. Each of the cosine terms in the truncated Fourier series is in turn expanded by the slowly converging power series. The cosine expansions must contain enough terms so that error given by truncation of the original Fourier series is not significantly increased. A large value of M is thus required to obtain a reasonably accurate approximation. Any constant term in the resulting expansion should be eliminated so that the approximation produces zero output in response to zero input.

Appropriate coefficient values can also be computed using a minimum mean squared error (MMSE) criterion. The mean squared approximation error is given by:

$$\varepsilon_M = \int_{-R}^{+R} \left[|y| - \sum_{m=1}^{M} a_m y^{2m} \right]^2 dy. \qquad \text{(B.14)}$$

To obtain the value of any particular coefficient a_i which minimizes the error for $i=1,2,\ldots,M$:

$$\partial \varepsilon_M / \partial a_i = 0$$

$$= -4\int_0^R (y - \sum_{m=1}^{M} a_m y^{2m}) y^{2i} dy. \qquad \text{(B.15)}$$

The solution is given by

$$1/(i+1) = \sum_{m=1}^{M} a_m R^{2m-1}/(m+i+.5), \qquad \text{(B.16)}$$

which generates M equations in M unknowns and thereby specifies a_m for m=1,2,...,M.

As an example, let R=1 and M=7. Solving Equation B.16 yields:

$$w_2(t) \cong 1.6746y^2(t) - .078942y^4(t) - .28032y^6(t) - .72214y^8(t)$$

$$+ .024750y^{10}(t) - .0013560y^{12}(t) - .0088671y^{14}(t), \qquad \text{(B.17)}$$

which is the MMSE approximation for $w_2(t)=|y(t)|$ on the interval $-1 \leq y(t) \leq 1$ when seven terms are used. Evaluation of Equation B.17 with y(t)=1 yields $w_2(t) \cong .608$, which is a poor approximation. Repeating the procedure with M=10 yields:

$$w_2(t) \cong 5.8239y^2(t) - 34.0175y^4(t) + 108.3705y^6(t) - 156.0335y^8(t)$$

$$+ 74.8383y^{10}(t) - 16.8961y^{12}(t) + 115.5607y^{14}(t)$$

$$- 127.7208y^{16}(t) + 7.3678y^{18}(t) + 23.7314y^{20}(t). \qquad \text{(B.18)}$$

Evaluation of Equation B.18 with y(t)=1 yields $w_2(t) \cong 1.025$, which is a better approximation.

Note that a large number of terms must be used in order to obtain reasonable results. Thus, to determine the spectrum of $|y(t)|$ from the spectrum of $y(t)$, the modulation property of Fourier transforms must be applied many times. Since the coefficients must be accurate to many significant digits and a high degree of precision must be maintained in all computations, this approach is mainly of theoretical interest and has limited practical value.

B.3.3 RELATIONSHIP BETWEEN FULL WAVE AND HALF WAVE PIECEWISE LINEAR DETECTORS FOR ARBITRARY INPUTS

Under certain conditions, outputs from F/D subsystems using either full wave or half wave piecewise linear devices are the same (within a scale factor) for any arbitrary input signal. From Equations B.3 and B.4 it follows that:

$$w_3(t) = [y(t)/2] + [w_2(t)/2]. \tag{B.19}$$

The spectrum of $y(t)$ lies in the region $\Omega_c-\Omega_h<|\Omega|<\Omega_c+\Omega_h$ as shown Fig. B.4a. If the smoothing filter bandwidth is chosen so that $0<\Omega_{s3}<\Omega_c-\Omega_h$, then the bandpass component $y(t)/2$ is eliminated. Setting $\Omega_{s2}=\Omega_{s3}$ then results in F/D outputs $v_2(t)$ and $v_3(t)$ which differ by a factor of two:

$$v_2(t) = 2v_3(t). \tag{B.20}$$

B.3.4 RELATIONSHIP BETWEEN SQUARE LAW AND FULL WAVE PIECEWISE LINEAR DETECTORS FOR ARBITRARY INPUTS

The square law detector shown in Fig. B.3a lowpass filters the waveform $w_1(t)$ and takes the square root of the result to obtain output $v_1(t)$. The full wave piecewise linear detector of Fig. B.3b takes the square root of $w_1(t)$ and lowpass filters the result to obtain output $v_2(t)$. Since lowpass filter and square root operations are not interchangable, the outputs $v_1(t)$ and $v_2(t)$ are not equal in general. It will be shown, however, that given certain restrictions these two outputs are similar for a variety of different input waveforms $x(t)$. Note that while $v_2(t)$ is a bandlimited signal, $v_1(t)$ is not necessarily bandlimited. Therefore, a large smoothing filter bandwidth Ω_{s2} may be required in order that $v_2(t) \cong v_1(t)$.

B.3.5 NOISE RESPONSE

Let the input to each F/D subsystem of Fig. B.3, x(t), be a white Gaussian noise process with variance $S_x(\Omega)=4A$. If the window function h(t) is the impulse response of a unity gain ideal LTI lowpass filter with cutoff frequency Ω_h, the bandpass filter output y(t) will be a bandlimited Gaussian noise process with power spectral density

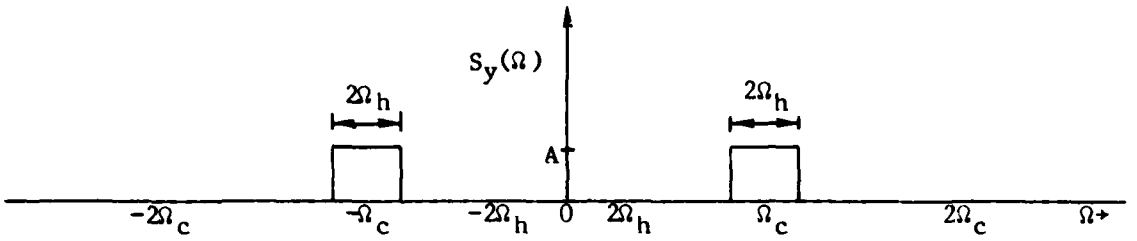
$$S_y(\Omega) = S_x(\Omega)|FT\{h(t)\sin(\Omega_c t)\}|^2$$

$$= A, \ \Omega_c-\Omega_h<|\Omega|<\Omega_c+\Omega_h$$

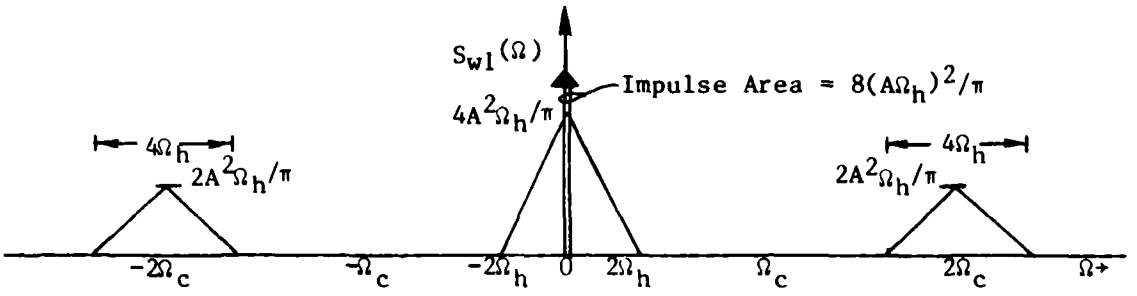
$$=0, \text{ otherwise.} \qquad \text{(B.21)}$$

Note that the spectral height of $S_y(\Omega)$ equals the input variance reduced by a factor of four, as shown in Fig. B.5a. Power spectral densities for noise processes at the output of each memoryless nonlinearity are shown in Figs. B.5b, c, and d (Davenport and Root [57]; Papoulis [58]).

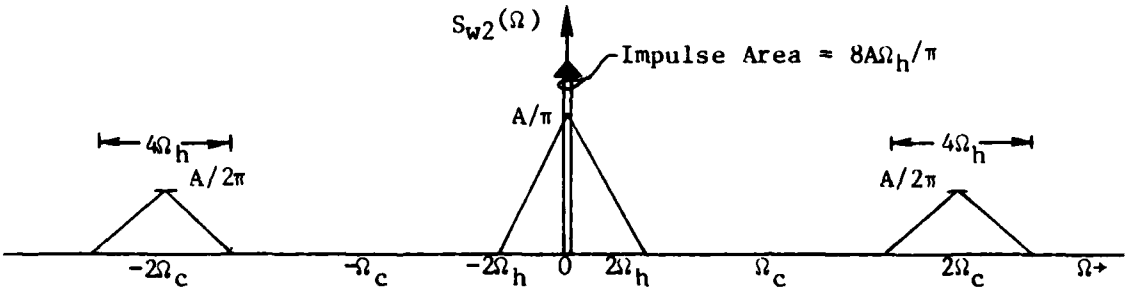
Given certain restrictions, comparable F/D noise responses can be obtained over a specified range of frequencies. To avoid loss of low frequency information while eliminating high frequency components, let $2\Omega_h<\Omega_{s1},\Omega_{s2}<2\Omega_c-2\Omega_h$ for the square law and full wave piecewise linear detectors while $2\Omega_h<\Omega_{s3}<\Omega_c-\Omega_h$ for the half wave piecewise linear detector. Under these restrictions, the full wave and half wave piecewise linear detector outputs will differ only by a scale factor. Note that the minimum smoothing filter bandwidth, $2\Omega_h$ in all cases, is twice the bandwidth normally used for detection of Amplitude Modulation (AM) signals (Siebert [55]). The wide bandwidth is required because the



