THE ROLE OF ZERO CROSSINGS IN SPEECH RECOGNITION AND PROCESSING

Lawrence Robert Morris

1970

A thesis submitted for the degree of
Doctor of Philosophy, in the Faculty of Engineering,
University of London

Department of Electrical Engineering,
Imperial College of Science and Technology,
University of London

ABSTRACT

The role of zero crossings in speech recognition and processing is twofold: zero crossings define the clipped speech waveform, and zero crossing interval sequences may yield objective estimates of certain speech features or form patterns representative of the original speech signal.

This thesis consists of four sections, two of which provide parallel treatment of the dual aspects of zero crossing phenomena.

First, topics concerning signal theory and the special nature of speech are considered. Included is a discussion of the philosophy and implications of machine classification as opposed to human perception of speech sounds.

Next, phenomena associated with the audition of clipped speech are reviewed and efforts to explain the high intelligibility of clipped speech are critically examined. The evidence which justifies the consideration of zero crossings as useful input parameters for automatic speech recognition is surveyed and interrelated.

Then, two experiments employing a measure of average rate of zero crossings and zero crossing interval histograms, respectively, in limited vocabulary, adaptive automatic speech recognition are described. The experimental results, though encouraging, reinforce the belief that a lack of understanding concerning the significance of zero crossings as parameters

representative of speech signals exists.

The final section approaches zero crossing-related speech phenomena from a unified, zero-based point of view. The concept of zero crossings as a subset of those zeros which are sufficient to completely specify a bandlimited periodic signal is introduced. It is shown that the clipping-bandlimiting operator effectively samples the speech waveform at the real zeros (zero crossings) and has limited ability to manipulate the complex zeros. A zero-based relationship connecting pre-clipping signal processing and post-clipping intelligibility is proposed and related to unexplained observations in psychoacoustic experiments. The sufficiency of zero crossings as objective waveform descriptors is then examined and it is argued that the zero crossings of highly structured signals such as vowels may implicitly contain sufficient information to almost completely reconstruct the signal's power spectrum.

The real problem in formulating
a mathematical model is to find
an adequate compromise
between realism and mathematical convenience.

I. J. Good, 1958

I can tell from your voice harmonics, Dave,
that you're badly upset. Why don't you
take a stress pill and get some rest?

HAL 9000 computer
in *2001: A Space Odyssey*,
Stanley Kubrick and Arthur C. Clarke

PREFACE

The research reported in this thesis constitutes a continuation of investigations into the *role of zero crossings in speech recognition and processing*. J.M. Dukes (1954), A.J. Fourcin (1959) and V.J. Phillips (1961), for example, have explored certain aspects of this subject in studies at the Imperial College Communications Laboratories.

The form of this thesis was dictated by several factors, one of which is that the thesis title implies that a comprehensive treatment of the subject is presented.

First, it is necessary to review briefly some aspects of signal theory in order to provide a firm basis for the establishment of certain results in zero-based signal representation. Similarly, various facts concerning speech and hearing in general and the time-frequency characteristics of speech sounds in particular must be established in order to provide a foundation for the understanding of the value of spectral features in human recognition (perception) and automatic recognition (classification). A common purpose of both these reviews is to *clarify time-frequency relationships in speech processing, analysis, and perception.*

Next, the philosophy of automatic speech recognition is discussed with the object of explaining the interactions among the three stages of the recognition process: parameterization, transformation of parameters, and decision making. This material includes several examples of recognition schemes and provides an introduction to our own experiments.

In reviewing the literature on clipped speech and zero crossing-related phenomena we reached at least one significant conclusion: the published reports in this area are scattered and relatively obscure. The lack of interrelationship among extant results is such that several unfounded myths have arisen regarding what *has* and what *has not* been shown regarding certain aspects of zero crossing-related speech signal phenomena. For this reason, two chapters are devoted to a detailed review and critique of research in this area with a view to explicitly establishing just what is known and understood in this field.

The final section of this thesis treats zero crossing-related speech phenomena from a zero-based viewpoint. That zeros can be regarded as *informational attributes of signals* (with zero crossings constituting a subset of the total zero array) was formally established by H.B. Voelcker in 1966. However, we expect that zero-based concepts will be essentially unfamiliar to most readers of this thesis. Therefore, a substantial amount of space is set aside to provide the background material *necessary* to create some feeling for these concepts and *essential* to the understanding of our zero-based treatment of speech clipping and zero crossing-related phenomena.

Zero-based signal theory may be considered novel and perhaps unrealistic for many signal analysis problems. However, the fact remains that vowels are most realistically represented over a pitch period as a finite Fourier series, and that *zero-based product representations specify periodic signals in terms of zero crossings and complex zeros*. Thus, although this thesis is ostensibly concerned with *zero crossings*, it is through the clarification of the significance of these unfamiliar *complex zeros* that the role of zero crossings in speech recognition and processing is deduced.

L. Robert Morris
June 1970.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Professor E.C. Cherry, for his initiation, encouragement, and support of this project, especially during the initial phases. Additionally, I am very grateful to Professor H.B. Voelcker, of the University of Rochester, both for the interest shown and advice given during his visit to Imperial College and for his invitation to visit the University of Rochester during the latter phases of this research. The scope of the work was greatly expanded due to Professor Voelcker's suggestions and guidance, and his careful criticism of the manuscript contributed significantly to improvements in its final form.

Many thanks are due to my colleagues at Imperial College for their advice, particularly Drs. R.L. Wiley, K. Patel and D.J. Goodman. I am particularly appreciative to E.V. Stansfield for much interesting discussion and for his assistance in locating material during the trans-Atlantic phase of this research. G. Lockhart provided invaluable advice concerning certain aspects of numerical analysis. A. White and B. Gurnhill assisted in equipment construction.

I am also indebted to Aristides A. Requicha who, during his sojourn at Imperial College and my visit to the University of Rochester, was the source of much valuable and enlightening advice. Professor E. Titlebaum, also at Rochester, was quite helpful and expressed interest in the work.

Financial assistance was provided by the British Board of Trade Athlone Fellowships for two years and then by the National Research Council of Canada. A research materials grant was donated by English Electric-Leo Marconi and support for research at the University of Rochester was provided by National Science Foundation grant GK 10,861.

GLOSSARY

Major Symbols and Definitions

Note: Arrangement in each alphabetical section is in order of usage with the section of first occurence given in parentheses.

$\{a_k\}$ - the Fourier series (cosine) coefficients of a periodic signal, s(t) (2.1)

$\{b_k\}$ - the Fourier series (sine) coefficients of a periodic signal, s(t) (2.1)

BL{ } - the bandlimiting operator (5.1.7)

comb - $\mathrm{comb}_{\mathcal{T}} s(t) \equiv \sum_{n=-\infty}^{\infty} s(n\mathcal{T})\cdot\delta(t-n\mathcal{T})$ (2.4.1)

C - the clipping operator. $C\,x \equiv \mathrm{sgn}[x]$ (5.1)

$\cos\phi(t)$ - the phase function of s(t) (5.1.7)

$\{c_k\}$ - the (complex) Fourier series coefficients of a periodic signal, s(t) (2.1)

CZ - complex zero (8.1.3)

$\{Cz_k\}$ - the (complex) Fourier series coefficients of $s_{CZ}(t)$ (8.1.3)

δ_{jk} - Kronecker delta (2.1)

$\delta(t)$ - delta function (distribution) (2.4.1)

E{ } - the expectation operator (5.2.1)

F_o - fundamental frequency of a signal periodic in T (T^{-1}) (2.1)

$\mathcal{F}\{\ \}$ - operation of Fourier transformation (2.2)

$\phi(t)$ - the phase of s(t) (2.3.3)

f_o - carrier frequency of a SSB signal (2.3.3)

Fn, F_n	- the n^{th} formant and its frequency (4.3.1)
$\phi'(t)$	- instantaneous frequency of s(t) (6.3.2)
$\overline{\phi'(t)}$	- average value of $\phi'(t)$ over a specified interval (6.3.2)
G(f)	- power spectrum (5.2.1)
$H\{\ \}$	- operation of Hilbert transformation (2.3.1)
Im[ ]	- imaginary part (8.1.1)
LHP	- lower half plane (8.2)
m(t),M(f)	- analytic signal [$m(t) = s(t) + j\,\hat{s}(t)$] and its Fourier transform (2.3.1)
$\lvert m(t)\rvert$	- the envelope of s(t) (2.3.3)
$m_{\omega_o}(t)$	- the analytic counterpart of the SSB translate of s(t) (2.3.3)
$2n_R$	- number of real zeros (zero crossings) per period in a periodic signal (8.1.1)
n_C	- number of complex zero pairs per period in a periodic signal (8.1.1)
2n	- number of zeros per period in a periodic signal (8.1.1)
Ω	- fundamental radian frequency of a signal periodic in T (2.1)
P[ ]	- Cauchy principal value (2.3.1)
rect	- rect[x] = 1 for $\lvert x\rvert \leq \frac{1}{2}$, and zero otherwise (2.3.1)
rep	- $\mathrm{rep}_T s(t) \equiv \sum_{n=-\infty}^{\infty} s(t-nT)$ (2.4.1)
$R(\tau),\rho(\tau)$	- autocorrelation function, normalized autocorrelation function (5.2.1)
ρ_o,ρ_m	- average time rate of zero crossings of a signal and its first derivative (6.2.1)
$\tilde{\rho}_o,\tilde{\rho}_m$	- average value of $\phi'(t)$ for a signal and its first derivative, respectively, measured over a specified interval (6.3.2)

Re[ ] - real part (8.1.1)

RZ - real zero (zero crossing) (8.1.3)

$\{Rz_k\}$ - the (complex) Fourier series coefficients of $s_{RZ}(t)$ (8.1.3)

$s(t), S(f)$ - general signal and its Fourier transform (2.2)

$\hat{s}(t)$ - Hilbert transform of $s(t)$ (2.3.1)

sgn - $\mathrm{sgn}[x] = 1, 0, -1$ as $x > 0$, $= 0$, or < 0, respectively (2.3.2)

$s_{\omega_o}(t)$ - single sideband translate of $s(t)$ (2.3.3)

$\tilde{s}(t)$, $\tilde{S}(f)$ - sampled version of $s(t)$ and $F\{\tilde{s}(t)\}$ (2.4.1)

sinc - $\mathrm{sinc}\ x = \sin \pi x/(\pi x)$ (2.4.1)

SSB - single sideband (5.1.2)

$s(t), \{c_k\}$ - a signal periodic in T and its complex Fourier series coefficients (8.1.3)

$s_{RZ}(t), \{Rz_k\}$ - the real zero component of $s(t)$ and its complex Fourier series coefficients (8.1.3)

$s_{CZ}(t), \{Cz_k\}$ - the complex zero component of $s(t)$ and its complex Fourier series coefficients (8.1.3)

T - period of a periodic signal (2.1)

$\mathfrak{T}$ - sampling interval for a sampled signal (2.4.1)

τ_i - location in time of the i^{th} real zero (8.1.1)

$\tau_\ell \pm j\sigma_\ell$ - location in time of the ℓ^{th} complex zero pair (8.1.1)

U - unit step, $U(x) = 1$, $x \geqslant 1$ and 0 otherwise (9.4.1)

UHP - upper half plane

W - signal bandwidth (2.4.1)

ω_o - carrier frequency of SSB signal (2.3.3)

w - the polynomial plane variable (8.1.1)

x(n),X(k) - sampled signal and its discrete Fourier transform (2.5.1)

z - the complex time variable ($z = t + j\,\sigma$) (8.1.1)

$\mathcal{z}$ - the $\mathcal{z}$-transform variable, $\mathcal{z} \equiv e^{-j2\pi f T}$ (2.5.2)

Miscellaneous

$x*y$ - convolution of x and y (2.3.1)

x^* - complex conjugate of x (2.1)

$\binom{n}{r}$ - $\binom{n}{r} \equiv n!/(n-r)!r!$ (8.4.1)

Phoneme Symbols and Key Words

Vowels		*Fricative Consonants*	
/i/	eve	/v/	vote
/I/	it	/ð/	then
/e/	hate	/z/	zoo
/ɛ/	met	/ʒ/	azure
/æ/	at	/f/	for
/a/	father	/θ/	thin
/ɔ/	all	/s/	see
/o/	obey	/ʃ/	she
/U/	foot	/h/	he
/u/	boot	*Stop Consonants*	
/ʌ/	up	/b/	be
/ɝ/	bird	/d/	day
Nasals		/g/	go
/m/	me	/p/	pay
/n/	no	/t/	to
/ŋ/	sing	/k/	key
Glides and Semi-Vowels			
/j/	you	/w/	we
/r/	read	/l/	let

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

1 INTRODUCTION

1.1 The Problem: Manifestations of Zero Crossings in Speech Recognition and Processing

This thesis is concerned primarily with the interpretation and clarification of two phenomena associated with clipped speech waveforms. First, clipped speech is highly intelligible. Secondly, the same zero crossing interval sequence which defines the clipped speech waveform can be manipulated so as to yield an objective estimate of certain speech spectral features.

The intelligibility of clipped speech is a subjective effect; it is a psychoacoustic phenomenon involving perception of speech using the human auditory system. In contrast, the use of zero crossings for extraction of information from the speech waveform must be cast in an objective, signal theoretic context. Nevertheless, speech signal analysis and human speech perception are not entirely unrelated.

Sections 1.2 and 1.3 are brief, introductory surveys describing clipped speech phenomena and the use of zero crossings as waveform descriptors, respectively. These ideas provide the motivation for this thesis and they will be expanded in later chapters.

Infinite clipping of speech results in a harsh sounding, but highly intelligible, acoustic signal. This phenomenon was first noted in 1947 by Licklider, Bindra and Pollack [L-13] who, in an investigation of questions related to the information carrying characteristics of speech, performed a classic set of experiments using clipping as a distorting operator on the speech waveform.

They found that removal of all amplitude information, except polarity, above one-tenth of peak waveform level resulted in discrete word articulation scores of 96% or more. Further elimination of amplitude information until the waveform was defined entirely by the times of polarity reversals (zero crossings) reduced the word articulation scores to an average of 70%; although for some listeners this score was as low as 50%, conversation could be carried on with little difficulty. Pre-clipping elimination of low frequency speech spectral components improved post-clipping intelligibility. Other tests, conducted with clipped and normal speech equal in peak amplitude and heard against a background of spectrally flat ('white') noise, demonstrated that for low speech to noise peak amplitude ratios the clipped speech was *more* intelligible than the original speech signal. The subjects' ability to understand clipped speech improved during the course of the experiments and the above scores are the maxima noted.

In another series of experiments [L-14], Licklider and Pollack examined the effects of pre- and post-clipping spectral tilting (differentiation and integration) on the intelligibility of the infinitely clipped speech signal. The figure below (from [L-14]) graphically describes the effects on word articulation of the various combinations of spectral manipulations.

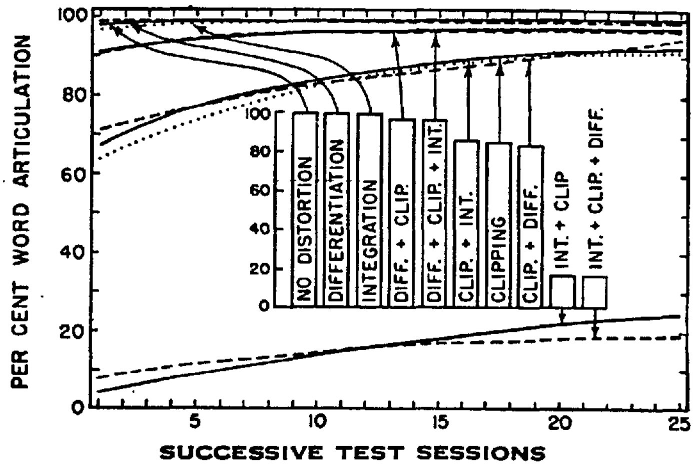


Fig. 1.1: The effects of various combinations of differentiation, integration and infinite clipping upon word articulation. The heights of the bars of the column diagram indicate the overall average for each of the ten arrangements. (From [L-14]).

Pre-clipping differentiation (6 db per octave positive spectral tilt) of the speech signal significantly improved the intelligibility of the clipped waveform while pre-clipping integration (6 db per octave negative spectral tilt) was severely deleterious under the same conditions. Post-clipping integration or differentiation produced only minor changes in per cent word articulation; however, the former operation lessened the subjective harshness of the clipped waveform while the latter operation accentuated it. Again, articulation scores improved with experience.

Finally, Licklider [L-15] showed that quantization of the times of zero crossings to the nearest 'x' milliseconds produced virtually unintelligible clipped speech if 'x' was greater than 0.2 milliseconds.

1.3 Objective Estimation of Speech Parameters

Automatic recognition--classification-- of speech sounds has been a primary research target for over twenty-five years.

The first step in machine recognition of speech usually involves a condensation of data so as to exclude "non-essential" information and preserve "invariant", or essential, data. The question as to what is "essential" for objective sound classification is central to the entire speech recognition problem.

For example, as we shall see, spectral features of certain speech sounds (principally vowels) are prominent and to some extent can characterize the sound; hence short-time estimates of amplitude spectra have often served as input data to speech recognition machines. Certain properties of zero crossing intervals and distributions may, after manipulation, yield an estimate of spectral parameters. In addition, histograms of zero crossing intervals have been found to possess prominent 'speaker invariant' features [B-5].

For these reasons, and perhaps due to the simplified hardware used for binary data processing, the infinitely clipped waveform (possessing only zero crossing information) has frequently replaced the original waveform as a data source to the primary feature extractor of speech recognition automata. We shall examine the implications of the use of zero crossing interval sequences as waveform descriptors and the significance of zero crossings as informational attributes of the original signal.

1.4 Unanswered Questions

Zero crossing interval sequences, evidently, carry sufficient information to construct a highly intelligible speech signal. They may also afford estimates of speech spectral features or, as first-order histograms, be regarded as distinctive attributes portraying the sound source. Yet, the exact significance of zero crossings as a representation of the original speech signal has been unclear. Good has conjectured, for example, that under certain circumstances zero crossings may completely specify, or in some cases overspecify, a signal source [G-9]. Finally, no convincing answer has been proferred to the question, "Is clipped speech intelligible because the original signal was speech, or because clipping is a special type of transformation, or are the two considerations inseparable?"

1.5 Zeros as Signal Descriptors: an Approach to the Role of Zero Crossings in Speech Recognition and Processing

In 1966 H.B. Voelcker showed formally [V-6] that zeros can be regarded as *complete* descriptions of bandlimited signal waveforms with the proviso that covert, or complex, zeros be included with the real zeros, or zero crossings, in the set of signal descriptors. He employed Analytic signal theory and zero-based concepts to unify many principles in the field of modulation theory.

We shall apply these ideas, amongst others in this thesis, to explore the role of zero crossings in speech processing and recognition. Specifically, we shall focus on the problem of accounting realistically for the high intelligibility of clipped speech, and of justifying and explaining the use of zero crossings as both an estimate of speech spectral features and a description of the waveform itself. We also describe two short experiments, carried out during the course of this research, concerning the computer implementation of limited vocabulary, zero crossing input speech recognition machines.

1.6 Organization of the Thesis

We conclude the introduction with a description of the thesis organization, by chapters.

2: This thesis is cast mainly in the language of the telecommunication engineer, but it should be useful to psychologists, physiologists and others concerned with speech phenomena. Therefore, in *chapter 2*, we briefly review the signal theory which provides the mathematical basis of the entire thesis.

3: *Chapter 3* is a survey of certain theories and experimental evidence which provide the necessary background for studies of speech and hearing. In particular, we examine the physiological and psychological aspects of theories of hearing, and the acoustic properties of speech sounds. Since we shall build a theory of post-clipping speech intelligibility upon a foundation of speech spectral characteristics, we examine the problem of whether static (time invariant) spectral information is sufficient for *human* recognition (perception) without such cues as transitions or context. In addition, we argue that accurate extraction of spectral parameters is not quite as straightforward as often implied.

4: *Chapter 4* is devoted to preliminary studies of *machine* recognition (classification) of speech sounds. We outline specific problems relevant to the implementation of automatic speech recognition machines. Brief descriptions of schemes using spectral information directly as input to the recognition machine are presented.

5: Psychoacoustic phenomena associated with audition of infinitely clipped speech are reviewed in detail in the first section of *chapter 5*. Attempts to justify analytically the intelligibility of clipped speech are then described and critically evaluated.

6: Zero crossings *per se* can be viewed as informational attributes of a signal. *Chapter 6* briefly outlines current knowledge concerning the statistics of zero crossings of random processes. Then, the use of zero crossings as an estimate of spectral parameters in speech signals is detailed. Single sideband modulation as a transformation affecting the zero crossings of the speech signal is described, and the effects on subsequent extraction of spectral parameters are noted. The chapter is terminated by a comprehensive review of automatic speech recognition schemes based on zero crossings as input parameters.

7: *Chapter 7* is a description of two experiments in machine recognition of speech carried out by the author. Both experiments relied upon zero crossing information as source data. The results of the experiments are discussed, together with the conclusions which resulted in the theoretical and experimental investigation of zero crossings as signal descriptors which constitutes the remainder of the thesis.

8: In *chapter 8* we elaborate upon a specific, quite general, zero-based signal model. We then apply zero-based concepts to speech signal models to construct a foundation, both theoretical and experimental, for certain postulates and conjectures concerning *clipped speech phenomena* and *zero crossings as waveform descriptors*.

9: *Chapter 9* explores the phenomena associated with *speech clipping* from a zero-based viewpoint. We discuss product formulations for the original and clipped waveforms and examine the relationship between low-pass and single sideband clipped speech. In conclusion, the effect of clipping on a signal's zeros, and hence its spectrum, is analyzed with some reference to critical band theories of hearing.

10: In *chapter 10* we examine the sufficiency of *real zeros as waveform descriptors*, and the relevance of this idea to the use of zero crossings as input to speech recognition machines. Methods of signal processing which ensure that the zero crossings almost completely describe the original signal are consolidated.

11: *Chapter 11* is dedicated to a summary of ideas developed throughout the thesis, a description of outstanding problems, and recommendations for further research.

2 TIME-FREQUENCY ANALYSIS

In the first five sections of this chapter we outline some of the basic analytical concepts of signal theory which have been adopted over the last 50 years as the primary tools of communication theory. The basis of these concepts is time-frequency, or Fourier, analysis.

Gabor, in a discussion of the physical significance of Fourier analysis methods, noted [G-1] that "if the word frequency is used in the strict mathematical sense which applies only to infinite duration wave trains, a changing frequency becomes a contradiction in terms as it is a statement involving time *and* frequency." That is, "Fourier's theorem makes of description in time and description in frequency two mutually exclusive methods." In order to resolve this anomaly, Gabor presented "a new method of analyzing signals in which time and frequency play symmetrical parts, and which contains 'time analysis' and 'frequency analysis' as special cases." Section 2.6 is devoted, therefore, to an outline of theories, including Gabor's, on the *interrelationship* of time and frequency in signal analysis.

Finally, we conclude the chapter by qualifying the use of Fourier methods in the study of psychoacoustic phenomena. We defer discussion of applications of time-frequency plane analysis in speech and hearing to chapter 3.

2.1 Fourier Series: Periodic Signals

Fourier series arise when the problem of describing a time function s(t) on an interval [0,T] is considered.

The general series expansion

$$s_e(t) = \sum_{k=1}^{N} s_k g_k(t) \ , \ 0 \leq t \leq T \qquad (2\text{-}1)$$

involves N coefficients $\{s_k\}$ which depend only upon s(t) and are not functions of time [S-3, p. 9]. The N functions of time, $\{g_k(t)\}$, are specified independently of s(t) and $s_e(t)$ is an approximation to s(t). In order to minimize the mean square error between s(t) and $s_e(t)$ for a given N, and have this error approach zero as N increases, for any finite energy signal

$$\text{i.e.} \quad \int_0^T |s(t)|^2 dt < \infty \ ,$$

it is necessary that [V-1, p. 170], [S-3, p. 12]

$$s_k = \int_0^T s(t) \ g^*_k(t) \ dt \quad . \qquad (2\text{-}2)$$

If the functions $g_k(t)$ are chosen so that

$$\int_0^T g_j(t) g^*_k(t) \ dt = \delta_{jk} = \begin{cases} 1 & j=k \\ 0 & j \neq k \end{cases} , \qquad (2\text{-}3)$$

they are orthonormal [S-3, p. 10]. δ_{jk} is the *Kronecker delta.*

The standard Fourier series form for signals periodic in T arises if one chooses

$$g_k(t) = \begin{cases} \cos(\frac{k-1}{2} \Omega t) & k \text{ odd} \\ & \qquad k=1,2,\ldots \\ \sin(\frac{k}{2} \Omega t) & k \text{ even} \end{cases} , \ \Omega = 2\pi/T. \qquad (2\text{-}4)$$

Since these g's are a complete set [S-3, p. 13], then over the interval $[0,T]$, $s_e(t) = s(t)$ in the sense that there is no energy in the error $\{s(t)-s_e(t)\}$, for $N = \infty$ in (2-1). Then

$$s(t) = a_0/2 + \sum_{k=1}^{\infty} (a_k \cos k\Omega t + b_k \sin k\Omega t), \quad 0 \leq t \leq T, \quad (2\text{-}5)$$

which, using Euler's identities, yields the complex form

$$s(t) = \sum_{k=-\infty}^{\infty} c_k \, e^{+jk\Omega t} \quad . \text{ [S-3, pp. 15-16]} \qquad (2\text{-}6)$$

Here

$$c_k = \frac{1}{T} \int_0^T s(t) \, e^{-jk\Omega t} \, dt \, . \qquad (2\text{-}7)$$

Note that c_k can also be written in the form

$$c_k = |c_k| \cdot e^{j\theta_k} \qquad (2\text{-}8)$$

where $|c_k| = \frac{1}{2}[a_k^2 + b_k^2]^{\frac{1}{2}}$ and $\theta_k = \tan^{-1}[-b_k/a_k]$. (2-9)
It follows that (2-5) can be expressed in the alternate form

$$s(t) = a_0/2 + 2 \sum_{k=1}^{\infty} |c_k| \cdot \cos(k\Omega t + \theta_k) \, . \qquad (2\text{-}5b)$$

$|c_k|$, θ_k, and c_k^2 represent, respectively, the amplitude, phase, and power of the k^{th} frequency (spectral) component of $s(t)$ [L-6].

2.2 The Fourier Transform: Aperiodic Signals

The periodicity, with T, of $e^{jk\Omega t}$ ensures that $s(t) = s(t+T)$ in (2-6). As noted in sec. 2.1, a periodic signal has a discrete line structure in the frequency domain. If the period $T \to \infty$, then the signal $s(t)$ becomes *aperiodic* and the spectral line spacing $\Delta f = \Omega/2\pi = 1/T$ tends to zero. That is, when $\Delta f \to 0$, $k\Delta f \to f$, a continuous independent variate.

Therefore, from (2-7)

$$\lim_{\substack{T \to \infty \\ k\Delta f \to f}} c_k \to c(f) \equiv S(f) = \int_{-\infty}^{\infty} s(t)\, e^{-j2\pi ft} dt \qquad (2\text{-}10a)$$

and

$$s(t) = \int_{-\infty}^{\infty} S(f)\, e^{\,j2\pi ft} df \qquad (2\text{-}10b)$$

so that s(t) and S(f) are a Fourier transform pair with (2-10) defining the members. That is,

$$s(t) \leftrightarrow S(f) \quad . \qquad (2\text{-}11)$$

The preceding approach through limits, while intuitively appealing, is not rigorous. In using the limiting conditions one does not define the conditions which are necessary for the existence and validity of (2-10a) and (2-10b) [P-2, p. 2]. In fact, satisfaction of either of the following restrictions on s(t) is the most important factor in assuring that S(f) exists and satisfies (2-11):

1. $\int_{-\infty}^{\infty} |s(t)|\, dt < \infty$[P-2, p. 9] (2-12)

2. $\int_{-\infty}^{\infty} |s(t)|^2 dt$ [S-3, p. 31] (2-13)

Hence, an alternative is to define the Fourier transform pair with their associated existence conditions and from them derive the Fourier series [P-2, pp. 42-45], [B-16, pp. 204-208].

We shall use the notation $F\{\ \}$ to signify the operation of Fourier transformation.

2.3 The Analytic Signal

If s(t) is a real, aperiodic signal then the real and imaginary parts of the complex spectrum S(f) are given by

$$S_R(f) = \int_{-\infty}^{\infty} s(t)\cdot\cos\omega t\; dt \tag{2-14}$$

and

$$S_I(f) = \int_{-\infty}^{\infty} s(t)\cdot\sin\omega t\; dt \tag{2-15}$$

respectively, where $\omega = 2\pi f$. Consequently, S(f) has real, even, imaginary, odd, symmetry about f=0. Thus, given S(f) for f>0, S(f) for f<0 can be defined by conjugation.

2.3.1 Definitions

For convenience, we can *define* a signal m(t) having a single-sided spectrum M(f) such that

$$M(f) = \begin{cases} 2S(f) & , \quad f > 0 \\ S(0) & , \quad f = 0 \\ 0 & , \quad f < 0\,. \end{cases} \tag{2-16}$$

It follows, (using Woodward's operational notation, [W-9]) that

$$M(f) = \lim_{W\to\infty} 2S(f)\cdot\mathrm{rect}[(f-W/2)/W] \quad . \tag{2-17}$$

Taking Fourier transforms, and using the "product-convolution" relationship [S-3, p. 45], one obtains

$$m(t) = s(t) + j\; s(t) * \frac{1}{\pi t} \quad , \quad [V\text{-}7] \tag{2-18}$$

where $s(t) * \frac{1}{\pi t} \equiv \hat{s}(t)$ is the *Hilbert transform* of s(t). (2-19)

That is, $$\mathcal{H}\{s(t)\} = \hat{s}(t) = P[\ \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{s(\tau)}{t-\tau}\, d\tau\ ], \tag{2-20}$$

where P[ ] denotes the Cauchy principle value [G-1].

The function

$$m(t) = s(t) + j\ \hat{s}(t) \tag{2-21}$$

is termed the *analytic signal* representation [P-1].

2.3.2 Hilbert Transformers

In *principle*, since the definition of Hilbert transformation involves a convolution, a Hilbert transformer could be realized by a linear, time invariant network with impulse response

$$h_H(t) = 1/\pi t\ . \tag{2-23}$$

Such a network would have a frequency response given by

$$H_H(f) = \mathcal{F}\{h_H(t)\} = -j\ \mathrm{sgn}[f]\ , \tag{2-24}$$

where $$\mathrm{sgn}[x] = \begin{cases} 1\ , & x > 0 \\ 0\ , & x = 0 \\ -1\ , & x < 0 \end{cases}\ .$$

This network does not affect spectral amplitudes but causes a phase shift of -90° or +90° for positive or negative frequencies, respectively.

In *practice*, such a network is unrealizeable because $h_H(t)$ is non-causal and undefined at t=0. In addition, $H_H(f)$ has infinite bandwidth. Implementations and limitations of Hilbert transformers are discussed by Gouriet and Newell [G-11], and by Voelcker [V-8].

2.3.3 Phase-Envelope Models

Since m(t) is a complex signal, it can be represented in the form

$$m(t) = |m(t)| \cdot e^{j\phi(t)} \quad , \qquad (2\text{-}25)$$

where $$|m(t)| = \sqrt{s(t)^2 + \hat{s}(t)^2} \qquad (2\text{-}26)$$

is the *envelope* of s(t)

and $$\phi(t) = \tan^{-1}[\hat{s}(t)/s(t)] \qquad (2\text{-}27)$$

is the *phase function* of s(t) [D-15].

The real part of the analytic signal, s(t), can be expressed in the form

$$s(t) = |m(t)| \cdot \cos\, \phi(t) \quad , \qquad (2\text{-}28)$$

the phase-envelope formulation for a bandlimited signal.

If a positive frequency translation, $f_o = \omega_o/2\pi$, is applied to m(t), then

$$m_{\omega_o}(t) = m(t) \cdot e^{j\omega_o t} \quad , \qquad (2\text{-}29)$$

and the real part of $m_{\omega_o}(t)$ is

$$\begin{aligned} s_{\omega_o}(t) &= s(t) \cdot \cos\, \omega_o t - \hat{s}(t) \cdot \sin\, \omega_o t \\ &= |m(t)| \cdot \cos[\omega_o t + \phi(t)] \quad . \qquad (2\text{-}30) \end{aligned}$$

This, a model for a real *single sideband* signal (upper sideband form), differs from (2-28), s(t), only in the addition of $\omega_o t$, the frequency translator. It follows that the phase and envelope of the signal are analytic signal attributes which are not affected by frequency translations and (2-29) is a suitable model for studying such processes. Phase-envelope relationships in speech signals will be discussed in sec. 6.4.

2.4 Sampling Theory

In 1928 Nyquist demonstrated that the number of "signal elements" (i.e., telegraphic 'dots') which can be transmitted per unit time over a bandlimited line is a function of the bandwidth[N-5]. Eighteen years later Gabor stated as the fundamental theorem of communications that [G-1]: "In whatever ways we select N data to specify a signal in the interval τ, we cannot transmit more than a number $2(f_2-f_1)\tau$ of these data, or of their independent combinations by means of the $2(f_2-f_1)$ independent Fourier coefficients." Here f_1 and f_2 were the limits of the frequency range in which the bandpass signal was to be defined. Gabor's proof was based on Fourier series expansions and he noted that "*it leaves a sense of dissatisfaction.*"(Italics mine.)

In the next three sections we briefly review the fundamental concepts of sampling theory in order to provide a framework for our work on specification via zeros. These ideas constitute a development and rigorization of Gabor's "fundamental theorem."

2.4.1 Lowpass Sampling

A conventional approach to lowpass sampling is via Fourier transform theory, again referring to Woodward [W-9].

The sampled version of s(t) can be represented as

$$\tilde{s}(t) = \mathrm{comb}_{\mathfrak{T}} s(t) = \sum_{n=-\infty}^{\infty} s(t)\cdot\delta(t-n\mathfrak{T}) \qquad \text{(2-32a)}$$

where $\mathfrak{T}$ is the sampling interval in seconds. The Fourier transform of $\tilde{s}(t)$ is

$$\tilde{S}(f) = \frac{1}{\mathfrak{T}}\, \mathrm{rep}_{1/\mathfrak{T}}\, S(f) = \frac{1}{\mathfrak{T}} \sum_{n=-\infty}^{\infty} S(f-n/\mathfrak{T}). \text{ [W-9]} \qquad \text{(2-32b)}$$

If $S(f) = 0$ for $|f| > W$, and if $\text{Ŧ} < 1/2W$, then $S(f)$ can be recovered from $\tilde{S}(f)$ by filtering since the repeated $S(f)$'s which constitute $\tilde{S}(f)$ will not overlap. That is,

$$S(f) = \mathrm{rect}(f/2W)\cdot \mathrm{rep}_{1/\text{Ŧ}}S(f), \quad \text{Ŧ}<1/2W \ . \tag{2-33}$$

Using the convolution-product theorem, and taking Fourier transforms of both sides of (2-33) one obtains,

$$s(t) = \sum_{n=-\infty}^{\infty} s(n\text{Ŧ})\cdot \sin[2\pi W(t-n\text{Ŧ})]/2\pi W(t-n\text{Ŧ}) \ . \tag{2-34}$$

Or, using $\mathrm{sinc}\ x = \sin\pi x/\pi x$,

$$s(t) = \sum_{n=-\infty}^{\infty} s(n\text{Ŧ})\cdot \mathrm{sinc}\ 2W(t-n\text{Ŧ}) \ . \tag{2-35}$$

Hence $s(t)$ can be completely recovered from (an infinite number of) its samples, taken every Ŧ seconds, by interpolation with sinc functions [W-9], [K-10].

If $s(t)$ is periodic in T, as well as bandlimited to $\pm W$ Hz, then

$$s(t) = \sum_{n=0}^{n_1-1} s(nT/n_1)\cdot \frac{\sin\,[\pi(n-n_1 t/T)]}{n_1\cdot \sin\,[\pi(n/n_1-t/T)]} \ , \tag{2-36}$$

where $n_1 = 2WT-1$ [G-8]. Thus, a periodic signal having a finite number of Fourier coefficients requires only a finite number of samples for complete determination. In sec. 2.4.3 we will show that these samples need not be taken at uniform time intervals.

2.4.2 Bandpass Sampling

If the signal spectrum occupies the band $f_o \leq |f| \leq f_o+W$ then only in special circumstances (i.e., when $f_o=cW$, $c=0,1,\ldots$) is it possible to reconstruct s(t) from its samples at 2W equispaced points per second. Generally, a minimum uniform sampling rate R_{min} --where $2W \leq R_{min} \leq 4W$--is required. The actual value of R_{min} depends upon the relationship of f_o and W.

Second order sampling, which involves two interlaced sequences of W equispaced sampling points per second, may be used but the interpolation functions corresponding to this mode of sampling are quite complicated [K-10], [L-17].

However, uniform sampling of a bandpass signal and its Hilbert transform, at a rate $\geq$ W times per second, suffice to uniquely determine that signal [L-17].

2.4.3 Noniniform Sampling

J. L. Yen considered the problem of nonuniform sampling of lowpass signals. He showed [Y-1] that if the signal s(t) is band-limited to ±W Hz, then it is uniquely determined by (and can therefore be completely reconstructed from) its values at recurrent sets of N sample points taken at

$$\tau_{pm} = t_p + mN/2W, \qquad p = 1,2,\dots.N$$
$$m = \dots,-2,-1,0,1,2,\dots$$

That is,

$$s(t) = \sum_{m=-\infty}^{\infty} \cdot \sum_{p=1}^{N} s(\tau_{pm}) \cdot \Psi_{pm}(t) \ , \tag{2-38}$$

where

$$\Psi_{pm}(t) = \frac{\prod_{q=1}^{N} \sin \frac{\Omega}{2}(t-t_q) \cdot (-1)^{mN}}{\prod_{\substack{q=1 \\ \neq p}}^{N} \sin \frac{\Omega}{2}(t_p-t_q) \cdot \frac{\Omega}{2}(t-t_p-mN/2W)} \tag{2-39}$$

is the interpolating function. If s(t) is also *periodic* in T=N/2W, then, $s(\tau_{pm}) = s(t_p)$ for all m and only one set of N nonuniform samples is required for complete signal determination.

2.4.4 Uniform vs Nonuniform Sampling

A major difference between the interpolating function for uniform and nonuniform sampling of bandlimited signals should be emphasized.

For uniform sampling the maximum value of the *sinc* interpolating function occurs at the sample point and this value is unity. For nonuniform sampling, however, the maximum value of the interpolating function $\Psi_{pm}(t)$ does not necessarily occur at the sample point. While the value of the interpolating function at its particular sample point *is* unity, its maximum value may become very large due to bunching of sampling points [Y-1].

We shall examine the phenomenon of signal growth due to "bunching of sampling points" in chapter 9.

2.5 Finite Sample Sets: the Discrete Fourier Transform

Signal analysis using the digital computer as a tool--either as a sophisticated calculator or as a simulator of a communication system--requires that all signals be both *sampled* and *quantized*; that is, defined only at specific instants of time or values of frequency and specified only to some finite degree of accuracy.

As discussed in sec. 2.4, bandlimited signals may be completely specified by sampling at uniform rates exceeding twice the highest frequency present in the waveform. However, quantization implies introduction of noise. Quantization error is analyzed by Gold and Rader [G-4, ch. 4], Papoulis [P-4], and Widrow [W-7].

We are concerned primarily with the properties of the transform pair which apply to signals represented by finite sets of discrete samples in both the time and frequency domain. In the next section, therefore, we briefly describe the *discrete Fourier transform*--DFT--for sampled signals and in sec. 2.5.2 we discuss some of its properties.

2.5.1 Formulation of the Discrete Fourier Transform

If a continuous signal x(t) is sampled every $\mp$ seconds, the sampled signal $\tilde{x}(t)$ can be represented as

$$\tilde{x}(t) = \sum_{n=-\infty}^{\infty} x(n\mp)\cdot\delta(t-n\mp) \quad . \tag{2-40}$$

If $\tilde{x}(t)$ is defined only for $t \geqslant 0$ and we consider only a finite number of samples--N--, then, taking Fourier transforms of both sides of (2-40),

$$\tilde{X}(f) = \sum_{n=0}^{N-1} x(n\mp)\cdot e^{-j2\pi fn\mp} \quad . \tag{2-41}$$

Since $e^{-j2\pi fn\mp}$ is a periodic function of f, the sampling operator has "folded" the frequency axis so that frequencies greater than $1/2\mp$ Hz are discriminated only as *aliases* of themselves. Therefore, it is imperative for accurate sampled signal representation that s(t) be effectively bandlimited to $\pm W$ Hz, where $W = 1/2\mp$ [B-9, pp. 31-33].

We evaluate (2-41) at equispaced intervals of Ω Hz. That is, let $f = k\Omega/2\pi$ where, by definition,

$$\Omega \equiv 2\pi/N\mp = 2\pi(W/0.5N) \quad . \tag{2-42}$$

Then, from (2-41),

$$\tilde{X}(k\Omega/2\pi) \equiv \tilde{X}(k) = \sum_{n=0}^{N-1} x(n\mp)\cdot e^{-j\Omega nk} \quad . \tag{2-43}$$

It can be shown that $\tilde{X}(k) = \tilde{X}(k+pN)$, p an integer [G-4, p. 163]. Therefore (2-43) yields a periodic sequence of complex numbers with period N.

Letting $\tilde{X}(k) \rightarrow X(k)$ and $x(n\mp) \rightarrow x(n)/N$, then, using (2-42) in (2-43) we obtain the DFT of the time sequence x(n):

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \cdot e^{-j2\pi nk/N}, \quad k=0,1,\ldots N-1. \tag{2-44}$$

The inverse discrete Fourier transform maps the {X(k)} back into the {x(n)} and is given by

$$x(n) = \sum_{k=0}^{N-1} X(k) \cdot e^{\,j2\pi nk/N}, \quad n=0,1,\ldots N-1. \tag{2-45}$$

That (2-45) is the inverse of (2-44) can be shown by substitution [G-4, p. 165]. Equations (2-44) and (2-45) are the *discrete Fourier transform pair*. That is,

$$\{x(n)\} \leftrightarrow \{X(k)\} \ .$$

2.5.2 Nature of the Discrete Fourier Transform

Summarizing, the discrete Fourier transform of a finite sampled signal {x(n)} is a finite sampled complex spectral series {X(k)}. Both series are periodic in N in their respective domains, due to the cyclic nature of $e^{j2\pi nk/N}$.

For {x(n)} real and N even, X(N/2) is real and represents the amplitude of the real part of the highest frequency component of {x(n)}, while X(0) represents the average value of the sampled time function. {X(k)} possesses real even, imaginary odd symmetry about X(N/2) with positive frequency complex Fourier coefficients indexed by k=1,2,...N/2-1, increasing in frequency with increasing k and the negative frequency complex Fourier coefficients indexed by k=N/2+1, ... N-1, decreasing in frequency with increasing k.

<u>Spectral Component</u> (·) <u>k</u> (•)

·	$X(0)$	$X(\Omega)$	$X(2\Omega)$	...	$X([N/2-1]\Omega)$	$X(N\Omega/2)$	$X^*([N/2-1]\Omega)$	...
•	0	1	2		$N/2-1$	$N/2$	$N/2+1$	

.....	$X^*(2\Omega)$	$X^*(\Omega)$	·
	$N-2$	$N-1$	•

Fig. 2.1 Discrete Fourier transform output array.