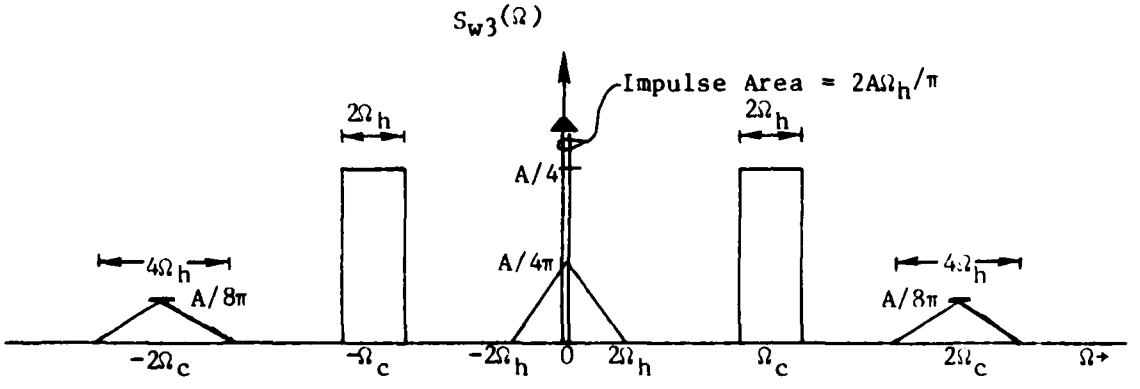
(a) Bandpass Filter Output Power Spectrum



(b) Square Law Device Output Power Spectrum



(c) Full Wave Piecewise Linear Device Output Power Spectrum



(d) Half Wave Piecewise Linear Device Output Power Spectrum

Figure B.5: Filter/Detector Noise Response

input signal is not generally AM in nature, and has neither the carrier nor the symmetry inherent in AM signals.

Given the stated restrictions, noise responses of F/D subsystems using square law and full wave piecewise linear devices are similar in many ways. From Figs. B.5b and B.5c it is apparent that the noise processes $w_1(t)$ and $w_2(t)$, and therefore $v_0(t)$ and $v_2(t)$, have comparable power spectral density shapes. However, total area under the square law device power spectral density curve is proportional to the square of the input variance, while the area under the curve for the full wave piecewise linear device is directly proportional to the input variance. Due to the square root device shown in Fig B.3a, the zero frequency component of $v_1(t)$ is proportional to the input variance. The zero frequency component of $v_2(t)$ is also proportional to the input variance. Thus, in applications where a F/D subsystem is used to measure noise process characteristics, the zero frequency component of the F/D output is often the only quantity of interest (see Sections D.2 and D.5 for such applications).

B.3.6 IMPULSE RESPONSE

Consider the F/D subsystem of Fig. B.3a which uses a square law device. If the input is an impulse, $x(t)=\delta(t)$, then

$$w_1(t) = h^2(t)[1 - \cos(2\Omega_c t)]/2. \tag{B.22}$$

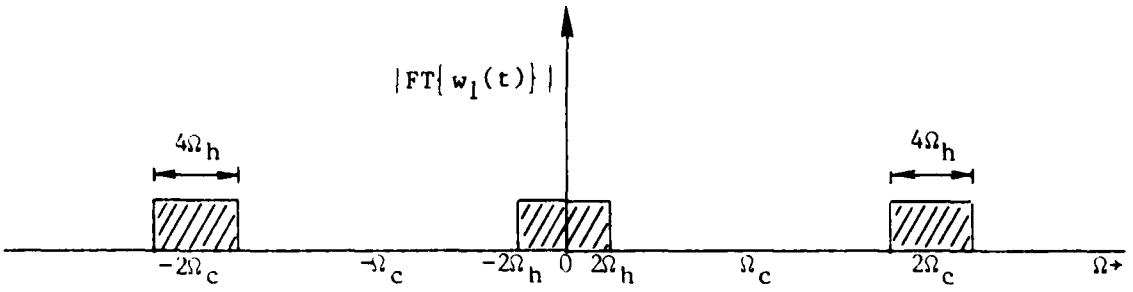
Frequency regions occupied by $FT\{w_1(t)\}$ are shown in Fig. B.6a. If $2\Omega_h<\Omega_{s1}<2\Omega_c-2\Omega_h$ then

$$v_0(t) = h^2(t)/2 \tag{B.23}$$

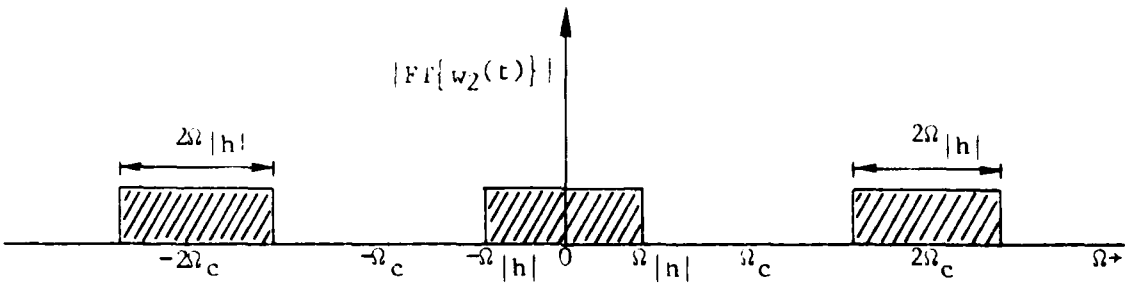
and

$$v_1(t) = |h(t)|/\sqrt{2}. \tag{B.24}$$

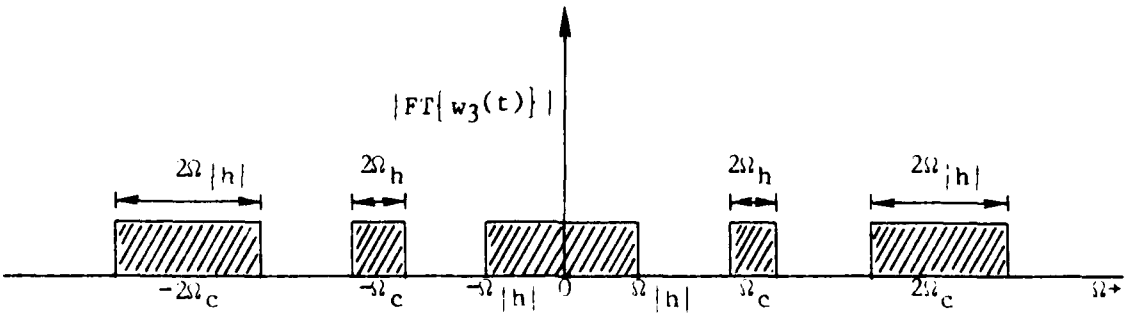
The function $|h(t)|$ may or may not be bandlimited. For convenience, assume that $FT\{|h(t)|\}$ is for all practical purposes lowpass bandlimited to some frequency $\Omega_{|h|}$. For example, if the window function is always positive, then $h(t)=|h(t)|$ and $|h(t)|$ is bandlimited to $\Omega_{|h|}=\Omega_h$. When $|h(t)|$ is bandlimited, $v_1(t)$ as given by Equation B.24 can be compared with the bandlimited functions $v_2(t)$ and $v_3(t)$.



(a) Regions Occupied by the Spectrum of $w_1(t)$



(b) Regions Occupied by the Spectrum of $w_2(t)$



(c) Regions Occupied by the Spectrum of $w_3(t)$

Figure 8.6: F/D Impulse Response Frequency Domain Characteristics

Now consider the F/D subsystem of Fig. B.3b which uses a full wave piecewise linear device. If the input is an impulse then, by Fourier series expansion (Spiegel [59]),

$$w_2(t) = |h(t)||\sin(\Omega_c t)|$$

$$= |h(t)|\{(2/\pi) + (4/\pi)\sum_{n=1}^{\infty}[\cos(2n\Omega_c t)]/(1-4n^2)\} \qquad (B.25)$$

Frequency regions occupied by $FT\{w_2(t)\}$ are shown in Fig. B.6b. If the lowpass smoothing filter has cutoff frequency ω_{s2} such that $\omega_{|h|} < \omega_{s2} < 2\omega_c - \omega_{|h|}$, then the F/D output is:

$$y_2(t) = 2|h(t)|/\pi \qquad (B.26)$$

Finally, consider the F/D subsystem of Fig. B.3c which uses a half wave piecewise linear device. If the input is an impulse, then

$$w_3(t) = [y(t)/2] + |y(t)/2|$$

$$= [y(t)/2] + |y(t)|/2$$

$$+ [y(t)/2] + (2/\pi)\sum_{n=1}^{\infty}[\cos(2n\Omega_c t)]/(1-4n^2) \qquad (B.27)$$

Frequency regions occupied by $FT\{w_3(t)\}$ are shown in Fig. B.6c. Let the lowpass smoothing filter have cutoff frequency ω_{s3}, where $\omega_{s3} > \omega_c - \omega_{|h|}$ and $\omega_{s3} < 2\omega_c - \omega_{|h|}$. The half wave piecewise linear detector output is then proportional to the full wave piecewise linear detector output.

Under the stated restrictions and assumptions, the impulse responses of all three F/D subsystems differ only by a scale factor.

The results of this section can be applied to determine the impulse response of a F/D subsystem using a half wave square law device, as used in Chapter 2. It follows from Equations B.2, B.3, B.5, and the modulation property that

$$FT\{w_4(t)\} = [FT\{w_1(t)\}]/2 + [FT\{y(t)\}*FT\{w_2(t)\}]/4\pi. \qquad (B.29)$$

Spectral regions occupied by $[FT\{w_1(t)\}]/2$ are shown in Fig. B.6a. The lowest frequency spectral region occupied by $[FT\{y(t)\}*FT\{w_2(t)\}]/4\pi$ is $\Omega_c-\Omega_{|h|}-\Omega_h<|\Omega|<\Omega_c+\Omega_{|h|}+\Omega_h$, as can be seen by convolving Figs. B.4a and B.6b. Thus if a smoothing filter with bandwidth Ω_{s4} is chosen such that $2\Omega_h<\Omega_{s4}<\Omega_c-\Omega_{|h|}-\Omega_h$ and $2\Omega_h<\Omega_{s4}<2\Omega_c-2\Omega_h$, the half wave square law F/D output is $h^2(t)/4$.

B.3.7 SINUSOIDAL RESPONSE

Let the input to each F/D subsystem, x(t), be a sinusoidal waveform. Because the bandpass filter is LTI, the filter output y(t) is also sinusoidal. The waveform will be changed in amplitude and phase if the bandpass filter is non-ideal. Assume

$$y(t) = A_1 \sin(\Omega_1 t), \tag{B.30}$$

where $\Omega_c - \Omega_h < \Omega_1 < \Omega_c + \Omega_h$. Then

$$w_1(t) = (A_1)^2(1-\cos 2\Omega_1 t)/2. \tag{B.31}$$

For $0 < \Omega_{s1} < 2\Omega_c - 2\Omega_h$,

$$v_0(t) = (A_1)^2/2 \tag{B.32}$$

and

$$v_1(t) = |A_1|/\sqrt{2}. \tag{B.33}$$

From the Fourier series expansion of Equation B.25, it follows that

$$w_2(t) = |A_1|\{(2/\pi) + (4/\pi)\sum_{n=1}^{\infty}[\cos(2n\Omega_1 t)]/(1-4n^2)\} \tag{B.34}$$

and, for $0 < \Omega_{s2} < 2\Omega_c - 2\Omega_h$,

$$v_2(t) = 2|A_1|/\pi. \tag{B.35}$$

Similarly,

$$w_3(t) = (A_1 \sin\Omega_1 t)/2 + |A_1|\{(1/\pi) + (2/\pi)\sum_{n=1}^{\infty}[\cos(2n\Omega_1 t)]/(1-4n^2)\}. \tag{B.36}$$

Thus, for $0 < \omega_{s3} < \omega_c - \omega_h$,

$$v_3(t) = |A_1|/\pi. \tag{B.37}$$

Under the stated restrictions and assumptions, the sinusoidal responses of all three F/D subsystems differ only by a scale factor.

B.3.8 SINUSOIDAL PAIR RESPONSE

Let $x(t)$ be a sinusoidal pair. Because the bandpass filter is LTI, $y(t)$ is also a sinusoidal pair. Assume

$$y(t) = A_2[\cos(\Omega_1 t) - \cos(\Omega_2 t)] \tag{B.38}$$

where $\omega_c - \omega_h < \Omega_1, \Omega_2 < \omega_c + \omega_h$ and $\Omega_2 < \Omega_1$. Thus

$$w_1(t) = (A_2)^2[1 - \cos(\Omega_1 - \Omega_2)t - \cos(\Omega_1 + \Omega_2)t + (\cos 2\Omega_1 t + \cos 2\Omega_2 t)/2]. \tag{B.39}$$

Let $2\omega_h < \omega_{s1} < 2\omega_c - 2\omega_h$. Then

$$v_0(t) = (A_2)^2[1 - \cos(\Omega_1 - \Omega_2)t] \tag{B.40}$$

and

$$v_1(t) = \sqrt{2}|A_2 \sin[(\Omega_1 - \Omega_2)t/2]|. \tag{B.41}$$

The waveform $v_1(t)$ of Equation B.41 is not strictly bandlimited. However, an effective bandwidth Ω_e may be chosen such that, for practical purposes, frequency components of $v_1(t)$ in the region $\Omega_e < |\omega|$ are negligible. Since

$$|\sin[(\Omega_1-\Omega_2)t/2]| = \{(2/\pi) + (4/\pi)\sum_{n=1}^{\infty}[\cos(\Omega_1-\Omega_2)nt]/(1-4n^2)\}, \tag{B.42}$$

the average power level of the component at a frequency of $6(\Omega_1-\Omega_2)$ is -40dB relative to the zero frequency component. Since $6(\Omega_1-\Omega_2)<12\Omega_h$, an effective bandwidth choice of $\Omega_e=12\Omega_h$ is reasonable.

Equation B.38 can be rewritten as:

$$y(t) = -2A_2\sin[(\Omega_1-\Omega_2)t/2]\sin[(\Omega_1+\Omega_2)t/2]. \tag{B.43}$$

Thus

$$w_2(t) = 2|A_2\sin[(\Omega_1-\Omega_2)t/2]|\{(2/\pi)+(4/\pi)\sum_{n=1}^{\infty}[\cos(\Omega_1+\Omega_2)nt]/(1-4n^2)\}. \tag{B.44}$$

If $\Omega_e<\Omega_{s2}<2\Omega_c-2\Omega_h-\Omega_e$ then

$$v_2(t) = (4/\pi)|A_2\sin[(\Omega_1-\Omega_2)t/2]|. \tag{B.45}$$

Similarly,

$$w_3(t) = A_2[\cos(\Omega_1 t) - \cos(\Omega_2 t)]/2 + |A_2\sin[(\Omega_1-\Omega_2)t/2]|\{(2/\pi)+(4/\pi)\sum_{n=1}^{\infty}[\cos(\Omega_1+\Omega_2)nt]/(1-4n^2)\}. \tag{B.46}$$

If $\Omega_{s3}<\Omega_h-\Omega_m$ and $\Omega_e<\Omega_{s3}<2\Omega_c-2\Omega_h-\Omega_e$, then

$$v_3(t) = (2/\pi)A_2\sin[(\Omega_1-\Omega_2)t/2]. \tag{B.47}$$

[illegible] the system equations and assumptions, the sinusoidal pair [illegible] systems differ only in [illegible].

B.4. CONCLUSION

In this appendix, F/D subsystem components have been described and three common F/D subsystems have been investigated in detail. For arbitrary input signals, the response of a F/D using a square law device is easily determined. Responses of detectors using full and half wave piecewise linear devices are not easily determined in general. It has been shown that, under certain conditions, the outputs of detectors using full and half wave piecewise linear devices differ only by a scale factor. It was also shown that a F/D subsystem using a square law detector can be turned into a F/D subsystem using a full wave piecewise linear device by interchanging square root and lowpass filter operations (see Fig. B.3). Thus, the outputs of these subsystems are not the same in general. Under certain restrictive conditions, however, the subsystems have similar responses to noise, impulse, sinusoid, and sinusoidal pair inputs. These results will be used in Section D.3 to relate spectrograms with the spectrogram-like representations generated from Short-Time Fourier Transform magnitude.

APPENDIX C

GENERALIZED SHORT-TIME FOURIER TRANSFORM COMPUTATION

C.1 GSTFT ANALYSIS USING FIR WINDOWS

Assume that each window function is the Finite-duration Impulse Response (FIR) of a lowpass filter. Let the set of window functions $h_k(n)$ be zero outside the range $0 \leq n \leq M_k - 1$. Note that each window function may have a different duration M_k. Window functions can be defined by an equation, as for a Hamming window, or values may simply be defined on a point by point basis.

When a FIR window is used, the GSTFT magnitude squared is given by:

$$|X_n(e^{j\omega_k})|^2 = [\sum_{m=0}^{M_k-1} x(n-m)h_k(m)\cos(\omega_k m)]^2$$

$$+ [\sum_{m=1}^{M_k-1} x(n-m)h_k(m)\sin(\omega_k m)]^2, \qquad (C.1)$$

where k=1,2,...,K. One of the set of K F/D subsystems is shown in Fig. C.1. In this figure, unit delays are denoted by z^{-1} and amplifier symbols (triangles) indicate multiplication by a constant. The F/D of Fig. C.1 is a discrete-time version of Fig. 2.7b with the filters drawn in detail to show their FIR structure.

The F/D implementation shown in Fig. C.1 (or 2.7b) is of special interest when the data, x(n), has been quantized to one bit. Such a situation arises when speech data has been encoded using linear Delta Modulation (Steele [60]). In this case the bandpass filters can be implemented without use of multiplication; ie., multiplication by zero or one is trivial. Such a structure is therefore suitable for real-time speech analysis systems implemented with microcomputers (Anderson [61]). Note that the same computational efficiency is not achieved by the system of Fig. 2.7a where the data is modulated prior to filtering.