The $\{X(k)\}$ can also be regarded as the output at time $(N-1)\mp$ of a linear digital filter whose unit sample response is

$$h_k(n\mp) = \begin{cases} [e^{-j2\pi/N}]^{(N-1-n)}, & n=0,1, \ldots N-1 \\ 0 , & \text{otherwise} , \end{cases} \qquad (2\text{-}46)$$

and whose input is the sequence

$$..0,0,....0,x(0),x(\mp),x(2\mp),...x([N-1]\mp),0,......... \qquad [B\text{-}22] .$$

From (2-41), the Fourier transform of $h_k(n\mp)$ is

$$\tilde{H}_k(f) = \sum_{n=0}^{N-1} [e^{-j2\pi/N}]^{(N-1-n)k} \cdot e^{-j2\pi fn\mp} . \qquad (2\text{-}47)$$

Letting $e^{-j2\pi f\mp} = z$, then, following some manipulation,

$$\tilde{H}_k(f) \equiv \tilde{H}_k(z) = \frac{z^{-N}-1}{z^{-1}-[e^{-j2\pi/N}]^k}, \qquad (2\text{-}48)$$

which has N zeros located at $z_m = e^{j2\pi m/N}$, m=0,1, ... N-1

and one pole at $z = e^{j2\pi k/N}$

which cancels the k^{th} zero.

Then, evaluating (2-48) as a function of f, with $f_s = 1/T = 2W$,

$$|\tilde{H}_k(f)| = |\tilde{H}_k(z=e^{j2\pi f/f_s})| = |\sin(\pi N f/f_s)/\sin[\pi(f/f_s-k/N)]|. \qquad (2\text{-}49)$$

$|\tilde{H}_k(f)|$ is shown in Fig. 2.2 for N=8 and k=0.

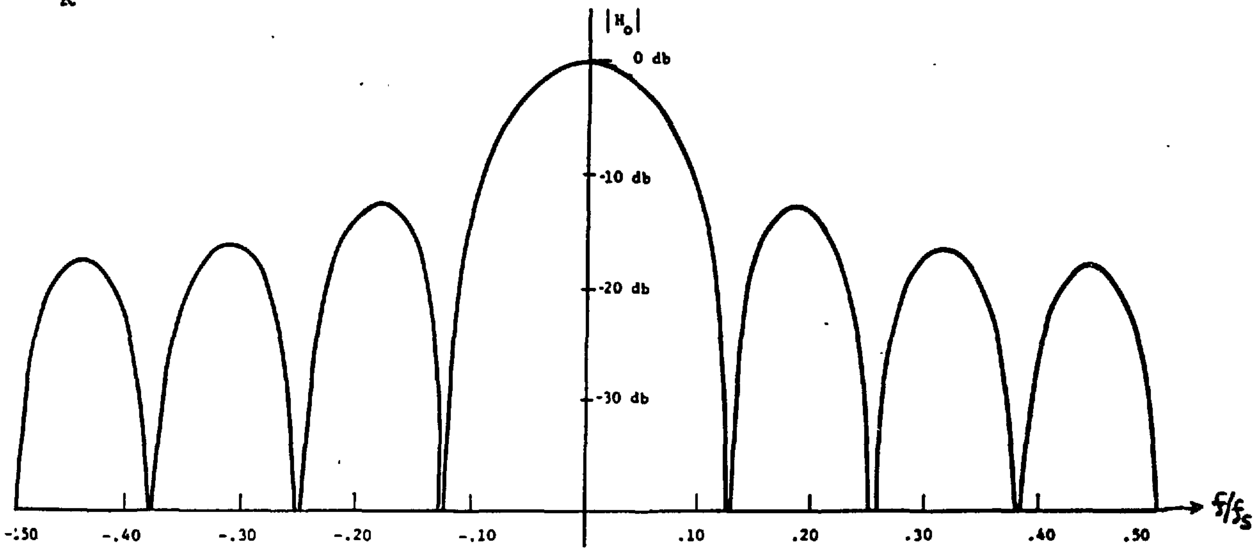


Fig. 2.2 $|\tilde{H}_k(f/f_s)|$ for N=8 and k=0 . (From [B-22].)

Therefore, the discrete Fourier transform corresponds to filtering the input signal $x(n\mathrm{T})$ with N filters having center frequencies $f_c = (W/N)[2k-N]$, $k=0, 1, \ldots N-1$, and frequency responses of the form $\sin Nx/\sin x$. The outputs of the filters at time $(N-1)\mathrm{T}$ are the Fourier coefficients [B-22]. Figures 2.3 and 2.4 illustrate the spectral distortion introduced by time and frequency sampling of aperiodic and periodic signals, respectively, and by truncation of aperiodic signals.

Direct evaluation of the N complex frequency coefficients $\{X(k)\}$ requires a number of operations (complex additions and multiplications) proportional to N^2. The Fast Fourier transform, or FFT [G-4, pp. 173-201], [M-14], enables computation of the DFT in a number of operations proportional to $N \log_2 N$ if $N = 2^M$, M a positive integer. Much of the computer analysis of speech waveforms described in chapter 9 was made economically feasible by using the FFT to evaluate the DFT. We postpone description of some uses of the FFT algorithm until section 8.5.1.

2.6 Energy Distribution in the Time-Frequency Plane

The time and frequency descriptions of signals can be represented by orthogonal coordinates on a time-frequency plane [G-1]. A continuous sine wave, for example, exists for all time and is represented on the positive frequency axis by a delta function at its frequency of oscillation; conversely, a time domain delta function exists for a vanishingly short time but has equal energy at all frequencies. Gabor suggested that the problem of describing the frequency spectrum of a truncated sine wave be resolved by reference to the response to such a waveform of a physical system, a bank of tuned reeds, for instance. Such systems, he proposed, divide the time-frequency plane into approximately rectangular areas whose

APERIODIC SIGNALS

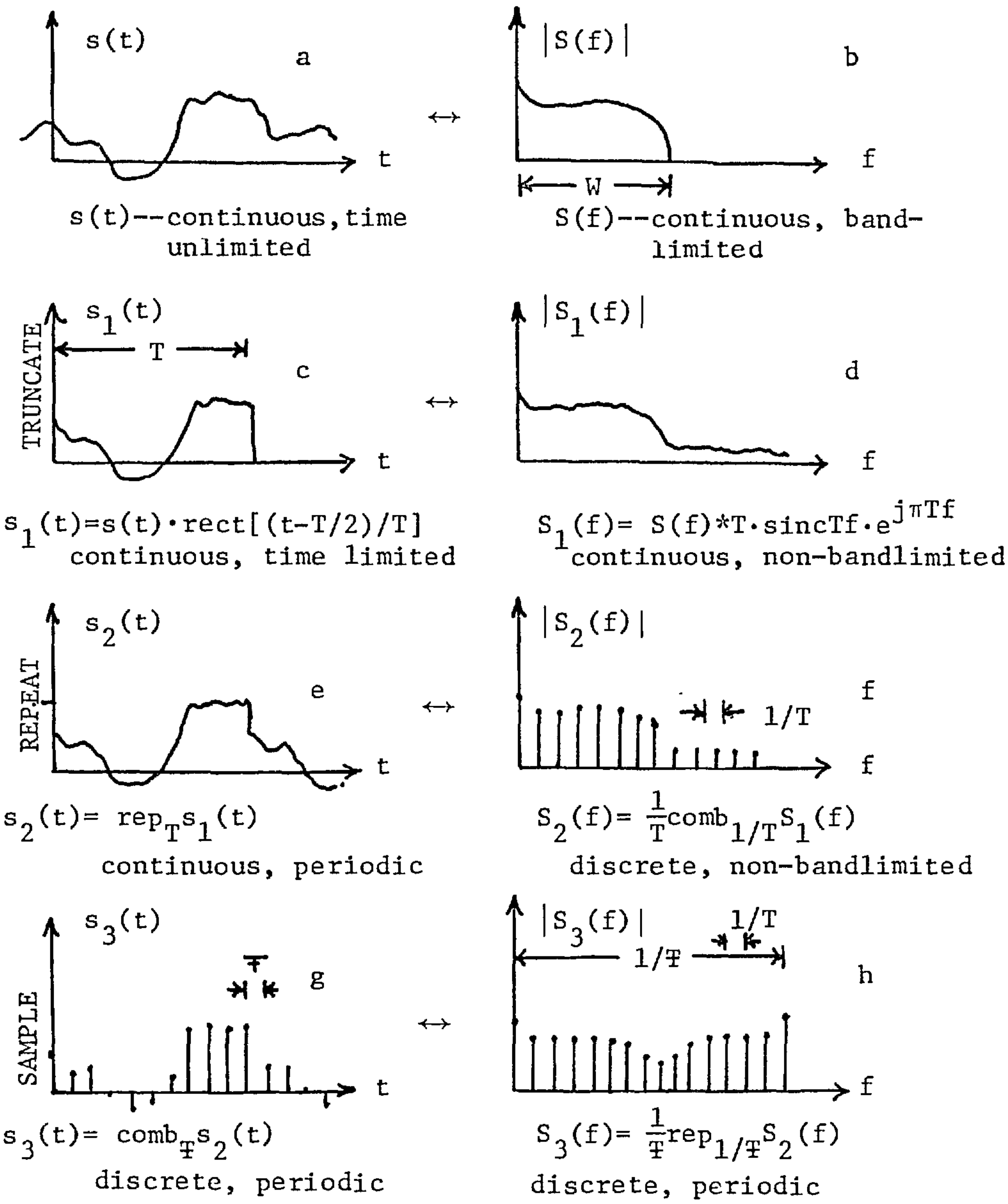


Fig. 2.3 The errors introduced by time and frequency sampling of aperiodic signals.

i) Truncation implies spectral smearing via convolution (c,d).

ii) Sampling in time domain causes spectral overlap of non-bandlimited spectra (g,h).

iii) Analogue to digital conversion causes quantization error.

Note: Only positive spectral frequencies are shown.

PERIODIC SIGNALS

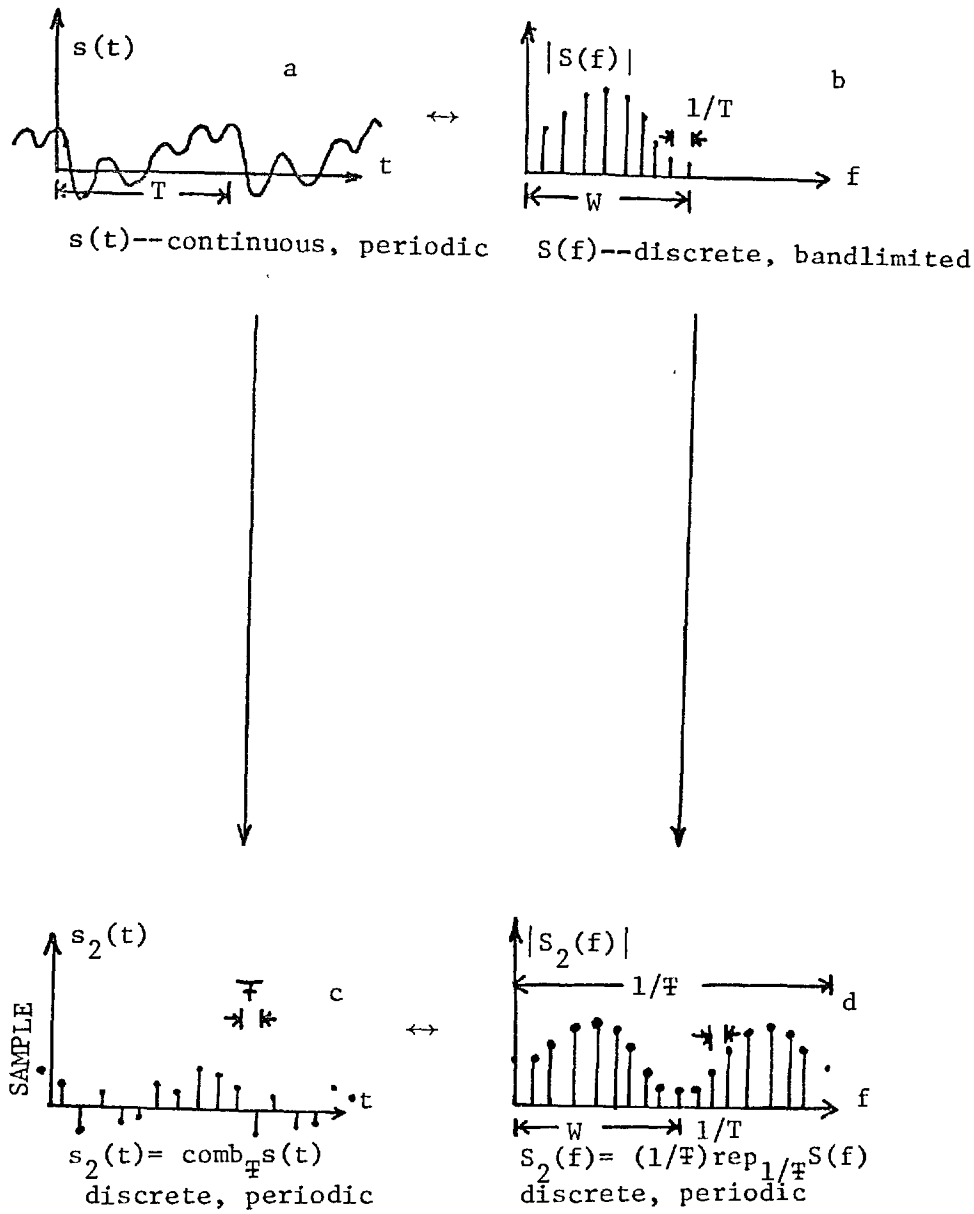


Fig. 2.4 The errors introduced by time and frequency sampling of periodic signals

i) Sampling in time domain implies no spectral overlap if $1/Ŧ > 2W$.

ii) Analogue to digital conversion causes quantization error.

Note: Only positive spectral frequencies are shown.

shapes are dependant upon the nature of the system, with the restriction that no more than $2(f_2-f_1)\tau$ independant data can be obtained from the 'occupied' area, $(f_2-f_1)x\tau$, of the plane.

He argued that by making a function of time *or* frequency a function of both time *and* frequency an arbitrarily exact analysis with respect to either, but not both, of the variables could be made. The product of the 'uncertainty of measurement' in time and frequency is [G-1]

$$\Delta t \cdot \Delta f \geqslant \tfrac{1}{2} , \qquad (2\text{-}50)$$

where $\Delta t = \sqrt{2\pi} \cdot D_t \quad , \quad \Delta f = \sqrt{2\pi} \cdot D_f \quad .$

Here $$D_t = [\overline{(t-\bar{t})^2}]^{\frac{1}{2}} , \qquad (2\text{-}51)$$

and $$D_f = [\overline{(f-\bar{f})^2}]^{\frac{1}{2}} \qquad (2\text{-}52)$$

are the rms deviation of t or f from the mean epoch, $\bar{t}$, or frequency, $\bar{f}$, of a signal. Equation (2-50), rather than expressing a true 'uncertainty' effectively places bounds on the 'duration' of a signal and the bandwidth of its Fourier transform. The definition of 'duration' and 'bandwidth' is usually dependent upon the nature of the signal being studied [P-2, pp. 62-74].

Gabor found that the signal which makes (2-50) an identity is

$$s(t) = \text{Re}\ [e^{-\alpha^2(t-t_o)^2} \cdot e^{j(\omega_o t+\phi)}] , \qquad (2\text{-}53)$$

a sinusoidal signal with a Gaussian [normal] shaped envelope and Fourier transform

$$S(f) = e^{-(\pi/\alpha)^2(f-f_o)^2} \cdot e^{-j[2\pi t_o(f-f_o)+\phi]} . \qquad (2\text{-}54)$$

α, Δt, and Δf are related by

$$\Delta t = \sqrt{\pi/2}\ /\alpha \quad \text{and} \quad \Delta f = \alpha/\sqrt{2\pi} \quad . \qquad (2\text{-}55)$$

Each elementary signal occupies an area of ½ unit, called a logon, and an arbitrary signal could be expanded, approximately, in terms of the elementary signal. However, since the elementary signals are not orthogonal, this process is inconvenient. Slepian and Landau (see [P-2, pp. 67-74]) generalized Gabor's uncertainty principle and showed that the prolate spheroidal wave functions are the orthogonal, time-limited signals which squeeze the most energy into a given bandwidth.

Recently, Rihaczek derived an analytic expression for the energy distribution of an arbitrary signal [R-11]. He showed that the complex energy density function (on the time-frequency plane) of a signal s(t), with analytic representation m(t), is defined by

$$e_c(t,f) \equiv m(t)\cdot M^*(f)\cdot e^{-j2\pi ft} \qquad (2\text{-}56)$$

with the real form given by

$$e(t,f) = s(t)\cdot \mathrm{Re}[S(f)\cdot e^{-j2\pi ft}] \ . \qquad (2\text{-}57)$$

This equation can be used, for example, to interpret Gabor's question regarding a truncated sinusoid. If

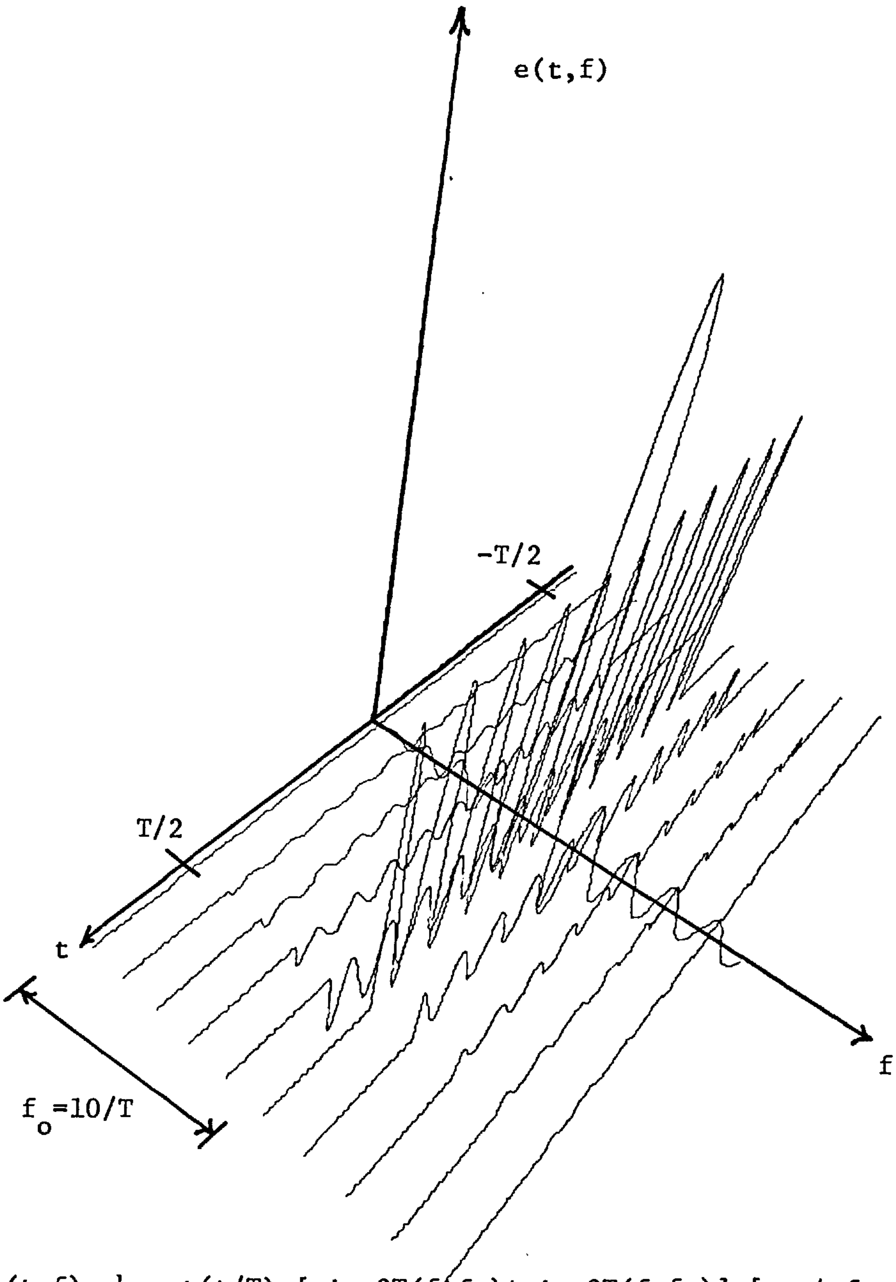
$$s(t) = \mathrm{rect}(t/T)\cdot \cos 2\pi f_o t \quad ,$$

then

$$e(t,f) = \tfrac{1}{2}\cdot \mathrm{rect}(t/T)\cdot[\mathrm{sinc}2T(f+f_o)+\mathrm{sinc}2T(f-f_o)]\cdot[\cos 4\pi f_o t+1].$$

This function is illustrated in Fig. 2.5.

Equation (2-57) satisfies the requirement that its integral over all t gives the signal energy density as a function of f--the energy density spectrum--and that integration over all f gives the energy density at time t--the energy density waveform. As expected, integration over the entire time-frequency plane gives the total energy in the signal, E_t. Furthermore, the total energy in a



$e(t,f) = \frac{1}{2}\mathrm{rect}(t/T)\cdot[\mathrm{sinc}\,2T(f+f_o)+\mathrm{sinc}\,2T(f-f_o)]\cdot[\cos 4\pi f_o t+1]$

Fig. 2.5 Energy density function for $s(t)=\mathrm{rect}(t/T)\cdot\cos 2\pi f_o t$, $f_o = 10/T$.

particular cell, centre t_o, f_o, of the t-f plane is given by

$$E_{T,B} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{rect}\left(\frac{t-t_o}{T}\right)\cdot\mathrm{rect}\left(\frac{f-f_o}{B}\right)\cdot m(t)\cdot M^*(f)\; e^{-j2\pi ft}dt\;df\;. \tag{2-58}$$

However, if T and $B \to 0$, the resultant point value for the energy $e(t_o,f_o)$ is *not* really a true measure of energy distribution since neighbouring points might have energy densities which are nearly equal in magnitude but opposite in sign and hence cancel. This implies that, as Gabor suggested, a cell of minimum dimensions should be used. Now (2-58) can be written as

$$E_{T,B} = \int_T\int_B |m(t)|\cdot|M(f)|\cdot e^{j[\phi(t)-\theta(f)-2\pi ft]}dt\;df. \tag{2-59}$$

For signals with strong phase modulation, (e.g., speech, as we shall see) this is an integral which fluctuates *rapidly* under a *slowly* varying envelope. Rihaczek noted that for these signals, the significant contributions to the integral come from the time-frequency areas where the phase,

$$\Phi(t,f) = \phi(t) - \theta(f) - 2\pi ft\;, \tag{2-60}$$

is stationary; that is, where its derivative goes through zero. Then

$$\frac{\partial\Phi(t,f)}{\partial t} = \phi'(t) - 2\pi f = 0 \text{ when } f = \phi'(t)/2\pi \equiv f_i(t)$$

and $$\frac{\partial\Phi(t,f)}{\partial f} = -\theta'(f) - 2\pi t = 0 \text{ when } t = -\theta'(f)/2\pi \equiv \tau_g(f).$$

Thus we have a concentration of energy, simultaneously, at $f_i(t)$ and $\tau_g(f)$ with the value of the energy dependent upon $m(\tau_g)$ and $M(f_i)$. If $\phi(t)$ is linear about the stationary point, $E_{T,B}$ increases linearly with T; similarly, for $\theta(f)$ linear, $E_{T,B}$ increases linearly with B. In both these cases, the linear variation of the term $2\pi ft$ is just offset. Rihaczek derived, using these concepts, an expression for T_r, the *relaxation time* (or interval within which the signal energy is concentrated at a particular time) of the signal and for B_d, the *dynamic signal bandwidth* (or frequency band within which the signal energy is concentrated) and showed that

$$T_r \cdot B_d = 1 \qquad (2\text{-}61)$$

and that the *shape* of the cell on the t-f plane depends upon the rate of change of $\phi'(t)$, the instantaneous signal frequency. This, effectively, is a more physically meaningful formulation of Gabor's 1946 "mathematical identity which is at the root of the fundamental principle of communication." (sec. 2.4)

2.7 Fourier Analysis in Speech Recognition and Processing

Gabor [G-1] explained the choice of sine waves in favour of other orthogonal functions for $g_k(t)$, (2-1), by noting that only simple harmonic functions transmit the same amount of information in equal time intervals. He also explained that only harmonic functions satisfy linear differential equations in which time does not figure explicitly and that it follows that these are the only ones which can be transmitted by linear, time invariant circuits.

However, we must justify the use of Fourier analysis in speech processing, specifically clipping, analysis. Helmholtz, in his classic work, *On the Sensations of Tone* [H-10, p. 35] pointed out that Fourier techniques give convenient, but without reference to auditory perception, arbitrary mathematical descriptions of sounds. The study of speech clipping, using Fourier techniques, might then appear to be a study of the phenomenon using an arbitrary mathematical description. However, when auditory perception is understood as a *form* of spectrum analysis, then Fourier techniques provide an analog description of the psychophysical process.

In the next chapter, we briefly review the nature and theories of speech and hearing and attempt to show that the description of hearing as a form of spectral analysis is compatible with both physiological evidence and psychophysical experimental results.

3 SPEECH AND HEARING

This chapter is intended to serve three purposes. First, it introduces some basic theories concerning speech and hearing. This material is directed primarily towards readers of this thesis familiar with signal processing concepts, but unacquainted with the distinctive characteristics of the speech signal source, the acoustic speech waveform and the human auditory system. Secondly, the role, if any, of spectrum analysis in the perception of speech sounds must be critically examined before we can discuss the effects of clipping as an operator on the speech spectrum. Finally, we provide a realistic physical basis for the adoption, in chapters 9 and 10, of certain mathematical models of speech waveforms for use in the study of the role of zero crossings in speech perception and automatic recognition.

In outlining the characteristics of speech sounds we shall make an important distinction between objective features and perceptual cues. First, we describe those features of the acoustic waveform which, either directly or indirectly (through a transformation), enable speech sounds to be *objectively* categorized--perhaps with reference to the ultimate mode of production. Next, we consider certain static and dynamic characteristics which have been shown to be important cues for *perceptual* discrimination among speech sounds. Finally, the role of various *objective features* as *perceptual cues* will be emphasized.

Similarly, in describing the nature of the auditory system we shall differentiate between *physiological characteristics* and *psychoacoustic phenomena and models.* The physical nature of the peripheral auditory system--and its response to external stimuli--is known through *objective* observations, notably those of Corti (see [H-10]) and Békésy [B-1]. Psychoacoustic phenomena are *subjective* effects--that is, subjectively reported responses (of the auditory system) to external stimuli. These phenomena often enable researchers to postulate--independently of structural detail--models which describe aspects of auditory system behaviour.

3.1 Auditory Perception as a Form of Spectrum Analysis

Helmholtz, in his classic work *On the Sensations of Tone* [H-10], investigated the physical nature of acoustic disturbances and the physiological aspects of the mechanical sensing of these disturbances in the ear. He began by exploring the physical characteristics and mathematical analysis of acoustic vibrations, in consonance with the following law of G.S. Ohm:

> Every motion of the air which corresponds to a composite mass of musical tones is capable of being analyzed into a sum of simple pendicular vibrations, and to each such simple vibration corresponds a simple tone, sensible to the ear, and having a pitch determined by the periodic time of the corresponding motion of the air.

Helmholtz proceeded to justify the correctness of this law by emphasizing, with reference to Fourier analysis, that "the multiplicity of vibrational forms produced by the composition of simple pendicular [harmonic] vibrations is not merely extraordinarily great: it is so great that it cannot be greater."

To demonstrate that the harmonics contained in complex tones can be physically detected *independently* of the human ear, Helmholtz introduced the idea of sympathetic resonance of physical bodies. He extended this to the use of external acoustic resonators acting as analyzers of sounds with the ear serving merely to detect whether or not the analyzer is excited, and to what degree. Next, he attempted to show [H-10, ch. 4] that the ear itself was capable of carrying out the analysis. In fact, he demonstrated that an experienced observer can detect the presence of harmonics in tones and speech, in some cases up to the sixteenth harmonic. In addition, Helmholtz emphasized that "the quality of the musical portion of a compound tone depends solely on the number and relative strength of its harmonics and in no respect to their differences of phase."

In further investigations [H-10, ch. 7,8], however, Helmholtz found that audible beats were produced by simple tones above a few hundred Hz when the frequency ratio of the tones is less than five to six. As Goldstein pointed out [G-5], on the basis of these and other experiments with interference of sound, Helmholtz suspected, but neglected to state explicitly, that there is the possibility of phase perception among tones *which are not separately resolved by the ear.*

Thus, on the basis of psychoacoustical experiments only, Helmholtz postulated a model describing sound perception as a form of continuous, parallel, spectrum analysis with limited frequency resolution.

3.2 The Nature of the Auditory System

In this section we briefly describe the structure of the ear and its relevance to Helmholtz' psychoacoustic model. We then examine more recent attempts to specify and describe the operation of the auditory system.

3.2.1 Physiological Structure

The components of the ear can be divided into three regions:

specifically, the outer, middle and inner ear (Fig. 3.1).

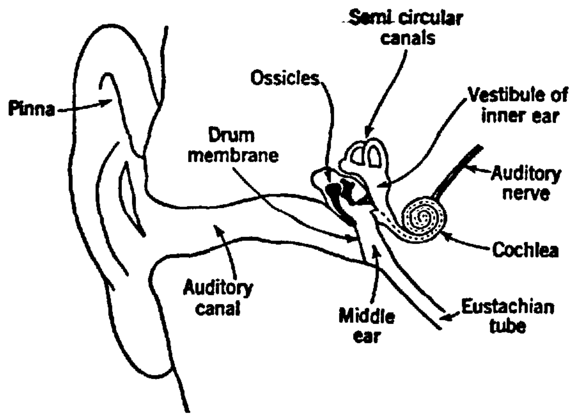


Fig. 3.1

Schematic diagram of outer, inner and Middle ear regions. Not to scale. (From [K-7].)

The outer ear, consisting of the visible pinna (or ear flap), surrounds and protects the entrance to the meatus, or external ear canal, which approximates a uniform tube. The meatus is about 2.7 cm in length and, hence, one-quarter wavelength at 3000 Hz for acoustic sounds. Near this frequency, the resonance effects provides a sound pressure increase of 5-10 db at the closed termination, the eardrum or tympanic membrane, over the value at the ear canal entrance [W-1].

The middle ear contains the ossicular bones, the malleus (hammer), incus (anvil) and stapes (stirrup). This coupled assembly--eardrum, hammer, anvil, stirrup--effects an upward acoustical impedance transformation from the low impedance of the air to the high impedance presented by the inner ear. A pressure transformation, as much as 15:1 [F-8, p. 78], is accomplished through the lever action of the ossicular chain and the large effective ratio of input-output (eardrum-stirrup) surface areas. Besides having the additional property of protecting the inner ear-by means of a change in mode of ossicular vibration--against very intense sounds, the middle ear possesses a low pass amplitude transmission characteristic [F-8, p. 80] whose effective "roll-off" frequency, though on the order of 1 kHz, is subject to much variation. Helmholtz described, qualitatively but accurately [H-10, pp. 129-135] the form and function of the outer and middle ear relying mainly on anatomical observations. It remained for Békésy, Zwslocki and Moller (see [F-8, p. 79]) to quantify this description nearly 100 years later.

The complex inner ear (described as "the labyrinth" by Helmholtz) consists of the vestibular apparatus, the cochlea and the auditory nerve terminations. The vestibular apparatus comprises three semi-circular canals used primarily in sensing spatial orientation. The cochlea (see Fig. 3.1) proceeds forward from the oval window and takes the form of a spiral "snail shell" filled with perilymph, a colourless liquid. The spiral is divided into two canals separated by a partition which is itself a channel (Fig. 3.2). This channel, the scala media, is bounded by a bony shelf and two membranes--the soft Reissner's membrane and the more rigid basilar membrane (Fig. 3.3). The two canals, the scala

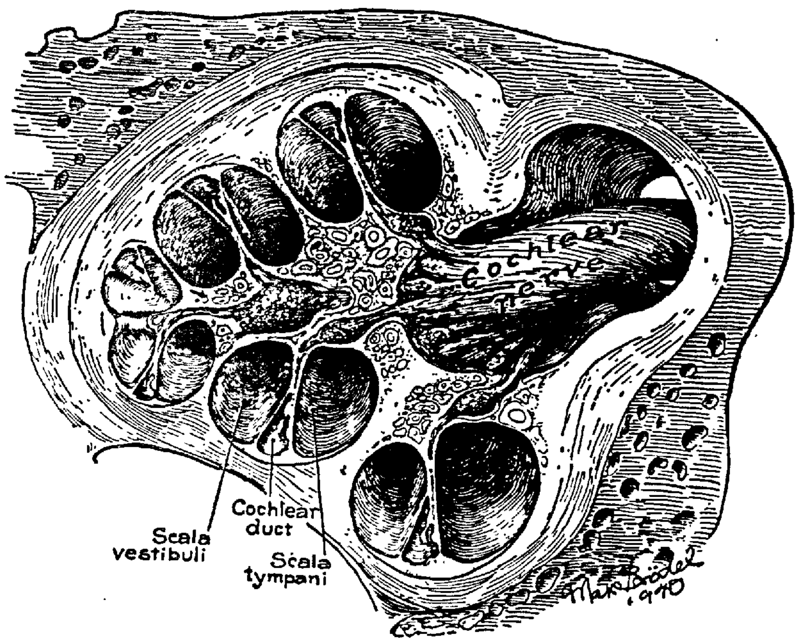


Fig. 3.2 A cross section of the cochlea. (From [W-6].)

vestibula and scala tympani, are connected only at the helicotrema (a small gap where the basilar membrane and bony shelf terminate just short of the spiral's end) and hence form a continuous, folded tube.

In operation, the stapes vibrates the oval window which, acting as a piston, produces a volume displacement of the cochlear fluid. This displacement is relieved by the compliant covering of the round window at the far end of the folded tube. The fluidic vibrations are transferred to the basilar membrane which, via the organ of Corti resting on the membrane [and containing over 30,000 sensory cells which terminate the auditory nerve] provides the mechanical to neural transduction. Therefore, it is the acousto-mechanical properties of the basilar membrane which provide the key to the first step in the analysis and perception of sounds.

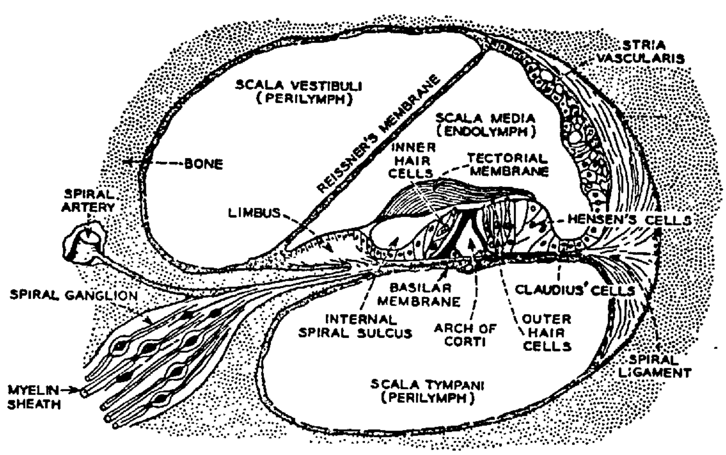


Fig. 3.3 Enlarged cochlear cross section. (From [F-8].)

Helmholtz believed that the basilar membrane was tightly stretched in its transverse direction (width), but rather limp along its length. He also knew that the width of the membrane increases about an order of magnitude from beginning (oval window) to end (helicotrema) [H-10, pp. 145-146]. Using this anatomical evidence, he hypothesized that the basilar membrane acts approximately as a system of parallel, independent, damped, stretched strings each having a slightly lower resonant frequency than the one before. Furthermore, he stated, the nerve cells in the organ of Corti "will be the means of transmitting the vibrations received from the basilar membrane to the terminal appendages of the conducting nerve."

Helmholtz claimed that his hypothesis "has reduced the phenomenon of hearing to that of sympathetic vibration and thus furnished a reason why an originally simple [compound] periodic vibration of the air produces a sum of different sensations and hence also appears as compound to our perceptions." [H-10, p. 148] The hypothesis accounted for beats which are produced by single tones "so near to each other in the scale that they both make the same elastic appendages of the nerves vibrate sympathetically." Helmholtz' hypothesis concerning the mechanism of the ear thus accounted for his formulations based on psychoacoustic experiments.

Békésy performed extensive investigations concerning the mechanism of the middle and inner ear, particularly the basilar membrane [B-1]. He demonstrated that the place of maximum membrane vibration in response to sinusoidal excitation of the stapes varies as a function of frequency with lower excitation frequencies causing maximum vibration at membrane locations further from the oval window. As to the mode of vibration, Békésy concluded (from observations of both models and actual membrane motion) that "during stimulation, a travelling wave is formed on the basilar membrane and not standing waves." [B-1, p.425] This behaviour results from the absence of reflections (at the helicotrema) which in turn is due to the gradual variation of the membrane structural parameters [F-8, p. 83].

Let each place on the basilar membrane be *identified* with the input sinusoidal excitation frequency causing maximum vibration amplitude at that place. Békésy found that, for an input frequency f_i, the amplitude of vibration at other membrane "frequency" places is analogous to the amplitude response of a broadly tuned bandpass filter with center frequency f_i. Place-amplitude response curves for various excitation frequencies have "bandwidths" (when places are identified with frequencies, as above) which are a constant percentage of the excitation frequency. The "frequency" resolution of the membrane is best, therefore, at the "low frequency" end (helicotrema) and the "time" resolution best at the "high frequency" end (oval window).

Békésy's findings regarding the place ("frequency") -amplitude response of the basilar membrane, to sinusoidal signals, were not in accordance with the auditory model postulated by Helmholtz for the following reason: the limited frequency resolution of the human auditory system, as evidenced by Helmholtz' experiments with beats, is much better than the mechanical "frequency" resolution of the basilar membrane (see Fig. 3.4). In the next section we first discuss attempts to quantify the frequency resolving power of the "auditory spectrum analyzer." We then mention some attempts to explain the discrepancy noted above.

3.2.2 Cochlear Analysis and Critical Band Theories

R. Plomp observed in 1964 [P-14] that the only quantitative statements concerning the audibility of harmonics date from a time when it was impossible to measure the objective strength of the tones. To rectify this situation, he performed a series of experiments to investigate the number of distinguishable "harmonics" of signals composed of a series of simple tones with integer *(harmonic)*, and non-integer *(inharmonic)* frequency ratios.

He found, for example, that with a fundamental of 250 Hz, integer harmonics of frequency less than 1625 Hz (the '6.5^{th}' harmonic) could be distinguished from an independent test tone 125 Hz away more than 75% of the time. This means that at 1625 Hz, two harmonics must be separated by a *critical frequency difference* of more than 250 Hz in order to be distinguished, or resolved. The experiments were duplicated for other fundamental frequencies and, to eliminate the possibility that observers could recognize frequency ratios, repeated for inharmonic tone complexes. Figure 3.4, from [P-14], illustrates that the critical frequency differences for both harmonic and inharmonic tone complexes agree quite closely. Note also the fairly close correspondance with the lower solid curve, which represents the critical bandwidth of the auditory system as determined by Zwicker *et al.*[Z-1].

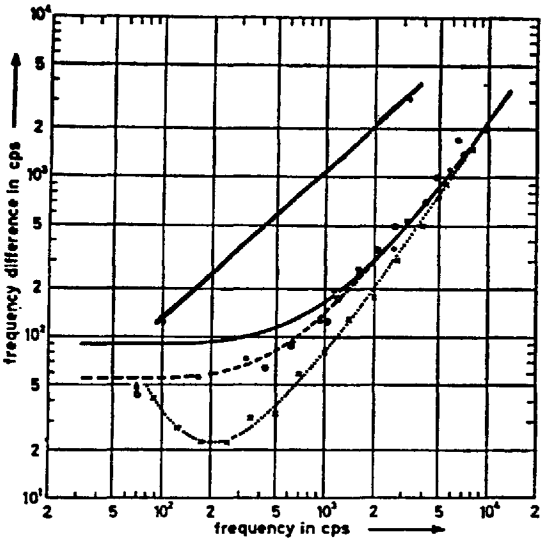


Fig. 3.4 Solid Line--Analyzing bandwidth of basilar membrane as determined by Békésy [H-29].

Solid Curve--Critical bandwidth of auditory system [Z-1].

Dashed Curve--Frequency difference between the partials of complex sounds required to hear them separately, as determined by Plomp [P-14] using harmonic (.) and inharmonic (o) tone complexes.

Dotted Curve--As dashed curve, but using two tone complexes for test signal (x). (From [P-14].)

The concept of *critical bandwidth* is used to describe the fact that the subjective response to auditory stimuli with a frequency spectrum exceeding a certain "critical" bandwidth is different from that when stimuli not exceeding this bandwidth are used. J.L. Goldstein briefly described [G-5, pp. 45-51] recent experiments which were carried out to quantitatively measure the actual size of the critical bands as a function of their centre frequency.

As Helmholtz had suggested, the frequency resolution of the auditory spectrum analyzer *is* limited and tones sufficiently close together (within the same critical band) excite "common areas" and give rise to anomalous perceptual phenomena, including beats. In fact, Goldstein demonstrated through his own experiments that, as Helmholtz had implied, perception of phase effects in monaural sound is possible as a consequence of this limited resolution. However, as we have seen, Helmholtz' belief that the "common areas" were "elastic appendages of the nerves" cannot, in the light of Békésy's findings, accound for the observed degree of resolution of "the auditory spectrum analyzer."

We now discuss one attempt to reconcile the anomaly of broad cochlear bandwidth apparently giving rise to acute perception of minimal pitch changes [S-15] and critical bandwidths as little as one-tenth of the cochlear bandwidth at the same frequency (see Fig. 3.4).

Huggins and Licklider [H-25] postulated mechanical and neural mechanisms for supplementing the mechanical "frequency" resolution of the basilar membrane. The several *mechanical hypotheses* mentioned show that mechanical processes interposed between the motion of the basilar membrane and the excitation of the auditory nerve could produce a resolution sharpening effect; conversely, or in addition, various neural sharpening mechanisms were proposed. In discussing the possible models for *neural* sharpening, Huggins and Licklider emphasized that

> one of the basic facts of neurophysiology is that the nervous system works despite a considerable amount of disarrangement of detail . . . Nevertheless it is important to keep in mind that a statistical interpretation of details is required. Thus, the hypothesis that the nervous system computes an exact derivative, as by a digital process, is hardly to be taken seriously. But the hypothesis that the nervous system performs, in its statistical and analogical way, an operation that may roughly be described as differentiation, and one that we may represent by differentiation in a mathematical model, seems to account economically for a considerable range of facts.

In an *analogical* sense, we might reasonably justify--in the light of"a considerable range of facts"--the performance of the combined ear-brain system (the human auditory system) as a form of "auditory spectrum analyzer." However, again quoting Huggins and Licklider: "The principle of diversity [i.e. that the peripheral auditory processes may present a number of "transforms" to the central nervous system, which may use one or all of them] suggests that *a simple description of the auditory process may not be possible because the process may not be simple.*" (Italics mine.)