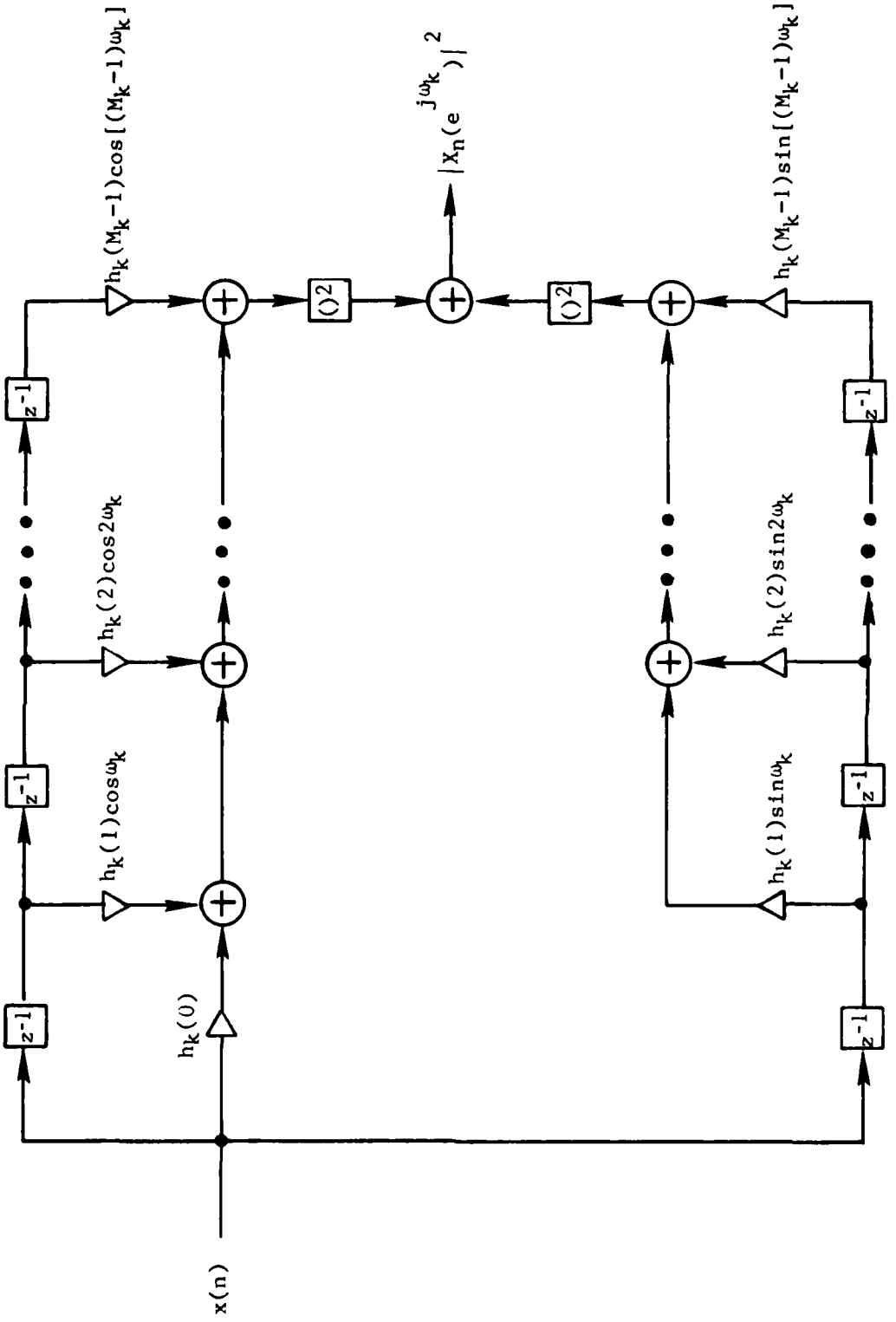


Figure C.1: GSTFT Magnitude Squared using FIR Windows

C.2 GSTFT ANALYSIS USING IIR WINDOWS

Assume that each window function is the Infinite-duration Impulse Response (IIR) of a lowpass filter. Let the set of window functions $h_k(n)$ be zero for $n<i$, where i is an arbitrary integer constant. For $n>i$,

$$h_k(n) = \sum_{\psi=1}^{\Psi_k} p_k(\psi) h_k(n-\psi) + \sum_{r=i}^{R_k} q_k(r)\delta(n-r) \qquad (C.2)$$

where the set of coefficients p_k and q_k are real constants with $q_k(i)\neq 0$. From Equation C.2 it can be seen that the window functions are right-sided sequences (Oppenheim and Schafer [31]) with first nonzero value $h_k(i)=q_k(i)$. This choice for relative time alignment of the window functions, although arbitrary, serves to simplify the synthesis equations of Chapter 3.

Each IIR window function described by Equation C.2 has a rational z-transform (see Appendix A):

$$H_k(z) = \frac{\sum_{r=i}^{R_k} q_k(r) z^{-r}}{1 - \sum_{\psi=1}^{\Psi_k} p_k(\psi) z^{-\psi}} \qquad (C.3)$$

The window function spectral zeros can be determined by factoring a polynomial involving the set of q_k coefficients, and poles are similarly obtained from the p_k coefficients.

Since a FIR filter has zeros but no poles in its system function, the IIR window function includes the FIR analysis window as a special case. To eliminate the poles, let $p_k(\psi)=0$ for all values of k and ψ. For convenience let $i=0$ and $R_k=M_k-1$. Equation C.2 then becomes

$$h_k(n) = \sum_{r=0}^{M_k-1} q_k(r)\delta(n-r), \qquad (C.4)$$

from which it follows that

$$h_k(n) = q_k(n), \quad 0 \leq n \leq M_k-1$$

$$= 0, \text{ otherwise.} \qquad (C.5)$$

The FIR window discussed in Section C.1 is thus a special case of the IIR window function defined by Equation C.2.

The recursive formula for the GSTFT, which results when the IIR window is substituted into the defining equation for the GSTFT (see Equation 2.29) is given by:

$$X_n(e^{j\omega_k}) = \sum_{\psi=1}^{\Psi_k} p_k(\psi) X_{n-\psi}(e^{j\omega_k}) + \sum_{r=i}^{R_k} q_k(r)x(n-r)e^{-j\omega_k(n-r)}. \qquad (C.6)$$

The recursion of Equation C.6 can be implemented using a variety of filter configurations (Oppenheim and Schafer [31]; Rabiner and Gold [56]). For example, a "direct form two" implementation can be obtained by defining an auxiliary sequence $L_k(n)$, where

$$L_k(n) = x(n)e^{-j\omega_k n} + \sum_{\psi=1}^{\Psi_k} p_k(\psi)L_k(n-i). \qquad (C.7)$$

The GSTFT is then computed by

$$X_n(e^{j\omega_k}) = \sum_{r=i}^{R_k} q_k(r)L_k(n-r). \qquad (C.8)$$

The required sine and cosine sequences can also be computed recursively, if desired, since

$$\cos \omega_k n = (\cos \omega_k)[\cos \omega_k(n-1)] - (\sin \omega_k)[\sin \omega_k(n-1)] \qquad (C.9)$$

and

$$\sin \omega_k n = (\sin \omega_k)[\cos \omega_k(n-1)] + (\cos \omega_k)[\sin \omega_k(n-1)]. \qquad (C.10)$$

It should be noted that the sine and cosine sequences computed via the recursion may become less accurate with increasing n. This problem can be overcome by periodically resetting the recursion variables to their correct values. Correct values for the reset operation may be obtained from a similar, but lower frequency, recursion (Gold and Rader [62]).

Fig. C.2 depicts a F/D subsystem using a direct form two filter implementation and recursive sine and cosine generation. Parameters for this subsystem are i=1, Ψ_k=3, and R_k=2. Note that the F/D of Fig. C.2 is a discrete-time version of Fig. 2.7a with the filters drawn in detail to show their IIR structure. This implementation is suitable for real-time speech analysis systems, and may be used in the system described in Chapter 2.

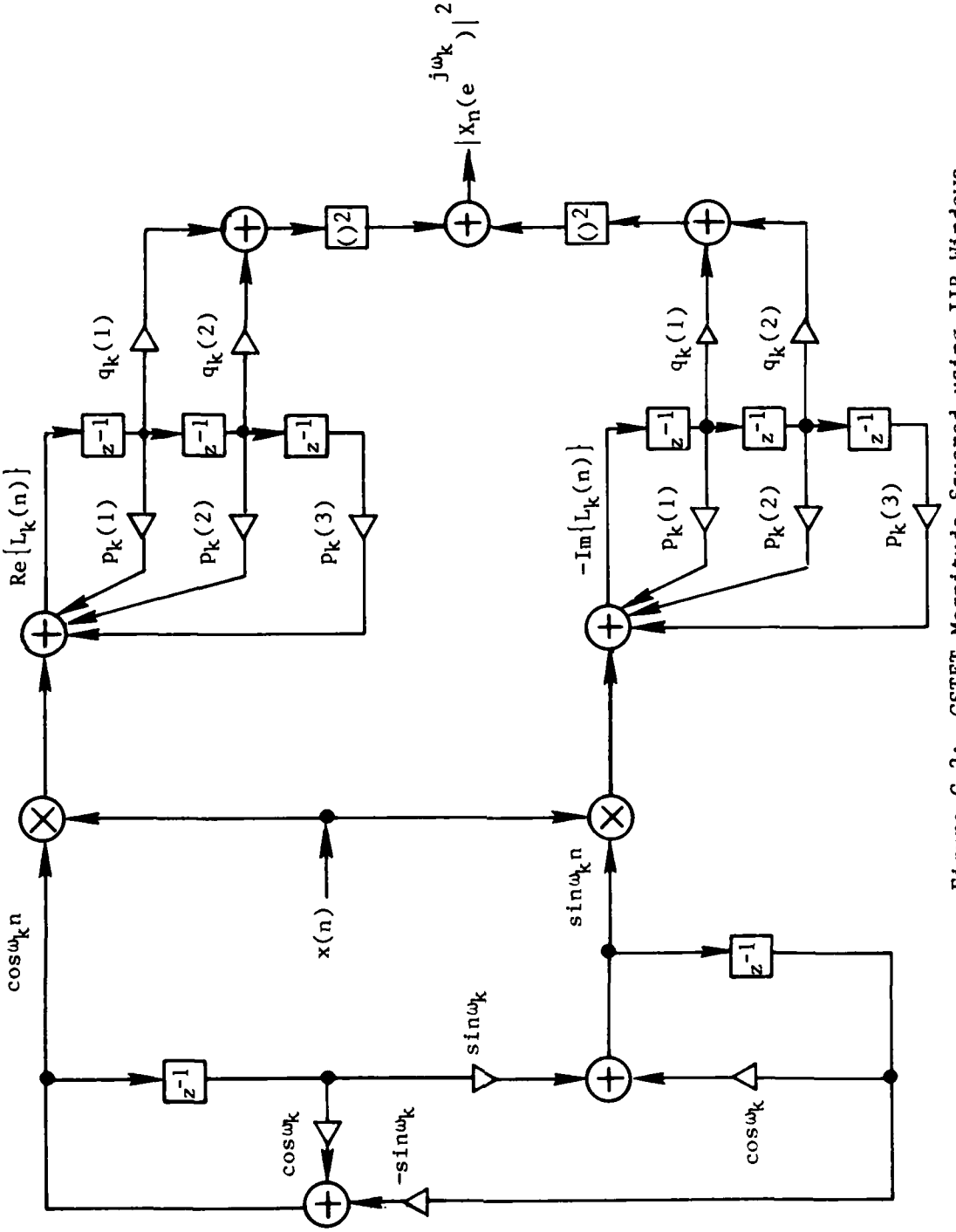


Figure C.2: GSTFT Magnitude Squared using IIR Windows

APPENDIX D

APPLICATIONS

D.1 INTRODUCTION

Filter/Detector (F/D) subsystems are used in a variety of speech analysis and synthesis systems. By varying the bandpass filter characteristics, memoryless nonlinearity type, and smoothing filter cutoff frequency, relationships between several speech processing techniques can be examined.