3.2.3 Auditory Analysis on the Time-Frequency Plane

If the human auditory system can be considered to effect a *form* of spectrum analysis, then using the principles reviewed in chapter 2, we should be able to quantify its action. Gabor [G-1], for example, applied the concept of information on the time-frequency plane in an attempt to calculate the minimum area on the time-frequency "information diagram" which could constitute a datum of information and to test the shape dependence of this threshold value. He analyzed the experiments of Shower and Biddulph (concerning frequency modulated signals) [S-15], and Bürck et al (concerning truncated sinusoids) [B-23], and concluded that below 1 kHz, the performance figure of the auditory system is such that only 50% of the available information is rejected. This discrimination is the maximum for an instrument, like the ear, which is effectively phase insensitive. At higher frequencies the efficiency is much less. In addition, he argued, the auditory system appears to have a variable time constant adjustable "at least between 20 and 250 milliseconds."

Finally, in order to explain the facility of the auditory system for accurately defining the relative pitch of a prolonged sinusoid (e.g. see [F-8, pp. 211-213]), Gabor stated that it is necessary to assume a second mechanism (besides the mechanical "analyzer" constituted by the basilar membrane) "which after about 10 milliseconds detaches itself from the mechanical resonator curve and locates the centre of the resonance region with a precision increasing with the duration of the stimulus." Cherry emphasized [C-7, p. 157] that if the action of the auditory system is to be modelled as a *form* of spectrum analysis, then the parameters of the analyzer (bandwidths, for example) would be expected to be variable, rather than fixed.

To conclude this section we ask the following question: "Is it possible that the inner ear, rather than the auditory system, effects a form of spectrum analysis?" Huxley recently pointed out [H-27] that when certain physical features of the cochlea are taken into account it becomes theoretically possible for a truly resonant oscillation, the position of which shifts with frequency, to occur in the cochlea. He showed that by taking into account both the spiral shape of the basilar membrane (hitherto ignored in mathematical models) and the prestressed condition of the bony structure which supports the membrane, it is possible to postulate a realistic model which incorporates a truly resonant mode of oscillation rather than the travelling wave solution formulated by Békésy to explain his observations. Huxley states that Békésy, by opening the cochlea and using an artificial stapes, may have altered the mechanical conditions sufficiently to convert a resonant mode which had existed during life into the travelling wave observed.

3.3 Speech Production

Speech is the product of a highly restricted mechanism--the human vocal system--which can be modelled as a linear, time varying acoustic system [F-2],[F-8],[F-9],[S-21]. Since the attributes of the vocal apparatus determine the character of its output, we begin with a short description of the system emphasizing properties responsible for the distinctive characteristics of the acoustic speech signal.

Some speech sounds (vowels, for example) are characterized by spectrally prominent features which are relatively speaker invariant. Other sounds, some consonants, for instance, are spectrally uninformative and may be perceptually unambiguous only in context. In section 3.4, therefore, we discuss the spectral characteristics of speech sounds, describe some methods of parameter measurement and classification, and evaluate the objective and subjective information conveyed by static and dynamic measures of spectral features. Sections 3.5 and 3.6, respectively, are reserved for a description of the statistical properties of speech waveform amplitudes and a discussion of alternate modes of speech perception and classification.

3.3.1 The Source

A basic outline of the speech production system is given in Fig. 3.5.

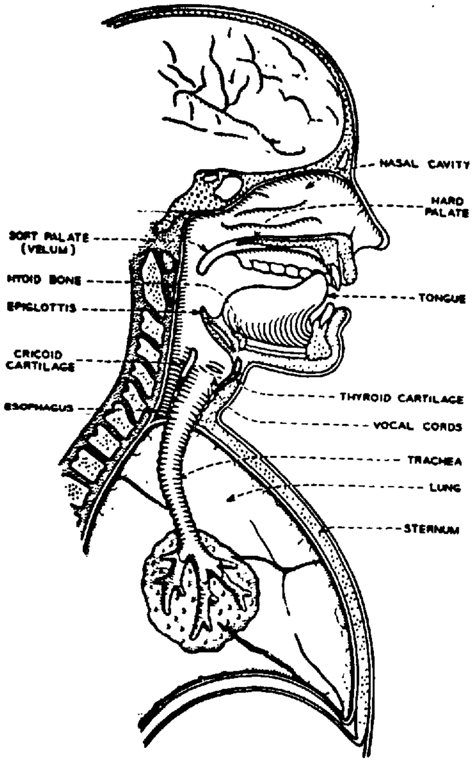


Fig. 3.5. The Human Articulatory System (From [F-8].)

The source excitation for *voiced* sounds (e.g. vowels) is the volume velocity output of the vibrating vocal cords, or folds. Miller suggested [M-11] that, based on his experimental observations, the most significant fact concerning the spectral structure of the glottal waveform is that "uniform harmonic distribution . . . is a rarity." However, the time variation of the glottal aperature *area* is most aptly described as a quasi-periodic 'triangular' wave. The fairly constant pressure supplied to the glottis by the lungs gives rise to a *volume velocity* wave which duplicates in form the area wave and hence, due to the spectral qualities of triangular waveforms, has a spectral envelope falling in amplitude as $1/f^2$.

The quasi-periodic nature of vowel acoustic waveforms results from exciting a linear system, the vocal tract, with quasi-periodic waves. The system output can therefore be calculated using time domain convolution. Since time domain convolution corresponds to frequency domain complex multiplication, the downward sloping (with increasing frequency) envelope characteristic of most vowel spectra directly reflects the nature of the glottal wave spectrum.

Two modes of glottal waveform time behaviour are observed: As the period of the waveform (pitch period) is varied, the wave-shape over the cycle may simply be uniformly stretched *or* the basic triangular pulse duration may be invariant with an increase in interval between pulse occurence. In the former case the spectral line components retain the same amplitude but their separation changes; in the latter case the envelope of the spectrum of the basic triangular pulse is sampled at different points. Therefore, as well as attenuating the vocal tract transfer function with increasing frequency, the time characteristics of the glottal waveform effectively specify the discrete frequencies at which the continuous tract frequency transfer function is sampled to give the line spectrum characteristic of a voiced sound [M-11].

Excitation for *unvoiced* sounds occurs not at the glottis but between glottis and lips [F-8, pp. 47-51] and is created by forcing air through a narrow constriction or across a barrier. The resultant turbulent airflow is characterized by a random pressure distribution which directly contrasts with the quasi-periodic, deterministic nature of voiced sounds. Stop consonants result from pressure buildup and rapid release at a constriction (e.g. teeth, lips) within the system.

3.3.2 The System

In 1928 Russell [R-17] accepted Alexander Graham Bell's suggestion (1907) that "The quality or 'timbre' of the human voice . . . is due in a very minor degree to the vocal cords and in a much greater degree to the shape of the passages through which the vibrating column of air is passed." Thirty years later Fant [F-2] reinforced the overall concept of the vocal tract system as a *filter* with his theoretical studies and practical confirmation (using X-ray studies of Russian articulation) of the nature of the vocal tract.

The human male vocal tract (Fig. 3.5) is about 17 cm. long and has its cross section varied in area by the movement of lips, jaw, tongue and velum--a small flap which connects the nasal side tract to the main tract. The frequency response or transfer function of the vocal tract is dominated by three or more marked resonances which are manifested as formants, or peaks, in the spectrum of voiced sounds. Finally, the radiation impedance which terminates the vocal tract contributes a radiation resistance directly proportional to frequency [F-8, pp. 33-34],[M-15].

3.4 Time-Frequency Characteristics of Speech Sounds

G.E. Peterson emphasized [P-10] that only a minimal amount of the information required for the *interpretation* of speech is in the signal itself and that "the listener who is able to interpret the speech of a particular language successfully has large quantities of information about that language stored in his central nervous system." However, the first step in any speech processor involves a reduction and extraction of information-bearing acoustical parameters from the waveform and knowledge concerning the nature of, and bounds on, these parameters is essential to proper analysis. In sections 3.4.2-3.4.8 we describe the information conveyed by the short-term amplitude spectrum of speech sounds. We begin by describing the process of short-term spectral analysis.

3.4.1 Short-term Spectral Analysis

The generalized short-term amplitude spectrum is *defined* as [G-5, p. 90],[F-8, p. 121] the amplitude spectrum of the Fourier transform of a signal weighted so as to eliminate future values of the signal and progressively attenuate past values.

That is,

$$S(t,f_o) = |F_y\{s(t-y)\cdot h(y)\}|_{f=f_o} \quad (3\text{-}1a)$$

$$= \left| \int_0^{\infty} s(t-y)\cdot h(y)\cdot e^{j2\pi f_o y} dy \right| \quad (3\text{-}1b)$$

where $\omega_o = 2\pi f_o$ and $h(t) = 0$ for $t < 0$ (Fig. 3.6).

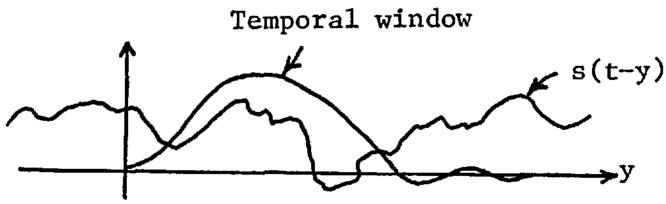


Fig. 3.6 'Time limiting by weighting with finite impulse response'

Expanding (3-1b),

$$\begin{aligned} S(t,f_o) &= |s(t)*h(t)\cos 2\pi f_o t + j\, s(t)*h(t)\sin 2\pi f_o t| \\ &= |F^{-1}\{S(f)\ [H_1(f) + j\ H_2(f)]\}| \end{aligned}$$

where $h_1(t) = h(t)\cos 2\pi f_o t \leftrightarrow H_1(f)$

and $h_2(t) = h(t)\sin 2\pi f_o t \leftrightarrow H_2(f)$.

$H_1(f)$ and $H_2(f)$ can be interpreted as the frequency characteristics of phase-complementary bandpass filters centered at f_o [G-5, p. 92], [F-8, p. 123]. (See Fig. 3.7) $S(t,f_o)$ can be *regarded* as the detected temporal response of $H_1(f)$ *and* $H_2(f)$, found by taking the square root of the sum of the squared responses of the filters. In practice, for economy, $S(t,f_o)$ is approximated by detecting the temporal *envelope* response of a bandpass filter having a frequency characteristic identical to $H_1(f)$.

Goldstein noted [G-5, p. 93] that unless the relative bandwidth of the bandpass filter $H_1(f)$ is very small, the temporal envelope response of the filter is not identically equal to $S(t,f)$ as defined above. The temporal envelope response of $h_1(t)$ to a signal $s(t)$ is

$$E(t,f_o) = |F^{-1}\{S(f)\ [H_1(f) + j\ H_3(f)]\}| \quad ,$$

where $\quad jH_3(f) = \text{sgn}(f)H_1(f)$.

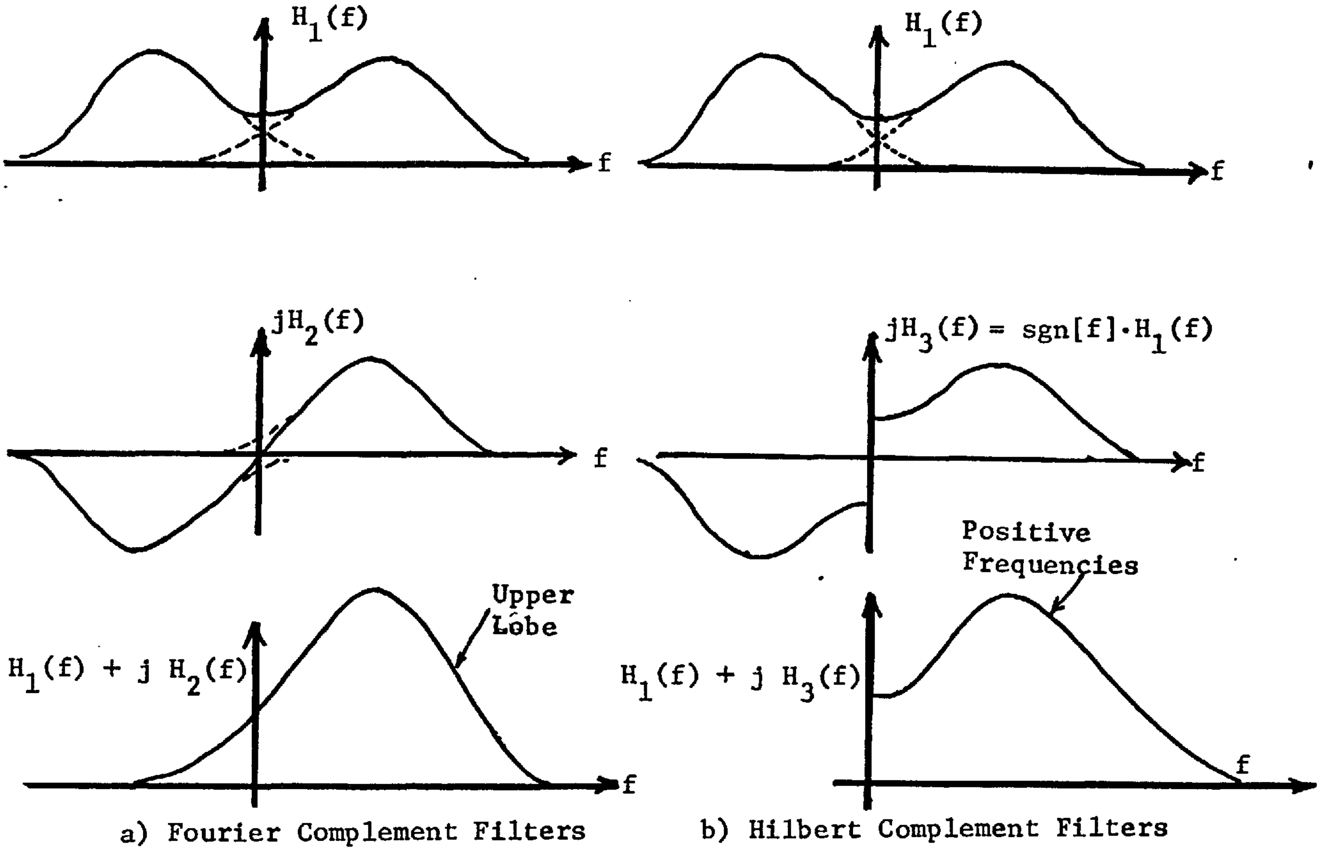


Fig. 3.7 'Fourier and Hilbert Complement Filters' (From [G-5].)

If $jH_3(f) = \text{sgn}(f)\cdot H_1(f)$ (Fig. 3.7), then $h_3(t)$ is equal to $h_1(t)*1/\pi t$ and hence time-unlimited. Therefore $h_3(t)$ cannot be the impulse response of a realizeable filter [G-5, p. 95]. However, for a narrow band $H_1(f)$, the short-term amplitude spectrum $S(t,f_o)$ closely approximates the *temporal envelope response* of a bandpass filter centre frequency f [G-5, p. 93], [F-8, p. 123]. Hence the analysis performed by a *model* of the auditory system as a continuous parallel spectrum analyzer with limited frequency resolution (i.e. a set of contiguous bandpass filters) and that implemented by the instrument called the Sound Spectrograph--a short-term spectral analyzer, using envelope detection--are approximately the same.

3.4.2 Vowels: Their Acoustic Nature and Physiological Correlates

The term *"visible speech"* [P-15] has become synonomous with the short-term speech spectrogram (Fig. 3.8a). Spectrograms of vowels are dominated by a number of formants, or spectral peaks, which are characterized by location, magnitude and bandwidth parameters. Figure 3.8b, from [F-8, p. 131], shows the average of the first 3 formant frequencies (F_1,F_2,F_3), and amplitudes (A_1,A_2,A_3)

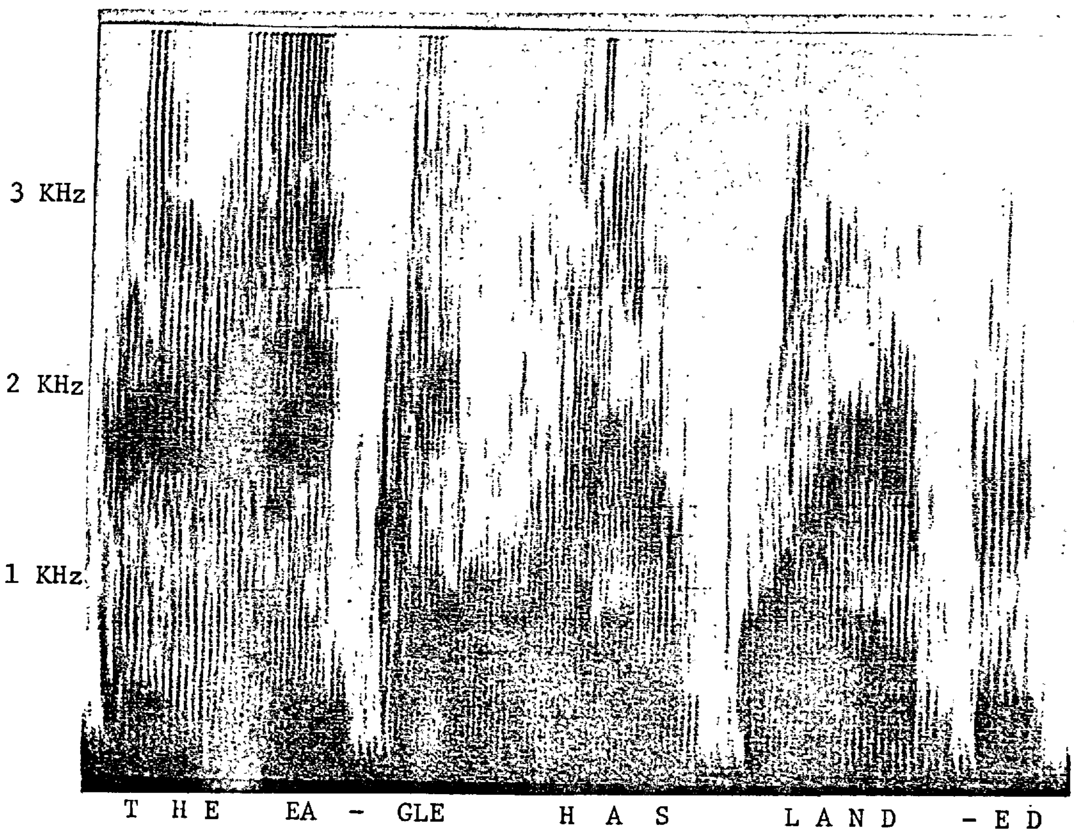


Fig. 3.8a Short-term speech spectrogram, made on Kay Sonograph. Speaker-NAA (American)

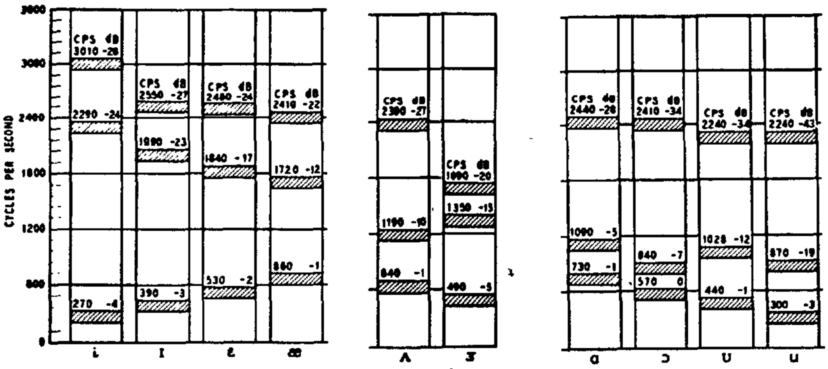
referred to the first formant of /ɔ/, for 33 men speaking the common English vowels in a /h-d/ environment [P-11].

Fig. 3.8b

English Vowels

i) Formant freq. and Amplitudes re F1 of /ɔ/.

ii) Physiological correlates



eve it bet at / up bird/ father all foot boo

Tongue Hump Position . . FRONT CENTRAL BACK

Degree of Constrict.. HIGH MED LOW MED HIGH LOW MED HIGH

Plots of F_1 vs F_2 for different vowels as spoken by *one person* reveal a characteristic closed loop on the F_1-F_2 plane; in addition the areas occupied on the F_1-F_2 plane by different vowels uttered by *various speakers* are generally non-overlapping [P-16, Fig. 5], [P-11, Fig. 8]. These graphic phenomena suggested to some researchers that articulatory interpretations might be accorded to the frequency locations of the first three formants. In fact, Delattre showed [D-7] that degree of maximum constriction in the vocal tract and position of the tongue hump possess striking formant position correlates. For a given tongue hump position, decreasing the degree of constriction raises the first formant position; for a given degree of constriction, the further back the tongue hump the lower the frequency of the second formant (see Fig. 3.8b). Other relationships were noted and later quantitatively analyzed by Fant [F-2].

3.4.3 The Information Conveyed by Vowel Spectra

Speech sounds convey linguistic information--needed for word identification purposes--as well as social-linguistic and personal information [L-2]. The apparent objective vowel classification afforded by formant location and magnitude parameters [P-11] has suggested that these parameters alone might be sufficient for conveying the linguistic information of vowels. However it has been debated whether static formant information alone is sufficient to convey any linguistic information. Moreover, if formant parameters are used as perceptual cues, are these cues contained in the absolute values of certain formant properties (especially frequencies) or in the relationship between these properties and the values for other vowels pronounced by the same speaker? Finally, the relative importance of each of the first three formants as carriers of information has been questioned. We attempt to illuminate these problems in the following subsections:

i) The Intelligibility of Sustained Vowels

A. Jones remarked [J-1] that "if any chosen vowel is sung steadily for some time, the lack of contrast soon makes the vowel less easy to recognize . . ." Siegenthaler devised a set of experiments [S-16] designed to test this assertion by answering the following questions:

> 1. To what extent can . . . [subjects] . . . identify vowels of English as spoken in isolation when the usual elements of initiation and conclusion are eliminated, and when all vowels are sustained for the same period of time?
>
> 2. Are certain sustained vowels more easily recognized than others?

He found that experienced listeners showed an average correct perception of 57.6% for sustained, isolated vowels with the vowel /i/ having the greatest intelligibility and the vowels /U/ amd /e/ being the least accurately recognized. For naive subjects the score dropped to 47.2%. Most vowels incorrectly identified were mistaken for vowels in close proximity, from a physiological and hence spectral viewpoint, to the presented vowels. The arrangement shown in Fig. 3.8b minimizes articulatory steps between adjacent vowels.

W. Tiffany approached the same problem from another viewpoint [T-8]. He noted that vowels in connected speech vary continuously in fundamental frequency, are surrounded (and presumably influenced) by adjacent sounds, and have varying durations. He attempted to determine whether the specification of the physical nature of a vowel *solely* in terms of its acoustic spectrum over a few pitch periods was possible. "To what extent," he asked, "are variations inherent in the contextual speech pattern required for a complete specification of the physical characteristics of vowel phonemes?" Tiffany's results showed a mean rate of 71% to 77% correct recognition for uninflected, *electronically* isolated short vowel segments and a rate of 86% for short vowel segments *spoken* in isolation. His findings that duration and context did influence the recognition rate precluded any hypothesis that vowels are physically specified *solely* in terms of spectra over a few pitch periods. He also noted that some vowels are more stable than others, and hence better understood, possibly because they represent 'limit' positions of the articulatory mechanism [P-11]. He suggested, therefore, that "standardization of [enunciated] phonemes is a much more difficult task than might be supposed."

Lehiste and Peterson, in a more recent study [L-7], showed that sustained vowels can be recognized correctly between 90 and 100% of the time *with training*. Siegenthaler and Tiffany allowed no training period in their experiments; we note here, that as mentioned in the introduction, a training period is required to achieve maximum comprehension of clipped speech.

We conclude, that on the basis of the preceding experimental evidence, nearly 100% recognition of *sustained, uninflected vowels*--i.e. on the basis of time-invariant spectral parameters--is possible with training. Ordinarily, in running speech, the availability of other cues obviates the need for this training. Nevertheless, the high rates obtained without any learning period whatsoever demonstrate that the spectral parameters are doubtlessly a very important factor in vowel perception.

ii) The Importance of Formant Structure

Ladefoged and Broadbent [L-2] attempted to discover whether, as Joos had suggested [J-2], the information conveyed by a vowel depends on the relationship between the formant frequencies of a particular vowel and the formant frequencies of other vowels pronounced by the same speaker rather than the absolute values of their formant frequencies. Using synthesized sentences varying in formant frequency ranges, followed by an unaltered reference word, they showed that the auditory context greatly affected the identification of the fixed word. Thus, Joos' theory was verified and the authors concluded that "it is, therefore, only of limited service to look for common points in the acoustic structure of equivalent vowels spoken by different speakers." The consequences of this statement will become evident when we discuss the use of spectrograms for speech recognition in chapter 4. However, Haggard [H-2] cautioned that "the hypothesis that relationships, not absolute values, determine vowel quality . . . does not imply that vowel quality will be unaffected by octave frequency transpositions [translations], because human perception does not work with the mathematical precision of a slide rule."

Lehiste and Peterson [L-7], and Carterette [C-2], attempted to define further the information conveyed by vowel spectra by investigating, using lowpass and highpass filtering techniques, the importance of *individual* formants in sustained vowel perception. They concluded [L-7] that "one or more of the first 3 formants is essential to the recognition of each vowel" and that their data "did not support the thesis that any arbitrary portion of the vowel spectrum is adequate for identification of *all* vowels."

iii) The Influence of Vowel Duration

Tiffany [T-8] also studied the relation between vowel duration and recognition. Using vowel segments ranging from 0.08 to 8.0 seconds in length, he found that the nearer a given vowel is to its 'natural duration' in connected speech (e.g. [Fig. 1, H-20]) the better the recognition score for that vowel. Nevertheless, "differences in recognition attributed to duration were found to be [statistically] significant for [only] four [of the twelve] vowels" and the average recognition rate for uninflected vowel segments 0.08 seconds long (<8 pitch periods) was 70% rising to 78% for a hundred-fold increase in duration.

3.4.4 Indirect Extraction of Vowel Spectral Parameters

The use of short-term Fourier analysis (or banks of band-pass filters, [T-7]) as a starting point for estimating formant frequencies, amplitudes and bandwidth (all *system* properties) is quite common. Pinson [P-13] and Dunn [D-17], [D-18] stressed, however, that there are effects which limit the accuracy of this method.

First of all, little information is available about the spectrum between the spectral lines caused by the periodic source (sec. 3.3.1) so that spectral peak and bandwidth estimation require interpolation. Secondly, as Miller suggested (sec. 3.3.1 and [M-11]), the envelope of the glottal waveform spectrum is a rapidly varying function of frequency. If this spectrum has zeros near resonances of the vocal tract then bandwidth estimation, in particular, is difficult. These, and other problems--such as deciding how to define a measure of formant amplitude [F-3]--, have prompted investigators to adopt other, more indirect methods of measurement.

The *analysis-by-synthesis* technique, for example, is an attempt to specify parameters for a vocal tract which will best synthesize a spectrum to "match" the sample spectrum [M-8], [B-2], [P-8]. Suzuki *et al.*[S-28] extracted formant frequency parameters by calculating *spectral moments*. *Synthesis of a waveform* to fit the sample signal has also been tried with damped sinusoids [M-3], [P-13] and Gaussian (normal) shaped waveforms [H-22] among the fundamental signals proposed. This type of analysis is often "pitch synchronous" and requires accurate extraction of pitch parameters, a difficult task [G-3], [H-6], [N-4], [W-4], [S-11].

Autocorrelation has also been suggested and used [F-1], [H-24], [M-4], [S-23], [S-7], [P-6] as both a representation of the speech sound and as a means of obtaining the power, and hence amplitude, spectrum [L-6]. In addition, Kleinrock showed [K-8] that the repeated autocorrelation of a signal eventually yields a pure sine wave whose frequency corresponds to the location of the maximum peak of the original signal spectrum. He demonstrated the use of this method in accurate formant frequency estimation.

These indirect methods of spectral parameter estimation, developed to overcome the deficiencies of short-term spectral analysis, will be contrasted with methods using zero crossings in chapter 6.

3.4.5 Nasal Consonants

The nasal consonants /m/, /n/ and /ŋ/ (*sing*) are voiced but differ from vowels in two ways: first, the nasal side passage is coupled to the vocal tract during production and second, the nasals are all associated with dynamic movement of the articulatory system [N-2], [F-18]. The former condition causes zeros in the system transfer function at frequencies for which the transmission to the nasal cavity is short-circuited by a zero impedance oral cavity. The latter condition is responsible for the time variation of nasal consonant spectra. The portion of a nasal consonant during which the oral cavity is closed at a point is termed the nasal 'murmur'.

Fujimura found [F-18] that the nasal murmurs of /m/, /n/ and /ŋ/ are spectrally characterized by low (750-1250 Hz), medium (1450-2200Hz) and high (>3000 Hz) positions of the spectral anti-formant (zero), respectively. The cluster of the 2^{nd} and 3^{rd} (/m/), or 3^{rd} and 4^{th} (/n/), formants with the spectral zero generates a flat spectral null between, roughly, 800 and 2300 Hz. The first formant, he noted, is always low in frequency (≈300 Hz) and all formants are relatively highly damped.

Nakata [N-2] confirmed the importance of the wide band-width of the first formant of nasals as a *perceptual* cue. He also demonstrated, using a synthesizer, that the trajectory and frequency of the second formant, often obscured by the spectral zero, is quite informative, perceptually. Therefore, he concluded, second formant transitions to the adjacent vowel play an important part in human perception of nasal consonants.

3.4.6 Stop Consonants

The stop consonants, /b/,/d/,/g/,/p/,/t/,/k/, are produced when, with the nasal cavity closed, "a rapid closure and/or opening is effected at some point in the oral cavity. Behind the point of closure a pressure is built up which is suddenly released when the closure is released." [H-4] If, during the closure, the vocal cords vibrate, a voiced stop (/b/,/d/, or /g/) is produced; if not, a voiceless stop (/p/,/t/, or /k/) results. However, Halle *et al.* warned [H-4] that in English the *essential* difference between these two classes of stops is that the /p/,/t/,/k/ group result from a more intense pressure buildup causing a higher intensity burst than obtains with the other group.

Acoustically, stops involve rapid changes in the short-term amplitude spectrum preceded or followed by a fairly long (≈0.07 sec.) period devoid of all energy above the voicing component. When a stop consonant is adjacent to a vowel, three cues--silence, burst, transition or transition, burst, silence--are present of which the *silence* is a necessary, and--with *either* a transition *or* a burst--a sufficient, cue for stop perception. For example, in the /k/ of 'tack'both transition and burst are present; in that of 'task' only the burst is present; while in that of 'tact' the transition alone is present [H-4].

Halle *et al.*,after investigating the spectral properties of the stop spectral bursts, stated that the three classes of stops (/b,p/,/d,t/,/g,k/), each associated with a different point of articulation, have the following *spectral* characteristics:

> /p/ and /b/, the labial stops, have a primary concentration of energy in the low frequencies (500-1500 Hz).
>
> /t/ and /d/, the postdental stops, have either a flat spectrum or one in which the higher frequencies (above 4000 Hz) predominate, aside from an energy concentration in the region of 500 Hz.
>
> /k/ and /g/, the palatal and velar stops, show strong concentrations of energy in intermediate frequency regions (1.5-4.0 Khz).

Using observed spectral features only, the authors could classify correctly *and* objectively 95% of their sample sounds.

The very complex role of formant transitions and loci (defined as a formant transition source or target frequency) as acoustic cues in stop perception was also thoroughly investigated by Delattre *et al.*[D-8], Harris *et al.*[H-8] and Hoffman [H-17], all at Haskins Laboratories. Halle theorized on the nature of transitions as follows [H-4]:

> When a [system] resonance is changing in frequency, the formant bandwidth increases. The more rapid the movement, the broader the bandwidth. In the limiting case of instantaneous movement, the bandwidth is infinite; . . . the burst can therefore be considered as an extreme case of *transition* in which changes in the short-term energy-density spectrum are very rapid and the organization of the energy in the frequency domain [as in vowels] is replaced by organization in the time domain . . . . Formant transitions might then be intermediate structures whose assignments to the vowels or to the consonants is a function of their bandwidth, which in turn is dependent on their rate of change.

Summarizing, the cues for stop perception are quite complex. However, short-term spectral structure--i.e., the burst alone--is sufficient both for accurate classification, and--as Halle *et al.* found--for a high rate of recognition of /p/,/t/,/k/ in *perceptual* tests, with training [H-4, p. 108].

3.4.7. Fricative Consonants

The English fricative consonants, together with their place of maximum constriction ('articulation'), are shown in table 3.1.

Table 3.1 Fricative Consonants

Place of Articulation	Voiced	Voiceless
Labio-dental	/v/ vote	/f/ for
Dental . .	/ð/ then	/θ/ thin
Alveolar . .	/z/ zoo	/s/ see
Palatal . .	/ʒ/ azure	/ʃ/ she
Glottal . .		/h/ he

Fricative consonants are produced by a constant-pressure noise source located in the vocal tract (sec. 3.3.2). Since the poles of the vocal tract response are system properties and do not depend upon the location of the excitation [H-9], [F-8, p. 63-64], the energy density spectrum of fricatives, although continuous, may exhibit resonance peaks resembling those of vowels of similar articulatory configuration. In addition, spectral zeros appear at frequencies for which the impedance, looking back from the source towards the glottis, is infinite [H-9], [F-8, p. 64].[2] Poles (resonances) and zeros (anti-resonances) of the system may cancel; but the average spacing of the zeros is greater than that of the poles and, therefore, the cancellation is not present throughout the entire audio spectrum [H-9].

Hughes and Halle noted [H-26] that unvoiced fricatives have little energy below 700 Hz. Conversely, above 1 KHz the spectra of cognate[1] fricatives do not differ appreciably. By means of a set of *objective* spectral measurements, they were able to achieve 85% correct classification of unvoiced fricatives into three categories, each associated with a distinct point of articulation. In addition, using isolated 50 msec. portions of /s/,/f/ and /ʃ/ Hughes and Halle showed that 71% of the stimuli could be correctly *perceptually* classified with little training. They emphasized that the perceptual errors were highly correlated with the errors which occurred using the objective spectral methods of classification. The *physiological* correlates of fricatives and their spectra were investigated in detail by Strevens [S-27]. He showed that the bandwidth of voiceless fricatives (i.e., low, medium, high) was correlated with the place of articulation (i.e., front, back, middle).

---

[1] *Cognates* are pairs of consonants produced with the same articulatory configuration, but with different modes of excitation.

[2] This applies to the series excitation model of the vocal tract.

Heintz and Stevens, in another study of voiceless fricatives [H-9], demonstrated that "simplified versions of fricative consonants generated in accordance with the theory [of pole-zero transfer functions] are demonstrated to elicit responses that are in agreement with the results of the *spectral analyses* [of actual fricatives]." (Italics mine.)

Finally, the role of transitions in fricative *perception* was clarified by Harris, who showed [H-7] that transitions in fricative-vowel syllables are important for differentiating /f/ and /θ/ from their voiced cognates, /v/ and /ð/.

3.4.8 Glides and Semi-vowels

Physiologically, the glides /j/ (*y*ou) and /w/ (*w*e), and semi-vowels /r/ (*r*ed) and /l/ (*l*et), differ from the stops and fricatives in the lesser degree of oral stricture present and from the nasals in the absence of nasal coupling [O-1]. Phonetically, only /w,j,r,l/ can constitute the third member of an initial three-term consonant cluster--for example, splint, skew, square. In other consonantal clusters these consonants *must* occupy the position immediately before (b*r*ead, s*l*ow) or after (me*l*t, bi*r*d) the vowel [O-1].

O'Connor *et al.* [O-1] attempted to discover whether, *spectrally*, these sounds were distinctive among phonemes. They found, using spectrum synthesis and analysis, that the formants of /w,j,r,l/ begin, as do those of voiced, final stops, at loci or frequency starting points. However, they demonstrated, using synthesized phonemes in psychoacoustic tests, that--in contrast to the stops--the /w,j,r,l/ formants must, if confusion with other phonemes is to be eliminated, *remain* at the loci frequencies for 30 (/w,j/) to 50 (/r,l/) milliseconds before proceeding to the steady-state positions in the following vowel (see also [L-8]).

Discrimination among /w,j,r,l/ is accomplished using the transition directions and extents of the second and third formants. The first two formants of /r/ and /l/ have identical loci frequencies so that a third formant is required to remove the ambiguity. In contrast, /w/ and /j/ have different second formant loci so that two formants suffice for unambiguous synthesis and perception. Briefly, the low (600 Hz), medium (1200 Hz) and high (2400 Hz) frequency of the loci for the second formant of /w/, /r,l/ and /j/, respectively, distinguish among these sounds; the low (1500 Hz) locus of the third formant of /r/ contrasts to the high (2900 Hz) locus of /l/'s third formant.

3.4.9 Spectral Specification and Perception of Speech Sounds: an Overview

In section 3.4 we have examined, briefly but in some detail, the use of spectral features as *descriptors* of speech sounds.

We have shown, using experimental evidence, that steady-state spectral parameters are sufficient for vowel discrimination--i.e., that sustained, uninflected isolated vowels are highly intelligible, especially with training. Furthermore, we have seen that perception of nasals, stops, fricatives and glides/semi-vowels is greatly dependant upon their frequency domain structure; manipulation of certain spectral features of these sounds is directly reflected by a change in perceived identity of the sound.

We do not underestimate the importance of speech dynamics, especially transitions [L-16], [S-25]. Indeed, as noted in 3.4.6, certain stop consonants require a minimal period of virtually zero energy for correct perception! Neither do we fail to recognize the importance of contextual cues, especially under non-ideal (e.g., noisy) conditions. Our reference to Peterson's work (sec. 3.4, introduction) emphasized the relevance of the linguistic store.

What we have firmly established is that preservation of overall spectral structure is necessary, sometimes sufficient, and in any case desirable, for retention of high intelligibility.

3.5 The Statistical Properties of Speech Sounds

In section 3.4 we observed that certain portions of speech waveforms, (vowels, for example) are quasi-periodic. However, in general, extended observation of a speech signal does *not* permit prediction of its future behaviour, on a long-term basis. In this sense, speech is the result of a random or stochastic process. Moreover, if the time during which the speech signal is observed is not so long as to permit a fundamental change in the character of the speech source (e.g., fatigue) then stationarity (time invariance) of the stochastic process may be assumed.

If these postulates--set forth by Davenport [D-3]--are accepted, and their conditions of validity satisfied, then it is possible to describe speech, *on a long-term basis*, as a stationary, stochastic process [D-3; p. 4].

With these criteria in mind, Davenport made measurements of *long-term*, first-order and conditional speech waveform instantaneous amplitude distributions. In the next two subsections we consider briefly his findings and those of later investigations concerned mainly with Russian speech sounds. This section provides the necessary background material for the discussion, in chapter 5, of certain aspects of speech clipping.

3.5.1 First-order Density Functions

The first-order probability density function $f_X(x)$ is defined, for a stochastic process, as [D-3; p. 4]

$$f_X(x,t) = \lim \Delta x_1 \to 0 \quad P\{x_1 \leq x(t) \leq x_1 + \Delta x_1\} / \Delta x_1 \quad . \qquad (3\text{-}2)$$

If stationarity is assumed, then the definition becomes independent of time.

Davenport measured $f_X(x)$ over extended durations of speech (one-half minute to ten minutes) by sampling signal amplitudes every 12 usec. He showed analytically that these measurements would suffice to define $f_X(x)$ using the relationship

$$f_X(x_1) \simeq \frac{1}{\Delta x_1}(n_1/n) \qquad (3\text{-}3)$$

where n = total number of samples taken and n_1 = the number of samples in which the event $\{x_1 \leqslant x(t) \leqslant x_1+\Delta x_1\}$ occurs, *if* n is sufficiently large *and* Δx_1 is sufficiently small. In these studies, $n \geqslant 2.5 \times 10^6$ and Δx_1 = 1/50 to 1/100 of the maximum peak-to-peak signal amplitude. The experimentally determined density distribution is shown in Fig. 3.9 for three different speakers, in an anechoic chamber.

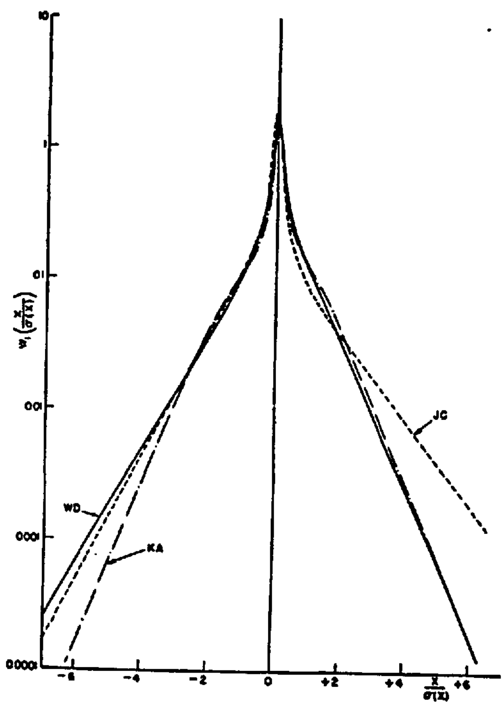


Figure 3.9 The first-order (normalized) probability density function for speech waveforms measured over long periods (½ to 10 minutes). Data for three speakers. (From [D-4].) Note: $W_1(x/\sigma(x)) = f_X(x)$.

By trial and error, Davenport derived an approximate expression for the graphical results. He hypothesized that "the spike" is due to both unvoiced sounds and system noise and that the "overall exponential character" is due to the vowels. Therefore, the vowels were modelled as an exponential distribution occurring 0.6 of the time and the unvoiced sounds and system noise as a Gaussian distribution occurring with probability 0.4. That is,

$$f_X(x) \simeq 0.6\left[\frac{1}{2\sigma_1} e^{-\sqrt{2}|x|/\sigma_1}\right] + 0.4\left[\frac{1}{\sqrt{2\pi}\sigma_2} e^{-x^2/2\sigma_2^2}\right], \quad (3\text{-}4)$$

and, by curve fitting procedures, $\sigma_1 = 1.23$ and $\sigma_2 = 0.118$. Similar measurements on Russian speech [F-6], [R-13], [V-3] yielded distributions quite close to those of (3-4).

A. Rimskii-Korsakov proposed [R-13] an extension of the idea that the long-term probability density function of speech waveform amplitudes is the sum of individual densities, each occurring for some proportion of time. He hypothesized that, if, over a long period of time (at least 2.5 minutes, according to Fersman [F-6]) each *different* speech sound [vowel] has a Gaussian distribution defined by a variance σ_T *and*, if each of these sounds is present for a proportion of time defined by another distribution, then the long-term probability density function for speech waveform amplitudes would be

$$f_X(x) = \int_0^\infty f_T(x)\cdot f_\sigma(\sigma_T)\, d\sigma_T\ , \quad (3\text{-}5)$$

where $$f_T(x) = \frac{1}{\sqrt{2\pi}\sigma_T} e^{-x^2/2\sigma_T^2}\ , \quad (3\text{-}6)$$

a Gaussian density function with variance σ_T^2. Furthermore, if the distribution of the variances, $f_\sigma(\sigma_T)$, is Rayleigh, that is

$$f_\sigma(\sigma_T) = (\sigma_T/\sigma_o^2)\cdot e^{-\sigma_T^2/2\sigma_o^2}\ , \quad \sigma_T \geqslant 0, \quad (3\text{-}7)$$

then substitution of (3-6) into (3-5) yields

$$f_X(x) = \frac{1}{2\sigma_o} e^{-|x|/\sigma_o} . \qquad (3\text{-}8)$$

Equation (3-8), after substituting $\sigma_1/\sqrt{2}$ for σ_o, is equal (except for a multiplicative constant) to the experimentally determined exponential distribution for vowels observed in (3-4). "In other words," he suggested, "there are strong bases [*sic*] for assuming that speech . . . signals are similar in their [long-term] statistical properties to a stationary random [Gaussian] process modulated in amplitude by other random processes [e.g., Rayleigh]."