In this appendix, the new relationship between Short-Time Fourier Transform (STFT) magnitude squared and F/D subsystems, as derived in Chapter 2, is used to describe channel vocoder operation. The relationship is also used to explain similarities between spectrograms

and the spectrogram-like representations generated from STFT magnitude, give a new F/D interpretation to the FFT magnitude, demonstrate the equivalence between the discrete-time Welch method of power spectral estimation and results produced by continuous-time power spectral estimation methods, and to examine several approaches to variable bandwidth analysis. Digital (discrete-time) as well as analog (continuous-time) systems will be discussed.

D.2 CHANNEL VOCODERS

Channel vocoders are analysis/synthesis systems which model a speech signal as being either voiced (having a periodic pitch) or unvoiced (noise-like). The analyzer typically includes a voiced/unvoiced (V/UV) decision subsystem, a pitch extractor to determine the fundamental frequency of voiced signals, and a F/D bank. The synthesizer contains a pitch generator, noise source, V/UV selector switch, modulators, and bandpass filters.

A channel vocoder analyzer described by Rabiner and Gold [56] uses sixteen bandpass filters with nonuniform center frequency spacing to analyze the .3-3 KHz frequency range. The bandwidth for the lowest frequency filter is 125 Hz while a bandwidth of 400 Hz is used for the highest frequency filter. The smoothing filter bandwidth is 25 Hz, and is the same for all channels regardless of the bandpass filter characteristics. Thus, the F/D bank measures only quasi-stationary aspects of the speech signal such as input signal variance (see Section B.3.5).

Since the smoothing filter has narrow bandwidth, F/D outputs are similar for a variety of different memoryless nonlinearities. Half wave and full wave piecewise linear devices are most often used in analog channel vocoders as they are easily implemented with diodes. A square law device is often used in digital channel vocoders. The increased dynamic range requirements are offset by the fact that the square law device produces bandlimited signals which can be represented digitally with little aliasing. In certain digital channel vocoder applications the square law device yields superior results compared to the full wave linear device (Sondhi, et al [54]).

The speech analysis/synthesis system based on perception, described in Chapters 2 and 3, can be viewed as a channel vocoder which does not require pitch extraction. The data rate for such a system, however, is much higher than that normally associated with channel vocoders. The data rate can be reduced by placing additional lowpass filters at each F/D output. However, it is clear from the results of Fig. 4.56 that high quality speech cannot be reconstructed from such lowpass filtered F/D outputs alone, and additional information is required. One method for obtaining such information, which corresponds to a form of pitch extraction, is described in Section 5.3. Note that, contrary to comments by Rabiner and Gold [56], a channel vocoder analyzer does not preserve the Short-Time Fourier Transform (STFT) magnitude, but instead preserves a lowpass filtered version of a generalized form of the STFT magnitude.

D.3 SPECTROGRAMS

The Sound Spectrograph machine [63] employs a measurement system which is similar to the F/D of Fig. B.3b. The machine uses a diode rectifier to implement the full wave piecewise linear device. Two types of analysis can be performed. A wideband analysis uses a bank of bandpass filters each having an effective 300 Hz bandwidth, while a narrowband analysis uses 50 Hz bandwidth filters. Filter center frequencies are 20 Hz apart (Flanagan [1]) and the frequency range .05-7 KHz is analyzed. Each lowpass smoothing filter has an effective bandwidth of several hundred Hertz, and is sufficiently wide to pass any envelope frequencies which may be present at the output of a wideband filter due to beating of adjacent harmonic pitch components.

It was shown in Section B.3.4 that since lowpass filters and square root devices are not interchangable, outputs of the F/D subsystems depicted in Figs. B.3a and B.3b are not the same in general. However, parameters for spectrogram generation are such that similar results may be produced by both F/D subsystems for a variety of input signals, as shown in Section B.3. Furthermore, since the STFT magnitude can be used to implement the F/D of Fig. B.3a, the STFT magnitude can also be used to roughly simulate Sound Spectrograph machine operation (Oppenheim [64]; Wood and Oppenheim [65]; Rabiner and Schafer [3]). Note that, contrary to results given by Flanagan [1], the correct F/D approximation to STFT magnitude is shown in Fig. B.3a and the F/D of Fig. B.3b is applicable only under the restrictive conditions discussed in Section B.3.

D.4 SLIDING DFT IMPLEMENTATION OF THE STFT

In this section, it will be shown that the STFT can be computed by performing the Discrete Fourier Transform (DFT) on segments of a long data sequence. The DFT approach is attractive since it can be efficiently implemented via the Fast Fourier Transform (FFT) algorithm.

The DFT is given by (Oppenheim and Schafer [31]):

$$Y(e^{j\omega_k}) = \sum_{m=0}^{M-1} x(m)h'(m)e^{-j\omega_k m} \tag{D.1}$$

where

$$\omega_k = 2\pi k/M \tag{D.2}$$

and k=0,1,2,...,M-1. Note that the analysis frequencies are uniformly spaced. The DFT window function h'(m) is finite length, and is zero outside the range $0 \leq m \leq M-1$.

A sliding DFT analysis is defined by:

$$Y_n(e^{j\omega_k}) = \sum_{m=-\infty}^{\infty} x(n+m-M+1)h'(m)e^{-j\omega_k m}. \tag{D.3}$$

Although the summation limits have infinite range, terms in the summation are nonzero only on the interval $0 \leq m \leq M-1$ due to the finite length DFT window function h'(m).

From Equation D.3 it can be seen that the sliding DFT segments the data sequence x(n) into sections of length M and performs a DFT on

each segment. For example, $Y_{M-1}(e^{j\omega_k})$ is the DFT of the first M signal points x(0), x(1), ..., x(M-1). Other definitions of the sliding DFT (Oppenheim [64]) use a time index such that the sliding DFT value at n=0 is the DFT of the first M signal points. Many variations are possible, but the resulting differences are unimportant and the definition of Equation D.3 is chosen for convenience.

The sliding DFT need not be computed for every time n. Often, to decrease computation time and data storage requirements, only samples of the sliding DFT are desired. In this case, it becomes the hopped DFT discussed by Rabiner and Gold [56].

To relate the sliding DFT to the STFT, define a new window function h(m) to be a time-reversed and delayed version of h'(m); ie.,

$$h'(m)=h(M-1-m) \tag{D.4}$$

for all m. Since many window functions used in conjunction with the DFT are symmetrical, the time-reversed and delayed window is often the same as the original window. By substitution into Equation D.3:

$$\begin{aligned} Y_n(e^{j\omega_k}) &= \sum_{m=-\infty}^{\infty} x(n+m-M+1)h(M-1-m)e^{-j\omega_k m} \\ &= \sum_{m=-\infty}^{\infty} x(n-m)h(m)e^{-j\omega_k(M-1-m)} \\ &= e^{j\omega_k(n+1)} X_n(e^{j\omega_k}), \end{aligned} \tag{D.5}$$

where $X_n(e^{j\omega_k})$ is the discrete-time STFT (a special case of the Generalized STFT) evaluated at a fixed frequency ω_k, as given by Equation 2.21. This result leads to the following procedure for computing the STFT via the sliding DFT:

1. Form a time-reversed and delayed version of the STFT window function, h(n), and call it h'(n).
2. Pre-multiply a data segment by h'(n).
3. Perform a DFT on the windowed segment by using the FFT algorithm.
4. Post-multiply the results by a time-varying complex exponential.

The complex exponential post-multiplication step converts sliding DFT outputs, which are bandpass functions, into lowpass STFT results. If only magnitudes are computed, then step #4 is unnecessary since:

$$|Y_n(e^{j\omega_k})| = |X_n(e^{j\omega_k})|. \qquad (D.6)$$

From Equation D.6 it is evident that a F/D interpretation can be placed on the sliding DFT magnitude as well as on the STFT magnitude. Since it is common practice to investigate the spectrum of a signal by examining the DFT magnitude of a signal segment, the F/D interpretation can be used to obtain insight into spectral behavior as a function of time. Although it is well known that the sliding DFT can be used to implement a filter bank (Rabiner and Gold [56]), the nature of the detection process brought about by the magnitude operation has not been adequately discussed in the literature. Therefore, a complete example is given in the remainder of this section.

For convenience, the Hamming window (Oppenheim and Schafer [31]) will be used both as a window function and also for lowpass filtering purposes. The Hamming window of length M, normalized for unity gain in the frequency domain, is:

$$h'(n) = [.54 - .46\cos(2\pi n/M-1)]/(.54M), \quad 0 \leq n \leq M-1,$$

$$= 0, \text{ otherwise.} \qquad (D.7)$$

The one-sided main lobe bandwidth of a Hamming window of length M is $\omega_{h'}=4\pi/M$.

As a specific example, assume that a continuous-time signal is sampled at a 10 KHz rate and a 12.8 millisecond (128-point) segment is selected for analysis. Let the data segment be denoted by $x(n)$, $0 \leq n \leq 127$. An example sequence is shown in Fig. D.1.

Using a 128-point Hamming window, the DFT magnitude squared is computed. The DFT is given by Equation D.1, and frequency spacings are given by Equation D.2, where $k=0,1,\ldots,64$. Since $x(n)$ is real, it is not necessary to compute values for $k=65,\ldots,127$. The DFT magnitude squared for the sequence of Fig. D.1 is shown in Fig. D.2.

A bank of discrete-time F/D subsystems of the type shown in Fig. 2.8 is now implemented, and the output of the F/D bank is sampled at a specific time. Since the Hamming window is symmetric, the DFT window function $h'(n)$ is the same as the STFT window function $h(n)$. Thus, in

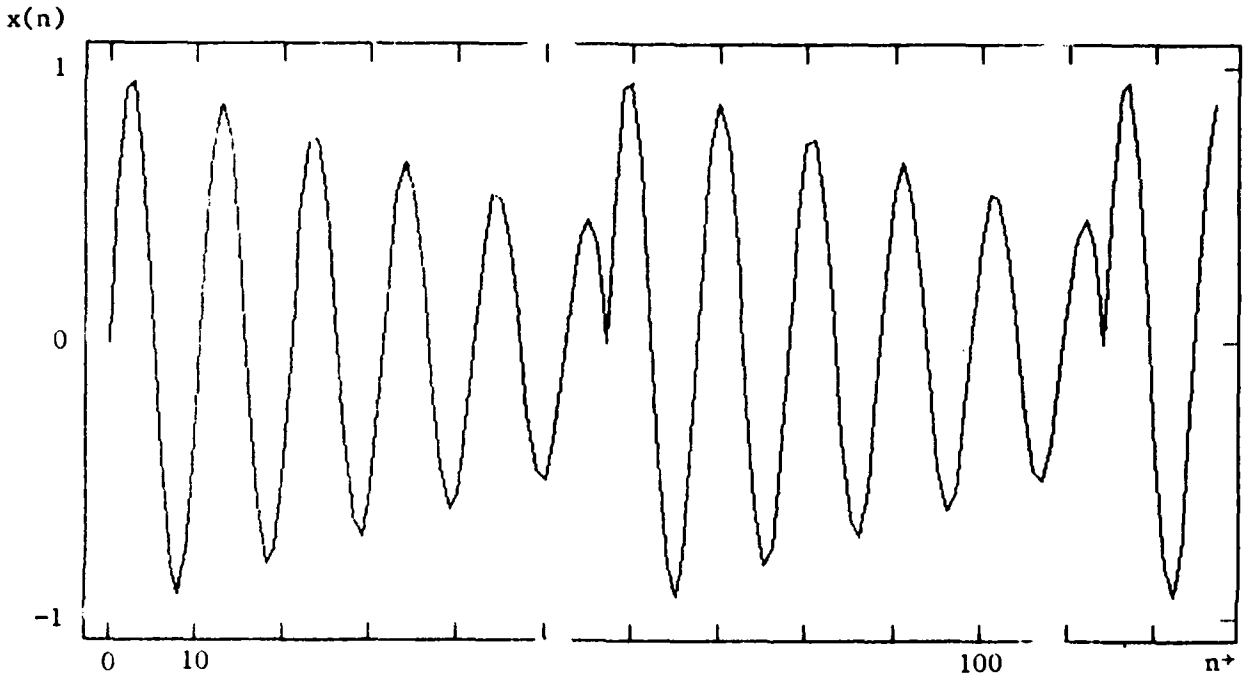


Figure D.1: Example Sequence

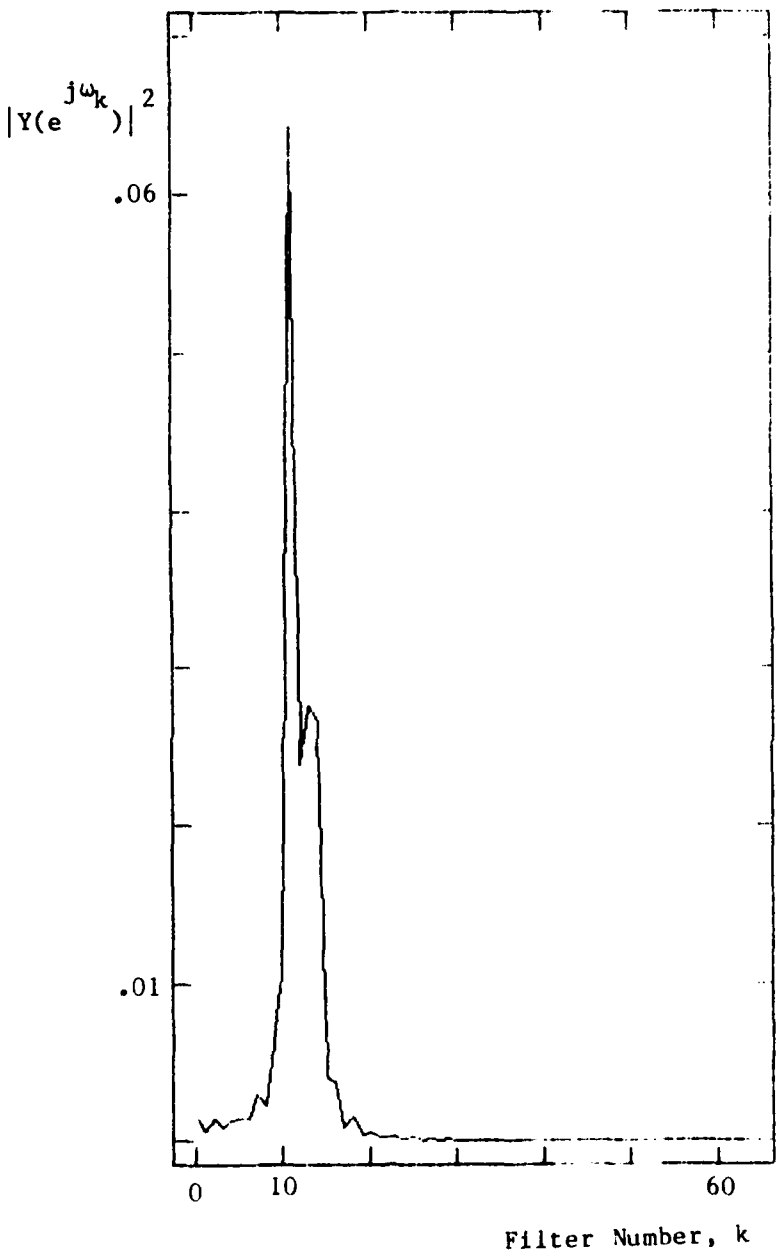


Figure D.2: DFT Magnitude Squared

Fig. 2.8, h(n) is a 128-point Hamming window, $\omega_c=\omega_k$, and θ is arbitrarily chosen as zero. Since the lowpass smoothing filter must have the same bandwidth as the bandpass filter, a 63-point Hamming window is used for $h_{sl}(n)$. The smoothing filter, therefore, introduces a 32-sample delay. For convenience, define $x(n)=0$ for $n<0$ and $n>127$. Let the output of each F/D subsystem be denoted by $2v_k(n)$. To approximate the DFT results, $2v_k(158)$ is computed:

$$2v_k(158) = 2 \sum_{i=0}^{62} [\sum_{m=0}^{127} x(158-i-m)h(m)\cos(\omega_k m)]^2 h_{sl}(i). \qquad (D.8)$$

The results are plotted in Fig. D.3, and are comparable to those of Fig. D.2.

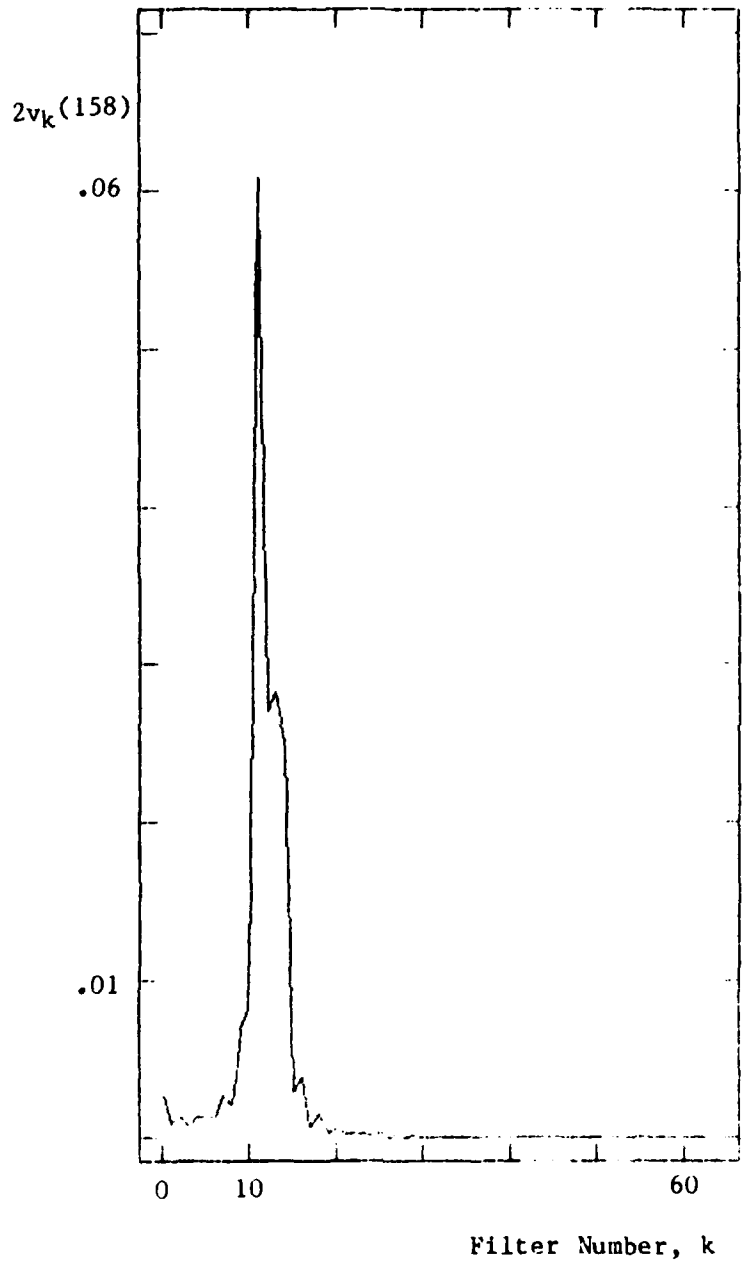


Figure D.3: F/D Bank Output Sample

D.5 AVERAGE POWER SPECTRUM ESTIMATION

In certain applications, reconstruction of high quality speech signals directly from spectral magnitude data is of interest. As demonstrated in Chapter 4, a high degree of time-domain detail in the data is essential for such applications. In many other applications, however, exact signal reconstruction is not required and the time-domain detail can be eliminated by averaging the spectral magnitudes.