3.5.2 Conditional Density Functions

Davenport also investigated the long-term conditional density distribution of speech waveform amplitudes. For a stationary stochastic process the conditional density function is defined, using Bayes' theorem, as

$$f_{X|Y}(x_1|x_2;\tau) = f_{XY}(x_1,x_2;\tau)/f_X(x_1) \quad , \; f_X(x_1)\neq 0 \qquad (3\text{-}9)$$

where $$f_{XY}(x_1,x_2) = \lim_{\substack{\Delta x_1\to 0\\ \Delta x_2\to 0}} \frac{P\{x_1\leq x(t)\leq x_1+\Delta x_1 ; x_2\leq x(t+\tau)\leq x_2+\Delta x_2\}}{\Delta x_1\cdot\Delta x_2} . \qquad (3\text{-}10)$$

Davenport showed [D-3, p. 26] that, for small Δx_1 and Δx_2,

$$f_{X|Y}(x_1|x_2;\tau) \simeq P(x_1|x_2;\tau)/\Delta x_2 \qquad (3\text{-}11)$$

where $P(x_1|x_2;\tau) = P(x_1,x_2;\tau)/P(x_1), \quad P(x_1)\neq 0.$

$P(x_1,x_2;\tau)$ and $P(x_1)$ are, respectively, the numerator of (3-10) and of (3-2). Therefore, for small Δx_1 and Δx_2, the conditional density function is $$f_{X|Y}(x_1|x_2;\tau) \simeq \frac{1}{\Delta x_2}(n_2/n_1) \quad , \qquad (3\text{-}12)$$

where n_1 = number of samples in which the event $\{x_1 \leqslant x(t) \leqslant x_1 + \Delta x_1\}$ occurs and

n_2 = number of samples in which the events $\{x_1 \leqslant x(t) \leqslant x_1 + \Delta x_1\}$ *and* $\{x_2 \leqslant x(t+\tau) \leqslant x_2 + \Delta x_2\}$ occur.

Davenport measured the conditional probability $P(x_1|x_1;\tau)$ for three different values of x_1: $x_1 = -0.33\sigma$, -0.65σ, and -1.3σ, where σ is the rms speech waveform amplitude. These experimental probability distributions are shown in Fig. 3.10. Note that, in

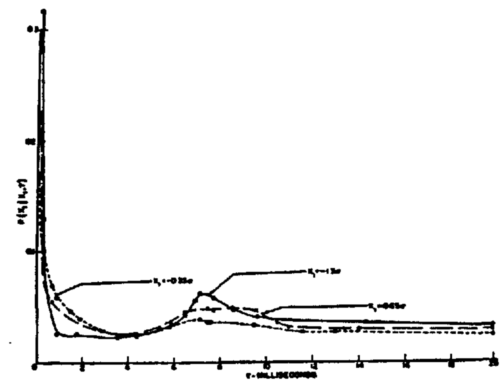


Fig. 3.10 The conditional probability $P(x_1|x_1;\tau)$. For three different values of x_1. Single speaker in anechoic chamber. (From [D-5].)

Fig. 3.10, a peak occurs in the distribution for $\tau \simeq$ a pitch period, and that the peak height is proportional to $|x_1|$. This peak reflects the quasi-periodic nature of the voiced sounds which account for most of the higher amplitude excursions in speech waveforms. Davenport also measured $f_{X|Y}(x_1|x;\tau)$ for $x_1 = -0.65\sigma$ as a function of x for several values of τ. The results of these measurements are shown in Fig. 3.11; note the change of vertical scale among the diagrams.

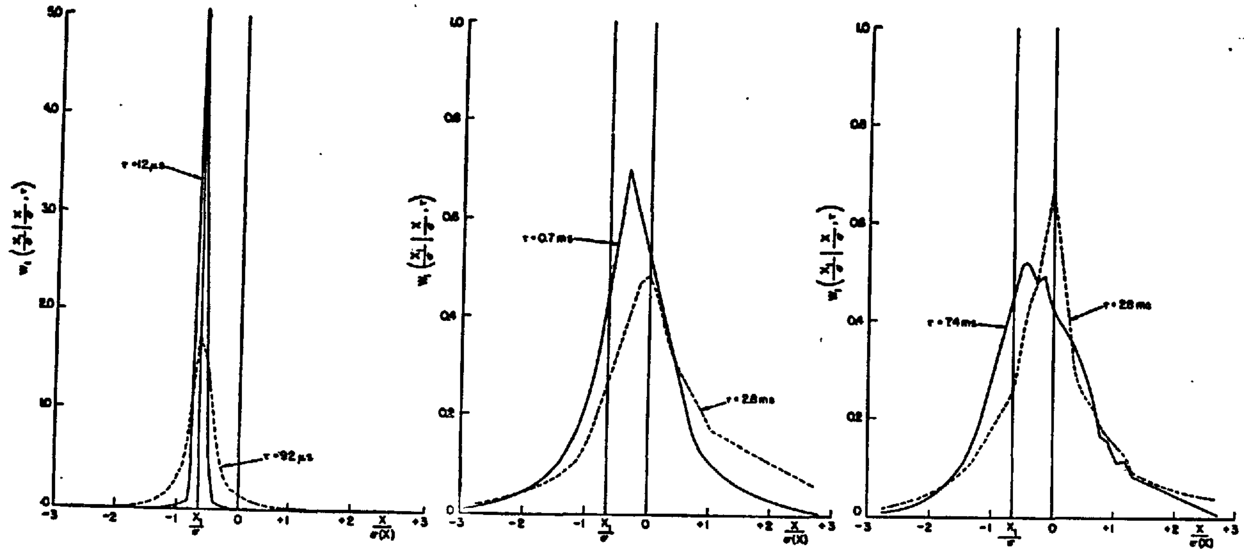


Fig. 3.11 The conditional probability distribution $f_{X|Y}(x_1|x;\tau)$. Measured data for six different values of τ for a single speaker in an anechoic chamber. (From [D-5].) Note: $W_1(x_1/\sigma|x/\sigma,\tau)=f_{X|Y}(x_1|x;\tau)$.

Davenport showed analytically that

$$f_{X|Y}(x_1|x;\tau) \rightarrow f_X(x) \quad \text{as } \tau \rightarrow \infty \qquad (3\text{-}13)$$

and

$$f_{X|Y}(x_1|x;\tau) \rightarrow \delta(x-x_1) \text{ as } \tau \rightarrow 0. \qquad (3\text{-}14)$$

Equation (3-13) obtains because, as τ increases, the amplitudes of the two points on the speech waveform tend to become statistically independent. Note that, in Fig. 3.11, the locus of the intersection of the line $x/\sigma(x) = x_1/\sigma(x)$ with $f_{X|Y}(x_1|x;\tau)$ as a function of τ is--except for the proportionality constant $1/\Delta x_2$--equal to $P(x_1|x_1;\tau)$, Fig. 3.10.

3.5.3 Joint Probability Density Functions

A. Rimskii-Korsakov, in conjunction with Lui Yung-Ts'un, experimentally determined [R-13] the long-term joint probability density function for speech waveform amplitudes, i.e.,

$$f_{XY}(x_1, x_2; \tau) \quad . \qquad (3\text{-}15)$$

Using the same criteria as Davenport used to derive (3-14)--that of independence of waveform amplitudes for large τ -- Rimskii-Korsakov argued that

$$f_{XY}(x_1, x_2; \tau) \simeq f_X(\, x(t)\,) \cdot f_Y(\, x(t+\tau)\,) \qquad (3\text{-}16)$$

for large τ. Therefore, using (3-4), for τ large,

$$f_{XY}(u, u_\tau; \tau) \simeq \frac{1}{2\sigma_1^{\,2}} \, e^{-[\sqrt{2}(|u|+|u_\tau|)/\sigma_1]}, \qquad (3\text{-}17)$$

the product of two exponential distributions. Constant density contours of this function occur for $|u|+|u_\tau|$ = a constant; i.e., squares with vertices on the u and u_τ axes. Fig. 3.12, from [R-13], shows that for τ>30 milliseconds, the distribution is close to that predicted. The sharper corners of the experimental distribution result from the Gaussian component of (3-4) predominating at small signal amplitudes. Rimskii-Korsakov showed explicitly that the elliptical character of the equal density contours for small values of τ can be explained "if we assume that the signal once again can be considered as a complex random process, randomly modulated in amplitude." [R-13]

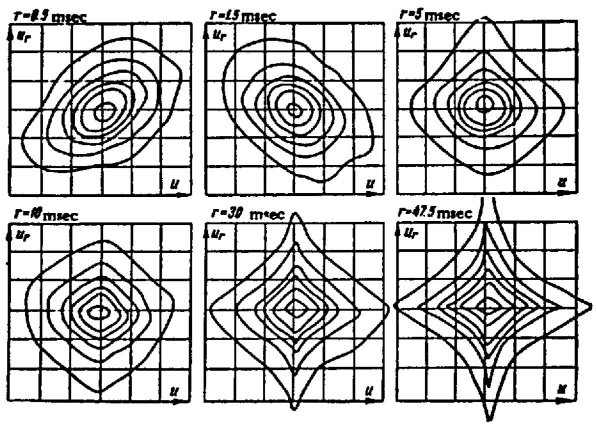


Fig. 3.12 Constant joint probability density contours, experimentally determined for varying τ, for Russian speech. (From [R-13].)

3.5.4 Summary

The agreement between Rimskii-Korsakov's experimental results and theoretical predictions for large τ tends to confirm what Davenport emphasized: that on a *long-term basis* speech can be considered to be a stationary stochastic process *only* in the sense that prediction of future speech sounds is not possible and that the characteristics of the source are "invariant" if the long-term period is short enough so that fatigue etc. does not occur.

We shall see (in sec. 5.3) that Davenport's models have often been misinterpreted and misused in an attempt to apply the powerful tools of stochastic signal theory to the analysis of speech signals which exhibit formant structure. As we emphasized in sec. 3.4, vowels--over the analysis period necessary to reveal formant structure--are *not* stochastic processes but quasi-periodic waveforms.

4 AUTOMATIC SPEECH RECOGNITION

What is the motivation behind attempts to realize automatic speech recognition machines? What is the value of such automata? What is their function? Are such machines simply an attempt to duplicate the human facility of speech perception? Motivation, value, function and method are important quantities to be considered in respect to automatic speech recognition.

This chapter, therefore, is concerned mainly with the philosophy of automatic speech recognition. Note that we do not propose to generate a model for a general purpose speech recognition system of the type described in some of our references. Instead, we wish to outline some of the conceptually important ideas which provide the foundation upon which such systems are constructed. The purpose of this chapter, then, is to establish a framework for the speech recognition experiments presented in chapter 7 and a perspective concerning the role of signal processing in automatic speech recognition.

4.1 Whither Speech Recognition?

J.R. Pierce, in a recent letter [X-2, October 1969] asked, "Whither speech recognition?" He implied that it is "not clear" that speech is desireable mode for man-machine communication. "In fact," he emphasized, "we do very well with keyboards, cards, tapes and cathode-ray tubes." After presenting some indication of the extreme difficulties associated with automatic speech recognition, Pierce noted that an undeniable justification for speech recognition research is "that through such work we [can] learn something about speech." He observed that this will be the case only if the "learning" is made an immediate goal rather than one of a number of means to a more important end. More often than not, he pointed out, the investigation of the nature of speech becomes subservient to "rapture for computers and for unproven schemes . . . for recognition." D.B. Fry expressed the same sentiment when he stated [F-16] that "It is disquieting to note the number of people in various parts of the world who have embarked upon the task of devising a speech recognizer without having learned anything at all . . . ."

Thus, although the immediate value of a speech recognition machine, *per se*, is questionable, the knowledge gained in the investigations which *should* provide the prelude to, and basis of, such ventures is invaluable. Unfortunately, the increase in fundamental knowledge which can be attributed to reported attempts at automatic speech recognition is small; furthermore, these schemes have been--until very recently--comparatively fruitless. "Why have two or more decades of intensive research concerning automatic speech recognition been rewarded with such apparent lack of success?" [Hill; H-12] We shall attempt to provide some answers in the next section.

4.2 The Philosophy of Automatic Speech Recognition

4.2.1 Function

In 1958 Fry and Denes described [F-17] the function of a mechanical speech recognizer as "recognition of linguistic elements on the basis of the acoustic input and the re-encoding of this sequence of elements in the form of a letter sequence." In essence, the recognition automaton serves to replace the human subject as a transcription device. The availability of the phonemic string in discrete, coded form is therefore inherent in the concept of a *phonetic typewriter*. It is important to note that in 1958 the bandwidth saving effected by a phonemic encoder was considered as important, perhaps, as the recognition aspect itself [F-17]. The string of phonetic symbols could be transmitted over a narrow bandwidth channel and a speech signal synthesized using a voice encoder, or "vocoder" [S-6].

4.2.2 Speech Specification via Articulatory Parameters

The modelling of the human auditory system as a form of spectrum analysis--and the success of short-term spectral analysis in revealing certain physically meaningful features in speech sounds--has prompted many researchers to adopt spectral analysis as a first step in the recognition process (sec. 4.3). Nevertheless, as early as 1950, Huggins proposed [H-23] that the auditory mechanism may effectively analyze *not* the acoustic waveform but the system [vocal tract] transfer function. "As far as the response of the basilar membrane . . . is concerned, the mouth and ear may be combined into a single linear system. In effect, *the speaker's mouth is part of the listener's ear*." [H-23] This idea, that a human perceives sounds (at one stage) by "reference" to the vocal tract configuration which produced the sounds, was formalized in 1960 as the *motor theory of perception*. N. Lindgren summarized the essence of this theory as follows [L-18], [L-19]: "Because perception seems to follow articulation rather than sound, the speculation arose that the relation between phoneme and articulation might be more nearly one-to-one than the relation between phoneme and acoustic unit."

4.2.3 Analysis, or Analysis-by-Synthesis?

We recall that an alternative to speech waveform analysis by direct extraction of spectral parameters is the synthesis of a pole-zero system whose transfer function approximates the amplitude spectrum of the incoming signal. (sec. 3.4.4) K. Stevens proposed a *model* for a speech recognition system which, in effect, involves the synthesis of a spectrum to match that of a particular speech spectrum in terms of articulatory parameters. He argued that [S-24]" . . . the analysis that leads to the articulatory description can be performed without reference to the particular language or dialect of the speaker. Since the output of this analysis stage provides, in effect, a description of vocal tract configurations . . . results of the analysis preserve sufficient information [so] that the original speech signal can be approximately recreated."

At a further stage in the analysis, articulatory configurations are expressed in terms of phonetic symbols. A matching process is used to select the phoneme which 'most likely' produced the articulatory configuration which, as noted in the previous paragraph, is determined to have produced the input speech spectrum. Both matching processes necessarily incorporate feedback loops.

Stevens justified the choice of spectral parameters as primary data by reaffirming the belief that "a . . . process similar to spectrum sampling . . . exists in the auditory mechanism." The use of an intermediate articulatory representation reflects the possibility that "a similar representation may likewise exist at some stage during the . . . process of speech recognition." Finally, D.M. MacKay summarized the arguments for the use of analysis-by-synthesis models in speech analysis as follows:

> . . . three distinct arguments are possible for the usefulness of 'active matching' or 'analysis-by-synthesis' in speech perception.

The first is that since speech is the product of a generative process with few degrees of freedom, and the ear, being a general purpose organ, converts it into a representation with many degrees of freedom, it would be economical to represent speech internally by a model of the generative process rather than the product. As it stands, however, this argument could equally apply to the perception of non-speech sounds with few generative degrees of freedom.

This leads to the second argument, that since speech is something *we* produce, we have a suitable internal generator ready made and can economically use it. Moreover 'delayed feedback' experiments have shown the existence of the necessary coupling from the ear to the organizing system for speech.

The third argument is of a different kind. In perceiving speech *as such* we are concerned not only with the classification of phenomena, nor even with the internal imitation of sounds. Our object, in part at least, is to discover *what the originator is up to,* as another agent like ourselves. Here, I suggest, is the chief reason for entertaining seriously the idea that perception of speech (as speech) requires the running of an internal active organizer matching that of the speaker in relevant respects; for it is, I think, the success of this ongoing enterprise that constitutes 'following' him. [M-1]

4.2.4 Segmentation: the Gating Problem

Speech is a continuous process. Yet the output of a speech recognition machine must be a series of discrete symbols. Speech is produced by a vocal track which has inertia. Thus, phonetic transitions are generally gradual rather than abrupt. J. Damman noted that [D-2] "one of the fundamental contrasts between the phonemic sequence and its physical manifestation is that, while the former is discrete, the latter is quasi continuous." In continuous speech, furthermore, the target configuration representing a certain phoneme is barely reached before motion towards the next is initiated; hence a given configuration may be the result of a motivation to produce more than one phoneme [H-3] and it may be impossible to establish a one-to-one correspondence between an acoustic utterance and a phoneme. P. Denes emphasized that [D-10; 1963]

> the basic premise of . . . automatic speech recognition . . . has always been that a one-to-one relationship exists between the acoustic events and the phonemes . . . . there was a deep seated belief that if only the right way of examining the acoustic signal was found, then the much sought-after one-to-one relationship would come to light. Only more recently has there been a wider acceptance of the view that these one-to-one relations do not exist at all . . ."

Indeed, experiments have shown that human recognition of phonemes may be dependant upon cues derived from several acoustic segments [F-17].

Segmentation--and the related problem of time scaling and normalization due to variability of speech rate [B-10]--is a major hindrance to successful automatic speech recognition. But, assuming that segmentation is somehow possible, the choice of acoustic unit (i.e. phoneme, word) presents a series of formidable, interrelated decisions [S-14]. For example, phonemes may not be combined in *any* order to form syllables [D-19]. Therefore the longer linguistic units (e.g. words) incorporate linguistic constraints which should make identification easier. Yet recognition presumably depends on matching a pattern derived from the incoming acoustic unit with one of a set of reference patterns; if so, the number of word patterns that would require storage seems prohibative. And even the largest practical store would not prevent forced, erroneous decisions on unknown words. Phonemes, however, would presumably form a compact, inclusive set [F-16].

4.3 Automatic Speech Recognition: An Overview

We discussed--in section 4.2--some concepts directly relevant to the implementation of automatic speech recognition machines. Specifically, we dealt with some aspects of speech production and perception which, to many researchers, seem desirable to imitate in automatic speech recognition machines.

The extraction of "patterns" from source data--or parameterization of the signal--is a central problem in this thesis. In particular, we will consider the role of zero crossings as a representation of the signal for speech recognition purposes. However, we believe that before this can be done a review of some actual implementations of (non-zero crossing) speech recognition machines should be presented. This review will serve a number of purposes.

First, most of the schemes described parameterize the speech signal via a well known and physically meaningful method--the features revealed in a short-term speech spectrogram. For this reason, the nature and purpose of processing applied subsequent to the initial parameterization, which we will define as *pre-processing*, should be reasonably clear. In contrast, the nature of the estimate of the source afforded by zero crossings is, at this point, somewhat obscure. This subject will be discussed in detail, and clarified, in chapter 6.

Secondly, the review will be logically organized in that we will describe, in turn, attempts at vowel, word and continuous speech recognition. In this manner the difficulties and limitations associated with the recognition of each speech unit should become apparent. Similarly, the complexity of the system associated with each mode should become clear. The brief description of the system used in each case should, we hope, provide some idea of the actual processes which may constitute a speech recognition machine.

Finally, we wish to demonstrate a key concept in automatic speech recognition. Hill argued that the lack of success in machine recognition of speech "is not due to a lack of means of analysis for the acoustic signal." [H-12] "What is difficult," he claimed, "is telling the machine what to do with the results of the analysis." We contend that, as these examples will demonstrate, the recognition phase--telling the machine what to do with the results of the analysis--*may* fail not through lack of technique but because the signal parameterization does not provide a sufficient basis for signal classification. Note that we do not claim that correct signal parameterization is *the* key to successful automatic speech recognition. However, correct parameterization is vital in the following sense: Mechanical speech recognition can be divided into three phases--measurement (or parameterization), transformation of measurements or parameters, and decision making (or recognition). The *decision* is made on the basis of information extracted from the signal via *measurement* or parameterization and presented to the decision function[1] through the *transformation*. We shall see that information is often lost or obscured when the parameterization is neglected in favour of premature excursion into the recognition stage without sufficient attention being given to transformations.

4.3.1 Vowel Recognition

J.W. Forgie and C.D. Forgie based their recognition system upon "the interpretation of the two-dimensional patterns of amplitude and frequency which exist during steady state portions of . . . vowels."[F-11] The envelope detected outputs of a bank of 35 contiguous bandpass filters covering the 115-10,000 Hz region were sampled 180 times per second and quantized versions fed into a computer. The vowels were extracted from a /b/-/t/ context; energy considerations provided the basis for a vowel-consonant decision.

---

[1] The role of decision theory in pattern recognition will be discussed in chapter 7.

The first operation in the detailed analysis of each frequency sample array was to estimate, *roughly*, the locations of F_1 and F_2. The experimenters noted that:

> outstanding among the problems encountered in attempting to set up a formant-tracking programme were (1) a voicing harmonic which was high enough in frequency to be F_1 and higher in amplitude than F1, (2) a low frequency F2 which was confused with F1 because the former was higher in amplitude than F1, and (3) an F1-F2 combination peak which might appear as F1 only or F2 only.

A somewhat complicated subdivision of the F_1-F_2 plane, into rectangular regions, resulted in which "as many as six vowels could have the same F_1 and F_2 locations." In order to eliminate confusion among vowels having similar F_1-F_2 configurations, sets of "confusion-elimination" operations were devised using empirically determined thresholds based upon ratios of areas under arbitrary regions of the spectral cross-sections. These measurements were an attempt to more accurately determine the formant frequencies and the authors remarked that "information about true formant locations can be obtained more reliably from measurements of the type used here than from measurements of peaks using a formant tracking technique." The final technique was to locate F_1 and F_2 approximately, resolve confusions associated with 9 of the 11 F_1-F_2 combinations and hence identify the unknown vowel. The overall performance of the system for 21 subjects (11 male and 10 female) was 88% correct classification. Application of vowel duration information (e.g. [H-20]) raised the average score to 93% correct. The Forgies concluded that "the development of the recognition process in the form of a tree, where rough operations are followed by more detailed ones 'tailor made' for the particular confusions which remain, results in a comparatively efficient program since only applicable operations need be executed in any particular case."

A question which might have been relevant to this investigation is "Can a precise F_1-F_2 mapping provide sufficient information for accurate vowel recognition?" In other words, the transformation stage has been overlooked completely and the decision process appears needlessly arbitrary and unjustifiably confusing.

J.D. Foulkes questioned the sufficiency of raw F_1-F_2 data for automatic classification of vowels [F-12]. He noted that Welch and Wimpress had shown [W-5] the necessity of retaining data concerning F_o, the voicing frequency, and F_3, the third formant frequency if maximum separability using objective techniques is desired. Foulkes therefore applied a series of transformations to the raw data. Figure 4.1, from [F-12], shows the scatter diagram of F_1 vs F_2 for isolated vowels as measured by Peterson and Barney [P-11]. Foulkes observed that the dotted lines in Fig. 4.1

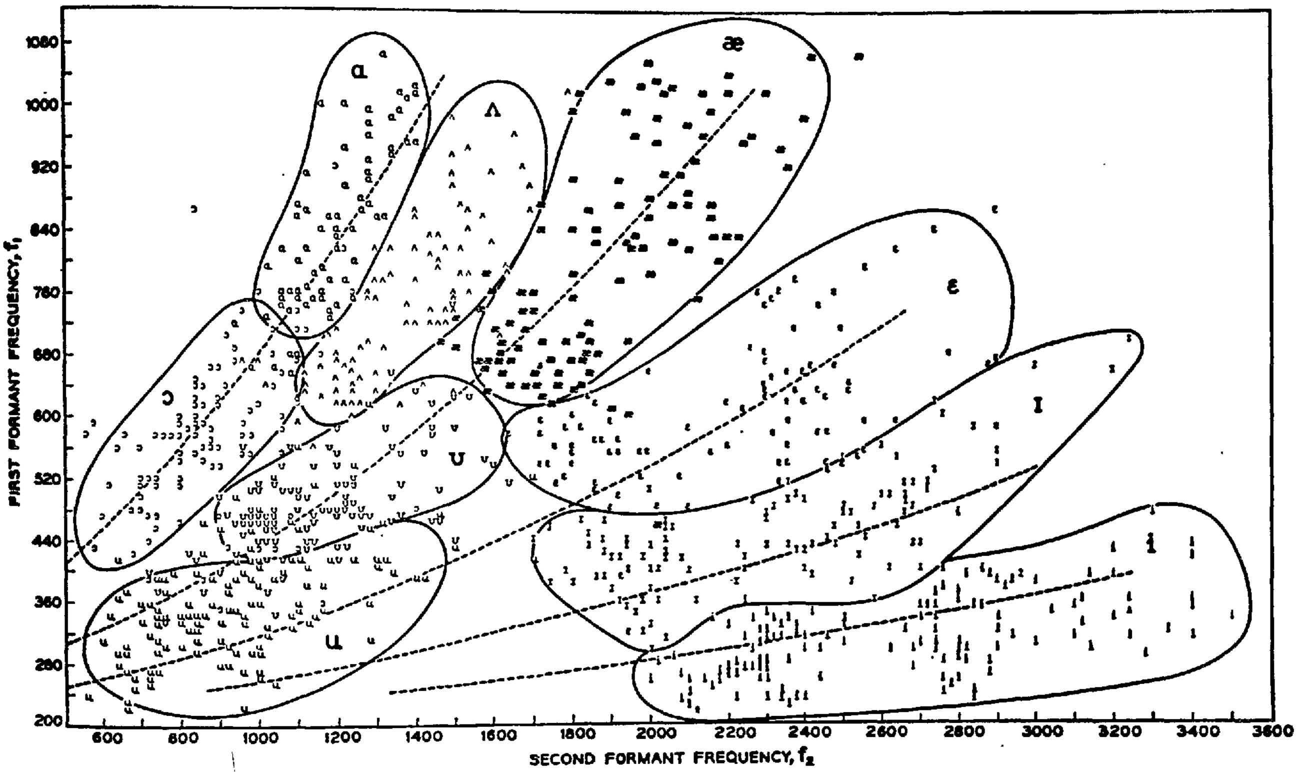


Fig. 4.1 Plot of f_1 [F_1] vs f_2 [F_2] for nine vowel types. From [F-12], using the data obtained by Peterson and Barney [P-11].

are members of a one parameter family of parabolas with a common origin at $F_1 = 200$ Hz and $F_2 = -500$ Hz. Using the coordinate translation

$$x = F_1 - 200 \quad \text{and} \quad y = F_2 + 500$$

he transformed x and y into a and b, as shown in Fig. 4.2.

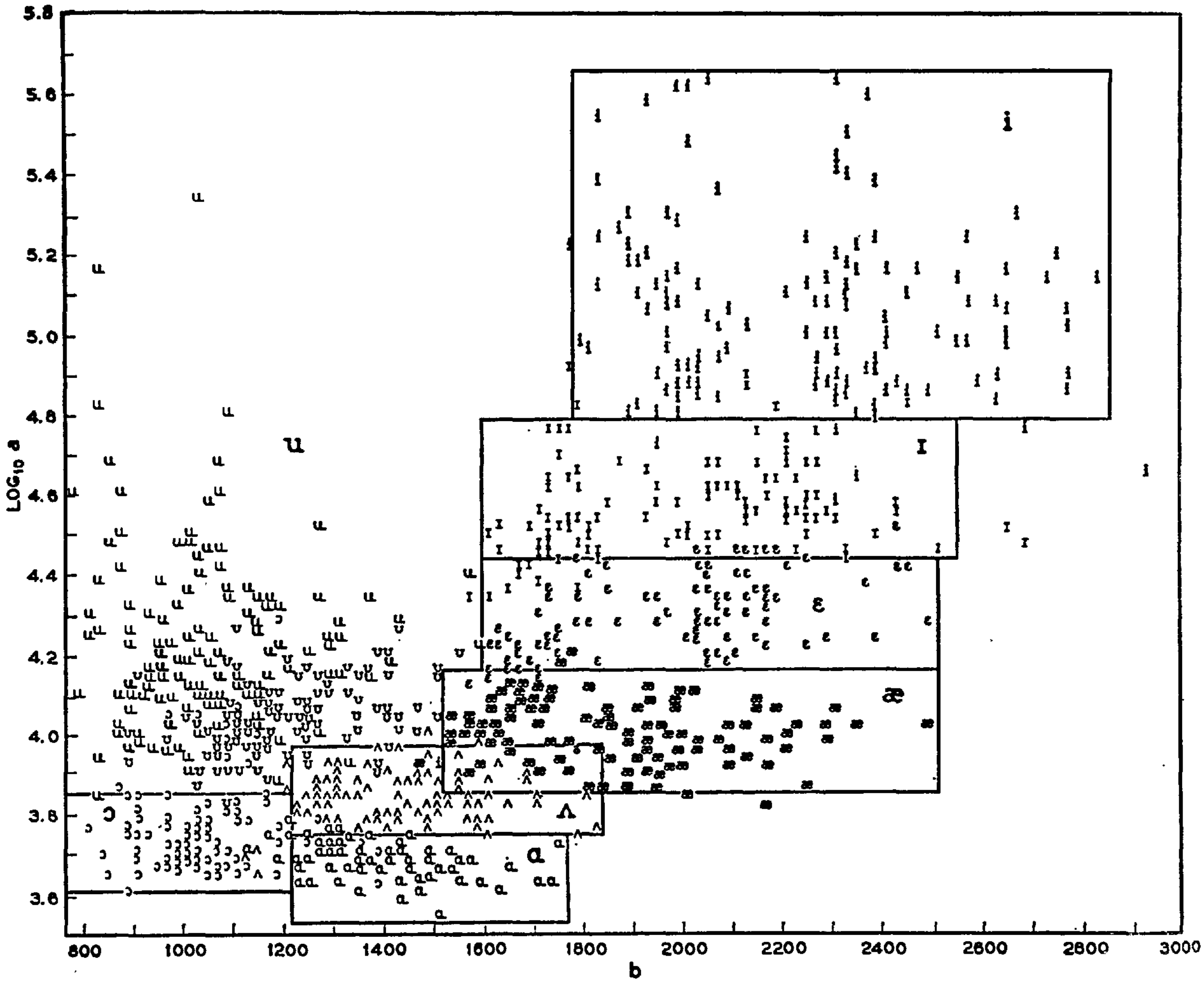


Fig. 4.2 Plot of $\log_{10} a$ vs b for nine vowel types. (From [F-12].)

We note that the isophonemic regions of Fig. 4.1 have become roughly rectangular in Fig. 4.2. However, there is still overlap, especially of /ʌ/ and /æ/. Foulkes therefore used data concerning F_o to apply a correction, transforming b to B. The result, shown in Fig. 4.3, eliminates most of the overlaps.

Finally, F_3 data can be used as a further correctional factor in a manner similar to that used to incorporate F_o. The total effect of the transformations is to substitute a simple matrix representing the boundaries in Fig. 4.3 for the extremely large table which would be required to describe those in Fig. 4.1. The penalty paid is the time required to effect the transformation.

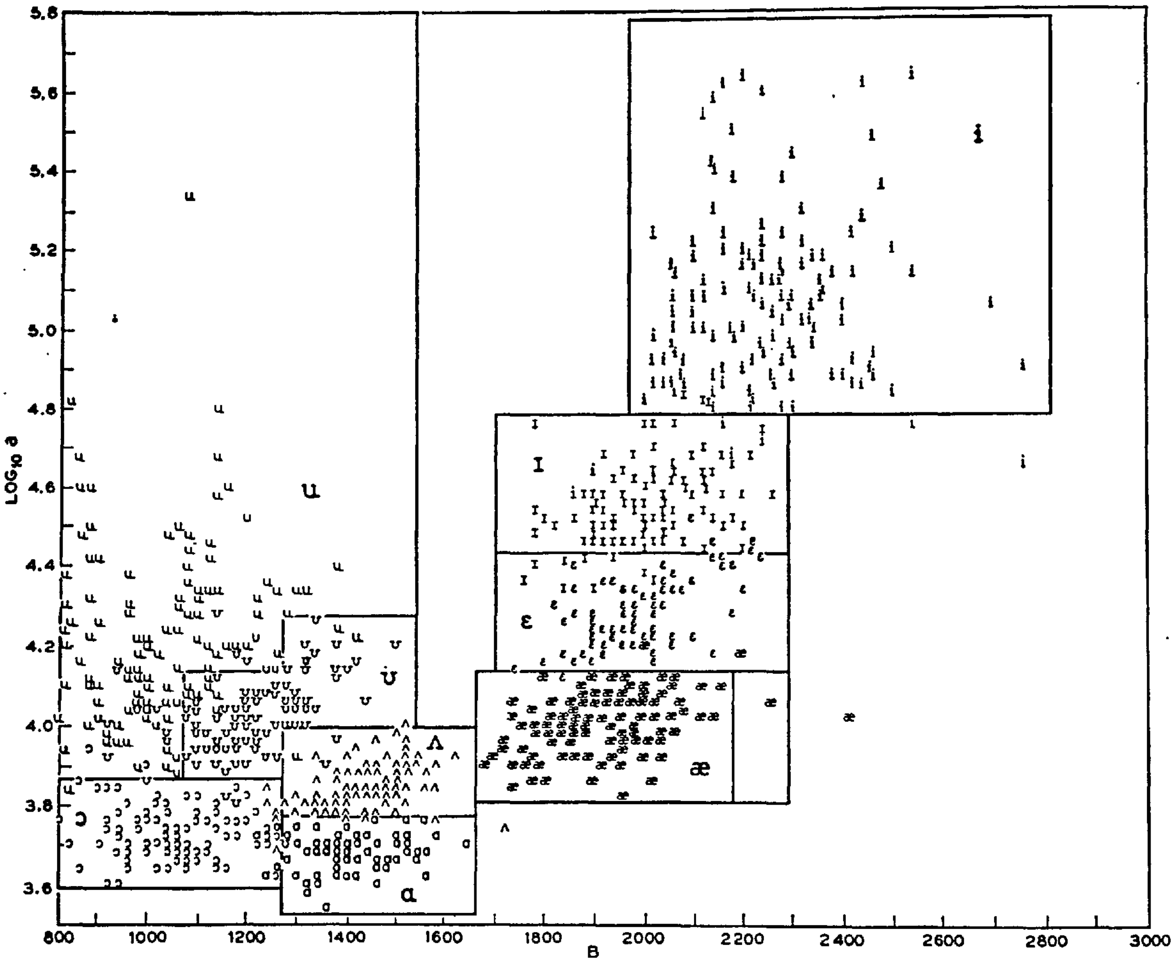


Fig. 4.3 Plot of $A = \log_{10} a$ vs $B = 800 + 320(b-800)/(F_o+120)$ for nine vowel types. (From [F-12].)

Recognition results using the transformed data were 88% correct. This exchange of computing time for store is, Foulkes noted, the "sole justification for the transformations . . . and the temptation . . . to speculate on the subjective significance [of the transformed parameters] . . . is worth resisting."

We have outlined the techniques employed by Forgie and Forgie, and by Foulkes, in order to emphasize the difference between recognition procedures using *raw data* and those using *transformed data*. Forgie and Forgie considered the preprocessing to end with spectral analysis. Their classification program was relatively complicated and the amount of data storage space required quite large. Foulkes transformed the input data and employed a relatively simple final classification criterion. Both efforts yielded precisely the same average rate of correct classification.

4.3.2 Word Recognition

H. Dudley and S. Balashek described a word recognition machine conceived as an extension of *Audrey*, an early (1952) spoken digit recognizer [D-6] which--since it used zero crossing information--will be reviewed in chapter 6.

Dudley and Balashek initiated their analysis with a set of 10 contiguous bandpass filters [D-14]. The filter outputs were envelope detected and the "patterns" thus generated then effectively cross correlated with a set of stored reference patterns derived by prior experiment. A continuous indication appeared at the output of this "phonetic pattern recognizer" and indicated which of six vowels /i,I,ε,a,o,u/, a semi-vowel /r/, a nasal /n/, or two fricatives /f,s/ was present. The next stage was a "word pattern recognizer". During a learning phase, the duration of each of the ten phonetic patterns was observed as each of the ten digits was spoken repeatedly. A resistor matrix was constructed--taking account of the variation in nominal duration among the spoken digits but not the sequence of appearance of phonemes within each spoken digit--such that the actual digit spoken is correctly identified via a capacitor charging operation. Dudley noted that in actual tests "the operation was invariably successful [i.e. correct more than 90% of the time] when the apparatus had been adjusted to the speaker's voice and he was careful to utter the digits just as he did in setting up the memory patterns."

We wish to emphasize two points concerning the results of this experiment:

First, in contrast to the vowel classifiers described in the previous section, this scheme was successful only for a single speaker--the speaker whose voice set up the machine. It seems probable that the explanation of this discrepancy (all schemes use spectral data) lies in the fact that, while the Forgies concentrated on defining differences and similarities between *significant spectral features* (e.g., formants), Dudley attempted a more generalized approach which seems to ignore spectral structural detail except in a general sense. That is, the resistor matrices treat all spectral areas with equal priority.

Secondly, the attention given to the duration information is not justified in view of the insufficient analysis performed to discover phoneme identity at a spectral level.

P. Denes and M. Mathews were among the first to programme the classification phase of an automatic speech recognition machine [D-11]. Although they were aware that "automatic speech recognition is probably possible only by a process that makes use of information about the structure and statistics of the language being recognized" they felt that "by restricting the library of words . . . to the relatively small number of 10, the acoustic redundancy of the speech waves will be increased to a level where linguistic information is no longer required for successful recognition."

The source data for their recognizer also consisted of the envelope detected output of a filter bank, 17 channels in this case. Sixty sweeps of the filter bank outputs (≈.85 sec) yielded 1020 analogue samples, each subsequently quantized and represented by a 10 bit number. Reference patterns were formed by adding together corresponding array points (after time normalization) from a group of utterances of the same digit and then normalizing so that the sum of the squared point values in each reference array equals unity. Recognition was accomplished by cross correlating input patterns with each reference pattern. The results were quite similar to those of schemes previously mentioned: correct recognition (classification) of words spoken by the speaker whose utterances were used to form the reference pattern set averaged greater than 90% while the rate of errors increased to 33% for other speakers.

P. Sholtz and R. Bakis dispensed with all analogue apparatus and inserted digitized speech directly into a computer [S-13]. However, the first computed operation was simulation of a filter bank (40 channels) giving a spectral cross section output every 10 milliseconds. The next step, the first in the recognition process, involved a vowel--non-vowel decision using energy considerations. Segmentation into phoneme strings was accomplished by observing changes in the spectral cross sections. Those segments deemed 'non-vowels' were further classified by means of an elaborate tree structure which incorporated many of the known time-frequency characteristics of speech sounds (chapter 3). Vowels were similarly separated into one of 11 categories using spectral energy measurements, time variation of spectral information and durational characteristics. Following word termination, the sequence of classified segments were referenced to a "dictionary", constructed during the learning phase, and the word identified or rejected.

The overall performance of this system was 96% correct recognition, 1.7% incorrect classification and 1.8% rejection. The authors emphasized that it is difficult to draw conclusions from these comparatively successful results but note that their procedure seems to be "more tolerant of interspeaker variations than other . . . procedures previously reported."

A final example of spoken word recognition using spectral primary data is the experiments of King and Tunis [K-6]. They claimed that their work "extends the results existing in the literature in that it deals with significantly larger sample sizes than have commonly been used, with a limited number of different vocabularies, and with the effect of transformations of the primary measurement space on recognition performance."

This scheme also commenced with envelope detection of the outputs of a set of (fifteen) contiguous bandpass filters. However, prior to sampling by a computer, an analogue ANDing operation sensed peaks in the spectral corss sections. The result was a record of the formant positions only. A separate highpass circuit detected energy associated with unvoiced sounds. The training and recognition algorithm used was a basic linear, adaptive decision function; this class of recognition algorithms will be considered in chapter 7.

King and Tunis are unusual in that they actually explicitly presented a rationale for their methods. "The hypothesis has been made," they stated, "that the spectrum analysis of a speech waveform provides measurements that contain, if they are not themselves, statistically invariant measures of the spoken words." The correct recognition rate for each of two 15 word vocabularies was greater than 97% for testing using the same speaker during algorithm training and recognition phases. An attempt to recognize words spoken by a person other than the 'trainer' resulted in a drop in correct recognition to 55% and 85% in two separate tests. Mixed training (samples from two speakers) raised the recognition rate to 99%.

We now summarize the results of the experiments described: Features extracted from short-term spectral analyses of speech appear to be sufficient only for recognition of a limited vocabulary. Training, or setting up of the machine, requires a vocabulary sample drawn from more than one speaker if multiple speaker recognition is to be successful. Recognition can be accomplished through cross-correlation with a set of master patterns [D-14], [D-11], decision trees based upon known time-frequency characteristics of speech sounds [S-13] or via adaptive classification algorithms [K-6].

To close this section, we note that W. Hillix achieved a high rate of spoken digit recognition using "nonacoustic measures" of speech information. These nonacoustic measures include lip and jaw movements and "wind velocity" in the vicinity of the mouth [H-13], [H-14].

4.3.3 Automatic Recognition of Continuous Speech

At this time (1969) only one significant attempt at continuous speech recognition has been reported in the literature. D.R. Reddy first described one solution to the problem of achieving primary sementation of continuous speech [R-4]. His techniques were determined "in an *ad hoc* way by the visual inspection of the waveform." The speech waveform--sampled, quantized and inserted directly into a computer--was divided into a succession of *minimal segments* using the variation or stability of sound intensity levels, with zero crossing counts used as an aid in resolving ambiguities and in error correction[2]. Minimal segments of similar characteristics were later combined to form larger segments and these, in turn, could be classified as sustained or transitional segments.

In a later paper [R-5], Reddy reiterated the problems encountered when a one-to-one correspondence between phonemes and their acoustic representation is attempted. He noted that Sanskrit grammarians often consider allophones (variant forms) of certain phonemes to fall into different phoneme classes. In English, for example, /f/ and /θ/ are often acoustically closer to stops than to fricatives. This occurs when the turbulent airflow is deemphasized. And, as noted in section 3.4.5, except for the coupling of the nasal passage the vocal tract configuration for nasal murmurs is close to that of stops. It is therefore imperative, Reddy noted, that "any grouping scheme for automatic speech recognition that is mainly dependent on the acoustic parameters for its classification cannot require that a given phoneme belong to one and only one phoneme group" and that "the grouping should be such that the acoustic parameters required for associating segments with a phoneme group are few and easily obtainable." Reddy's scheme was to group the sounds into four nonmutually exclusive subsets--stoplike sounds, fricativelike sounds, nasal-liquidlike sounds and vowellike sounds. The actual method of classification into the subsets was quite complicated and was based upon intensity and zero crossing measurements. We emphasize that the criteria incorporated in the flow graph which constitutes the classification system were *ad hoc* derived from the known characteristics of speech sounds. The main value of this phase of Reddy's automatic recognition system is undoubtedly in his interpretation of nonexclusive phoneme grouping.

[2] This phase will be elaborated upon in chapter 6.

Reddy's complete system for computer recognition of connected speech (single speaker) was described in detail in 1967 [R-6]. He noted that "any attempt at simulating the approaches that require the use of filters would have required excessive computer time[3]" and that he therefore sought "new and different solutions to the problems of speech processing." The prime objective of the system was to obtain a phoneme string from continuous speech.

The system is an extension of the segmentation method described in his earlier papers. Spectral analysis aids in classifying the segments; formant amplitude and frequency are among the spectral parameters extracted. Zero crossing information supplemented the spectral information (sec. 6.3). Classification within each of the four subsets (stop-, fricative-, nasal-liquid- and vowel-like) was accomplished using a tree-like flow net. The criteria for branching within the nets were, as before, based upon observations concerning the time-frequency characteristics of speech sounds. The results of a test on 287 phonemes gave 81% correct segmentation and classification.

Reddy's system was based on an extensive knowledge of speech characteristics and judicious application of these properties to the design of flow (decision) nets. No fundamentally new methods of speech processing were used. Nevertheless, this scheme, above all others in the literature, seems to hold the most promise for success in the near future.

---

[3]The fast Fourier transform algorithm (sec. 2.5 and 8.5), which obviates this problem was published in 1965. A lag of nearly two years in adoption of FFT techniques followed.

4.4 Barriers to Successful Automatic Speech Recognition

We have briefly examined a number of partially successful attempts at automatic speech recognition. Most of these systems made use of the envelope detected output of a bank of contiguous bandpass filters (or a variation thereof) as the source of primary data. For narrow bandwidth filters this method of processing approximates *short-term spectral analysis* (sec. 3.4.1). This reveals features which can be interpreted in a physiologically meaningful and conceptually attractive manner. However, except for a single speaker, spectral features do not seem to possess sufficient invariance to serve as a *useful* measure of the acoustic waveform in automatic speech recognition machines. By *useful* we imply successful.

We now explore one of the major problems in this chapter: should we expect any automatic speech recognition machine to be successful on the basis of acoustic information alone? Fry warned [F-16] that, "It is no use . . . looking . . . for acoustic invariants which characterize each sound that occurs in a given language. A language is a system of relations, at the level of acoustic recognition as at other levels, and what characterizes a sound depends *entirely upon what other sounds it has to be distinguished from.*" (Italics mine.)

4.4.1 The Contextual Problem

D.B. Fry has repeatedly emphasized the inadequacy of acoustic information for automatic speech recognition. "In the case of the human listener," he explained [F-16], "the classifying is done on the basis of a vast store of knowledge about the language system, and such is the degree of redundancy of natural languages that the weight the listener attaches to the incoming acoustic information is low compared with the weight given to the stored linguistic information. It is only in this way that we are able to make sense of running speech." He observed that limited vocabulary speech recognition schemes are somewhat successful only because in such a small ensemble acoustic information may be significantly more important than linguistic constraints. Thus, the design criteria for a word recognition machine would certainly be a function of the number of words in the vocabulary.

Contextual relationships--a knowledge of language statistics in general and the sequential probabilities of phonemes in particular--appear to be a key to the human facility of continuous speech perception under conditions involving varying speakers and conditions. A mechanical speech recognizer incorporating a linguistic store and able to simulate the use of statistical information at various levels would "undoubtedly work successfully even if its acoustic recognition was far from perfect." (Fry and Denes; [F-17]) Why then is a large portion of this thesis (chapters 6,8,9 and 10) devoted to the investigation and clarification of the significance of a particular type of *acoustic* signal processing (zero crossing extraction) to automatic speech recognition?

Fry and Denes answered this question by stating [F-17]:

"It is clear that a certain level of accuracy in acoustic recognition is necessary if the use of a sequential probability is not to lead to an increase rather than a decrease in errors . . ." (Italics mine.)

4.4.2 The Future of Automatic Speech Recognition

In a discussion of problems relating to the study of language, N. Chomsky recalled the situation which prevailed in the speech recognition field only a few years after the introduction of the speech spectrogram [C-10]:

> The interdisciplinary conferences on speech analysis of the early 1950's make interesting reading today. There

> were few so benighted as to question the possibility, in fact the immediacy, of a final solution to the problem of converting speech into writing by available engineering techniques . . .
>
> . . . there is little trace today of the illusions of the early postwar years.

Chomsky feels that as far as "automata-theoretic models" for language use (and related problems in perception) are concerned, there is a fundamental inadequacy in the systems of concepts and principles that have been advocated. He cautioned that

> 'extrapolation' from simple descriptions of language processes cannot approach the *reality of linguistic competance*; mental structures are not simply 'more of the same' but are qualitively different from the complex networks and structures that can be developed by elaboration of the concepts that seemed so promising to so many scientists just a few years ago. *What is involved is not a matter of degree of complexity but rather of quality of complexity*. Correspondingly, there is no reason to expect that the available technology can provide significant insight or understanding or useful achievements; it has noticeably failed to do so . . . (Italics mine.)

If Chomsky is correct, then the possibility of immediate, large-scale success in automatic speech recognition using conventional analysis techniques seems remote indeed. Nevertheless, the task of knowledgeably exploiting the only easily accessible evidence of human speech communication--the acoustic waveform--requires that the significance of any measure of information extracted from the waveform be fully understood. Therefore, the remainder of this thesis--with the exception of two experiments in automatic speech recognition described in chapter 7--is concerned with exploring and clarifying the role of zero crossings in speech recognition and processing.

5 CLIPPED SPEECH I: PSYCHOACOUSTIC PHENOMENA

The central theme of this thesis is "the role of zero crossings in speech recognition and processing." "Recognition" is intended to encompass both human recognition--*perception*--and machine recognition--*classification*--; "processing" signifies those *operations* on the speech signal which precede the "recognition" phase. In order to provide a foundation for these investigations, we have devoted the introductory portion of this thesis to a review of the more fundamental concepts of signal theory (chapter 2), a detailed description of some aspects of the nature of speech and hearing (chapter 3) and an outline of ideas, problems and experimentation in automatic speech recognition (chapter 4).

We now propose to establish the link between *zero crossings* and *perception-classification* which provides the basis for the direction and parallel structure of this and the next chapter.

A rectangular waveform which switches polarity at each zero crossing (instant of zero pressure) of a speech waveform is intelligible. In this chapter we describe in detail the key experiments which established this result and delineate certain phenomena associated with the intelligibility of "clipped speech". We then review some attempts--using conventional signal theoretic ideas--to account for these phenomena. Zero crossings *per se* can be, and have been, considered as informational attributes of signals.

In chapter 6, after a brief discussion of the zero crossings of random processes, we review some of the key papers concerned with the value, nature and use of zero crossings in speech processing. Then, after establishing the basic characteristics of a zero-based signal model in chapter 8, we will apply this model to speech clipping phenomena (chapter 9) and the use of zero crossings as waveform descriptors (chapter 10).

5.1 Experiments Concerning the Intelligibility of Clipped Speech

We have seen that certain prominent spectral features (e.g., formants) appear to contribute to the intelligibility of speech in the following sense: manipulation of these features causes a change in the perceived identity of a speech sound. Shortly after the introduction of the speech spectrograph as a tool for speech analysis, J.C.R. Licklider, D. Bindra and I. Pollack[1] asked [L-13] the following questions: "Upon what characteristics of the speech-wave does intelligibility depend? Are certain characteristics of the speech-wave of paramount importance for intelligibility? Are other characteristics perhaps irrelevant insofar as intelligibility is concerned?" Licklider proposed to operate upon the speech waveform in an effort to eliminate irrelevant characteristics and thus reveal essential features. Peak clipping was chosen as the primary operator.

Mathematically, an infinitely clipped signal C s(t) can be defined in terms of the original signal s(t) by the following relationship:

---

[1]For convenience, we shall refer to *Licklider* as the investigator in describing the papers by Licklider, Bindra and Pollack [L-13], Licklider and Pollack [L-14] and Licklider alone [L-15].

$$C\ s(t) = \mathrm{sgn}\ [s(t)], \tag{5-1}$$

where

$$\mathrm{sgn}[x] = \begin{cases} 1, & x>0 \\ 0, & x=0 \\ -1, & x<0 \end{cases} .$$

That is, a rectangular waveform of absolute value unity and having the same polarity as the original signal is *interpolated* through the zero crossings of the original signal. Practically, we speak of degrees of peak clipping. The term *infinitely clipped* is applied to a signal which has undergone some minimum degree of peak clipping.[2] The degree of peak clipping, or clipping level in decibels, may be defined as

$$C = 20 \log_{10}(P_1/P_2) \tag{5-2}$$

where P_1 = peak value of original waveform

and P_2 = level of original waveform at which clipping takes place.

Progressive peak clipping of a signal is illustrated in Fig. 5.1.

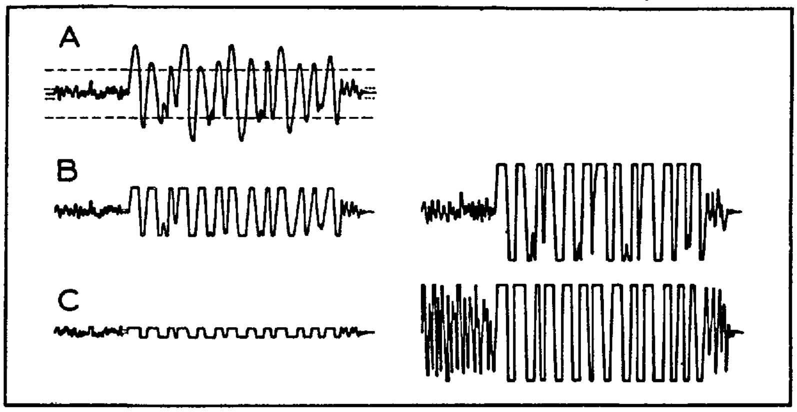


Fig. 5.1 Progressive peak clipping: A) original signal B,C) clipping at progressively lower signal levels. (From [L-13]).

[2] In practice, the highly peak clipped waveform is transformed into a truly rectangular waveform by a non-linear circuit (e.g., Schmidt trigger).

5.1.1 Licklider's Experimental Observations

Licklider's first experiments were designed to study the intelligibility of discrete words after the application of progressive peak clipping. For peak clipping less than 20 db, the articulation scores--the percentage of discrete words correctly identified--were greater than 96%. As the clipping level was increased, the articulation scores decreased; for clipping levels greater than 60 db ($[P_1/P_2] = 1000$), the 'word articulation score' vs 'peak clipping level' curve approached a minimum or flattened out (L1)[3]. This minimum varied from 50% for more difficult words to about 75% maximum. Licklider noted that 50% word articulation corresponded to about 90% sentence intelligibility for his tests, and that under these conditions, conversations could be carried on with little difficulty [L-13].

In order to prevent interword system noise from appearing at the output as clipped noise, a 25 KHz bias signal was added to the speech signal prior to clipping. The strength of this bias was such that clipped circuit noise was replaced by a 25 KHz inaudible square wave, and the speech signal was, ostensibly, unaffected.

Further tests involved the addition of white noise to the clipped speech signal. For comparison purposes, the original and clipped signals were made equal in peak amplitude. Figure 5.2 shows per cent articulation scores for various speech-to-noise ratios. It is apparent from these results that for low speech-to-noise ratios the clipped speech is *more* intelligible than the original speech (L2).

[3]For future reference, certain observations associated with observed phenomena will be labelled. The letter identifies the experimenter.

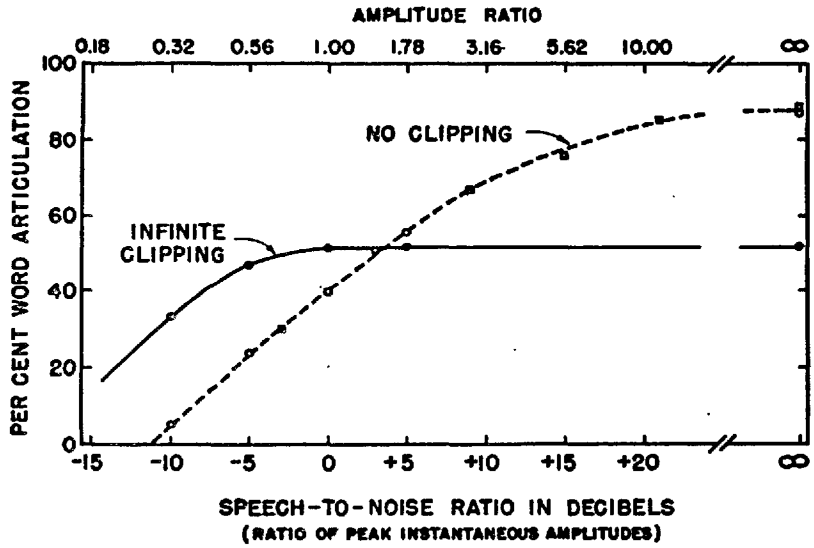


Fig. 5.2 Effect of added white noise upon the intelligibility of speech and clipped speech. (From [L-13].)

Licklider also noted that the frequency response of his record-playback system was uniform within ±5 db from 250 to 7000 Hz and that "severe peak clipping appears to be less deleterious if the low frequency components are supressed . . ." before clipping (L3).

During the conduct of these tests, over a period of 30 days, Licklider observed that the percent word articulation scores for both unclipped and clipped speech gradually increased (L4). The values for percent word articulation in Fig. 5.2 were the maximum noted. Although some of the improvement was attributed to the finite set of recorded words repeatedly used, introduction of new word sets showed that about 66% of the 'learning' (roughly 20 percentage points on the articulation scale) was indeed an increased ability to understand clipped speech. Licklider's analysis of the results also showed that the deleterious effects of clipping were least for more experienced subjects. In addition, the learning factor for the original speech plus noise was only apparent for intermediate noise levels.

In a further series of experiments [L-14], Licklider introduced "frequency-selective circuits" into the speech channel at various points. Specifically, a differentiator or integrator could be used to operate upon the original or clipped waveform. The differentiator introduced a 6 db per octave positive spectral tilt to frequencies between 1 and 16 KHz and the integrator a 6 db per octave negative spectral tilt to frequencies above 16 Hz. The following arrangements were used in word articulation tests:

1) No distortion--original speech
2) Differentiation only
3) Integration only
4) Differentiation + clipping
5) Differentiation + clipping + integration
6) Clipping + integration
7) Clipping
8) Clipping + differentiation
9) Integration + clipping
10) Integration + clipping + differentiation

A total of 250 word articulation tests were made: 25 with each of the 10 arrangements, 10 with each of 5 scramblings of 5 phonetically balanced (PB) word lists. The results of these experiments are summarized in Fig. 5.3 (a repeat of Fig. 1.1) and can be divided into four operational groups:

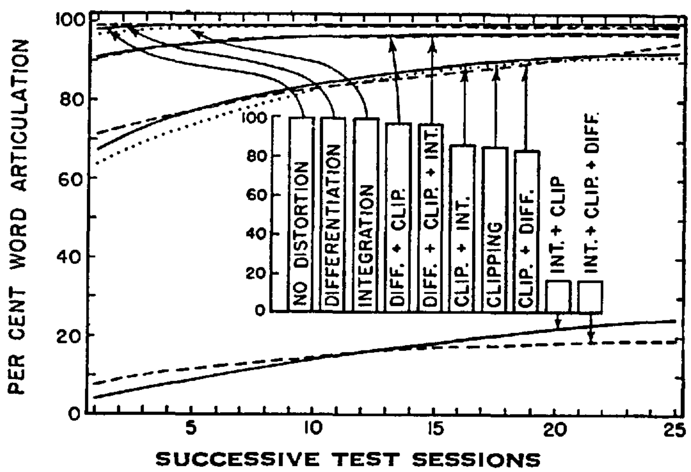


Fig. 5.3 The effects of various combinations of differentiation, integration and infinite clipping upon word articulation. The heights of the column diagram indicate the overall average for each of the ten arrangements. (From [L-14].)

Speech processed using the arrangements of the first group, (1,2,3)--all of which do *not* involve clipping--had virtually 100% intelligibility. However, Licklider emphasized that "this result concerning their intelligibility is in marked contrast to the observations concerning their naturalness and timbre. Differentiation, because it greatly emphasizes the fricative consonants and weakens the low pitched vowels makes the speech sound overly crisp. Integration emphasizes the low pitched vowels, weakens the consonants, and makes the speech sound muffled and 'boomy'."

The second group of operations (4,5)--both members involving differentiation *before* clipping--resulted in articulation scores of over 90%, "even for unpractised listeners (L5)." The effect of post-clipping integration was to improve intelligibility slightly (L6).

Group three (6,7,8) all involved clipping as the initial distorting operation. We record Licklider's impressions of the quantitative results shown in Fig. 5.3: ". . . it is evident that the process that follows clipping has but little effect on *intelligibility* (L6) and again it is true that the articulation scores fail to reflect differences in quality and timbre that are quite striking to the listener. The . . . integrator makes the effect of infinite clipping *sound* less noticeable . . . the differentiator made the clipped speech *sound* even worse . . . ." (Italics mine.)

The final group (9,10), involved pre-clipping integration and produced such subjective distortion that it was pronounced "incompatible with clipping." (L6)

The same learning effect observed in Licklider's first experiments appeared here. He noted that "the skill developed by the listeners during the tests is . . . only in part specific to the words of the test vocabulary. It is to a considerable extent a general skill, an ability to identify words correctly despite severe distortion." In discussing the value of an ultrasonic bias in eliminating interword noise, Licklider cautioned that "if the intensity of the speech is not well above that of the ultrasonic tone, there is danger that a spurious effect, a 'duty-cycle modulation' of [the] ultrasonic rectangular waves, would make the rectangular waves [clipped speech] more intelligible than they would be with infinite clipping *per se*." (L7)

In a final set of experiments [L-15], Licklider investigated the effects of quantizing the time scale in clipped speech. This process allows the rectangular waveform to switch polarity only at "predetermined instants." The following switching rules were formulated: The output waveform (rectangular) switches polarity at the end of a time interval if, during the interval, the input speech waveform has--rule A--one or more zero crossings or--rule B--an odd number of zero crossings. Word articulation tests were carried out using pre-clipping differentiation and post-clipping integration. The results of the tests are shown in Fig. 5.4. Licklider noted

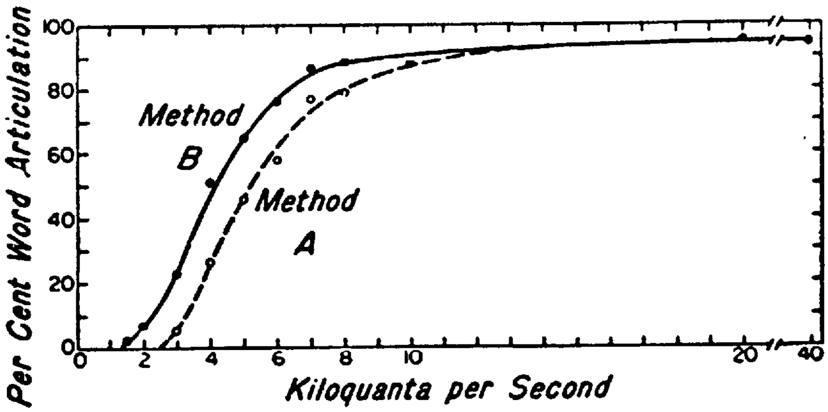


Fig. 5.4 Results of articulation tests. For the two methods of time scale quantization, average word articulation scores are plotted against the number of thousands of quanta per second. (From [L-15]).

that "with fewer than 2000 quanta per second, the listeners understood essentially nothing. The quantized speech sounded like an impure tone in the case of method A or like static in the case of method B (L8) . . . with either method vowels were the first to become intelligible . . . .

. . . the amplitude and time quantized speech sounded worse than the articulation scores suggest . . . and considerable training was required before . . . the level of proficiency . . . [observed was attained]."

We will now summarize Licklider's most significant experimental observations regarding the intelligibility of clipped speech:

L1. Progressive clipping: Increasing the clipping level on a speech waveform results in decreased word articulation scores. For infinite clipping (C>60 db) minimum word articulation scores of 50%, corresponding to 90% sentence intelligibility, were observed.

L2. Addition of noise: *In the presence of white noise*, clipped speech is more intelligible than the original speech for small speech/clipped speech-to-noise ratios (< 4db).

L3. Highpass filtering: Severe (e.g., infinite) peak clipping is less deleterious to intelligibility if the original speech is filtered so as to remove low frequency components.

L4. Learning: Repeated exposure to clipped speech enhances a subject's ability to understand it.

L5. Pre-clipping differentiation: Pre-clipping speech differentiation results in higher word articulation scores (> 90%), even for unpractised listeners.

L6. Post-clipping integration or differentiation: Integration of the clipped waveform has little effect on intelligibility but greatly improves the quality of the signal. Similarly, differentiation of the clipped waveform has little effect on intelligibility but worsens the quality.

L7. Ultrasonic bias: Unless the level of an ultrasonic bias--applied to the speech waveform before clipping--is "small" compared to the speech signal level, the resultant clipped speech will be more intelligible than it would be *per se*.

L8. Time quantization: For rule A or B, a quantization interval of 0.1 millisecond or less does not impair intelligibility of the clipped waveform but does cause degradation in quality. Quantization intervals less than 0.5 milliseconds results in an "impure tone" (Rule A) or "static" (Rule B) for speech input.

5.1.2 Licklider's Conclusions

Licklider offered explanations for some of the observed clipped speech phenomena:

L1. Progressive clipping: Licklider stated that "instead of asking why infinitely clipped speech is not as unintelligible as its wave-form would suggest, it is probably better to compare an intensity-frequency-time pattern [i.e., short-term speech spectrogram] of infinitely clipped speech with a corresponding pattern of normal speech." He did this and observed that "although many details of the pattern are changed by infinite peak clipping, the general . . . structure . . . is by no means rendered unrecognizeable. . . . only the details of the intensity-frequency-time pattern are modified."

L3. Addition of noise: Licklider asked, "What characteristic of square speech gives it an advantage over normal speech at low speech-to-noise ratios?" He quite rightly noted that clipping--by virtue of its rectangular interpolating waveform--distributes the power equally among the consonants and vowels, whereas in normal speech the consonants are relatively weak and therefore easily masked by noise. However, as the speech-to-noise ratio increases, the power advantage of clipped speech is balanced by the deleterious effects of distortion and, since more of the weak consonants pass the masked threshold, the ordinary speech becomes the more intelligible.

L8. Time quantization: Licklider noted that for long quantization intervals the probability that the speech waveform has at least one zero crossing approaches unity whereas the probability that it has an odd number of zero crossings is "in the neighbourhood of 0.5." Therefore rule A yields an impure tone (one output polarity change per time quanta) and rule B yields a "noise" (probability of polarity change in time quanta 0.5). The degradation in quality of time-quantized clipped speech over clipped speech, even for small quantization intervals, was--Licklider suggested--probably due to the fact that the reciprocal of the time quantization interval is usually unrelated to the fundamental frequency of voiced sounds.

In summary, Licklider suspected that the high intelligibility of clipped speech could be attributed to *overall preservation of the speech amplitude-power spectrum structure*. He offered no explanation for this preservation nor did he prove that it always did occur. Explanations for the other phenomena (L2,L4,L5,L6,L7) were not suggested.

5.1.3 Ahmend and Fatechand

R. Ahmend and R. Fatechand extended Licklider's experiments by examining the intelligibility (percent articulation) of vowel and consonant *segments* after differentiation or differentiation *and* clipping [A-2]. We shall list the effects observed:

A1. Initial consonant suppression: The removal of the initial consonant of a consonant-vowel-consonant (CVC) word has little effect on *vowel* recognition for either the normal or clipped versions.

A2. Final consonant suppression: Provided the initial part of the vowel portion of a vowel-consonant (VC) word is present (≃ 40 msec. gives 80% articulation of unclipped VC words), the presence of the final consonant does not materially alter the articulation of the original, or the clipped, vowel. In all cases, the articulation of the clipped vowels was less than that of the unclipped vowels.

A3. Initial part of vowel suppressed: If the initial part of a VC word is suppressed (for less than 100 msec.), there is little impairment of the percent *vowel* articulation. As the suppression time increases, anomalous effects are noted. Both clipped and unclipped /a/ and /ɔ/ remain highly intelligible until almost the entire vowel is deleted. The articulation of /o/,/u/ and /i/, however, falls rapidly even while a reasonable portion of "vowel" remains. We note, for future reference, that (see Fig. 3.8b) /a/ and /ɔ/ are the only vowels having substantially less than an entire octave between F_1 and F_2 while /u/ and /i/ have, respectively, 1½ and 3 octaves between F_1 and F_2. Ahmend and Fatechand concluded that, since the first 40 msec. of a VC word always provides high intelligibility (A2) "it would seem that the ends of the vowels, as modified by the final consonants [including transitions], provide much poorer recognition clues than 'pure' portions of equivalent lengths."

A4. Clipped consonants: The experimenters found that clipped *initial consonants* are not only less intelligible, but are also more susceptable to a degradation of intelligibility due to duration shortening. Clipped *final consonants* also appear to contain less redundant information than their unclipped counterparts.

The experimental evidence presented by Ahmend and Fatechand suggests, therefore, that clipping may cause *both* a decrease in the intelligibility of speech sounds *and* a decrease in the resistance of the speech sounds to degradation of intelligibility by alteration of durational cues. Clipped consonants, particularly, appear to lack some perceptual cues which, though normally of little use, are needed for identification purposes when durational information is destroyed.

5.1.4 Ainsworth

W. Ainsworth augmented Licklider's findings by investigating the intelligibility of *transforms* of clipped speech [A-3]. These transforms include:

1) the clipped waveform itself

2) pulses (delta function approximations) which indicate the occurence and direction of each zero crossing

3) pulses of the same polarity at all zero crossings

4) pulses which indicate only the zero crossings in one direction

5)-8) same as 1)-4) but using the zero crossings of the differentiated waveform.

Following the convention established in section 5.1, we can represent the signals used by Ainsworth as:

$$1)\ s_1(t) = C\ s(t) \tag{5-3}$$

$$2)\ s_2(t) = \pm\ s_1'(t) = \pm\{\sum_i \delta(t-\tau_i)\cdot(-1)^i\} \tag{5-4}$$

$$3)\ s_3(t) = \pm|s_2(t)| = \pm\{\sum_i \delta(t-\tau_i)\} \tag{5-5}$$

and

$$4)\ s_4(t) = \pm\{\sum_{i\ \text{odd}} \delta(t-\tau_i)\}\ \text{or}\ \pm\{\sum_{i\ \text{even}} \delta(t-\tau_i)\}\ . \tag{5-6}$$

Signals 5)-8) parallel signals 1)-4) with s(t) replaced by s'(t). Here $C\ s(t) = \text{sgn}\ [s(t)]$ and τ_i is the time of occurence of the i^{th} zero crossing. Figure 5.5 summarizes Ainsworth's results using standard PB word lists. The signals which retain zero crossing position and 'polarity' (i.e., signal goes from + to - or from - to + at a zero crossing) information (group 2) are the most intelligible, while the signals retaining only positional information (group 3) are the least intelligible. Signals consisting of pulses only at alternate zero crossings (group 4) have a percent word articulation between that of groups 2 and 3. The transformed signals derived from the differentiated speech are, in most cases, more intelligible than their counterparts derived from the original signal.

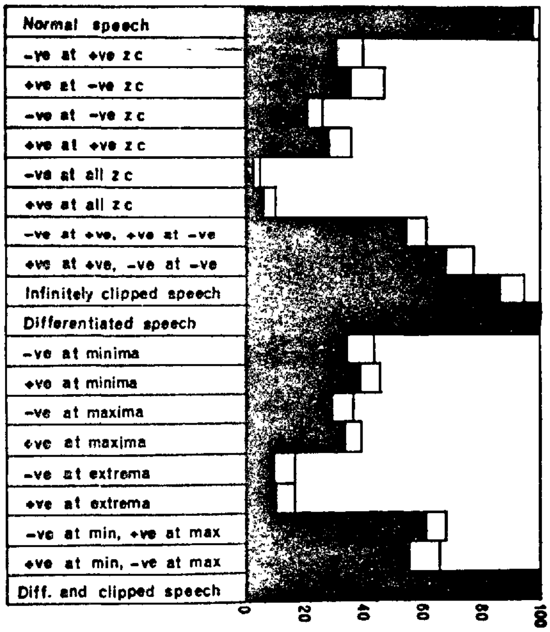


Fig. 5.5 Average (black bars) and standard deviation (white bars) of percent word articulation for normal and differentiated speech, and their clipped versions. *-ve at +ve zc* etc. means negative pulse at positive going zero crossing. (From [A-3].)

Ainsworth interpreted his results by analytically demonstrating that, if s(t) is a *sine wave*, then the clipped signal (a square wave) contains only odd order harmonics, $s_2(t)$ and $s_4(t)$ contain both odd and even order harmonics, $s_3(t)$ contains only even order harmonics and lacks a fundamental. A ranking according to number and/or type of harmonics correlates with the intelligibility results. For example, $s_2(t)$ and $s_4(t)$ have the most harmonic distortion and therefore should be least intelligible. The applicability of this analysis to speech clipping is somewhat dubious.

Finally, Ainsworth presented the results of experiments showing the confusion among clipped phonemes. He did not use an ultrasonic bias to prevent clipped noise in the silent intervals of stop consonants and he stated that this factor could have contributed to the observed confusion of voiced stop consonants and semi-vowels. Generally, in these experiments, vowels were least often confused with other sounds. However, Ainsworth further stated that "clipped vowels heard in isolation are not at all easy to recognize." Since the results of Ahmend and Fatechand are not referenced, we must assume that Ainsworth was unaware of these contrary findings (A2).

5.1.5 Thomas

I. Thomas' experiments were an investigation of the influence of F1 and F2 on the intelligibility of clipped speech [T-4]. He passed speech through one of two bandpass filters and clipped the resultant signal. One filter had minimum attenuation at the centre of the second formant frequency range for a male adult, ≈ 1500 Hz. Thomas noted that, for this filter, "spectrograms of the resulting clipped speech . . . show . . . that the first formant and voicing bands are entirely missing." However, the dynamic range of a spectrogram is only 12 db [P-17] and Thomas' "second formant filter" is 12 db down at 1200 and 2000 Hz; thus his claim that "only the second formant band and higher bands identifiable as its harmonics are present in the [filter] output" is highly suspect. Similarly, the observation that speech filtered by the "first formant bandpass filter" (centre frequency 500 Hz, attenuation ≈ 60 db per decade away from centre frequency) and then clipped, revealed only "occasional presence of [a] residual second formant" in *spectrograms* is inconclusive.

Thomas' two filters resulted in the following changes in formant amplitudes:

> <u>First formant filter</u>: F1, unchanged; F2, ≈ 20 db down; F3, ≈ 30 db down.
>
> <u>Second formant filter</u>: F1, ≈ 20 db down; F2, unchanged; F3, ≈ 20 db down.

His results showed an average word articulation score of 7.6% for speech passed through the *first formant filter* and then clipped. Speech passed through the *second formant filter,* and then clipped, yielded average word articulation scores of 71.1%. Thomas summarized his findings as follows:

> It is evident that [clipped] speech in which all formants but the second have been suppressed is still highly intelligible . . . . It is equally evident that speech in which all formants but the first have been suppressed is virtually unintelligible . . . it is [therefore] reasonable to attribute the high intelligibility of differentiated [then] clipped speech to the survival of the second (and possibly higher) formant frequency information through the clipping operation.

We remark here that, as will be noted in subsection 5.3.3, Vilbig [V-5] showed that clipping a predominantly F1 speech signal model yields distortion products which *must* fall in the frequency band below or within the F3 region--and therefore may mask any F2 or F3 present--whereas a predominantly F2 signal, when clipped, produces distortion products below the F3 region (≈ 3000 Hz) *only* for the vowels /ɔ/, /U/ and /u/. Considering the nature of the filtered, unclipped signals (i.e., first formant filter gives one formant 20 db down and one formant *30* db down while second formant filter gives two formants 20 db down) *and* the location of the distortion products produced by predominantly F1 or F2 signals, Thomas' conclusions regarding the importance of the second formant, *per se,*--"that the overall intelligibility of speech which has been subjected to amplitude distortion, frequency distortion . . . is largely determined by the extent to which second formant frequency information survives the distortion process"--are not justified.

Thomas attempted to further justify his "second formant theory" by referring to other experimental results. He noted, for example, that "intelligibility of speech which has been passed through either a lowpass or a highpass filter should [and does; see [L-7], for example] change from a very low value to a very high value as the passband of the filter is increased to *include* the entire second formant frequency range."(Italics mine.) However, such a signal then includes *both* F1 *and* F2 or F2 *and* F3. In another experiment [K-11], Thomas noted, "for a single bandpass filter of 500 Hz bandwidth, the highest articulation score is obtained when the passband extends from 1250 to 1750 Hz for a male speaker." Thomas neglected to point out that, first, this "highest articulation" is only 37% and second, that the *articulation* vs *centre frequency of passband* curve is double-peaked, with another peak of 32% occurring for passband 500 to 1000 Hz.

L.R. Focht noted [F-10], in describing a set of experiments in which the perceptual response of "all possible combinations of [three] formant amplitudes and frequencies were studies," that two formants were required to specify the perceptual value of a vowel and that "these two formants were not always the same pair but depending upon the perceived vowel jumped between combinations of the first, second and third formants." Therefore, although the second formant may be relatively important, we prefer to recall the results of Lehiste and Peterson's experiments on filtered volwels [L-7], that "one or more of the first three formants was found essential to . . . recognition."

In a further set of experiments [T-5], Thomas carried on Licklider's work on the perception of clipped speech in a noisy environment. Thomas showed that suppression of F1 prior to clipping increases post clipping intelligibility in the presence of noise. However, in this paper, Thomas correctly concluded that a predominant F1 is deleterious in that it creates clipping distortion products in the F2-F3 region. Thus, the high intelligibility of this clipped speech results from the suppression of F1 *relative* to F2 (below 700 Hz, Thomas' filter for these experiments is essentially a triple differentiator with a positive slope of 20 db per octave) rather than being due to the preservation of an ostensibly "most important" second formant. We shall clarify the correlation between spectral features and the intelligibility of clipped speech in chapters 8 and 9.

5.1.6 Rose

H. Rose's investigations were concerned with achieving maximum performance in clipped speech communication channels by determination of optimum combinations of spectrum shaping and clipping level as a function of relative levels and spectral shape of ambient noise at the speaker and listener positions [R-14].

For example, can we predict the percent articulation of a speech plus Gaussian noise signal which is clipped at an arbitrary level? To answer this question, Rose defined N_w as the average noise at the clipper output which can be attributed to the addition of noise to the speech signal *prior* to clipping. Physically, N_w is the output of a lowpass filter fed with the *difference* between the clipped *speech* signal and the clipped *speech plus noise* signal. By plotting S/N--the signal-to-noise ratio at clipper input--vs S_{out}/N_w--where S_{out} is the output signal level, and S/N vs AI, the articulation index (known from Licklider's experiments), a new curve of S_{out}/N_w vs AI can be determined. For other (non-infinite) clipping levels, Rose *claimed* that measurement of S_{out}/N_w will enable the articulation index to be predicted. He assumed here that both noise created by clipping and noise due to perturbation of speech zero crossings before clipping are effectively Gaussian. He did not, however, document his reasons for believing that "it is known that clipping . . . [of voiced sounds] . . . creates so many intermodulation products . . . that the added IM [intermodulation] noise power is essentially Gaussian . . . "

Rose also investigated the effects of pre-clipping differentiation and noise at the listening position on the intelligibility of clipped speech. The results presented may be valuable for predicting performance levels in clipped speech communications but do *not* offer any insight into the basic problem of explaining clipped speech-intelligibility phenomena.

5.1.7 Marcou and Daguet

P. Marcou and J. Daguet applied the tools of analytic signal theory to speech clipping-zero crossing studies [M-5]. They asked the following question:

If the phase-envelope representation of a speech signal is considered,

$$\text{i.e.,}\quad s(t) = |m(t)| \cos \phi(t) \tag{5-7}$$

then what perceptual information can be attributed to $|m(t)|$, the signal envelope, and to $\cos \phi(t)$, the phase function?

They presented a conceptually simple scheme for physically analyzing $s(t)$ into $|m(t)|$ and $\cos \phi(t)$: $s(t)$ is translated in frequency by a carrier of frequency ω_o using single sideband modulation. That is, as in (2-30), we consider

$$s_{\omega_o}(t) = |m(t)| \cos [\omega_o t + \phi(t)] \quad . \tag{5-8}$$

If $\omega_o >> 2\pi W$, where $s(t)$ is bandlimited to $\pm$ W Hz, then envelope detection of $s_{\omega_o}(t)$ yields $|m(t)|$ [S-3; p. 155] and infinite clipping of $s_{\omega_o}(t)$, followed by bandpass filtering--elimination of all frequency components for which $\omega_o - 2\pi W > |\omega| > \omega_o + 2\pi W$--and demodulation, yields cos $\phi(t)$. Marcou and Daguet reasoned that the latter result obtains since SSB modulation, with $\omega_o > 4\pi W$, assures that all harmonics created by clipping fall outside of the translated speech band. That this procedure does indeed yield cos $\phi(t)$ is shown by Sakrison [S-3; pp. 171-172]. Therefore,

$$D[BL\{C\ s_{\omega_o SSB}(t)\}] \simeq \cos\phi(t)\ ,\ \text{for}\ \omega_o >> 2\pi W\ . \qquad (5\text{-}8b)$$

D, BL and C are the demodulation, bandlimiting and clipping operators, respectively.[4]

Marcou and Daguet implemented this system and found that, for speech signals, "If $|m(t)|$ is used to drive a loudspeaker, an output is obtained which is essentially made up of a succession of loud and soft auditory impressions. If cos $\phi(t)$ drives the loudspeaker, the output gives *essentially the same aural sensation as the original signal* . . . ." (Italics mine.) That is,

> M1. "Single sideband clipping": The envelope of a speech signal is not perceptually recognizable as speech; the phase function of a speech signal is, perceptually, essentially the *same* as the original signal.

We shall complete our description of Marcou and Daguet's experiments with single sideband modulation in section 6.4.

---

[4]The approximation is due to the fact that cos $\phi(t)$--being an FM signal--is *not* strictly bandlimited [D-15], [S-3; p. 168]. Thus the bandlimiting operation necessary to eliminate clipping harmonics results in a deviation of the envelope of cos $\phi(t)$ from its nominal value of unity. The approximation becomes progressively better if the carrier frequency is increased so that it is much greater than twice the width of the band of frequencies over which the spectrum of cos $\phi(t)$ is appreciable in magnitude.

5.2 The Mathematics of Clipping as a Spectral Operator.

Licklider suggested that the high intelligibility of clipped speech is due to the overall preservation of the short-term speech power spectrum structure. In the next section, we will examine several attempts to quantitatively justify this statement. First, however, we will survey the methods used to predict the effects of clipping on the power spectrum of random and deterministic signals in general.

5.2.1 Random Processes

J.H. Van Vleck and D. Middleton remarked [V-2] that "the problem of determining the intensity spectrum of a disturbance subject to extreme clipping is closely related to that of finding the zero crossing points of the [time] axis) . . ." They showed that if a signal $s(t)$ is subjected to a limiter of transfer characteristic as shown in Fig. 5.6, then, if $s(t)$ is a wide sense stationary, Gaussian, random process with zero mean, autocorrelation function $R(\tau)$, and normalized autocorrelation function

$$\rho(\tau) = R(\tau)/R(0) \quad , \tag{5-9}$$

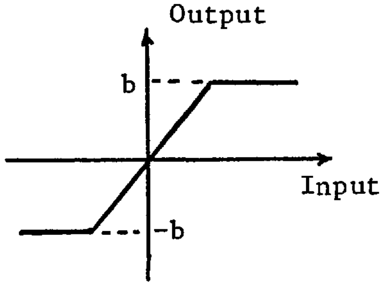


Fig. 5.6 Transfer function of a progressive clipper. (From [V-2].)

then the autocorrelation function of the output of the limiter is

$$R_y(\tau) = \rho^2(\mathrm{erf}(b/\sqrt{2})) + \sum_{n=1}^{\infty} \frac{\rho^{2n+1}(\tau)}{(2n+1)!} \left[H_{2n-1}(b)\cdot e^{-b^2/2}\right]^2, \quad (5\text{-}10)$$

$$\text{where} \quad \mathrm{erf}(x) = \sqrt{\frac{2}{\pi}} \int_0^x e^{-x^2}\, dx \quad (5\text{-}11)$$

and $H_n(x)$ is the Hermite polynomial,

$$e^{x^2/2}\cdot(-1)^n\cdot d^n(e^{-x^2/2})/dx^n. \quad (5\text{-}12)$$

If $b \to \infty$, then the limiter is a linear amplifier and $R_y(\tau) \to \rho(\tau)$, as expected.[5] If $b \to 0$, then

$$R_y(\tau) \to \frac{2}{\pi} b^2\cdot \sin^{-1}\rho(\tau) . \quad (5\text{-}13)$$

Normalizing (5-13) to unity mean square amplitude after clipping[6] gives

$$R_y(\tau) = \frac{2}{\pi} \sin^{-1}\rho(\tau) \text{ , the } \textit{arcsine law}. \quad (5\text{-}14)$$

The power spectrum G(f) of the output of the clipper is then obtained by using the relationship [W-2], [P-2, p. 240]:

$$G(f) = \mathcal{F}\{R(\tau)\} . \quad (5\text{-}15)$$

Van Vleck and Middleton applied (5-10), (5-14) and (5-15) to examine the effect of clipping on the shape of various input signal power spectra. For example, when a Gaussian process with a rectangular power spectrum of centre frequency ω_c and bandwidth $\pm\omega_a/2\pi$ Hz (white noise) is clipped by the limiter of Fig. 5.6 (b=0),

---

[5] The mean square amplitude of s(t) is normalized to unity before clipping.

[6] Divide (5-13) by $R_y(0) = b^2 - \sqrt{2/\pi}\cdot b\cdot e^{-b^2/2} + (1-b^2)\cdot \mathrm{erf}(b/\sqrt{2})$. $R_y(0) \simeq b^2$ for small b .

the output power spectrum is given by

$$G(f) = 2 \int_0^\infty \frac{2}{\pi} \sin^{-1}[(\sin \omega_a t/\omega_a t) \cos \omega_c t] \ . \quad \cos \omega t \, dt \ . \quad (5\text{-}16)$$

The shape of the "fundamental" component of G(f) after s(t) is clipped is given in Fig. 5.7, for various values of b. The

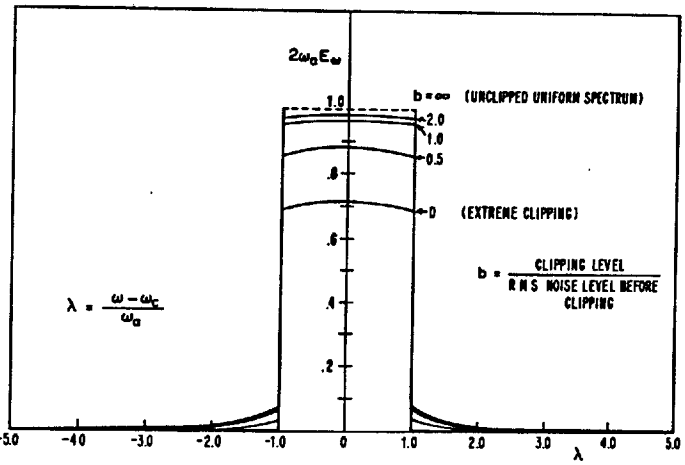


Fig. 5.7 The fundamental component of the post-clipping power spectrum of a Gaussian process. The total power in the spectrum in normalized to unity before and after clipping. (From [V-2].)

qualitative effect of infinite clipping (b = 0) on the original power spectrum is to diffuse a certain amount of power--31%--outside the limits $\lambda = \pm(\omega-\omega_c)/\omega_c$ and to make the power spectrum less uniform within these limits. Twelve percent (12%) of the diffused power is located in the "wings" of the "fundamental" power spectrum component (see Fig. 5.7).[7] The other 19% is located in harmonics

[7] L.R. Wilson has recently (1969) investigated the asymptotic behaviour of the "tails" of the power spectrum of the output of an infinite clipper when the power spectrum of the Gaussian input signal can be expressed as a rational fraction [X-3]. See also [X-4] and [C-1].

of the fundamental band, the first two of which are shown in Fig. 5.8.