For example, since one noise sequence may sound the same as many others, retention of information for exact signal reconstruction is unnecessary. For data reduction purposes it is more efficient to characterize the random process which originally created the data. Synthesis is then accomplished by generating a new data sequence from a random process which has the same characterization as the original data sequence. Since a random process is often described in terms of its power spectral density, average spectral magnitudes are useful for estimating random process characteristics. This approach is generally employed by the channel vocoders described in Section D.2.

The continuous-time F/D subsystem of Fig. B.3a, without the square root device, can be used to measure the average power spectrum of speech (Dunn and White [66]). The speech signal is decomposed into a number of frequency bands by a bank of bandpass filters. The mean squared power in each band is computed by placing a square law device and smoothing filter at the bandpass filter outputs. The smoothing filter time constant may range from 125 milliseconds for short-time measurements to more than a minute for long-time analysis. A long-time analysis can also be obtained by averaging many short-time measurements.

Digital techniques can also be used for power spectrum estimation. For example, a popular technique known as the Welch method can be described in terms of a digital F/D bank. The Welch spectrum estimate, as discussed by Oppenheim and Schafer [31], is computed by sampling the sliding DFT in a manner equivalent to hopping with no overlap. This is done in an attempt to ensure statistical independence of the measurements, and yields an undersampled representation. Hopping with overlap has also been discussed in the literature (Welch [67]), but will not be considered here. Magnitude squared samples are averaged for each frequency, and weighted by a constant which depends on the window function. The Welch spectrum estimate is therefore equivalent to sampling F/D bank outputs and averaging the samples in each channel to determine the power spectral density of the input noise process. The Welch method thus obtains a long-time measurement by averaging short-time measurements.

In precise terms, the Welch spectrum estimate is given by:

$$B_{xx}(e^{j\omega_k}) = (1/PQ) \sum_{p=1}^{P} |Y_{pM-1}(e^{j\omega_k})|^2 \qquad (D.9)$$

where

$$Q = \sum_{m=0}^{M-1} [h'(m)]^2, \qquad (D.10)$$

h'(m) is the DFT window function, $Y_n(e^{j\omega_k})$ is the sliding DFT given by Equation D.3, and the data sequence is x(n), $0 \leq n \leq MP-1$. It can be seen from Equation D.6 that the STFT can also be used to compute the Welch spectrum estimate, as long as the STFT window function is finite length and the window time-reversal and delay are taken into account.

D.6 NONUNIFORM BANDWIDTH ANALYSIS

Although a critical bandwidth filter bank is useful for perception-based speech analysis, such filter banks are not always readily available in the form of existing electronic equipment or computer programs. The most common type of digital filter bank consists of many narrow bandpass filters which are uniformly spaced in frequency, all filters having the same bandwidth. These filter banks are often implemented by the sliding DFT, since the sliding DFT can be efficiently computed via the FFT algorithm (see Section D.4). Such filter banks must be modified for perception-based analysis, allowing the filters to have a bandwidth which varies with center frequency. Modifications generally involve combining the outputs of several narrowband filters in order to simulate a single filter of broader bandwidth. Although such modifications can be used in dealing with Linear Time-Invariant (LTI) systems, effects of the nonlinear detection process which follows the filter bank must also be taken into account.

This section presents several approaches to variable bandwidth analysis which can be implemented by modifying narrowband filter banks. Unfortunately, if these approaches achieve the desired result at all, they do not approach the computational efficiency of the Generalized Short-Time Fourier Transform (see Appendix C). Nonetheless, since the approaches presented in this section are commonly used in practice, it is worthwhile to investigate the problems associated with each method.

D.6.1 SUMMATION OF FILTER/DETECTOR OUTPUTS

The sliding DFT is often used to implement a bank of many narrow bandpass filters (Rabiner and Gold [56]). The DFT magnitude can thus be interpreted as a time sample of a narrowband F/D bank output. The narrowband analysis may be broadened as required by adding together two or more F/D outputs, where the filters are adjacent in frequency. Although this approach broadens the steady state sinusoidal response, it will be shown that all outputs have the same form of impulse response. Since it is desirable to have shorter impulse response duration on the high-frequency wide-bandwidth F/D subsystems, as shown in Fig. 2.11, usefulness of this approach is diminished.

To demonstrate, let ω_a and ω_b be two analysis frequencies of interest. Define a broadened F/D output as:

$$Z_1(n) = |Y_n(e^{j\omega_a})|^2 + |Y_n(e^{j\omega_b})|^2 \qquad \text{(D.11)}$$

The two sliding DFT components thus implement a pair of narrowband F/D subsystems which are added together to form a broadened F/D. When the input is an impulse, $x(n)=\delta(n)$, the broadened F/D output is $Z_1(n)=2h^2(n)$. The broadened F/D output thus has the same form of impulse response as either of the two original narrowband F/D subsystems.

Adding together F/D outputs to decrease frequency resolution fails to give a corresponding improvement in time resolution. Important temporal information may be lost due to this "smearing" effect. It can easily be shown that the same result holds whether the F/D subsystems are implemented via the sliding DFT or implemented directly by using individual bandpass filters, memoryless nonlinearities, and lowpass smoothing filters.

D.6.2 SUMMATION OF FILTER OUTPUTS PRIOR TO DETECTION

Filter broadening is commonly accomplished by adding together the outputs of several adjacent (in frequency) filters prior to the detection process. Although the desired filter broadening is achieved, it will be shown that undesirable components may appear in the impulse response of the resulting F/D subsystem.

Consider the impulse response of the directly implemented F/D shown in Fig. D.4a. When $x(n)=\delta(n)$ the output is:

$$Z_2(n) = 2h^2(n)[1 + \cos(\omega_a - \omega_b)n]. \tag{D.12}$$

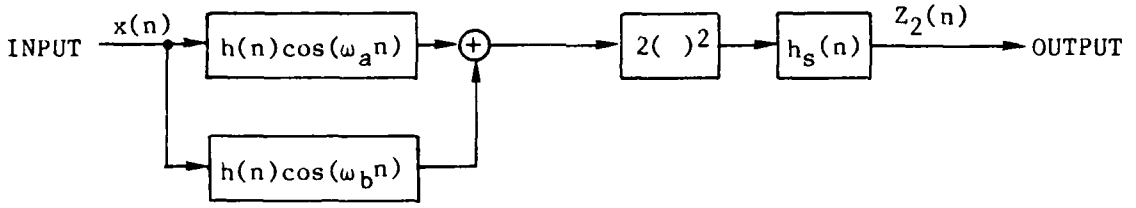
Note the presence of a high level beat frequency component.

Beat frequencies are also present when the sliding DFT is modified by adding adjacent complex results. Define a broadened F/D output as:

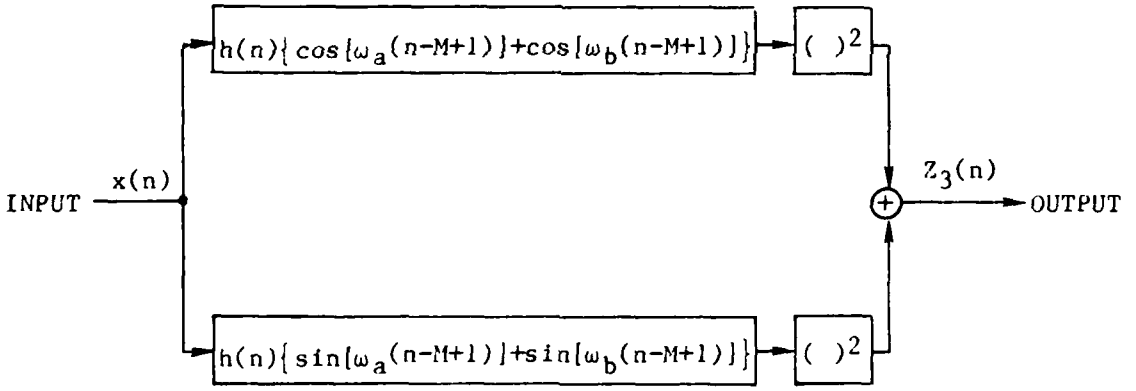
$$Z_3(n) = |Y_n(e^{j\omega_a}) + Y_n(e^{j\omega_b})|^2 \tag{D.13}$$

$$= |Y_n(e^{j\omega_a})|^2 + |Y_n(e^{j\omega_b})|^2 + 2\mathrm{Re}\{Y_n(e^{j\omega_a})[Y_n(e^{j\omega_b})]^*\},$$

where the asterisk denotes complex conjugation. The block diagram for this subsystem is shown in Fig. D.4b. It is easily seen that the bandpass filters have the desired broadened characteristics.



(a) Direct Implementation



(b) Sliding DFT Implementation

Figure D.4: Summation of Filter Outputs Prior to Detection

To investigate the dynamic characteristics of the subsystem shown in Fig. D.4b, let $x(n)=\delta(n)$. The output then becomes:

$$Z_3(n) = 2h^2(n)\{1 + \cos[(\omega_a-\omega_b)(n-M+1)]\}. \qquad (D.14)$$

The impulse response of the new broadened F/D subsystem thus contains an undesirable beat frequency term.