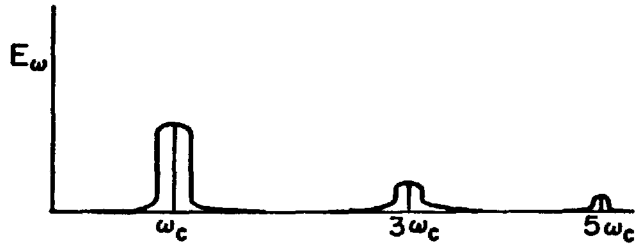


Fig. 5.8 The gross power spectrum structure of clipped white Gaussian noise. (From [V-2].)

In fact, it is shown [V-2; p. 14] that the distribution of energy among the harmonic bands and the fundamental band is exactly the same as occurs if a sine wave of frequency ω_c is clipped.

Derivation of the autocorrelation function of the output of a non-linear device--a clipper, for example--involves evaluation of the definite integral

$$
\begin{aligned}
R_y(t_1,t_2) &= E\{y(t_1),\ y(t_2)\} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x_1)\cdot h(x_2)\cdot f_{XY}(x_1,x_2)\ dx_1 dx_2\ , \quad (5\text{-}17)
\end{aligned}
$$

where $h(x)$ is the transfer function of the non-linear device ($h(x) = h_c(x) = \mathrm{sgn}[x]$ for an infinite clipper) and $f_{XY}(x_1,x_2)$ is the joint density function of the input signal.[8] The arcsine law, (5-14), for example, results from manipulation of (5-17) with $f_{XY}(x_1,x_2)$ a jointly Gaussian density function.

[8]For a wide sense stationary process $R_y(t_1,t_2) = R_y([t_1-t_2] = \tau)$.

In certain applications, where the transfer function h(x) has a simple Fourier transform, it is convenient to evaluate $R_y(t_1,t_2)$ using the "characteristic function method." It can be shown [T-6; pp. 284-285] that

$$R_y(t_1,t_2) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi_{AB}(\omega_1,\omega_2) \cdot H(\omega_1) \cdot H(\omega_2) \; d\omega_1 d\omega_2 \quad . \quad (5\text{-}18)$$

Here

$$\Phi_{AB}(\omega_1,\omega_2) = E\,\{e^{j(\omega_1 x_1 + \omega_2 x_2)}\}$$

$$= F^{-1}\{f_{XY}(x_1,x_2)\}$$

and $H(\omega{=}2\pi f) = F\{h(x)\}$. (5-19)

For the infinite clipper, $H_c(f) = -j/\pi f$. [S-10] . (5-20)

It follows immediately from (5-19) that the output of a non-linear device with input x can be expressed as

$$h(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(f) \cdot e^{j2\pi fx} \, dx \ ,$$

which, using (5-20), gives

$$h_c(x) = \frac{2}{\pi} \int_{0}^{\infty} \sin 2\pi fx/f \, df \quad (5\text{-}21)$$

$= \text{sgn}[x]$, for an infinite clipper [S-18].

5.2.2 Deterministic Signals

Equation (5-21) defines the output of an infinite clipper in terms of an infinite *integral* involving an arbitrary input "x". "x" may represent--as x(t)--a periodic signal, for example. F. Vilbig noted [V-5] that, for $|x| \leq \pi$, the output of an infinite clipper can also be expressed as an infinite *series*:

i.e.,
$$h_c(x) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin\,(2n-1)x}{2n-1} \quad (5\text{-}22)$$

W. Solfrey analyzed [S-18] the effect of clipping on the *members* of either a three-tone (two of equal amplitude) or a four-tone (amplitudes equal in pairs) complex under the assumption that the tone frequencies are incommensurable.[9] The signal model used was

$$s(t) = a \cos(\omega_1 t+\theta_1) + b\,[\cos(\omega_2 t+\theta_2) + \cos(\omega_3 t+\theta_3)]$$

where ω_1, ω_2, and ω_3 are incommensurable. This is inserted into (5-21) and expanded using Bessel functions. Solfrey's results showed that for b/a small (weak double input component), the output single component (at $\omega=\omega_1$) tends to amplitude $4/\pi$ while the output double component amplitude vanishes as $2b/\pi a$. For b/a large (weak single component) the double component amplitude tends to $8/\pi^2$ while the single component amplitude vanishes as $(2a/\pi^2 b)\cdot(\log_{10} 16b/a + \frac{1}{2})$. If b=a=1 both single and double output components have output amplitudes of 0.67. Thus, as a single component becomes greater in amplitude than the double component at the *input*, the effect at the output is to rapidly suppress the relative amplitude of the double component.

Using $s_{out}(t) = c\cos(\omega_1 t+\theta_1)+d[\cos(\omega_2 t+\theta_2)+\cos(\omega_3 t+\theta_3)]$, Solfrey defined a suppression ratio "γ",

$$\text{where } \gamma_1 = \frac{d/c}{b/a}, \qquad a/b > 1,$$

$$\text{and } \gamma_2 = \frac{c/d}{a/b}, \qquad a/b < 1.$$

He showed that γ_1 tends to a value of 2 for very large a/b. For b>a the suppression ratio γ_2, for large b/a, tends asymptotically to $\gamma_2[\text{db}] = -20 \log_{10}[0.818+0.576 \log_{10}(b/a)]$. This suppression

[9] A three-tone model for a vowel *could* satisfy this requirement provided that the fundamental, voicing frequency is not F_1.

is a *negative* suppression as it grows much more slowly than b/a. For example, when $b/a = 10^6$ [120 db], γ_2 is only -12.5 db. In effect, clipping then *enhances* the weaker single component with respect to the stronger double component for large b/a.

5.2.3 Summary

The methods available for analysis of the spectral effects of clipping appear to lack the power and generality desired for predicting, *qualitatively*, the spectral consequences of clipping. This is especially true for deterministic signals. In the next section, we review some attempts to apply these methods to speech clipping.

5.3 Why is Clipped Speech Intelligible?: Some Contemporary Viewpoints

This section is devoted to a detailed review of three attempts to explain the intelligibility of clipped speech in terms of Licklider's suggestion of overall power spectrum feature preservation.

5.3.1 Dukes

J.M. Dukes explained that the object of his paper [D-16] was "to examine to what extent the spectral content of the [clipped, and differentiated, then clipped] speech waveform . . . is similar *on the average* to that of the original signal. More important still, however, is the degree of coherence between the two spectra under consideration, i.e., the extent to which corresponding regions of the spectrum [spectra] are phase-related in a fixed rather than a random manner." Dukes stressed that his method is "only valid in so far as it relates to *averages over long periods of time*" and that "further work is still required to show what are the important invariants in the case of individual sounds." (Italics mine.)

In the first section of his paper, Dukes treated time-quantized, strictly-stationary, random signals completely specified by their first order probability density function--what he terms a *totally random signal*--and having zero mean. Dukes calculated the normalized crosscorrelation function between the input and (normalized) output of an infinite clipper and showed that it is

$$\begin{aligned} \rho_{xy}(\tau) &= \frac{2A_x}{\sigma_x}\frac{(\Delta t-|\tau|)}{\Delta t} && \text{for } |\tau| \leq \Delta t \\ &= 0 && \text{for } |\tau| > \Delta t \end{aligned} \tag{5-23}$$

where $$A_x = \int_0^\infty x \cdot f_x(x)\ dx \ ,$$

$$\sigma^2{}_x = \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\ dx \ ,$$

and Δt is the quantizing interval. Dukes noted that the cross correlation function $\rho_{xy}(\tau)$ is "a measure of the average in-phase energy of the two signals [input-output]" and defined $|\rho_{xy}(\tau)|_{max}$ as "the first coherence coefficient"-- μ. If A_x and σ_x are considered for the Gaussian and exponential distributions (representing, respectively, the long-term amplitude density functions for consonants and vowels, as noted in sec. 3.5) then

$$\mu_{xy} \text{ Gaussian} = \rho_{xy}(0) = \sqrt{2/\pi} = 0.798 \tag{5-24}$$

and $$\mu_{xy} \text{ exponential} = \rho_{xy}(0) = \sqrt{1/2} = 0.707. \tag{5-25}$$

Note that these results are independent of Δt, the quantizing interval. Dukes further showed that the coherence coefficient for the differentiated *clipped waveform* is

$$\mu_{xz} = \sqrt{2}\, A_x/\sigma_x \quad . \tag{5-26}$$

That is, post-clipping differentiation reduces the coherence coefficient by a factor of $\sqrt{2}$. Finally, the relationship between the normalized autocorrelation functions at clipper input $[\rho_{xx}(\tau)]$ and output $[\rho_{yy}(\tau)]$, the cross-correlation function $[\rho_{xy}(\tau)]$, and the first coherence coefficient can be expressed as follows:

$$\rho_{xy}(\tau) = \mu_{xy}\, \rho_{xx}(\tau) = \mu_{xy}\, \rho_{yy}(\tau) \quad , \; |\tau|<\Delta t \quad . \tag{5-27}$$

Therefore, the two autocorrelation functions and the cross-correlation function are identical in shape (for $|\tau|<\Delta t$) and differ only by a proportionality constant--μ_{xy}. Note that as $\Delta t \to 0$ (the continuous case), $\rho_{xy}(\tau)$, $\rho_{xx}(\tau)$ and $\rho_{yy}(\tau) \to 0$ except for $\tau=0$. Thus, nothing is *really* stated about the post-clipping shape of $\rho_{yy}(\tau)$.

The second section of the paper treats *partially constrained* time-quantized random signals; that is, signals whose density function at a point is conditioned by the preceding sample. Dukes noted that since the instantaneous output of a clipper is a function only of the instantaneous input, the first coherence coefficient μ is independent of statistical constraints between successive values of the input signal and is therefore unchanged. However, (5-27) no longer obtains and, in general, clipping a partially constrained signal may modify its power spectrum considerably.[10]

---

[10] R. Luce showed [L-25] that signals for which $\int_{-\infty}^{\infty} x_1 \cdot f_{XY}(x_1, x_2; \tau)\, dx_1 = P(x_1) \cdot Q(\tau)$ satisfy the relationship $\rho_{xy}(\tau) = k\, \rho_{xx}(\tau)$.

In his conclusions, Dukes emphasized that "the results only have significance in respect of very long samples [of speech sounds] and that with this formulation nothing can be deduced about the intelligibility of individual sounds, except . . . that deviations below the average must be relatively infrequent." He remarked that although *the values of the coherence coefficients calculated are near the overall intelligibility of clipped speech, "the principle difficulty is the unknown relationship between the coherence coefficients and intelligibility."* (Italics mine.)

5.3.2 Fawe

A. Fawe's paper [F-4] purports to include "a theoretical study of the phenomena [phenomenon] [that severely clipped speech is intelligible]." Yet Fawe almost immediately states that "whispered, as well as normal speech is intelligible after clipping; in this study we shall consider them only "since" voiced sounds that appear in *normal* speech are not easily described." (Italics mine.) In fact, Fawe strongly implied that he used whispered rather than normal speech for a model because he wished to apply the statistical theory of signals.

Fawe commenced his analysis of whispered speech by expanding the arcsine law, equation (5-14), into a power series and taking the Fourier transform of the result. That is

$$G(f) = \mathcal{F}\left\{R_y(\tau) = \frac{2}{\pi}\sin^{-1}\left[\frac{R(\tau)}{R(0)}\right]\right\}$$

$$= \frac{2}{\pi}\sum_{n=0}^{\infty}\frac{(2n)!\quad \underline{G}_{2n+1}(f)}{(n!)^2\cdot 2^{2n}\cdot(2n+1)\cdot[R(0)]^{2n+1}}\ , \quad (5\text{-}28)$$

where $$\underline{G}_{\underline{m}}(f) = \mathcal{F}\{[R(\tau)]^{\underline{m}}\} = \int_{-\infty}^{\infty}[R(\tau)]^{\underline{m}}\, e^{-j2\pi f\tau}d\tau\ . \quad (5\text{-}29)$$

Note that the first term in the summation--n = 0--in (5-28) is simply $\underline{G}_1(f) = \mathcal{F}\{R(\tau)/R(0)\}$, the power spectrum of the original signal. From this Fawe correctly concluded that "the infinite clipper adds a [spectral] noise and [also] suppresses the dynamics of the [Gaussian, random model for the speech wave." He also stated that "the [spectral structure of the] noise due to the clipping operator is very similar to [that of] the input signal; indeed, since $\underline{m}$ is odd, $R(\tau)^{\underline{m}}$ is like $R(\tau)$ and $\underline{G}_{\underline{m}}(f)$ like $\underline{G}_1(f)$." We agree in that $R(\tau)^{\underline{m}}$, $\underline{m}$ odd, has the same zero crossings and polarity as $R(\tau)$. Also, since the maximum value of $\rho(\tau) = R(\tau)/R(0)$ is unity, then $|[R(\tau)/R(0)]^{\underline{m}}| < |R(\tau)/R(0)|$.

Fawe gave an example for the clipping of *white, Gaussian noise*. Although the mathematics in (F-4) are very unclear, *some* valid conclusions were reached . He demonstrated that the real (apparent) noise power is only 15.8% of the expected noise power because "The [spectral structure of the] noise due to the clipping operator is very similar to [that of] the input signal [white Gaussian noise]."

He then extrapolated from these results for clipped, white Gaussian noise: "The [power spectrum] minimum [at $f=W_o$, 1.886] is about 5 percent below the [power spectrum] maximum [at f=0, 1.982]. Since the differential sensitivity of the ear for amplitude is 0.13 (or 0.26 for power) at a level of 40 db above threshold, the spectrum appears perfectly flat to the hearing mechanism, and *clipped* [whispered?] *speech* is highly intelligible." (Italics mine.) We would rather say that *clipped white Gaussian noise* might be perceptually indistinguishable from *white Gaussian noise*. Fawe further extrapolated by stating that "it is [now] evident that a flat spectrum is the optimum one, when the signal is passing through a nonlinear circuit and when the highest signal-to-noise ratio is desired an equalization of the mean speech power spectrum is required before clipping" and that, for speech, "a derivation [differentiation] of the signal before clipping . . . will be the best way to achieve the purpose." Pre-clipping differentiation, as Licklider noted (L5), does improve the intelligibility of clipped speech, but not--as we shall show in chapters 8 and 9--for the reasons Fawe extrapolates from a study of white Gaussian noise.

Fawe next noted that, as shown by Crater, clipping causes only small changes in the power spectrum of a Gaussian waveform with an original power spectrum resembling that of a *single* formant,[11]

$$\text{i.e.,} \qquad G(f) = \frac{R(0)}{2\pi}\left[\frac{a}{a^2+(f-F_1)^2} + \frac{a}{a^2+(f+F_1)^2}\right]. \tag{5-30}$$

He claimed that "this latest approach tends to prove that results for the ensemble of speech sounds are valuable for isolated utterances too."

Fawe then rederived the results of Dukes [equations (5-24,25)] concerning the coherence coefficients[12] and reworked Dukes' results with respect to Luce's theorem (see footnote 10; also [L-25]).

We do not believe that Fawe's final conclusion "that we have shown an infinite clipper has very little effect on the power spectrum when first flattened so that clipped [whispered?, see [M-9]] speech is highly intelligible" is justified.

---

[11] Velechin [V-4] apparently repeated Crater's experiments for Russian speech.

[12] Although Dukes' paper is referenced, direct credit for the results (5-24,25) is not given.

5.3.3 Vilbig

F. Vilbig used the expression (5-22)

$$h_c(x) = \frac{4}{\pi}\sum_{n=1}^{\infty} \frac{\sin(2n-1)x}{2n-1} \quad , \ |x| \leq \pi \quad , \qquad (5\text{-}31)$$

to examine the effect of clipping on two-tone speech models [V-5]. If $x = s(t) = a\cdot\cos\omega_1 t + b\cdot\cos\omega_2 t$, then

$$h_c(x) = \frac{4}{\pi}\sum_{\substack{m=1\\ m\ \text{odd}}}^{\infty} \frac{1}{m}[\ \sin(ma\cdot\cos\omega_1 t)\cdot\cos(mb\cdot\cos\omega_2 t) + \cos(ma\cdot\cos\omega_1 t)\cdot\sin(mb\cdot\cos\omega_2 t)\ ] \ . \qquad (5\text{-}32)$$

The Bessel function expansions can then be introduced: i.e.,

$$\sin(z\cdot\cos\theta) = 2\sum_{n=0}^{\infty}(-1)^n\cdot J_{2n+1}(z)\cdot\cos[(2n+1)\theta] \qquad (5\text{-}33)$$

and

$$\cos(z\cdot\cos\theta) = J_0(z) + 2\sum_{n=1}^{\infty} (-1)^n\cdot J_{2n}(z)\cdot\cos[(2n)\theta] . \qquad (5\text{-}34)$$

Unfortunately, expansion of these functions involves much calculation and the results are qualitatively unsatisfying.

Vilbig's graphical data concerning the frequency distortion caused by clipping *three-tone vowel models* probably represents the most comprehensive published data in this area (Fig. 5.9). He noted that when *one formant* is much larger than the other two, the distortion generated by clipping lies mainly at the third harmonic of this dominant formant--i.e., $3F_1$, $3F_2$, or $3\ f_3$. Bandlimiting the *clipped* signal to 3000 Hz eliminates clipping harmonics due to any but a dominant F1, or F2 of /ɔ/, /U/, or /u/. (see Fig. 3.8b) If two formants, Fm and Fn (m=1,2; n=2,3; m≠n) are approximately equal in amplitude and much larger than the other formant, then--using results from a two-tone model--the lowest frequency clipping distortion harmonics appear at two frequencies,

$$\omega_1 = 1.5\cdot(\omega_m+\omega_n) + 0.5\cdot(\omega_n-\omega_m) \qquad \text{and}$$

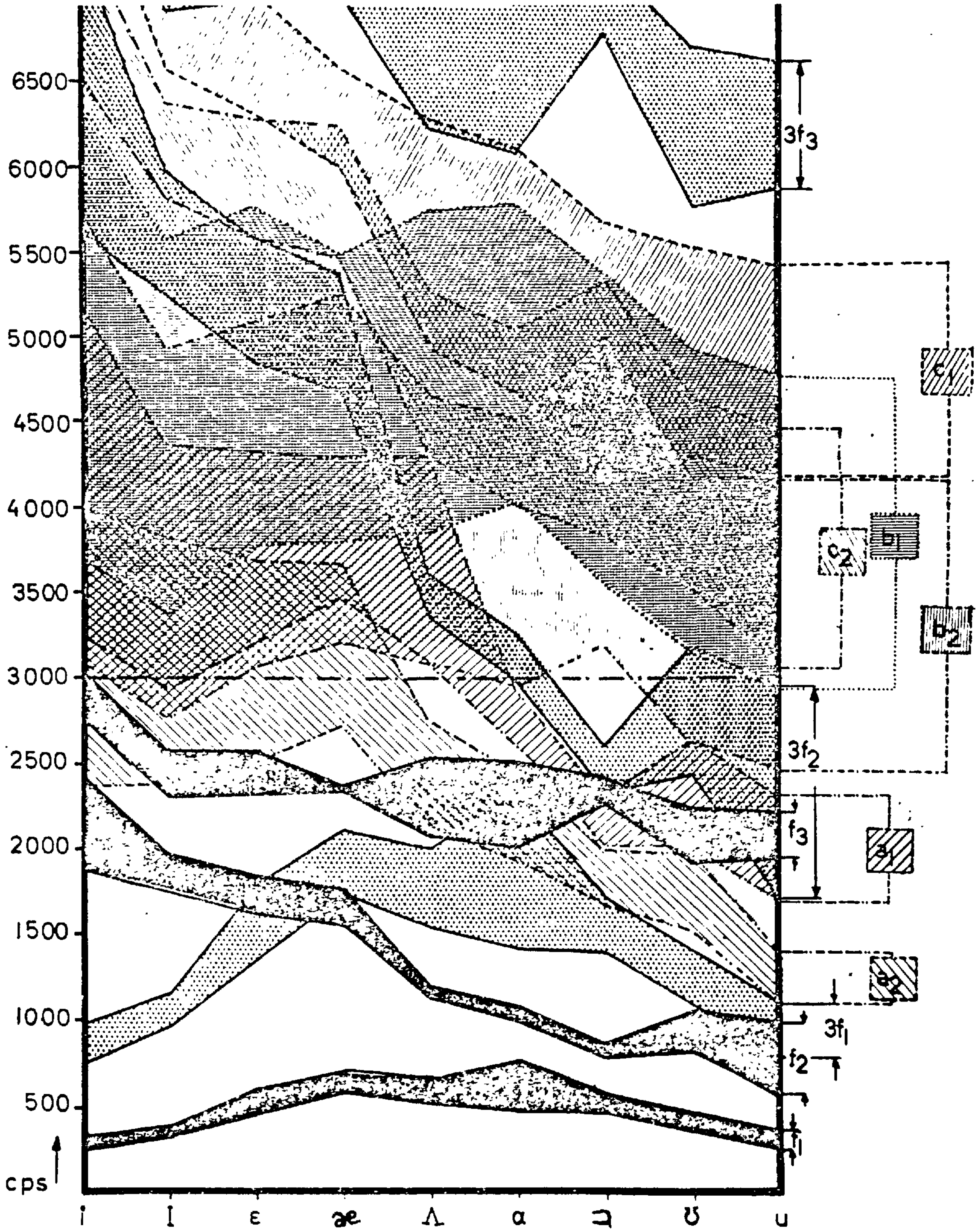


Fig. 5.9 Distribution of the formants of various vowels and position of the distortion frequencies created by the clipping process. (From [V-5].)

$$\omega_2 = 1.5 \cdot (\omega_m + \omega_n) - 0.5 \cdot (\omega_n - \omega_m) \quad .$$

The areas in which these frequencies may fall for three-tone models of vowels are represented in Fig. 5.9 as a_1 and a_2 (m=1, n=2), b_1 and b_2 (m=1, n=3), and c_1 and c_2 (m=2, n=3). In this case, only the ranges a_1 and a_2 fall within the 0-3000 Hz region.

In summary, only clipping harmonics from a dominant F1, a dominant F2 for /ɔ/, /U/ or /u/ *only*, or a dominant (equal amplitude) F1-F2 complex can fall within the 3 KHz passband for the three-tone model. Vilbig argued that the third harmonic of a dominant F1--falling between F_1 and F_2 or between F_2 and F_3, depending on the vowel--produces the most perceptually degrading distortion. The a_1-a_2 distortion regions interfere predominantly with F3. He added that "for vowels . . . all the newly created [distortion] frequencies are harmonics of the pitch frequency and . . . are less noticeable than if the frequency had been arbitrary." Finally, Vilbig stressed that pre-clipping attenuation of frequencies in the F_1 region will weaken the otherwise strong clipping harmonics of F1 ($3F_1$) and thus yield a less distorted clipped signal. Actual pre- and post-clipping spectral cross sections of actual vowels modified in this manner objectively support his assertion.

5.3.4 Summary

Dukes and Fawe, and Vilbig, provide arguments which support conjectures suggesting overall power spectrum preservation in clipped *random processes* and *three-tone periodic signal models*, respectively. However, the explanations proposed are somewhat unsatisfactory:

First, they do not satisfactorily explain why a process (infinite clipping) which ostensibly destroys all waveform amplitude information and preserves only zero crossing positional data does not yield changes of similar apparent magnitude in the frequency domain.

Second, there is no indication of whether the nature of the original *waveform* and the extent of post-clipping power spectrum preservation are correlated in any manner.

Finally, although Vilbig suggests a method for processing the speech signal *before* clipping in order to enhance post-clipping power spectrum preservation, the technique--although intuitively justifiable--is somewhat *ad hoc*.

We will show, in chapters 8 and 9, that certain types of waveform processing (and the spectral transformations associated with such processing) will produce signals of extremely high post-clipping intelligibility. Furthermore, we will produce arguments that certain waveform attributes are highly correlated with post-clipping power spectrum preservation. Finally, we will argue that clipping preserves other waveform attributes in addition to zero crossing information.

6 ZEROS I: ZERO CROSSINGS AND AUTOMATIC SPEECH RECOGNITION

6.1 Evidence for Consideration of Zero Crossings as Input Parameters for Automatic Recognition of Speech

Rectangular interpolation of speech waveform zero crossing sequences yields a highly intelligible signal. Can this sequence of zero crossing intervals be used *independently* of the auditory system to provide an estimate of the spectral features of the original signal? If so, then presumably, zero crossings could serve as input data for automatic speech recognition schemes. Furthermore, can zero crossing interval sequences be interpreted meaningfully without explicit reference to the frequency domain and, are such interpretations useful for automatic speech recognition purposes?

In this chapter we discuss these, and other closely related problems from the viewpoint of conventional signal theoretic ideas. The related problems include manipulation of zero crossing information via single sideband (SSB) modulation and, finally, examples of automatic speech recognition machines using zero crossing information.

6.2 The Zero Crossings of Random Processes

In our review of the acoustic properties of speech sounds (ch. 3) we noted that some speech sounds--unvoiced fricative and stop consonants--result from excitation of the vocal tract by a noise source. Davenport observed (sec. 3.5.1; [D-3,4]) that the amplitude distribution of these sounds could be represented by a Gaussian model. Spectrally, these sounds often resemble "white" noise bands with different frequency location and bandwidth parameters (secs. 3.4.6,7; [F-14],[H-4,9,26],[S-27]). It is imperative, therefore, to briefly state some results--derived by S.O. Rice [R-10]--concerning the characteristics of the zero crossings of random processes.

6.2.1 Average Rate of Zero Crossings

Rice showed that the expected number of zero crossings, per second, of a Gaussian random process is completely determined by knowledge of the power spectrum G(f) of the process:

i.e.,
$$\rho_o = 2 \left[\frac{\int_o^\infty f^2 G(f)\, df}{\int_o^\infty G(f)\, df} \right]^{\frac{1}{2}} . \qquad (6\text{-}1)$$

For bandpass white Gaussian noise such that

$$G(f) = \begin{cases} K , & 0 \leq f_a \leq |f| \leq f_b \\ 0 , & \text{otherwise} , \end{cases}$$

eq. (6-1) becomes

$$\rho_o = \frac{2}{\sqrt{3}} \left[f_a^{\,2} + f_a f_b + f_b^{\,2} \right]^{\frac{1}{2}} . \qquad (6\text{-}2)$$

When f_a=0 (lowpass, white Gaussian noise), (6-2) becomes

$$\rho_o = 2f_b/\sqrt{3} \quad . \tag{6-3}$$

In this case ρ_o is $\sqrt{1/3}$ times the Nyquist rate, $2f_b$.

Finally, it can be shown that for the m^{th} derivative of lowpass bandlimited white Gaussian noise,

$$\rho_o = 2f_b \sqrt{(2m+1)/(2m+3)} \tag{6-4}$$

$$\rightarrow 2f_b \quad \text{for } m \text{ large.}$$

These properties form the basis of much of our discussion of the zero crossings of speech signals. Extensions of Rice's work are detailed in Cramér and Leadbetter [C-11].

6.3 Zero Crossings as an Estimate of Frequency Information in Speech Signals

H. Dudley noted the possibility of extracting frequency information indirectly from the speech waveform in 1965 [D-13]. In an example, he reproduced a portion of an oscillogram of the vowel /a/ (Fig. 6.1) and analyzed it as follows:

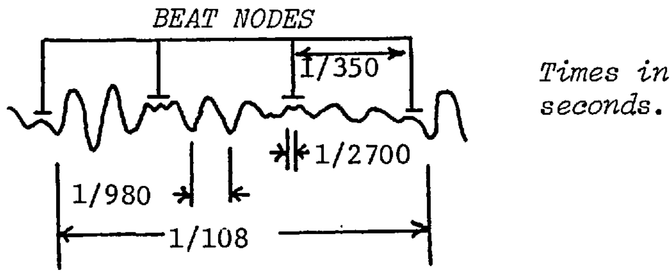


Fig. 6.1 Oscillogram of the vowel /a/. (From [D-13].)

"We note a high frequency ripple . . . [of approximately] 2700 cps for F_3. . . . A clear beat shows up separating strong sections . . . the separation corresponds to 350 times per second. If we measure a period of the strong wave itself we get a correspondence to 980 cps which is presumably F_2 and F_1 is then (980-350) or 630 cps." Dudley then tabulated some waveform characteristics which may be related either directly, or indirectly, to the spectral features of the speech sound. He emphasized that "there can be no change in the sound spoken and heard without a corresponding change . . . [in the waveform]."

Dudley's estimates of spectral information involved, indirectly, measures of zero crossing data. In the following subsections we shall review and evaluate attempts to directly use zero crossing information to estimate spectral parameters in speech signals.

6.3.1 Chang

S. Chang *et al.* [C-3],[C-4] considered the problem of "the representation of speech sounds and some of their statistical properties." They noted that, while the Fourier transform of a signal contains both amplitude and phase information, the *time* autocorrelation function, $R(\tau)$, defined as [T-6; p. 90]

$$R(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} s(t)\cdot s(t+\tau)\, dt \tag{6-5}$$

for a random process,[1] discards *phase* information. That $R(\tau)$ contains no phase information about $s(t)$ is made clear by noting that

$$R(\tau) = \mathcal{F}^{-1}\{G(f)\} = \int_{0}^{\infty} G(f)\cdot \cos 2\pi f\tau\, df \tag{6-6}$$

and

$$G(f) = |S(f)|^2 \ . \tag{6-7}$$

---

[1]The same definition applies to a periodic signal if $T\to T_o/2$, where T_o is the period of the signal [L-6, p. 11].

The usefulness of $R(\tau)$ as a representation of speech sounds is inferred from the relationship between $R(\tau)$ and $|S(f)|$, the amplitude spectrum, via $G(f)$. In fact, the short-term autocorrelation function can be defined (S-7) and used (B-8) for automatic word recognition.

Similarly, they noted, clipping--another time domain operation involving $s(t)$--discards *amplitude* information. The usefulness of zero crossings in obtaining estimates or representations of speech sounds is to be inferred, ostensibly, from the fact that clipped speech is intelligible. Chang[2] pointed out that more direct links between zero crossings and signal spectral features can be established. Rice's classic relationship for the average number of zero crossings per unit time--in a stationary, random process $n(t)$--can be written as [C-4]:

$$\rho_o = k_o \sqrt{\frac{\overline{n'(t)^2}}{\overline{n(t)^2}}} \qquad (6\text{-}8)$$

while the average number of zero crossings per unit time, of $n'(t)$ is

$$\rho_m = k_m \sqrt{\frac{\overline{n''(t)^2}}{\overline{n'(t)^2}}} \ . \qquad (6\text{-}9)$$

The value of k_o and k_m is $1/\pi$ when $n(t)$ is a Gaussian signal. The n^{th} moment of the power spectrum of $n(t)$, $G(f)$, is defined as

$$M_n = \int_0^\infty f^n \, G(f) \, df \ . \qquad (6\text{-}10)$$

Since $R(0) = M_o$ and $-R''(0) = 4\pi^2 M_2$ (from (6-6)), then, using

---

[2]For convenience, we shall refer to the authors Chang, Pihl and Essigman [C-4] and Chang, Pihl and Wiren [C-5] as "Chang".

(6-6) and (6-10) in (6-1), we can write

$$\rho_o = 2\pi k_o \sqrt{M_2/M_o} = k_o \sqrt{-R''(0)/R(0)} \qquad (6\text{-}11a,b)$$

and, similarly,

$$\rho_m = 2\pi k_m \sqrt{M_4/M_2} = k_m \sqrt{-R''''(0)/R''(0)} \quad . \qquad (6\text{-}12a,b)$$

In this manner the average (expected) rate of zero crossings per unit time can be related, through the autocorrelation function, to the power spectrum of the signal. However, as Chang pointed out [C-4], "application . . . [of these relationships] . . . to a speech sound assumes that it can be regarded as a stationary time series . . . and *the extent that this requirement is met can only be conjectured at the present time* [1950]." (Italics mine.)

Chang presented limited experimental results which implied that "there is a close similarity between the shapes of the ρ_o- and ρ_m-grams and the first two bars [formants] of the spectrogram." He explained that "since the frequency components in the first bar [formant] are usually strong enough to cause zero crossings, the ρ_o-gram is a close approximation of this bar [formant]. The frequency components in the second resonance region may not be strong enough to cause extra zero crossings, but they will affect the slope of the wave [-form s(t)] and may, therefore, contribute extra maxima and minima which are included in ρ_m."

6.3.2 E. Peterson

Soon after Chang's conjectures and limited experiments concerning the utility of the average time rate of zero crossings as an estimate of formant trajectories in speech spectrograms, E. Peterson published [P-9] an excellent experimental and theoretical study of such techniques. Peterson first described an accepted method of *estimating* ρ_o for a speech signal: an impulse is generated at each zero crossing of the signal and these impulses are averaged for a time interval *greater* than the fundamental period of voiced sounds (≈10 msec.) and *less* than the phonemic utterance rate (≈10 per second). [This type of *estimate* will be defined as $\tilde{\rho}_o$ to distinguish it from ρ_o, the *true average rate of zero crossings per second.*] In his experimental work, Peterson used a lowpass filter with a 30 Hz cutoff frequency to implement the averaging process. Experimental results for two-tone signals are shown in Fig. 6.2.

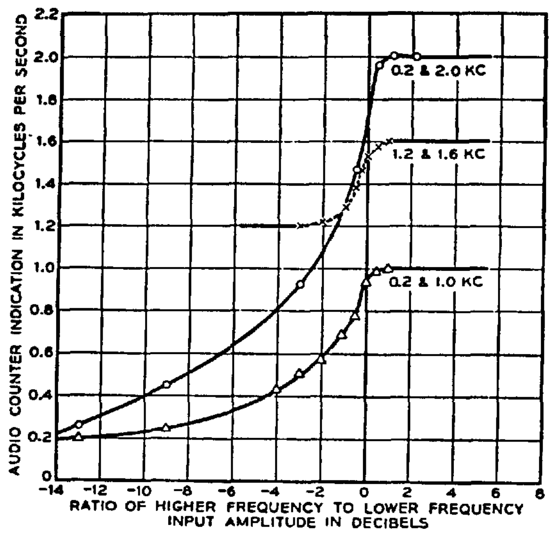


Fig. 6.2 Response of an impulse averaging $\tilde{\rho}_o$-meter to a two-tone input. Ordinate is the "counter" reading in KHz and abscissa is $20 \log_{10}(A_2/A_1)$, the ratio of the input amplitudes in db. The three curves apply to the pairs of input frequencies noted. (From [P-9].)

Note that Lobanov [L-24] derived an expression for the *true number of zero crossings per second* for a two-tone signal.

If

$$s(t) = A_1 \sin \omega_1 t + A_2 \sin \omega_2 t \ , \ \omega_2 > \omega_1 \tag{6-13}$$

then

$$\rho_o = \begin{cases} \frac{2}{\pi}(2F_2 - 2F_1)\cdot\sin^{-1}[A_1/A_2] + 2F_1, & 0 \leq A_2/A_1 \leq 1 \\ 2F_2, & A_2 > A_1 \end{cases} \quad (6\text{-}14)$$

where $F_1 = \omega_1/2\pi$ and $F_2 = \omega_2/2\pi$. Peterson's experimental results for ρ_o *estimates*, $\tilde{\rho}_o$, and Lobanov's expression for actual ρ_o both suggest that when the amplitude of the higher frequency signal dominates, then all (zero crossing) indication of the lower frequency tone is lost. However, when the low frequency tone has the larger amplitude, the indicated "frequency" lies between the two input frequencies over a very extended amplitude ratio range. Peterson emphasized that this anomalous behaviour is not due to the nature of the "counter"; it is, he showed, fundamental to operation of this type of "counter" in the *audio band*. His explanation was as follows:

The envelope of s(t), (6-13), is

$$|m(t)| = [A_1^2 + A_2^2 + 2A_1A_2\cos(\omega_1-\omega_2)t]^{\frac{1}{2}}, \quad (6\text{-}15)$$

the phase is

$$\phi(t) = \tfrac{1}{2}[\omega_1+\omega_2]t + \tan^{-1}\left[\left\{\frac{A_1-A_2}{A_1+A_2}\right\}\tan[\tfrac{1}{2}(\omega_1-\omega_2)t]\right], \quad (6\text{-}16)$$

and the instantaneous frequency, the time derivative of the phase, is

$$\phi'(t) = \tfrac{1}{2}[\omega_1+\omega_2] + \tfrac{1}{2}[\omega_1-\omega_2]\left\{\frac{A_1-A_2}{A_1+A_2}\right\}\left[\frac{1+\tan^2[\tfrac{1}{2}(\omega_1-\omega_2)t]}{1+\left\{\frac{A_1-A_2}{A_1+A_2}\right\}^2\tan[\tfrac{1}{2}(\omega_1-\omega_2)t]}\right]. \quad (6\text{-}17)$$

The value of $\phi'(t)$, averaged over a half-period, is

$$\overline{\phi'(t)} = \frac{2}{\pi}\int_0^{\pi/2}\phi'(t)\,d[\tfrac{1}{2}(\omega_1-\omega_2)] = \tfrac{1}{2}\left(\omega_1+\omega_2 + (\omega_1-\omega_2)\cdot\mathrm{sgn}\left[\frac{A_1-A_2}{A_1+A_2}\right]\right). \quad (6\text{-}18)$$

That is,

$$\overline{\phi'(t)} = \begin{cases} \omega_1 & , \quad A_1 > A_2 \\ \frac{1}{2}[\omega_1+\omega_2] & , \quad A_1 = A_2 \quad ^3 \\ \omega_2 & , \quad A_1 < A_2 \quad . \end{cases} \tag{6-19}$$

Fig. 6.3 shows a plot of $\phi'(t)$ for $\omega_1 = \omega$, $\omega_2 = 3\omega$ and $(A_2/A_1) = q$.

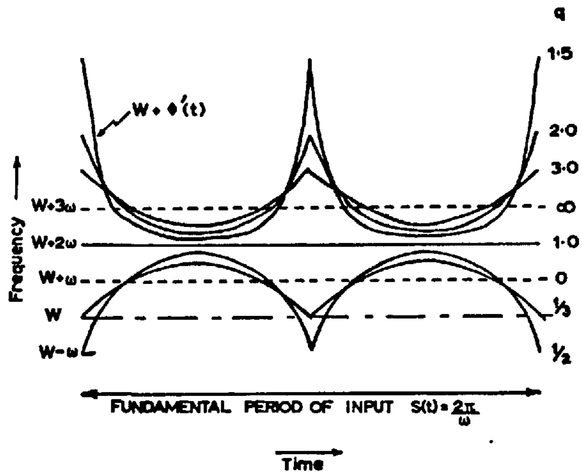


Fig. 6.3 $\phi'(t)$ waveforms for
$s(t) = \cos\omega t + q\cdot\cos 3\omega t$.
(From [C-9].)

Note that "W" is a frequency translator, considered equal to zero for our purposes. The average values of $\phi'(t)$, $\overline{\phi'(t)}$, are the dotted lines labelled "W+3ω", for q>1; "W+ω", for q<1; and the solid line labelled "W+2ω", for q=1. From Fig. 6.3 and equation (6-17), it is apparent that

$$\phi'(t) \quad \left.\begin{matrix} > \\ = \\ < \end{matrix}\right\} \quad \overline{\phi'(t)}\Big|_{q=1} \quad \text{as} \quad q \quad \left\{\begin{matrix} > & 1 \\ = & 1 \\ < & 1 \end{matrix}\right. , \tag{6-20}$$

and that for $\frac{1}{2}<q<2$, $q \neq 1$, the instantaneous frequency, $\phi'(t)$,

---

[3] That $\overline{\phi'(t)} = \frac{1}{2}[\omega_1+\omega_2]$ for $A_1 = A_2$ is shown by Cherry and Phillips [C-9, p. 1070]; it also follows directly from (6-17) with $A_1 = A_2$. This situation is very unstable.

exhibits very sharp peaks.

The problem is to show why the output of the $\tilde{\rho}_o$-meter,* shown in Fig. 6.2, does not indicate the readings predicted by (6-19). We assume here that this type of "meter"--i.e., impulse averaging--should indicate $\overline{\phi'(t)}$.

Peterson showed that the answer lies in the bandwidth required to transmit $\phi'(t)$, the instantaneous frequency function. From (6-17),

$$\phi'(t=\pi/[\omega_1-\omega_2]) = \tfrac{1}{2}[\omega_1+\omega_2] + \tfrac{1}{2}[\omega_1-\omega_2]\left\{\frac{A_1+A_2}{A_1-A_2}\right\} \quad , \qquad (6\text{-}21)$$

and $\phi'(\pi/[\omega_1-\omega_2])_{max} \to \pm\infty$ as $A_1 \to A_2$ or, equivalently, as $q \to 1$. Therefore, when $t = \pi/[\omega_1-\omega_2]$, $\phi'(t) \to \infty$ or $-\infty$ as q approaches unity from a value greater than 1, or less than 1, respectively. In the practical case, for q small, but greater than unity, the positive peaks of the $\phi'(t)$ function are attenuated due to the bandwidth limitations incorporated in the $\tilde{\rho}_o$-meter. This lowers the value of $\overline{\phi'(t)}$. Conversely, for $q<1$ but near unity, $\phi'(t)$ "attempts" to become very small and much less than ω, its theoretical average value. When $\phi'(t)$, a positive quantity, "attempts" to become negative, it is reflected positive. This substantially raises the average value of $\phi'(t)$, $\overline{\phi'(t)}$. These effects are both evident in Fig. 6.2. A solution to this system deficiency is to translate the audio band upwards (using SSB modulation) in order to eliminate the source of the greatest errors, the positive reflections of $\phi'(t)$ for $q<1$. Figure 6.4 shows the results of

---

* $\tilde{\rho}_o \equiv \overline{\phi'(t)}$; $\tilde{\rho}_m \equiv \overline{\phi'(t)}$ for $s'(t)$. Note that ρ_o is *average rate of zero crossings* and has dimensions of *sec*$^{-1}$ whereas $\tilde{\rho}_o$ or $\tilde{\rho}_m$ is *average value of instantaneous frequency* and has dimensions of *radians/sec.*

SSB modulating a 1 KHz--4 KHz tone complex with a 60 KHz carrier (W = 60 KHz in Fig. 6.3) and then measuring $\overline{\phi'(t)}$ via the $\tilde{\rho}_o$-meter.