The beat frequency in the sliding DFT becomes more pronounced (ie., more beat cycles are evident in the F/D impulse response) when $\omega_a-\omega_b$ is large. The effect is minimized if two adjacent filters are added. For addition of two adjacent filters, it follows from Equation D.2 that:

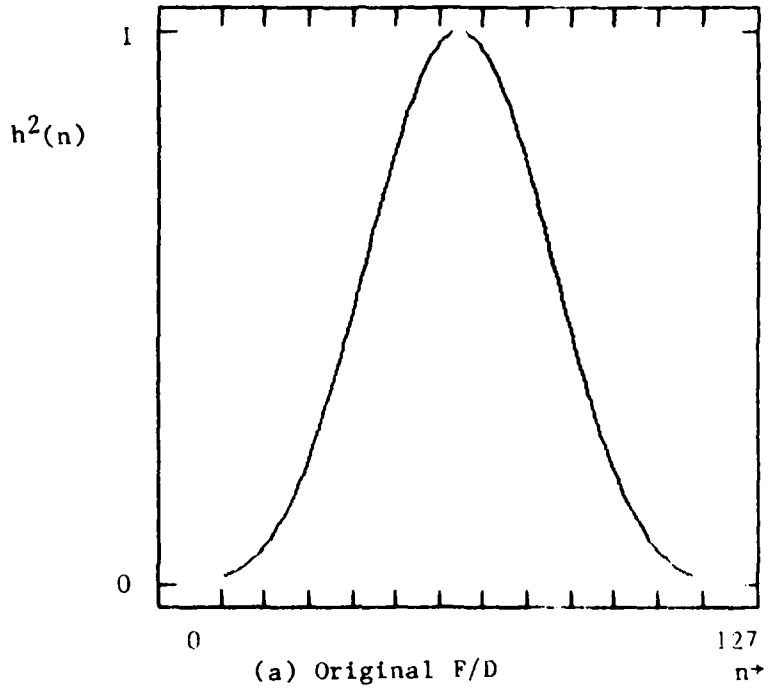
$$Z_3(n) = 2h^2(n)\{1 + \cos[2\pi(n+1)/M]\}. \qquad (D.15)$$

As a specific example consider a 128-point sliding DFT using a Hamming window; ie., M=128 and

$$h'(n) = .54 - .46\cos[2\pi n/(M-1)], \quad 0 \leq n \leq M-1$$

$$= 0, \text{ otherwise.} \qquad (D.16)$$

Since the Hamming window is symmetric it follows from Equation D.4 that $h'(n)=h(n)$. The impulse response of an original F/D subsystem, $h^2(n)$, is shown in Fig. D.5a. The impulse response of the broadened F/D, as given by Equation D.15, is shown in Fig. D.5b. In this example of a broadened F/D subsystem, a single impulse input results in two peaks at the output, which is generally an undesirable result.



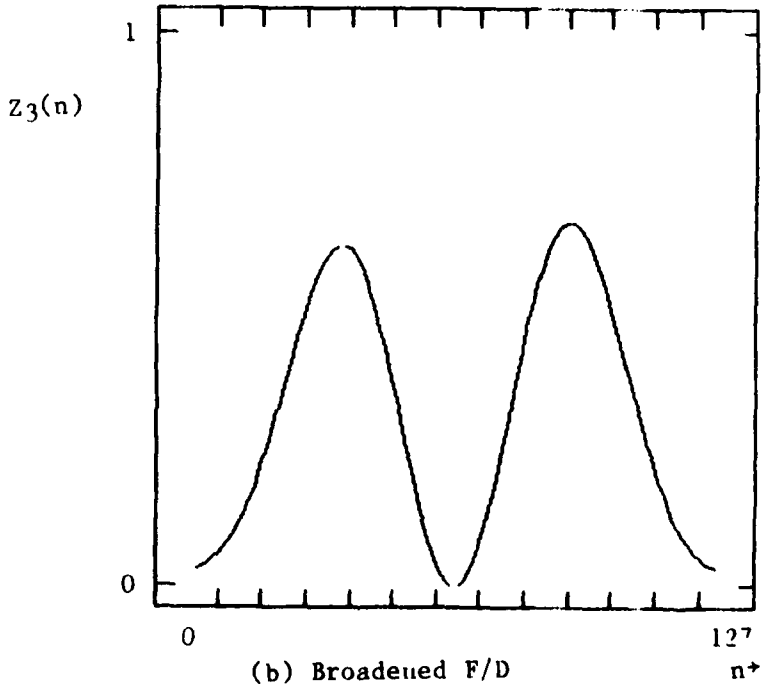


Figure D.5: F/D Impulse Responses

As noted by Rabiner and Gold [56], the equivalence between multiplication in the time domain and convolution in the frequency domain implies that windowing can be accomplished by a complex weighted summation of many adjacent (in frequency) values of the sliding DFT. A carefully chosen combination of weights can be used to modify the original analysis window, and can reduce or eliminate beat frequency effects. Thus, although it is possible to broaden filters by this approach, computational efficiency is sacrificed.

D.6.3 SUMMATION OF STFT COMPONENTS PRIOR TO MAGNITUDE

STFT results are lowpass functions, and are unlike the bandpass results produced by the sliding DFT. Thus, no beat frequencies will occur when adjacent (in frequency) complex STFT results are added and the magnitude squared is computed. Let the broadened STFT analysis be given by:

$$Z_4(n) = |X_n(e^{j\omega_a}) + X_n(e^{j\omega_b})|^2, \qquad \text{(D.17)}$$

where ω_a and ω_b are two STFT frequencies of interest. The subsystem block diagram is shown in Fig. D.6. When the input is an impulse at time m, $x(n)=\delta(n-m)$, the output is:

$$Z_4(n) = 2[1 + \cos(\omega_a-\omega_b)m]h^2(n-m). \qquad \text{(D.18)}$$

Thus the subsystem has a time-varying impulse response, which is clearly an undesirable result.

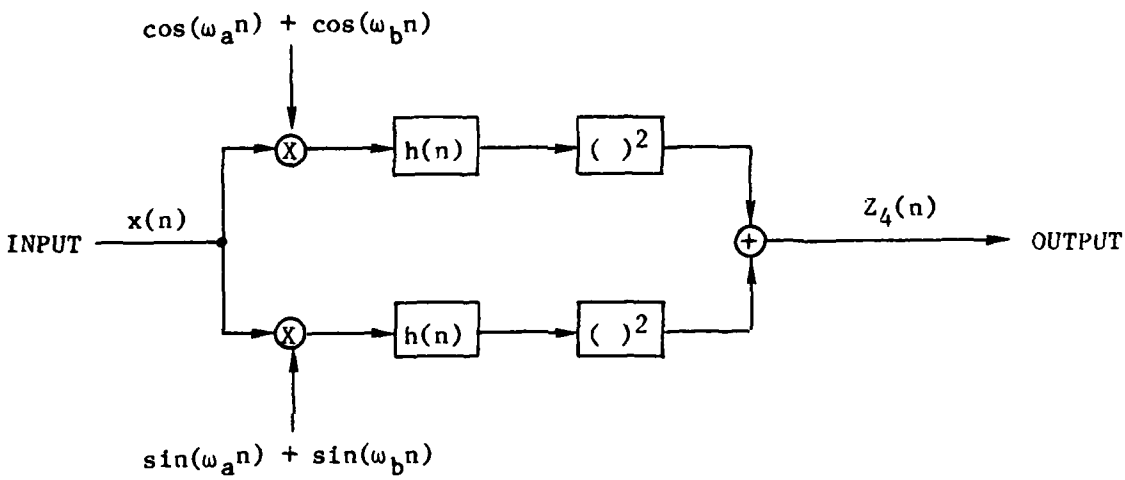


Figure D.6: Summation of STFT Components Prior to Magnitude

D.7 CONCLUSION

In this appendix, the new relationship between STFT magnitude squared and F/D subsystems was used to describe the characteristics of several speech analysis and synthesis systems. The relationship provides a common basis for understanding the operation of many systems, and can be used to indicate similarities and differences between various systems.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ESD-TR-84-048	2. GOVT ACCESSION NO. AD-A151320	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Speech Analysis/Synthesis Based on Perception		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER Technical Report 707
7. AUTHOR(s) James C. Anderson		8. CONTRACT OR GRANT NUMBER(s) F19628-85-C-0002
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173-0073		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Element Nos. 63250F
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Systems Command, USAF Andrews AFB Washington, DC 20334		12. REPORT DATE 5 November 1984 13. NUMBER OF PAGES 260
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* Electronic Systems Division Hanscom AFB, MA 01731		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

Submitted in partial fulfillment of Doctor of Philosophy degree from Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, September 1984.

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

auditory modeling
short-time Fourier transform
magnitude-only reconstruction
power spectrum estimation
perception
filter banks
speech recognition
spectograms
vocoders

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

A speech analysis system based on a combination of physiological and psychoacoustic results has been developed. The system contains a nonuniform Filter/Detector bank. A new relationship between Filter/Detectors and the Short-Time Fourier Transform magnitude is derived, and a generalized version of the Short-Time Fourier Transform magnitude is used to implement the analysis system. The new relationship is also applied to a discussion of channel vocoders, spectrograms, the sliding Discrete Fourier Transform, average power spectrum estimation, and nonuniform bandwidth analysis. Next, a new synthesis approach is used to reconstruct signals from the magnitude data produced by the nonuniform analysis. Apart from an overall sign factor, the analysis/synthesis system achieves exact reconstruction in the absence of data modification. The ability of the system to reconstruct signals from modified data is also demonstrated. Suggestions for further research, including data reduction and Automatic Speech Recognition applications, are given.

DD FORM 1 Jan 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

END

FILMED

4-85

DTIC