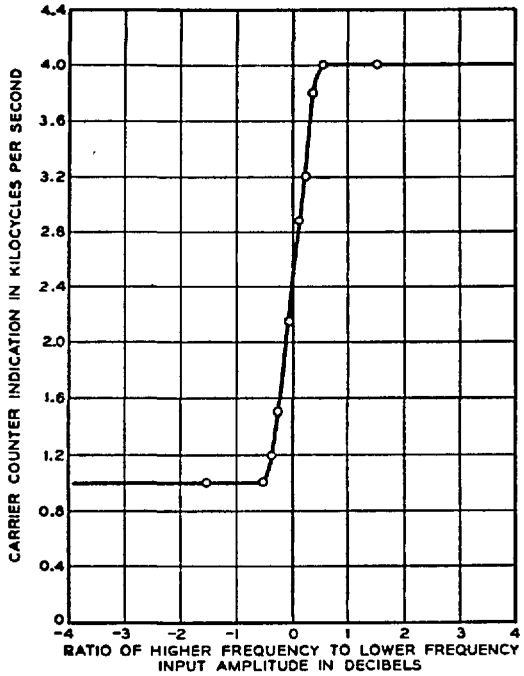


Fig. 6.4 Response of a $\tilde{\rho}_o$-meter to a SSB modulated 1 KHz-4 Khz tone complex. Carrier frequency is 60 KHz. (From [P-9].)

The transition region has been substantially reduced, especially for the $A_2<A_1$ (negative db) range.

Peterson summarized his analysis by stating that the audio band is badly situated for obtaining an accurate estimate of the average value of the instantaneous frequency of a two-tone signal and that SSB modulation *must* be used to insure an accurate indication. He concluded his investigation with experimental tests, using the SSB $\tilde{\rho}_o$-meter on speech waveforms. He found that the $\tilde{\rho}_o$ trajectory was generally higher than that of the first formant, F1, and was located between the first and second formant spectrogram bars. For differentiated speech, the $\tilde{\rho}_m$ trajectory closely paralleled, but was somewhat higher than, the second formant spectrogram bar.

He concluded that "the average axis crossing rates [as estimated by $\overline{\phi'(t)}$] cannot be trusted in general to follow specific [formant spectrogram] bars, whether the speech is normal or differentiated" and that "the [formant] bars higher in the spectrum affect the axis crossing averages." Finally, tests with SSB $\tilde{\rho}_o$- and $\tilde{\rho}_m$-meters using bandpass filtered speech provided a fairly accurate estimate of F_1 (bandpass = 0.2-1.0 KHz) and F_2 (bandpass = 1.0-4.0 KHz). Estimation of F_2 was made more accurate by introducing a 6 db per octave attenuation in the 1.0-4.0 KHz bandpass filter.

Three questions arise after consideration of Chang [C-4] and Peterson [P-9]:

1: Of what value are simple zero crossing measurements (e.g., precise $\tilde{\rho}_o$- or $\tilde{\rho}_m$- meters) in obtaining *accurate* estimates of formant frequencies?

2: Is there *any* zero crossing measurement which can provide *accurate* estimates of formant frequencies?

3: Is $\tilde{\rho}_o$ [$\equiv \overline{\phi'(t)}$] $= \pi \cdot \rho_o$? That is, is the *average value of the instantaneous frequency* proportional to the *average rate of zero crossings*?

Peterson and Hanne, Focht, and Scarr have provided some answers to questions 1: and 2:. Question 3: is considered in sec. 6.5.

6.3.3 Peterson and Hanne

We first consider Peterson and Hanne's answer to question 1:. They analyzed the ideal case where, by filtering, it is possible to isolate a single formant and, by deconvolution (e.g., [M-11]), the effect of glottal excitation may effectively be removed [P-12].

The transfer function of a resonator model for a *single* formant is [F-2, p. 53]

$$|H(f)| = \frac{F_1{}^2 + (B/2)^2}{\{[(F_1-f)^2 + (B/2)^2][(F_1+f)^2 + (B/2)^2]\}^{\frac{1}{2}}} , \quad f \geq 0 \tag{6-22}$$

where F_1 is the formant frequency[4] and B is the formant bandwidth. If $F_1 > B/2$ (usual for vowels), then $|H(f)|_{max}$ occurs for

$$f = [F_1{}^2 - (B/2)^2]^{\frac{1}{2}} . \tag{6-23}$$

The result of periodically exciting this resonator with an impulse (delta function) of period T is

$$s(t) = \sum_{n=0}^{\infty} U(t-nT) \cdot a \cdot e^{-\pi B(t-nT)} \sin[2\pi F_1(t-nT)+2\pi\Phi] , \tag{6-24}$$

where $$a = [1-2\cdot e^{-\pi BT} \cdot \cos 2\pi F_1 T + e^{-2\pi BT}]^{-\frac{1}{2}} ,$$

$$\tan 2\pi\Phi = [\sin 2\pi F_1 T/(e^{\pi BT} - \cos 2\pi F_1 T)] ,$$

and U(x) is the unit step, $$U(x) \begin{cases} = 1, & x \geq 0 \\ = 0, & x < 0. \end{cases}$$

Peterson and Hanne showed that for

$$\frac{2n-1}{F_1} \leq T < \frac{2n+1}{F_1} , \quad n \geq 1 , \tag{6-25}$$

s(t) will exhibit 2n zero crossings per period. Then the average *counted* rate of zero crossings per second is $\rho_o = 2n/T$. For $T = (2n+1)/F_1$, there is a discontinuity of magnitude $2F_o$ ($F_o=1/T$)

[4] F_1 represents an arbitrary formant frequency here, not necessarily the first formant. Our notation is that of [P-12] and Fig. 6.5.

in ρ_o. Figure 6.5 shows ρ_o (the zero crossing counter estimate of F_1 is $\rho_o/2$) vs T, the period of resonant cavity excitation. The envelopes of the ρ_o output are given by

$$E(T) = 2F_1 \pm F_o \quad , \quad F_o = 1/T \, . \tag{6-26}$$

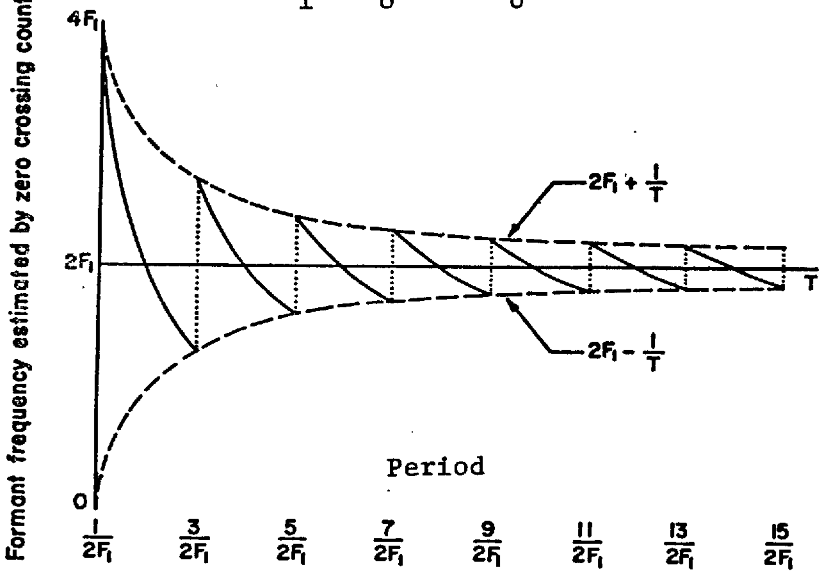


Fig. 6.5 Steady state zero crossing frequency estimation of a single formant ($f=F_1$) resonator as a function of the period of impulse excitation. (From [P-12].)

Therefore, the error in estimating formant frequency using an *accurate* zero crossing counter method can be as much as $F_o/2$. Furthermore, this estimate is the *nearest* harmonic of F_o to F_1 rather than--as is often suggested--the *strongest* harmonic of F_o. Since the resonance peak of $|H(f)|$ does not occur exactly at F_1 (see equation (6-23)), the strongest harmonic of F_o is *not* necessarily the nearest to F_1.

Peterson and Hanne also calculated the estimate of formant frequency afforded by a harmonic tracker which indicates the frequency of the strongest harmonic of F_o. They showed that, in contrast to the maximum frequency magnitude error of the *zero crossing counter*--

$0.5F_o$--the *strongest harmonic tracker* has a maximum frequency magnitude error which ranges from $0.550F_o$, for $F_1 = 2.55F_o$, to $0.515F_o$, for $F_1 = 8.5F_o$. Nevertheless, it turns out that the *strongest harmonic tracker* is a slightly better F_1 estimator on the basis of maximum percentage error.

In summary, both methods of formant frequency estimation yield approximately the same *large* percentage maximum error in estimating formant frequency as long as the actual formant frequency is less than about $17F_o$ (error = 3% in this case). Thus, even in these ideal circumstances (single formant, glottal waveform influence removed) a simple zero crossing formant frequency estimator is potentially as inaccurate as a more conventional "highest energy" harmonic tracker.

6.3.4 Focht

One answer to question 2: is provided by L. R. Focht [F-10]. In a study of the perceptual identity of various combinations of formant amplitudes and frequencies, for three-formant sounds, Focht found that only two formants (F1-F2, F1-F3, or F2-F3), depending on the particular vowel, were required to specify the perceptual value of a vowel. A plot of F_d (the frequency of the larger amplitude or *dominant* formant) vs F_r (the frequency of the lesser amplitude or *recessive* formant) revealed that *all* isophonemic areas on the F_d-F_r plane intersect the $F_d = F_r$ line. In other words, a different *single equivalent formant* (SEF) frequency can be specified to evoke the perceptual response of each vowel. The frequencies of vowel SEF's are shown in Fig. 6.6. Moreover, Focht stated that "it was observed that the zero-axis crossing period of the first excursion for the speech wave after glottal . . . excitation is proportional to the half-period of the largest amplitude formant. The value of the SEF was also found to follow closely the dominant formant frequency. Thus a reasonable approximation of the SEF parameter may be made by the measurement of the first zero-axis crossing period after each excitation pulse." In section 6.5 we shall describe a limited vocabulary speech recognizer based upon the SEF principle.

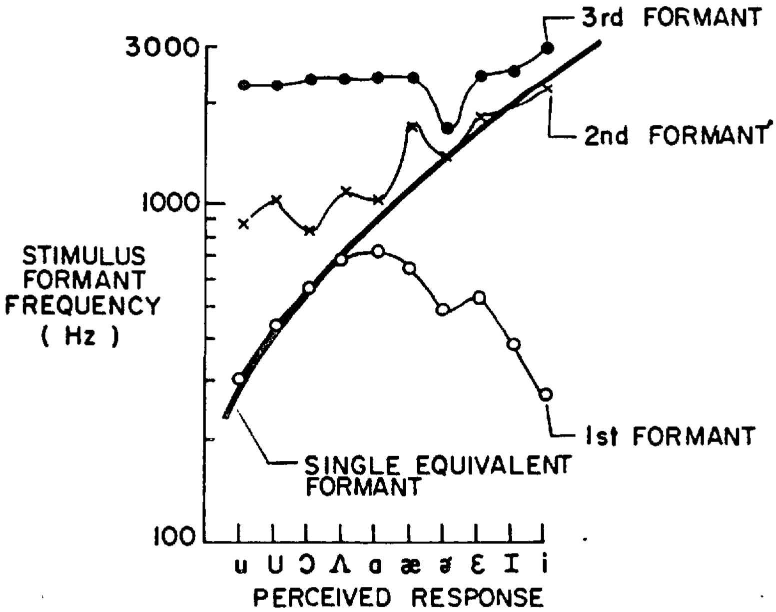


Fig. 6.6 Single equivalent formant (SEF) frequencies (heavy line) for English vowels. Conventional F_1, F_2, and F_3 are shown by light lines. (From [T-2].)

6.3.5 Scarr

R. W. Scarr's work [S-4] represents a theoretical and experimental extension of that of Peterson and Hanne [P-12] and Focht [F-10].

Equation (6-22) can be rewritten as

$$|H(f)| = \frac{1}{\{[1-(f/F_1)^2]^2 + [fB/F_1]^2\}^{\frac{1}{2}}} \qquad (6\text{-}27)$$

with

$$\arg [\ H(f)\ ] = \tan^{-1}\left[\frac{(f/F_1)}{Q[(f/F_1)^2 - 1]}\right] , \qquad (6\text{-}28)$$

and $Q = \sqrt{F_1^2 + (B/2)^2}/B$ *if* $(B/2)^2 << F_1$. This criterion is generally satisfied for English vowel formants.

Using this simplified version of the single formant model, Scarr analyzed the expected waveform zero crossing pattern when the excitation is a bandlimited sawtooth waveform, [5]

$$g(t) = K\,(\sin\Omega t + \sin 2\Omega t/2 + \sin 3\Omega t/3 + \ldots). \qquad (6\text{-}29)$$

"K" is an arbitrary constant and $F_o = \Omega/2\pi = 1/T$ is the excitation or voicing frequency. Scarr considered only the second, third and fourth harmonics of (6-29). The output of the resonator is then

$$\begin{aligned} s(t) = {} & A_2\sin 2\Omega t + A_3\sin 3\Omega t + A_4\sin 4\Omega t \\ & + B_2\cos 2\Omega t + B_3\cos 3\Omega t + B_4\cos 4\Omega t \quad , \end{aligned} \qquad (6\text{-}30)$$

where
$$A_n = \frac{K}{n}\,\frac{[\,1 - (nF_o/F_1)^2]}{\{[1-(nF_o/F_1)^2]^2 + [nF_oB/F_1^{\,2}]^2\}^{\frac{1}{2}}} \quad , \qquad (6\text{-}31)$$

$$B_n = \frac{K}{n}\,\frac{nF_oB/F_1^{\,2}}{\{[1-(nF_o/F_1)^2]^2 + [nF_oB/F_1^{\,2}]^2\}^{\frac{1}{2}}} \quad , \qquad (6\text{-}32)$$

and
$$\tan^{-1}[B_n/A_n] = \arg\,[\;H(nF_o)\;] \quad . \qquad (6\text{-}33)$$

Equation (6-30) was solved (iteratively) for $s(t) = 0$ for varying F_o, $130\text{ Hz} \leqslant F_o \leqslant 200\text{ Hz}$ with $F_1 = 500$ Hz. Figure 6.7 shows contours which represent the position where the zero crossings of $s(t)$ occur as a function of the phase angle ψ. The time interval Δt between any two adjacent contours separated horizontally by $\Delta\psi$ degrees is

$$\Delta t = (\Delta\psi/360)\cdot(1/F_o) = (\Delta\psi/360)\cdot T. \qquad (6\text{-}34)$$

---

[5] Peterson and Hanne [P-12] dealt only with the case of periodic delta function excitation of the resonator.

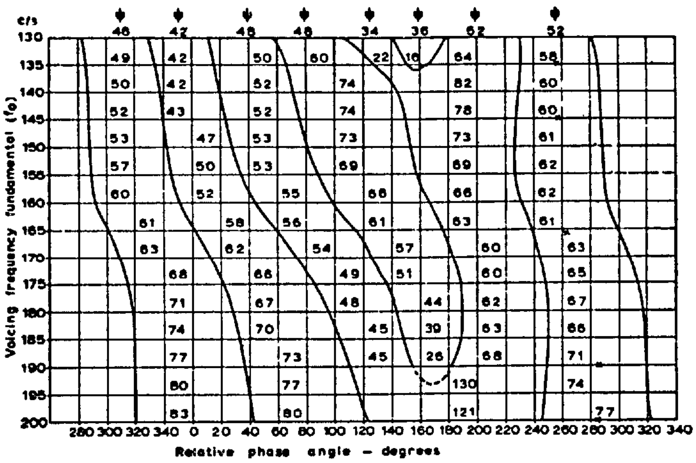


Fig. 6.7 Zero crossing pattern for equation (6-30) as a function of voicing frequency, for F_1 = 500 Hz. Each number represents $\Delta\psi$, in degrees, for the intersection of the contour lines (on either side of the number) with the horizontal line immediately below that same number. (From [S-4].)

Scarr calculated the frequency $\bar{f}$ represented by the *average* zero crossing interval $\Delta\bar{t}$,

$$\text{i.e., } \bar{f} = (2\Delta\bar{t})^{-1} \, . \qquad (6\text{-}35)$$

He also found that the frequency f_1 represented by the phase interval separating the two vertical contours at the left of Fig. 6.7,

$$\text{i.e., } f_1 = (360/\Delta\psi)\cdot(F_o/2) \quad , \qquad (6\text{-}36)$$

where the $\Delta\psi$-F_o pairs are 46°-130 Hz, 49°-135 Hz . . . 83°-200 Hz, gave the best estimate (of all pairs of adjacent contours) of F_1. His results showed that while $\bar{f}$ varied as much as +70 or -110 Hz from F_1 for varying F_o, f_1 remained within +10 and -65 Hz of F_1. Scarr also noted that f_1 varies smoothly with F_o. In contrast, $\bar{f}$, as well as being a poorer estimate of F_1, is a discontinuous function of F_o. The calculations also showed that the peak amplitude of s(t) always fell between the pair of vertical contours at the right of Fig. 6.7. In summary, Scarr stated that—*for this model*—a measure of frequency based upon the zero crossing interval *following* that interval containing the maximum value of s(t) is a better estimate of F_1 than that derived from the average zero crossing interval length.

Scarr also showed that the following conditions govern the number of zero crossings per second--ρ_o--of a slightly more complicated version of (6-13):

$$s(t) = A_1 \sin n t + A_2 \sin (m t+\theta) \;,\quad m > n. \qquad (6\text{-}37)$$

1: If $A_2 > A_1$, $\rho_o = 2mf$, $f = \omega/2\pi$.

2: If $A_2 = A_1$, ρ_o is usually 2mf but may be, depending on θ, (n+m)f.

3: If $m/n > A_1/A_2 > 1$, $\rho_o = 2pf$, where $n < p < m$.

4: If $m/n = A_1/A_2$, $\rho_o = 2nf$, including (m-n)f triple zeros if $\theta = 0$ or 2π.

5: If $A_1/A_2 > m/n$, $\rho_o = 2nf$.

Clearly, these results represent an extension and confirmation of those of E. Peterson and Lobanov. For example, note that if SSB modulation is applied to s(t) then, for a carrier frequency ω_o such that $\omega_o = k\omega >> m\omega$, $(m+k)/(n+k) \to 1$. Then, as Peterson noted, regions 3: and 4: are narrowed and

$$\rho_o \simeq \tilde{\rho}_o/\pi = \begin{cases} 2mf + \omega_o/\pi \;, & A_2 > A_1 \\ 2nf + \omega_o/\pi \;, & A_1 > A_2 \;. \end{cases} \qquad (6\text{-}38)$$

Scarr's experimental work consisted of a comparison of formant frequency estimations derived using the "second crossing interval"(equation (6-36)) of *bandpass filtered* speech sounds [passband = 250-1200 Hz, 950-1500 Hz *or* 1500-3000 Hz] with those extracted from a 13 channel third-octave filter bank [290-6000 Hz] by "peak-picking" techniques. The true formant positions were visually determined by inspection of speech spectrograms. Scarr summarized his results by referring to physiological vowel correlates (see Fig. 3.8b):

For "front" vowels (/i/,/I/,/ɛ/,/æ/) both methods gave good F_1-F_2 estimation and separation.

For "central" vowels (/ɔ/,/ʌ/,/a/,/ɚ/), F_1 and F_2 fall within the *same* (250-1200 Hz) region and the zero crossing estimate gave the average frequency of F_1 and F_2. This result is in overall agreement with Focht's SEF findings (see Fig. 6.6), and both Peterson's [P-9], Lobanov's and Scarr's predictions concerning two-tone signals.

For "back" vowels (/U/,/u/), having closely spaced F_1-F_2 and large F1, zero crossing estimates indicated the position of F1, F_1.

Generally, in close agreement with the analysis of Peterson and Hanne [P-12], both the zero crossing and "peak-picking" methods were subject to large errors and neither was entirely satisfactory.

6.3.6 Summary

Before closing this section, we note that Lavington demonstrated experimentally that--for synthesized speech sounds--the following zero crossing-formant frequency correlations can be observed [L-4]:

1) The number of zero crossings "T" per 10 msec. of the differentiated waveform shows a close correlation with the average value of F_2 and F_3, i.e., $\text{"T"} \simeq 0.05(F_2+F_3)/2 - 73$.

2) Plots of the number of waveform zero crossings per 10

msec., "Z", vs "T"--for various phonemes--ostensibly divided the Z-T plane into isophonemic regions.

However, the measurements seem quite arbitrary and no rationale is given for using them.

Finally, Ahmed showed [A-1] that if the number of zero crossings in a short time interval, "n", is plotted against the time interval duration, Δt, for a sustained vowel, then a straight line approximated by $n=k\Delta t$ results. The slopes "k" for *different* speakers uttering the *same* vowel are more similar than for one speaker uttering different vowels. This report, however, is not conclusive.

In summary, the use of zero crossings for formant frequency estimation is, theoretically, well founded and, experimentally, reasonably successful *if* prefiltering excludes other formants. If two formants are present then the frequency of either formant can be estimated closely by zero crossing methods *if* suitable pre-emphasis ensures that the amplitude of the desired formant is dominant, and SSB counting methods (e.g., sec. 6.3.2) are used.

6.4 Frequency Division by Zero Crossing Manipulation

We have already briefly discussed a specific type of speech signal transformation, single sideband modulation (SSB), and two phenomena associated with it:

1. Single Sideband Clipping (M1, sec. 5.1.7):

The envelope of a speech signal--$|m(t)|$--is not perceptually recognizeable as speech; the phase function of a speech signal--$\cos \phi(t)$--is, perceptually, essentially the same as the original signal. That is,

$$\cos \phi(t) \overset{P}{=} |m(t)| \cos \phi(t) \quad ,$$

where "P" denotes "perceptually."

2. <u>Single Sideband Frequency Estimation</u> (sec. 6.3.2):

Estimation of the average value of the instantaneous frequency--$\overline{\phi'(t)}$--of a two-tone signal (one approximation to ρ_o, the average time rate of zero crossings) is ambiguous for a wide range of tone amplitude ratios *unless* SSB modulation methods are used.

In summary, the phase of s(t), $\phi(t)$, yields *both* cos $\phi(t)$,

$$\text{and} \qquad \cos \phi(t) \overset{P}{=} s(t) \tag{6-39}$$

and $\phi'(t)$,

$$\text{where} \qquad \overline{\phi'(t)} \equiv \tilde{\rho}_o \ . \tag{6-40}$$

6.4.1 <u>Bandwidth Compression Techniques</u>

Marcou and Daguet reasoned that if the constant amplitude signal

$$s_{\omega_o SSB}(t) = \cos\,[\omega_o t + \phi(t)] \quad , \tag{6-41}$$

is frequency divided by "n" to yield

$$s_{n,\omega_o SSB}(t) = \cos\{[\omega_o t + \phi(t)]/n\} \ , \tag{6-42}$$

then, provided that $\phi'(t)_{max} < \omega_o$, "the spectrum of $\cos\{[\omega_o t+\phi(t)]/n\}$ will be effectively narrower than that of $\cos\,[\omega_o t+\phi(t)]$ by the factor *n*." They implemented this system and found that, for speech input, the signal obtained by frequency division, transmission over a channel, and frequency multiplication "was evidently of high intrinsic intelligibility but . . . difficulties are encountered *when it is required to pass the divided signal through a narrow band filter* which cuts off sharply." (Italics mine.) [M-5]

Cherry and Phillips explained this phenomenon. They noted [C-9] that equation (6-17), describing $\phi'(t)$ for a two-tone signal, can be rewritten as

$$\phi'(t) = \frac{\omega_1 + q^2\omega_2 + q(\omega_1+\omega_2)\cdot\cos(\omega_2-\omega_1)t}{1 + q^2 + 2q\cdot\cos(\omega_2-\omega_1)t} , \qquad (6\text{-}43)$$

with $q = A_2/A_1$. From (6-43), it is clear that $\phi'(t)$ is a periodic function, period $T = 2\pi/(\omega_2-\omega_1)$. In Fig. 6.3, for example, $T = \pi/\omega$. Therefore, although the frequency divided signal, $\cos\{[\omega_o t+\phi(t)]/n\}$, has its major component at frequency $(\omega_o+\omega_1)/n$--assuming that $A_1>A_2$--, the second harmonic occurs at

$$\omega = (\omega_o+\omega_1)/n + (\omega_2-\omega_1) . \qquad (6\text{-}44)$$

This demonstrates that although the major (i.e., greatest amplitude) signal component *is* divided down in frequency, the inter-tone spacing, $\omega_2-\omega_1$, is preserved. Frequency division is, therefore, a *bandwidth preserving transformation.* In the case of unvoiced sounds the argument given against bandwidth reduction by frequency division is less convincing.

Marcou and Daguet's alternative suggestion, that $\phi'(t)$ be divided and manipulated directly for bandwidth compression purposes, neglects the fact that--as shown by Peterson--$\phi'(t)$ is not band-limited.

R. Bogner experimentally confirmed that, for more complicated signals, frequency division has two major effects. First, it translates the entire signal spectrum downwards in frequency, with the spectral component of largest magnitude being the only component that is truly frequency divided [B-11].[5] Second, it suppresses [relatively] minor spectral components and tends to produce a spectrum which is symmetrical about the largest magnitude component. In addition, he demonstrated analytically that the recovery of the

---

[5]Bogner showed that, similarly, frequency multiplication produces an upward frequency translation of the signal spectrum about the largest amplitude frequency component. [X-1]

original signal after remultiplication depends upon the cancellation of several terms, so that very accurate preservation of the phase and amplitude characteristics of the frequency divided signal is important. Bogner also explained that the distortion noted by Marcou and Daguet when a lowpass filter was inserted into their frequency division system was undoubtedly attributable to phase errors incurred by frequency division during time periods of small minima of $|m(t)|$. Although "jumps" in phase of $2\pi/n$ (in an imperfectly frequency divided signal) are audible only as a series of faint clicks, insertion of a narrow bandwidth filter modifies the clicks so as to produce "chirps", following signal multiplication. Frequent chirps produce a characteristic "burbling" distortion.

Bogner noted that rooting the envelope as well as dividing the phase,

$$\text{i.e.,} \quad s_{1/n}(t) = |m(t)|^{1/n} \cos[\phi(t)/n] \qquad (6\text{-}44)$$

effects an apparent expansion of the dynamic range of the frequency divider and may make the system less amenable to phase errors.

J.L. Daguet had used (1963) signal rooting (n=8) in each of the three, separated speech formant ranges (300-700 Hz, 700-2000 Hz, and 2000-3400 Hz) to implement a practical bandwidth compression system [D-1] using SSB modulation. Schroeder, Flanagan and Lundry later [1967; S-8] simulated a four-channel, bandwidth compression system--using "signal rooting"--directly, without SSB modulation. They showed that, for n=2, (6-44) can be written as

$$s_{1/2}(t) = (1/2)^{1/2} \, [\,|m(t)| + s(t)]^{1/2} \, , \qquad (6\text{-}45)$$

and noted that the phase ambiguity inherent in taking the square root can be avoided by changing the sign of $s_{1/2}(t)$ whenever $\phi(t)$ goes through an integer multiple of 2π radians.

6.5 The Relationship between the Spectrum and the Instantaneous Frequency of a Signal

Both $\phi'(t)$, the instantaneous frequency, and $S(f)$, the Fourier transform or "spectrum", are derived from the same source--the signal $s(t)$; both constitute, in different senses, descriptions of that signal. In section 6.3.2, for example, we noted that for a two-tone signal, $\overline{\phi'(t)}$ indicates the spectral frequency of the tone having the larger amplitude. Can direct, more generalized, relationships be established between $\phi'(t)$ and $S(f)$? We provide some answers to this question in this section.

6.5.1 Fink's Theorems

We first define

$$\omega_I = \frac{\int_0^\infty \omega\, G(\omega)\, d\omega}{\int_0^\infty G(\omega)\, d\omega}, \qquad (6\text{-}46)$$

as the *mean frequency*, or centroid, of the power spectrum $G(\omega)$. Here, $G(\omega) = |S(\omega)|^2$, where $S(\omega) = \mathcal{F}\{s(t)\}$.

The *mean-square frequency* of $G(\omega)$ is defined as [B-16, p. 155]

$$\omega_{II} = \frac{\int_0^\infty \omega^2\, G(\omega)\, d\omega}{\int_0^\infty G(\omega)\, d\omega}, \qquad (6\text{-}47)$$

while the *mean-square width* of $G(\omega)$ is [B-16, p. 156]

$$(\Delta\omega)^2 = \frac{\int_0^\infty (\omega-\omega_I)^2\, G(\omega)\, d\omega}{\int_0^\infty G(\omega)\, d\omega} = \omega_{II}-\omega_I^{\,2}. \qquad (6\text{-}48)$$

L. Fink defined the following measures of the instantaneous frequency, $\phi'(t)$, of the signal $s(t)$ having envelope $|m(t)|$ and phase $\phi(t)$ [F-7]:

The *mean instantaneous frequency*:

$$\Omega_I = \lim_{T\to\infty} \frac{\int_{-T}^{T} \phi'(t)\cdot|m(t)|^2\, dt}{\int_{-T}^{T} |m(t)|^2\, dt} \qquad (6\text{-}49)$$

The *mean-square instantaneous frequency*:

$$\Omega_{II} = \lim_{T\to\infty} \frac{\int_{-T}^{T} [\phi'(t)]^2\cdot|m(t)|^2\, dt}{\int_{-T}^{T} |m(t)|^2\, dt} \qquad (6\text{-}50)$$

The *mean-square width* (or deviation from Ω_I) of $\phi'(t)$:

$$(\Delta\Omega)^2 = \lim_{T\to\infty} \frac{\int_{-T}^{T} [\phi'(t)-\Omega_I]^2\cdot|m(t)|^2\, dt}{\int_{-T}^{T} |m(t)|^2\, dt} \qquad (6\text{-}51a)$$

or

$$(\Delta\Omega)^2 = \Omega_{II} - \Omega_I^2 \quad . \qquad (6\text{-}51b)$$

Note that for signals periodic in T, all integrals need only be evaluated over [0,T].

Fink established the following results:

1: $$\Omega_I = \omega_I \qquad (6\text{-}52)$$

2: $$\Omega_{II} \leqslant \omega_{II} \qquad (6\text{-}53)$$

3: $$(\Delta\Omega)^2 \leqslant (\Delta\omega)^2 \qquad (6\text{-}54)$$

These results are subject to the existence of ω_I and ω_{II} and are valid for the signals we shall consider. It can be shown that equations (6-53) and (6-54) become equalities for $|m(t)|$ constant.

6.5.2 $\overline{\phi'(t)}$ and Ω_I

Fink's theorems establish direct links among Ω_I, $\phi'(t)$, $|m(t)|$, and $G(\omega)$. However, how do we interpret the definition of Ω_I? For example, is Ω_I related to $\overline{\phi'(t)}$? In order to establish a relationship between Ω_I and $\overline{\phi'(t)}$, we turn to a result of Hiramatsu *et al.* [H-16].

Specifically, they proved that the mean value of $\phi'(t)$, $\overline{\phi'(t)}$, over an arbitrary time T is:

$$\overline{\phi'(t)} = \mathrm{Re}[M^{(1)}] - \mathrm{Im}[M^{(2)}/2!]T - \mathrm{Re}[M^{(3)}/3!]T^2 + \mathrm{Im}[M^{(4)}/4!]T^3 \;. \;. \tag{6-55}$$

where

$$M^{(1)} = \frac{\int_0^\infty \omega\, S(\omega)\, d\omega}{\int_0^\infty S(\omega)\, d\omega} \;, \tag{6-56}$$

and

$$M^{(n)} = \frac{\int_0^\infty [\omega - M^{(1)}]^n\, S(\omega)\, d\omega}{\int_0^\infty S(\omega)\, d\omega} \;, \quad n>1 \;. \tag{6-57}$$

If $S(M^{(1)}+\Delta\omega) \simeq S^*(M^{(1)}-\Delta\omega)$, and if T is "small" (e.g., $T \simeq 30$ msec. for speech signals, as per sec. 6.3.2), then the contributions of the higher order moments, $M^{(n)}$, in (6-55) are negligible and

$$\overline{\phi'(t)} \simeq \mathrm{Re}[M^{(1)}] = \mathrm{Re}\left[\frac{\int_0^\infty \omega\ S(\omega)\ d\omega}{\int_0^\infty S(\omega)\ d\omega}\right] . \qquad (6\text{-}58)$$

Furthermore, the assumption that $S(\omega)$ possesses "symmetry" about its mean guarantees that

$$\mathrm{Re}\left[\frac{\int_0^\infty \omega\ S(\omega)\ d\omega}{\int_0^\infty S(\omega)\ d\omega}\right] = \frac{\int_0^\infty \omega\ G(\omega)\ d\omega}{\int_0^\infty G(\omega)\ d\omega} . \qquad (6\text{-}59)$$

Combining (6-46), (6-52), and (6-59) we find that, when the "symmetry" conditions on $S(\omega)$ are satisfied,

$$\overline{\phi'(t)} \simeq \omega_I = \Omega_I . \qquad (6\text{-}60)$$

In summary, Hiramatsu had shown analytically that the centroid, ω_I, of the power spectrum, $G(\omega)$, is a reliable estimate of $\overline{\phi'(t)}$ *only* when $S(\omega)$ satisfies the amplitude-phase symmetry criteria. Fink's first theorem tells us *why*:

i.e., for periodic signals,

$$\omega_I = \Omega_I \equiv \frac{\int_0^T \phi'(t)\cdot|m(t)|^2\ dt}{\int_0^T |m(t)|^2\ dt}$$

and for $\Omega_I = \frac{1}{T}\int_0^T \phi'(t)\ dt = \overline{\phi'(t)}$, it is sufficient (but not necessary) that the symmetry conditions be satisfied. Thus, when the symmetry conditions *are* satisfied, Fink's second and third theorems establish bounds on the *mean-square deviation* of the instantaneous frequency from its average value. More important perhaps, the first theorem implies that $\overline{\phi'(t)}$,--for signals which satisfy the required amplitude-phase criteria--is independent of the *absolute value* of the phase spectrum of $S(\omega)$ provided that the required phase symmetry is present.

For the simple example of the two-tone signal [$s(t) = A_1 \sin\omega_1 t + A_2 \sin\omega_2 t$] substitution of $|m(t)|$, (6-15), and $\phi'(t)$, (6-43), into (6-49) yields

$$\Omega_I = \frac{\omega_1}{1+q^2} + \frac{\omega_2}{1+q^{-2}} \quad , \tag{6-61}$$

where $q = A_2/A_1$. When the symmetry criterion is satisfied,

$$\text{i.e.,} \quad q \gg 1 \quad \text{or} \quad 0 < q \ll 1 \quad ,$$

(6-61) yields $\Omega_I \approx \overline{\phi'(t)} = \omega_2$ or ω_1, respectively. This is in agreement with (6-19). For example, when $q = 4$ or $1/4$, then $\Omega_I = 0.94\omega_2 + 0.06\omega_1$ or $0.94\omega_1 + 0.06\omega_2$, respectively. It should be noted that, although (6-49) is fairly difficult to evaluate directly, Fink's theorem 1, equation (6-52), enables (6-61) to be calculated by simple, direct reference to the power spectrum of $s(t)$.

As another example, the bound on the deviation of $\phi'(t)$ from Ω_I (for the two-tone signal) can be calculated directly using (6-48) and (6-54):

$$\text{i.e.,} \quad (\Delta\Omega)^2 \leqslant (\Delta\omega)^2 = [q/(1+q^2)]^2 \, (\omega_1 - \omega_2)^2 \; . \tag{6-62}$$

Then, for values of q such that $\Omega_I \approx \overline{\phi'(t)}$, we would expect that

$$[\Delta\phi'(t)]^2 \approx (\Delta\Omega)^2 \leqslant [q/(1+q^2)]^2 (\omega_1 - \omega_2)^2 \quad , \tag{6-63}$$

where $$[\Delta\phi'(t)]^2 = \overline{\phi'(t)^2} - \overline{\phi'(t)}^2 \tag{6-64}$$

is the mean-square deviation of $\phi'(t)$ from $\overline{\phi'(t)}$.

The *actual* value of $[\Delta\phi'(t)]^2$ is calculated by using $\phi'(t)$, (6-17), in (6-64). This yields, after much manipulation,

$$[\Delta\phi'(t)]^2 = [(|k|-1)^2/8|k|]\ (\omega_1-\omega_2)^2\ , \tag{6-65}$$

with
$$k = (A_1-A_2)/(A_1+A_2) \tag{6-66a}$$
$$= (1-q)/(1+q)\ . \tag{6-66b}$$

Rewriting (6-62) with $q = (1-k)/(1+k)$ gives

$$(\Delta\omega)^2 = [(1-k^2)/2(1+k^2)]^2\ (\omega_1-\omega_2)^2\ .$$

Hence, for the approximate range $[1/2<|k|<1]$--i.e., $[0<q<1/3]$ or $[3<q<\infty]$--$\overline{\phi'(t)}\approx\Omega_I$ and, from (6-63), we expect that

$$[\Delta\phi'(t)]^2 = (\omega_1-\omega_2)^2[(|k|-1)^2/8|k|]\lessgtr(\omega_1-\omega_2)^2[(1-k^2)/2(1+k^2)]^2 = (\Delta\omega)^2$$

or
$$[(|k|-1)^2/2|k|]\lessgtr[(1-k^2)/1+k^2]^2\ . \tag{6-67}$$

This is indeed the case.

The two-tone signal is, spectrally, somewhat simple, However, the single formant resonator satisfies the symmetry criteria (see sec. 6.3.5, equations (6-27) and (6-28)). We would therefore expect that the value of $\overline{\phi'(t)}$ would accurately approximate F_1, the frequency of maximum resonator output ($\approx$ formant frequency).

Experimentally, Hiramatsu found that the estimate of formant frequency afforded by $\overline{\phi'(t)}$ *was* invariably more accurate than that resulting from the calculation of ω_I, equation (6-46). For unfiltered vowels (i.e., multi-peaked spectra) the *maximum* error in estimating the frequency of the largest amplitude formant using $\overline{\phi'(t)}$ --provided the amplitude of this formant was at least 6 db above the others--was $F_o/2$, half the voicing frequency. This is exactly the

the maximum error predicted by Peterson and Hanne for measuring the formant frequency of a periodically pulsed single formant resonator (sec. 6.3.3) using "average rate of zero crossings"![6] In contrast, the maximum error of ω_I based estimates was F_o. When bandpass filtering was introduced to insure that only the first formant, F1, was present--thus satisfying the symmetry conditions--both the $\overline{\phi'(t)}$ and ω_I estimates had a *maximum* error of $F_o/2$. Generally, the $\overline{\phi'(t)}$ estimates were more accurate than the ω_I estimates.

In summary, we have shown that the average value of the instantaneous frequency $\phi'(t)$ --$\overline{\phi'(t)}$-- of a signal provides an accurate, reliable estimate of the centroid of the *power spectrum* $G(\omega)$ only when the *spectrum* $S(\omega)$ possesses amplitude-phase symmetry about the centroid ω_I. When the symmetry criteria *are not satisfied*, then

$$\omega_I = \Omega_I = \frac{\displaystyle\int_0^T \phi'(t)\cdot|m(t)|^2\,dt}{\displaystyle\int_0^T |m(t)|^2\,dt} \neq \overline{\phi'(t)}, \text{ by definition.}$$

Therefore, in the case when the symmetry criteria are satisfied (again using a periodic signal)--e.g., the single formant resonator--,

$$\omega_I = \Omega_I = \overline{\phi'(t)}$$

and, since $\omega_I = 2\pi f_I$, we obtain an expression for the average rate of zero crossings, ρ_o, using the results of Peterson and Hanne (sec. 6.3.3):

$$\rho_o = 2\tilde{f}_I \; . \tag{6-68}$$

Here, $\tilde{f}_I$ is the *nearest* spectral component to f_I. Therefore, for

---

[6]Hiramatsu was apparently unaware of this work.

the periodically excited single formant resonator,

$$\rho_o = 2\ \tilde{f}_I \simeq 2\ f_I = \tilde{\rho}_o/\pi = \overline{\phi'(t)}/\pi$$

or

$$\rho_o \simeq \overline{\phi'(t)}/\pi = \tilde{\rho}_o/\pi \quad . \tag{6-69}$$

It follows that (as Hiramatsu experimentally observed) the best *zero crossing estimate* of F_1, for an isolated formant, is

$$\overline{\phi'(t)}/2\pi = \tilde{\rho}_o/2\pi = f_I \quad . \tag{6-70}$$

6.6 Zero Crossing Interval Sequences as Descriptors of Speech Sounds

In section 6.3 we described methods of processing "zero crossings" so as to yield an objective estimate of speech formant frequencies. The interpretation of zero crossing *interval sequences* as patterns, without explicit reference to the frequency domain, is of great relevance to automatic speech recognition studies.

6.6.1 The Intervalgram

S. Chang proposed [C-5] that if the interval, Δt, between adjacent zero crossings of a speech waveform or its derivative is displayed as a function of time, then the number of points per unit area on the Δt-t plane--defined as the *intervalgram*--could be interpreted in a manner analogous to the spectral energy density displayed in a spectrogram. Figure 6.8 shows the method of generation of intervalgrams. Figure 6.9 a,b,c shows intervalgrams for the vowel /u/, while Fig. 6.10 shows an intervalgram for the words "one, two" spoken in succession.

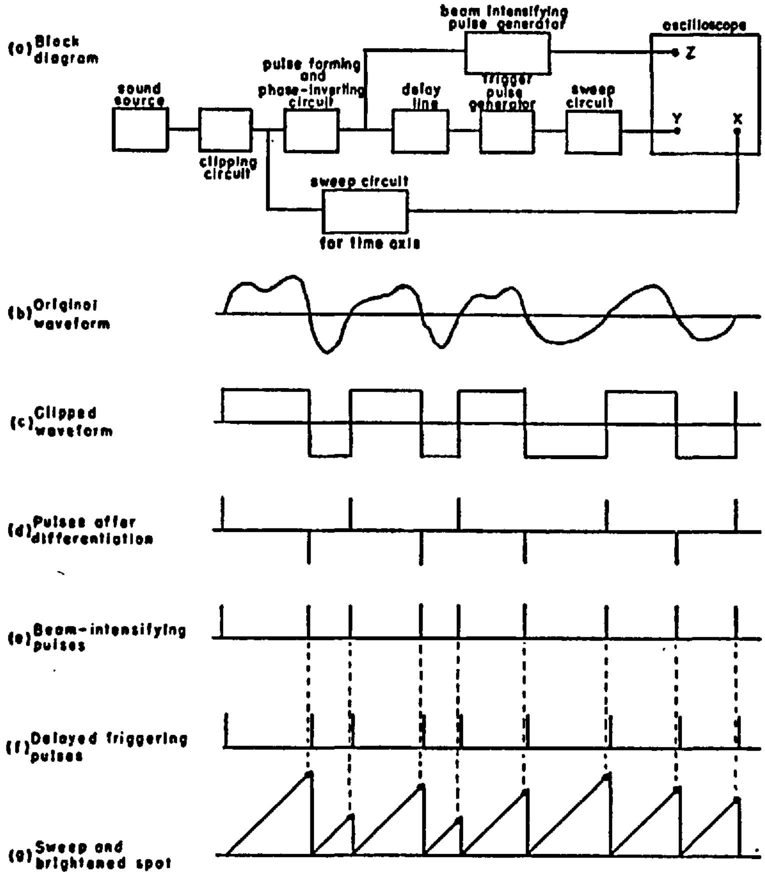


Fig. 6.8 Generation of "Intervalgram."
(From [C-5].)

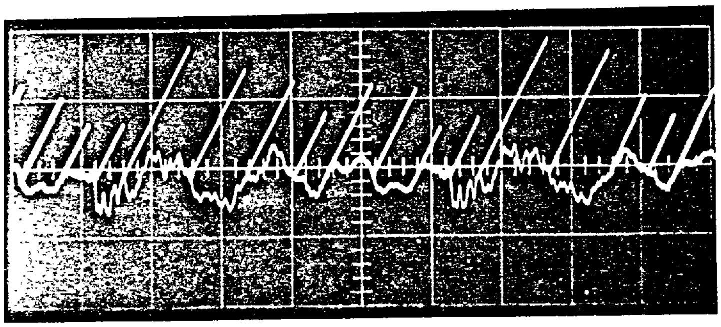


Fig. 6.9a Intervalgram for vowel /u/, speaker LRM.
Sweep = 2 msec/cm.

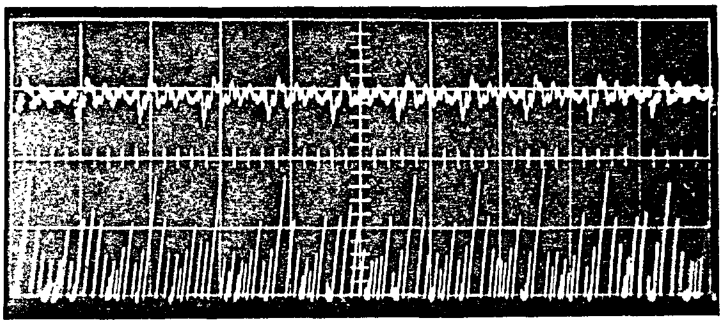


Fig. 6.9b Intervalgram for /u/.
Sweep = 10 msec/cm.

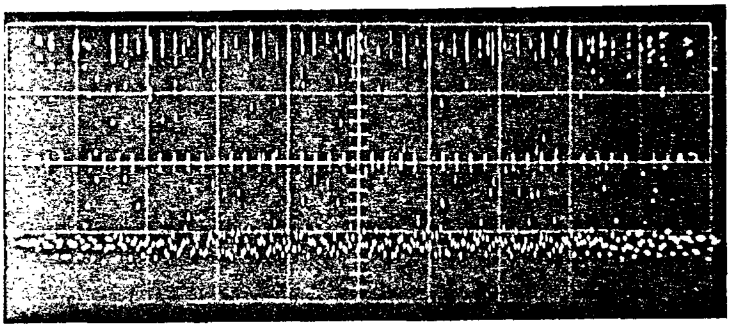


Fig. 6.9c Intervalgram for /u/.
Sweep = 50 msec/cm, with beam suppression as per Fig. 6.8g.

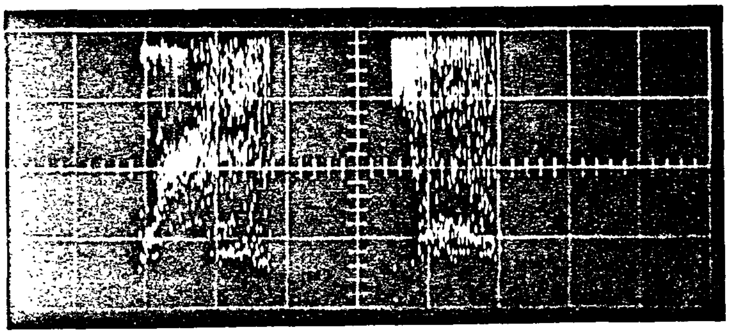


Fig. 6.10 Intervalgram for words "one, two" spoken in succession. Sweep = 500 msec/cm.

We recall here that the bandwidths chosen for the "analyzing filters" in the speech spectrograph are invariably a compromise between frequency resolution and time resolution. M. Lecours has shown, for example, that Cherry's suggestion (sec. 3.2.3) regarding variable bandwidth filters in models of the auditory system is applicable to automatic speech recognition [L-5]. Is it possible (as Chang suggests) that zero crossing intervals, which may be measured with arbitrary accuracy, are in some respects superior to short-term spectral analyses as an estimate of the speech waveform?

Chang also noted that other functions can be substituted for the linear ramp--which gives a vertical axis gradation linear with respect to time. A hyperbolic wave generator, which can be approximated by an exponential source, gives a vertical axis gradation linear in frequency. Finally, Chang argued that the centre of gravity of the intervalgram, with respect to the Δt scale, approximates the ρ_o function.

T. Sakai and S. Inoue suggested that the zero crossing intervals of speech waveforms be classified into a number of channels [S-1], each channel corresponding to a range of zero crossing interval lengths. This is equivalent to dividing the vertical axis of the intervalgram into a number of discrete, contiguous segments, or "bins", and projecting the "dots" representing the lengths of the zero crossing intervals occurring over some larger time interval--corresponding to a vowel, for example--horizontally, i.e., into the "bins". The array of numbers representing the number of interval lengths falling into each "bin", or channel, can be defined as a *zero crossing interval histogram*.

More specifically, Davenport defined a first-order density distribution associated with measurement of zero crossing intervals over a time interval T [D-3]:

$$f_t(\tau_{mi}) = \frac{1}{\Delta\tau_i} \frac{\tau_{mi}\; n_i}{T} \quad , \; i = 1,\ldots c \quad , \qquad (6\text{-}71)$$

where n_i is the number of zero crossing intervals falling into the i^{th} of c channels,

$\Delta\tau_i$ is the time interval difference between the upper and lower limits of the i^{th} channel,

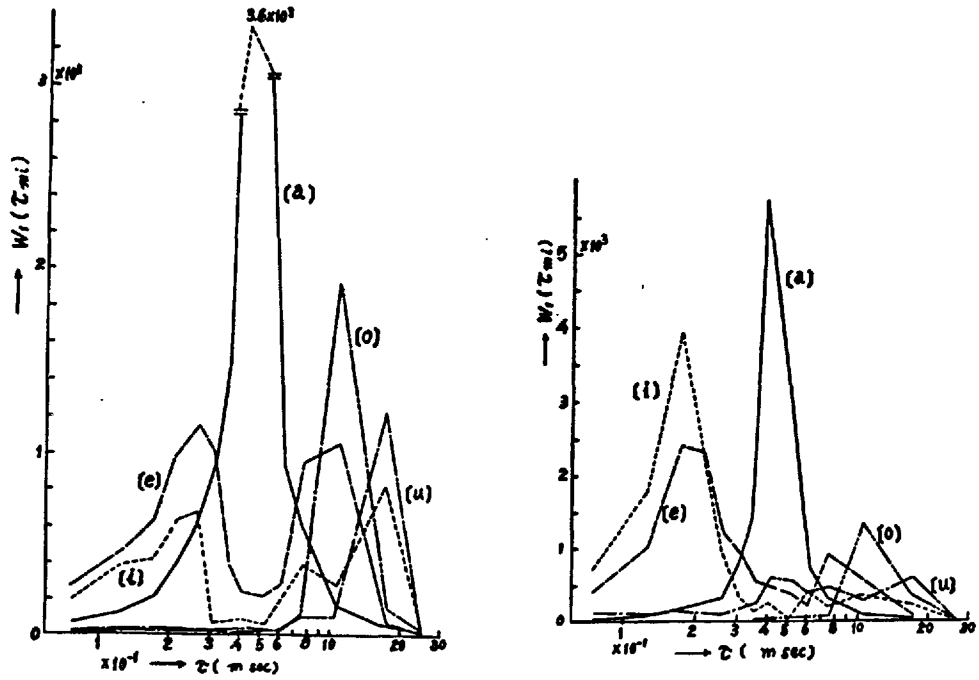
and τ_{mi} is the time representing the midpoint of the i^{th} channel.

In equation (6-71), $f_t(\tau_{mi})$ is defined by [D-3]

$$f_t(\tau_{mi})\cdot\Delta\tau_i = P(\tau_{mi}) \quad , \qquad (6\text{-}72)$$

where $P(\tau_{mi})$ is the probability that a given instant of time t, the duration of the zero crossing interval falls within the limits $[\tau_{mi}-\Delta\tau_i/2,\ \tau_{mi}+\Delta\tau_i/2]$. Strictly speaking, (6-71) obtains only for $T\to\infty$ and $\Delta\tau_i\to 0$.

Sakai and Inoue measured $f_t(\tau_{mi})$ for Japanese vowels and found that characteristic peaked distributions resulted (Fig. 6.11). They noted that "the peaks in longer intervals seem to correspond to the first formants of /i/, /e/, /o/, and /u/, but the peak of /a/ is probably a combination of the first and second formants. The peaks in shorter intervals are the second or third formants of /i/ and /e/." [S-1] They further observed that "if the peak in a [the] shorter interval is removed from /i/ by a low-pass filter, the distribution of the zero crossing wave turns to that of /u/. Such an /i/ is misheard by listeners as /u/ . . . It was found that filtered [low- and high-pass] vowels that generate similar distribution patterns were often confused with each other in listening tests made at the same time." It appears, therefore, that zero crossing interval distributions *implicitly* relate back to the spectral nature of the source signal.



a: original signal b: differentiated signal

Fig. 6.11 First-order density distribution $W_1(\tau_{mi})$ $[f_t(\tau_{mi})]$ vs τ $[\tau_{mi}]$, for a male Japanese speaker. (From [S-1].)

Histograms, or first-order distributions (which are really weighted histograms), fail to retain information concerning the sequence in which different zero crossing interval lengths occur. Experiments relating to the *digram structure* of the zero crossing intervals of speech waveforms (specification of the relative frequencies of occurence of different pairs of interval lengths in succession) have been carried out by MacKay *et al.* [M-2]. They found that digram displays discriminate among vowel sounds that generate almost identical histograms, and that articulator movements are reflected in "corresponding movements of major points of the display."

6.7 The Use of Zero Crossings in Automatic Speech Recognition: Some Examples

In section 4.3 we described in detail some schemes for using spectral speech data in automatic recognition. Besides considering the measure of spectral information used, we also briefly described the method of training the machine and carrying out the recognition phase. This description was intended to familiarize the reader with conventional methods prior to the introduction of adaptive algorithms in chapter 7.

Therefore, in this section the discussion of the use of zero crossings in automatic speech recognition will be limited to a description of the measure of zero crossing information used, and the rationale for the particular choice. We shall group the schemes according to the measure of zero crossing information used.

6.7.1 Average Rate of Zero Crossings

"Audrey"--an automatic digit recognizer--was one of the earliest attempts at automatic speech recognition [D-6, 1952]. Audrey used the average rate of zero crossings in 1) the 200-900 Hz band and 2) the 800-3000 Hz band (presumably detected by $\tilde{\rho}_o$-meters) as input to the X and Y axes of an oscilloscope. The time varying trajectory displayed for each spoken digit was then regarded as a pattern representative of that digit. Training and recognition was effected by measuring the time occupancy, of the trajectory, in each of 30 squares of a 6x5 grid superimposed upon the oscilloscope screen.

Subject to the criterion that only one formant was actually present in each of the two channels, Audrey's input consisted of zero crossing estimates of F_1 and F_2 of the type described in sec. 6.3.3 and 6.5. Effectively, then, this was an F_1-F_2 tracker and the results (≈98% correct classification for a single speaker) are about the same as those of other systems of the same genre(e.g., sec. 4.3.1).

Seventeen years after "Audrey", Ewing and Taylor [1969, E-4] suggested that "a display of averaged zero crossing rate of the original waveform versus that of the differentiated waveform should be of interest . . . it would be a pattern defined by the first and second formants. . . ."[6]

C. Howard also used a $\tilde{\rho}_o$-meter to track F_1 in bandpass filtered (300-1000 Hz) speech [H-21]. He then used this first $\tilde{\rho}_o$ estimate to tune an active filter so as to more accurately define the actual position of F_1. A second $\tilde{\rho}_o$-meter was applied to the active filter output and a final, ostensibly more accurate, F_1 estimate resulted. Finally, the accurate F_1 estimate was used to tune another active filter so as to ensure that no F1 energy entered the F2 $\tilde{\rho}_o$-meter.

Lobanov showed that average zero crossing rates could also be used to separate phonemes into various classes [L-24]. We recall his expression, equation (6-14), for the average number of zero crossings per second of a two-tone signal (equation (6-13)):

$$\rho_o = \begin{cases} \frac{2}{\pi}(2F_2-2F_1)\cdot\sin^{-1}[A_2/A_1] + 2F_1 \ , & 0\leq A_2/A_1\leq 1 \\ 2F_2 \ , & A_2>A_1 \ . \end{cases} \quad (6\text{-}73)$$

If this function is plotted and compared to the (imperfect) estimate of the average value of instantaneous frequency obtained by an audio band $\tilde{\rho}_o$-meter (Fig. 6.2), it is clear that the two curves are equal for $A_2>A_1$ and are identical in shape for $A_1>A_2$. However, Lobanov's

[6] They also claimed that "we have found no indication in the literature. . . to show that anyone else has attempted to verify Chang's conclusions [regarding the similarity of ρ_o and F_1, and ρ_m and F_2] for speech sounds." Peterson (sec. 6.3.2), of course, quantified Chang's conclusions [C-4] in the same year Chang's work was published.

function has a slower rate of fall for decreasing A_2/A_1.

Lobanov, and Howard, argued that *unvoiced fricatives* (/f/, /θ/, /s/, /ʃ/) can be modelled as a band of white Gaussian noises of "proper center frequency and bandwidth." [H-21] Such a signal, having bandwidth Δf and centre frequency f_o, has an average time rate of zero crossings given by (from equation 6-2):

$$\rho_f = 2\ [f_o{}^2 + \Delta f^2/12]^{\frac{1}{2}} \quad . \tag{6-74}$$

Finally, Lobanov suggested that an acceptable model for certain *voiced fricatives* (e.g., /z/) is a sine wave of having randomly distributed amplitude and phase (but not frequency) superimposed upon a white Gaussian noise background. In this case (see [B-3, p. 384]),

$$\rho_{vf} = 2 \left[\frac{f_s{}^2 \cdot \overline{Q}^2/2 + M_2}{\overline{Q}^2/2 + M_0} \right]^{\frac{1}{2}} \quad . \quad . \tag{6-75}$$

Here the sine wave is $r(t) = Q \cdot \sin(2\pi f_s t + \phi)$ and $r(t)$, Q, and ϕ have (respectively) Gaussian, Rayleigh, and uniform distributions. $\overline{Q}=E\{Q\}$ and M_0, M_2 are defined by equation (6-10). For $\overline{Q}\to 0$, (6-75)→(6-11); for $G(f)\equiv 0$, (6-75) = $2f_s$, as expected. If $A_2>A_1$, $\rho_o>\rho_{vf}$; if $A_1>A_2$, a value of $[A_2/A_1]$ can always be found such that $\rho_{vf}>\rho_o$.

Lobanov showed that by proper use of pre-emphasis, fricatives (both voiced and unvoiced) can be separated from vowels using average zero crossing measurements and equations (6-73), (6-74) and (6-75). For example, good separation of vowels from unvoiced fricatives is ensured by pre-emphasizing the first formant region; then $\rho_o \approx 2F_1$ while ρ_f is very large. However, since $f_s \approx F_1$ for speech sounds (equation (6-75)), this type of filtering can lead to a low value of ρ_{vf}. In summary, Lobanov found that--for Russian speech sounds--

a simple differentiating network with a time constant of 48 usec. produced maximum separation of vowels from voiced and unvoiced fricatives when the average zero crossing rate criterion was used.

A scheme somewhat analogous to that proposed by Lobanov had been used by Wiren and Stubbs [W-8] to separate *voiced-unvoiced* sounds in the first stage of a phoneme classification system based upon "distinctive features" (e.g., Cherry *et al.*, [C-8]). They generated a sawtooth voltage between zero crossings (see Fig. 6.8) and allowed only sawtooth peaks greater than some arbitrary height to be amplified and gated through to a "voiced-unvoiced" relay coil. This system depends upon the greater average zero crossing interval in voiced sounds predicted as a consequence of equation (6-73) and observed by Chang (sec. 6.3.1).

Histograms showing the total number of zero crossings for a large sample of unvoiced fricatives and stops suggested to Wiren and Stubbs that these sound classes might be objectively distinguished using a measure of average rate of zero crossings. In fact, an estimate of phoneme *energy* during the time required for the first 40 zero crossings was ultimately chosen. Unvoiced fricatives have *low* average energy and a *high* average zero crossing rate (in equation (6-74), f_o>2000 Hz as per sec. 3.4.7 and [H-9], [H-26]). In contrast, the unvoiced stops have greater average energy and more energy in the low frequency regions--hence a *low* expected zero crossing rate (sec. 3.4.6).

G. Tsemel found that the general features of the spectral noise structure of unvoiced Russian fricatives can be characterized by using measurements of the mean duration of zero crossing intervals during periods of ≈ 25 msec. and the variance of the interval lengths [T-10]. In an earlier paper [T-9], Tsemel had experimentally determined that, for the unvoiced stops (/p/, /t/, /k/) a plot of "n"-- the number of zero crossings in the first 10 msec. of the sound--

vs "t"--the total sound duration--divided the n-t plane into isophonemic regions.

H. Resnikoff discovered that the third-order moment (about the mean) of the reciprocals of the zero crossing interval lengths for /s/ and /z/ (alveolar fricative consonants) are negative; the same measure is positive for all other speech sounds [R-8, 9].

Finally, D. Reddy used the mean [R-4] and standard deviation [R-4, 5, 6] of zero crossing *counts* over 10 msec. periods to aid in resolving ambiguities in segmentation of speech sounds into sustained and transitional segments.

In summary, we note that the known acoustic properties of speech sounds (reviewed in chapter 3) enable models to be formulated which, in turn, suggest certain correlation of average zero crossing rates with spectral features. Additionally, experimentally determined characteristics of zero crossing interval lengths (and variance of interval lengths) have been used in creating tests for discrimination among phonemes.

6.7.2 Zero Crossing Interval Sequences

The simplest zero crossing interval measure is "The zero-axis crossing period of the first excursion in the speech wave after glottal . . . excitation." [T-2] The reciprocal of twice this zero crossing interval is a measure of the Single Equivalent Formant, or SEF, frequency (sec. 6.3.4.). C. Teacher, H. Kellet and L. Focht constructed a compact, limited vocabulary speech recognizer using three parameters: SEF frequency, SEF amplitude (maximum waveform amplitude during SEF zero crossing interval) and state-of-voicing. Performance of the system on the spoken digits, for members of the design or "teaching" group, averaged 90% correct classification [T-2].

W. Bezdel and H. Chandler carried out an exercise in sustained vowel recognition by measuring zero crossing interval histograms [B-6]. The histogram vector (row matrix) for the j^{th} vowel sample is defined by

$$\underline{\underline{X}}_j = [\sum_{i=1}^{c} x_i]^{-1} \cdot [x_1, x_2, \ldots x_c] \qquad (6\text{-}76)$$

where x_i is the number of zero crossing intervals of length τ_i such that $[\tau_{mi} - \Delta\tau_i/2 < \tau_i < \tau_{mi} + \Delta\tau_i/2]$. Here, as in (6-71), $\Delta\tau_i$ is the width, and τ_{mi} the midpoint, of the i^{th} channel. Equation (6-76) is an unweighted version of $f_t(\tau_{mi})$, equation (6-71).

During the learning phase, Bezdel and Chandler established reference sets, $\overline{\underline{\underline{X}}}_j$, for each vowel. Recognition involved comparison of unknown histogram vectors, $\underline{\underline{X}}$, with each reference vector by such methods as dot product [$C_j = \underline{\underline{X}} \cdot \overline{\underline{\underline{X}}}_j$, with j for C_j max identifying the unknown class] or weighted Euclidian distance [$D_{wj} = \underline{\underline{W}}_j \cdot (\overline{\underline{\underline{X}}}_j - \underline{\underline{X}})$, where $\underline{\underline{W}}_j = \overline{\underline{\underline{X}}}_j^{-1}$ or $\underline{\underline{S}}_j$ ($\underline{\underline{S}}_j = [\sigma_{1j}^{-1}, \sigma_{2j}^{-1}, \ldots \sigma_{cj}^{-1}]$, and σ_{ij} is the standard deviation of the i^{th} element of the j^{th} reference vector)]. For these tests, using c=16 and 5 different vowels (j_{max}= 5), the best recognition scores were 97, 95, and 94% for women, men, and mixed groups of speakers, respectively. These scores were obtained using the D_{wj} criterion, with $\underline{\underline{W}}_j = \underline{\underline{S}}_j$.

T. Sakai and S. Doshita extended the ideas presented in sec. 6.5.1 [S-1] by periodically measuring $f_t(\tau_{mi})$ for both the F1 (0-1500 Hz) and F2 (800-2500 Hz) regions of Japanese speech [S-2]. They argued that peaks in $f_{t_{LP}}(\tau_{mi})$ and $f_{t_{HP}}(\tau_{mi})$ should correlate with F_1 and F_2, respectively. A fairly complicated hardware system was provided for speech segmentation and phoneme identification. The recognition rate claimed was 90% for the vowel part, and 70% for the consonant part of Japanese monosyllables.

W. Bezdel and J. Bridle also used broad (LP, HP) filtering as a prelude to zero crossing analysis [B-4,5,7]. In their system, zero crossing intervals are sorted into different channels, as in other systems mentioned. However, the channel boundaries are moveable and a separate *digital interval filter* is used for each sound class to be detected. These filters are dynamically adjusted to maximize discrimination against sounds outside of the design class.

Finally, R. Purton implemented a limited vocabulary word recognizer using the autocorrelation functions of lowpass and highpass filtered, then clipped, speech (0-1 KHz, 1-4 KHz) as patterns to form master matrices for training and recognition.

6.8 Summary

In this chapter we have reviewed, related and evaluated some methods of extracting "useful" information from the zero crossings of speech signals. "Useful" implies that the measure of information extracted is valuable for automatic recognition of speech processing purposes. The relationship among the various techniques described is shown in Fig. 6.12.

s(t) → Clipper	→ Count Zero Crossings	:Lobanov[L-24]	6.7.1
		:Peterson and Hanne [P-12]	6.3.3
		:Reddy [R-4,5,6]	6.7.1
	→ Frequency Division	:Marcou and Daguet [M-5]	6.4
		:Cherry and Phillips [C-9]	6.4
		:Bogner [B-11],[X-1]	6.4
		:Daguet [D-1]	6.4
		:Schroeder *et al.* [S-8]	6.4
	→ $\tilde{\rho}_o$-meter	:Chang [C-4]	6.3.1
		:Peterson [P-9]	6.3.2
		:Cherry and Phillips [C-9]	6.3.2
		:Howard [H-21]	6.7.1
	→ Single Zero Crossing Interval	:Focht [F-10]	6.3.4
		:Scarr [S-4]	6.3.5
		:Teacher *et al.* [T-2]	6.7.2
	→ First Order Intervalgram	:Chang [C-5]	6.6.1
	↓ Histogram	:Bezdel and Chandler [B-6]	6.7.2
		:Wiren and Stubbs [W-8]	6.7.1
	↓ $f_t(\tau_{mi})$	:Sakai and Inuoue [S-1]	6.6.1
	→ Second Order Histogram	:MacKay *et al.* [M-2]	6.6.1
→ LP Filter → Clipper	→ ρ_o or $\tilde{\rho}_o$	:Peterson [P-9]	6.3.2
		:Scarr [S-4]	6.3.5
		:Lavington [L-4]	6.3.6
		:Davis *et al.* [D-6]	6.7.1
→ HP Filter → Clipper	→ ρ_m or $\tilde{\rho}_m$	:Ewing and Taylor [E-4]	6.7.1
	→ $f_{t_{LP}}(\tau_{mi})$	:Sakai and Doshita [S-2]	6.7.2
	→ $f_{t_{HP}}(\tau_{mi})$	:Bezdel and Bridle [B-5]	6.7.2
	→ Correlator	:Purton [P-19]	6.7.2
	→ Correlator		

Fig. 6.12 Zero crossings in automatic speech recognition and processing: Summary of papers reviewed.

7 EXPERIMENTS IN AUTOMATIC SPEECH RECOGNITION USING ZERO CROSSINGS

7.1 Motivation

This chapter is intended to give the reader some feel for the actual mechanics involved in implementing a speech recognition machine. To do this we will briefly review the literature associated with adaptive pattern recognition and then describe in more detail two different methods of pattern recognition, their structure and implementation. The vehicle for this description will be two short experiments in limited vocabulary speech recognition using zero crossing data. These experiments were originally intended to form the nucleus for the implementation of a large scale but limited vocabulary speech recognition machine. As will be noted in sec. 7.9, this goal was abandoned in order to carry on the studies concerning the nature of zero crossings as signal informational attributes which comprise the remainder of this thesis.

7.2 Pattern Recognition

We noted in Chapter 4 that the first step in recognition is parameterization of the signal. The analogy to parameterization in the jargon of pattern recognition is the receptor which "has as its input a physical sample to be recognized, and as an output a set . . . of quantities which characterize the physical sample. These quantities will be called *measurements* of the sample . . . " [H-11].

The output of the receptor is the input to the categorizer, which is "a device which assigns each of its . . . inputs to one of a finite number . . . of categories." [H-11] As Nilsson emphasized, adaptive pattern classifiers or learning machines are concerned with categorization only and that "we shall henceforth assume that the . . . measurements yielding the pattern to be classified have been selected as wisely as possible while remembering that *the pattern classifier cannot itself compensate for careless selection of measurements.*" [N-3]

Some methods of adaptive pattern recognition are taken from classical detection theory [G-12], [T-1], [V-1]. For example, if n classes--S_1, S_2, . . S_n-- are to be identified and thereby correctly categorized, a cost C_{ij} can be assigned to the decision that a member of S_i is identified as belonging to S_j [G-12]. That is, C_{ii} is the cost of correctly identifying a member of S_i whereas C_{ij}, $i \neq j$, is the cost of incorrectly identifying a member of S_i as a member of S_j. C_{io} could be the cost of rejection, or failure to assign a class when the pattern belongs to S_i. Generally, $C_{ij} > C_{io} > C_{ii}$. It can be shown ([G-12], [H-11], for example) that if the a priori probability of occurrence of a pattern of the class i, $1 \leq i \leq n$, is p_i, then the optimum Bayesian categorizer is the implementation of the decision function which minimizes the expected loss

$$C(\delta) = \sum_{i=1}^{n} \sum_{j=0}^{n} C_{ij} \cdot p_i \cdot f_{M|S}(m|S_i) \cdot \delta_{D|M}(d_j|m) \; dm \qquad (7\text{-}1)$$

where

$f_{M|S}(m|S_i)$ is the conditional probability that a certain measurement m will be made, given a pattern from class i at the receptor

and

$\delta_{D|M}(d_j|m)$ is the probability that the decision function or categorizer will make the decision d_j, $0 \leq j \leq n$, given the measurement m, with $j = 0$ corresponding to rejection.

If we let

$$Z_j(m) = \sum_{i=1}^{n} (C_{ij}-C_{io}) \cdot p_i \cdot f_{M|S}(m|S_i) \quad , 1 \leq j \leq n \qquad (7\text{-}2)$$

where $Z_j(m)$ measures the *excess* of the cost of identifying a pattern which gives rise to the measurement m as belonging to S_j over the cost of failure to make any identification ($Z_0(m) = 0$), then it can be shown that $C(\delta)$ is minimized by associating with m the class S_j for which $Z_j(m)$ is least: that is, let $\delta_{D|M}(d_j|m) = 1$ if $Z_j(m) \leq Z_i(m)$, $i \neq j$, and zero otherwise. If the cost of any error is equal and greater than the cost of rejection, and if the cost of correct recognition is zero, then minimizing the expected cost *is equivalent to minimizing the error rate for a given rejection rate* [H-11], [V-1, pp. 46-52] and this type of processor is called a *maximum a posteriori probability computer* [V-1].

However, as Highlyman pointed out [H-11], $f_{M|S}(m|S_i)$ is usually unknown to the designer of the machine and therefore "categorizers based on the optimum decision function are not, in general, practically realizable." Highlyman also asserted that a key factor in realizing a pattern classifier is economic feasibility. A possible procedure is "to make no assumptions about . . . the particular distributions involved but rather make certain restrictions on the structure of the categorizer. Then search through all possible structures of this type to find the categorizer which is optimum with respect to a sampling of patterns from the real world." Furthermore, he emphasized, "if the designer can limit his search to those structures which are economically feasible, and if the optimum structure in this class works well enough for the given purpose, then a technically feasible solution has been found."

7.2.1 Linear Decision Functions

Because the decision criterion is non-random--that is, every point in the measurement space is, effectively, preassigned to a particular category or rejected--the decision function can be represented by the boundaries of the regions which comprise the measurement space. If the measurement space is considered to be a vector space of dimension N (N measurements per sample), then a *linear decision function* is simply a partitioning of this hyperspace by one or more hyperplanes, each of dimension N-1. Then, "the effectiveness of a linear decision function in identifying a given family of patterns is contingent upon the possibility of specifying an adequate linear decision function in terms of *an economically reasonable number of hyperplanes.*" ([G-12], Italics mine.) We emphasize that the repeated reference to economy of implementation is vital primarily because published accounts of applications involving large-scale computer simulation of decision functions frequently overlook this factor either as a direct cost, or, because of complexity-time factors, as a barrier to real time implementation of the recognition scheme.

Highlyman noted that the question of whether or not a linear decision function is useful is partially answered by the fact that "for any categorizer based upon minimizing a Euclidean distance to a set of reference points there exists a categorizer based upon a linear decision function which is at least as good. This includes categorizers which maximize a normalized cross-correlation function . . . "

Linear decision functions are discussed in detail in [A-4], [D-12], [F-19], [N-3], [P-7], [R-15], and [S-9]. Piecewise linear decision functions [D-12], and higher order surface decision functions (e.g., quadratic) are similarly described in [B-18], [B-19], [N-3], [S-9], and [S-19]. Methods of establishing the positions of hypersurfaces--training the machines--are also detailed. We shall limit our description of training methods to those algorithms associated with the speech recognition machines we have implemented. A useful comparison of various recognition algorithms is given by Nagy [N-1].

7.3 Perceptual Units in Automatic Speech Recognition

The problem of deciding upon a size of perceptual element to utilize in practical speech recognition investigations is quite important. It is often tempting to work with the simplest units of speech--the phonemes--initially and then attempt to extend any progress in recognition to more complex units. Although all acoustic information must be channelled through the same set of physiological transducers, the method of processing or attending to the neural signals probably varies with the difficulty and/or the circumstances of the recognition task involved.

Certainly we can recognize nonsense syllables under varying sets of conditions; but the mode of recognition has been shown to vary greatly--there is no continuum for acoustic recognition.

For example, in one experiment (see [F-8], p. 228) four groups of stimuli, varying in their similarity to speech, were presented in isolation to listeners who were to learn to identify the sounds in a certain manner. The tests showed that the greater the dimensionality of the stimulus (the dimensions being frequency, amplitude, and time) the more rapid the learning. However, actual speech sounds were learned most rapidly of all, with the *least* speech-like of the other tri-dimensional sounds being the next most "learnable" of the group. It was concluded that *the method of identification of sounds which are not speech is completely different from the method utilized on actual speech.*

Thus, unless a sound is speech it will not elicit response from the mechanism which identifies speech. The fact that a stimulus is "speech-like" (as some of the experimental stimuli were designed to be) apparently is not taken into account in the recognition process until we are sure that it is actually speech. Probably, then, the first step in *human* speech recognition is that of deciding that the stimulus is speech. Once this decision is made the recognition process can make use of the enhanced efficiency it demonstrates when dealing with actual speech sounds.

It has also been suggested that the mechanisms involved in the processing of isolated stimuli, even isolated speech sounds, may be considerably different than the "running speech" recognition mechanism. Flanagan has stated that "items such as syllables, words, phrases and sometimes even sentences may have a perceptual unity" and that "attempts to recognize speech in terms of brief acoustic units may be of little or no profit." [F-8, p. 238].

In the experiments described in the following sections it was necessary to restrict the size of the vocabulary. The spoken digits were chosen for the limited vocabulary for two reasons. First, they contain 18 of the 40 English phonemes and therefore represent a non-trivial set as far as complexity is concerned [S-14]. Second, this set has been chosen for many published experiments in automatic speech recognition because its elements represent a useful restricted vocabulary for verbal machine instruction. Thus, some comparison may be made (in a restricted sense because of differences in data rates and parameterization) to published results.

7.4 Experiment I: Motivation

Experiment I constituted an initial attempt at limited vocabulary speech recognition using zero crossing information. The motivation for this undertaking was the set of experiments described in sec. 6.7.1 associated with average rate of zero crossings. We wished to combine this measure of information with a simple, *adaptive* type recognition algorithm in an effort to test the possibility of recognition at very low data rates.

The basic limitation on the experimental procedure at this time was data gathering and handling. The only "automated" data gathering system was a combination digital voltmeter, (DVM) eight-hole paper-tape punch capable of punching 50 two decimal digit (4 bits per digit) numbers per second. The paper tapes could be transcribed via the Atlas Computer at the University of London and cards punched for use in the Imperial College IBM

7090 computer. The direct data facilities into the 7090 (later 7094 Mk II) used for the experiments of Chapter 9 were not available until a later period.

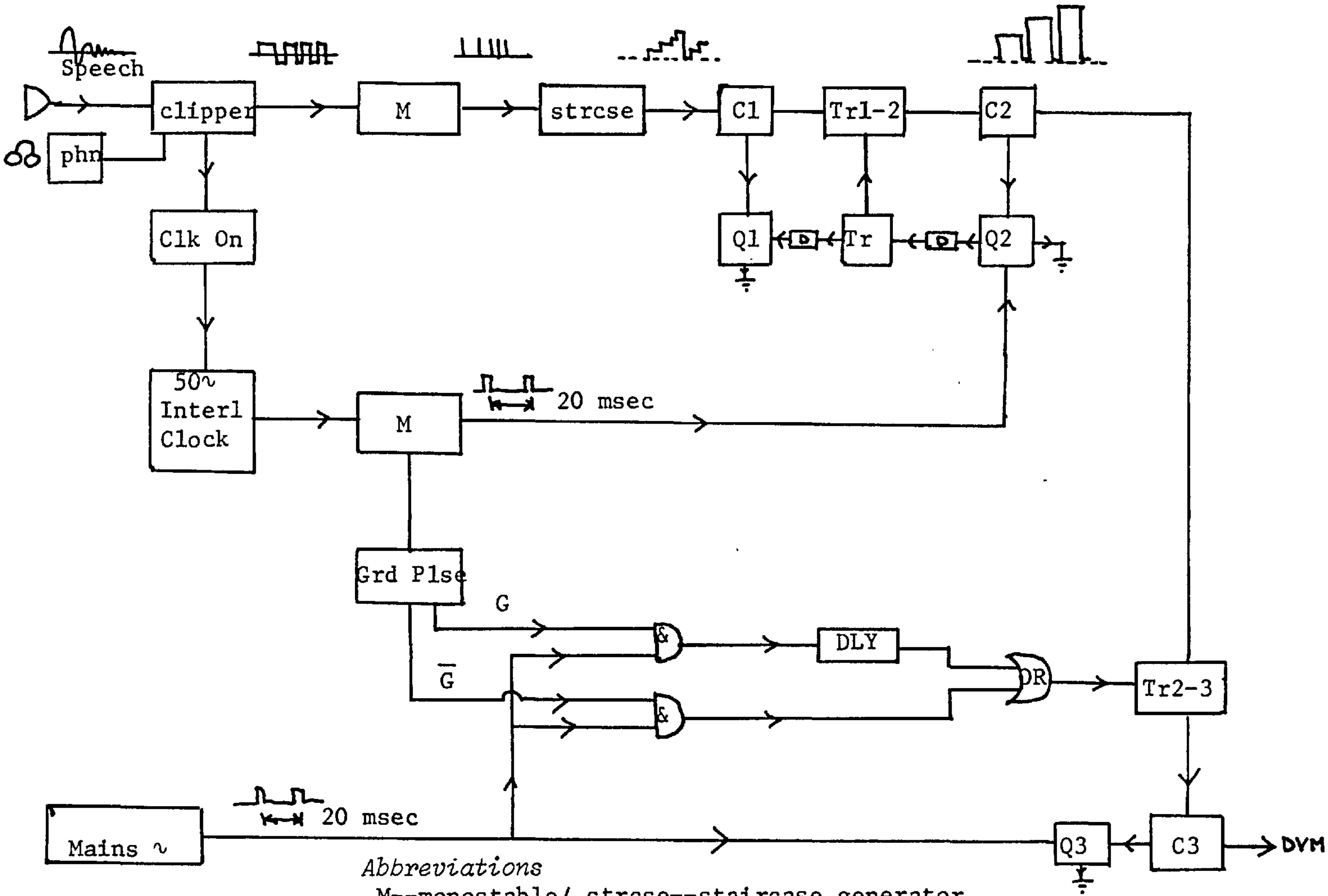
For these reasons, as a first attempt at *low data rate* (500 bits per second) adaptive speech recognition we chose the only zero crossing measurement compatible with the above limitations, a measure of average number of zero crossings per 20 msec. interval. The zero counting method chosen was a *staircase generator* incremented at each zero crossing and quenched to zero every 20 msec. This combined zero crossing counting with count-to-analog (voltage) conversion. A brief description of the data gathering assembly follows. A block diagram of the apparatus is shown in Fig. 7.1.

7.5 Experiment I: System Description

7.5.1 First Stage: Speech Clipper

The speech waveform is first "infinitely clipped." This action is accomplished by a modified Schmidt trigger which provides for adjustment of both the base level about which clipping occurs and the effective sensitivity. The base level is set so that clipping takes place about zero voltage; the effective gain adjustment is used to desensitize the device with respect to background noise. It is important that the position of the zero crossings be specified extremely accurately. However, due to the inevitable presence of noise it is obvious that "infinite" sensitivity of the zero crossing detector would entail noise induced clipping and hence erroneous zero crossing indication.

Previous investigators have tried to solve this problem



Abbreviations
M--monostable/ strcse--staircase generator
Tr--transfer circuit/ Q--quench/ D--delay/DVM--digital voltmeter
&--AND gate/ Clk--clock/ Grd Plse--guard pulse

Fig. 7.1 Zero crossing sampler of experiment I: block diagram.

in a number of ways. An ultra-sonic bias of amplitude "just greater" than that of the noise present in the system will ensure that noise does not actuate the clipper (sec. 5.1.1). However this preventative measure results in "distortion" and errors on low amplitude signals [F-13].

For this reason, noise interference was avoided by adjusting the level at which clipping occurred to be the minimum which would prevent the clipper from operating on noise, and by feeding in a speech signal which was of sufficient magnitude to ensure that clipping was effected very near, or at, the actual position of axis crossing. In the equipment used, the clipper responded only to voltages greater than 5 millivolts (peak). With a signal voltage of 5 volts (peak), a clipping ratio of 60 db is obtained. A certain amount of hysteresis with respect to actual zero crossing location is inevitable with this system. However, since the present experiment involves counting the number of zero crossings in intervals much greater than the period of the lowest frequency present, the errors due to hysteresis will be negligible and, more important, non-cumulative.

7.5.2 Second Stage: Zero Crossing Counting

The square wave output of the Schmidt trigger is fed to a gate which produces positive pulses of fixed duration and amplitude at each zero crossing. The duration of these pulses is constant and of length shorter than half the perod of the highest frequency speech component to be encountered. In the present apparatus the pulses are of magnitude 10 v. and 40 usec duration.

The positive pulses are fed into a linear staircase generator, each "step" of which is 0.05 volts. The staircase output is returned to zero (quenched) every 20 ms. For a sine wave frequency of 5 KHz the output is 10 volts, the maximum voltage desirable if 100 steps of 0.1 volts are to be "resolved" on the available digital voltmeter. The sample period of 20 msec (50 samples a second) was also chosen because of inherent limitations and characteristics of the digital voltmeter/readout combination.

7.5.3 Synchronization

In order to maintain maximum accuracy it is desirable that the first sample period should always terminate 20 milliseconds following the onset of each spoken digit.

7.5.4 Readout

The problems of readout into the digital voltmeter/punch device are two-fold.

First, the voltmeter requires that the voltage to be read is present for approximately 5 msec. Since the desired voltage--the peak (or final) voltage of the staircase generator--is present for a minimum time of approximately 40 usec, the shortest "possible" step, it is necessary to store this peak voltage for a delayed read/printout.

In this device, the staircase voltage is sampled just prior to quenching and stored for the next 20 msec in a capacitor designated capacitor 2 (C2).

Hence the procedure is:

(1) Quench capacitor 2.
(2) Transfer the voltage on the staircase store (cap. 1) to cap. 2.

(3) Quench cap. 1.

This sequence should take place as rapidly as possible so that cap. 1--the staircase store--is ready to receive the first zero crossing pulse of the new sample period immediately after it is quenched to zero. A chain of monostable delay elements provides the necessary sequencing.

Unfortunately, this storage facility is inadequate in that it is *synchronized with the voice input* whereas the digital voltmeter/punch is *synchronized with the mains* and may only sample at a specific point with respect to the 50 Hz mains waveform. Therefore it is probable that the digital voltmeter will often attempt to sample the voltage on cap. 2 when cap. 2 is being quenched, thus causing an erroneous "zero" reading. A second store was added to remedy this situation; synchronized with the mains, this store (capacitor 3 or C3) receives the reading from cap. 2 fifty times a second and is quenched at a time when the digital voltmeter is recycling for a new reading. This results in a store which always contains a reading at a time convenient fo the digital voltmeter. If the transfer [2-3] circuitry tries to operate when capacitor 2 is quenched, a "guard" pulse delays the transfer until cap. 2 contains a new reading.

7.5.5 Overall Operation

(i) Capacitor 1 is incremented by 0.05 volt at each zero crossing, and is quenched to zero 20 msec. after the first zero crossing of a speech sample and every 20 msec. thereafter for the duration of the spoken digit.

(ii) Capacitor 2 contains, for a period of 20 msec., a voltage equal to the peak voltage on capacitor 1 (i.e. the voltage present just before quenching) in the previous 20 msec. period.

(iii) Capacitor 3 contains, for a period of 20 msec. and in synchronism with the mains, a voltage equal to the voltage on capacitor 2 at the commencement of the 20 msec. mains period.

The overall "sine wave" transfer characteristic of the Zero Crossing Sampler (ZCS) is shown in the accompanying graph, Fig. 7.2. In practice, the upper limit to the output is determined by the ZCS voltage supply (10 v). Because of characteristics of the circuits used, a minimum input of 150 Hz (6 zero crossings per 20 msec. period) is necessary.

7.5.6 Speech Sample Recording Procedures

The subject (in the soundproof booth) records the desired speech sounds on the external tape recorder.[1] Following this, the data is played back via the line (600 ohm) output of the tape recorder into the Zero Crossing Sampler.

Subjects were instructed to speak the digits *zero to eleven* in sequence a number of times. The numbers zero and eleven were included to help eliminate the alterations in emphasis at the beginning and end of the "sentence." Only the numbers 1-10 were actually used. The subjects were asked to speak at a normal conversational level and to pause momentarily between digits. The microphone (AKG D19C) was positioned about 15-18" from the speaker's lips and a B & K voltmeter used to monitor speech recording level.

---

1 A more detailed description of the experimental recording apparatus will be found in Chapter 9.

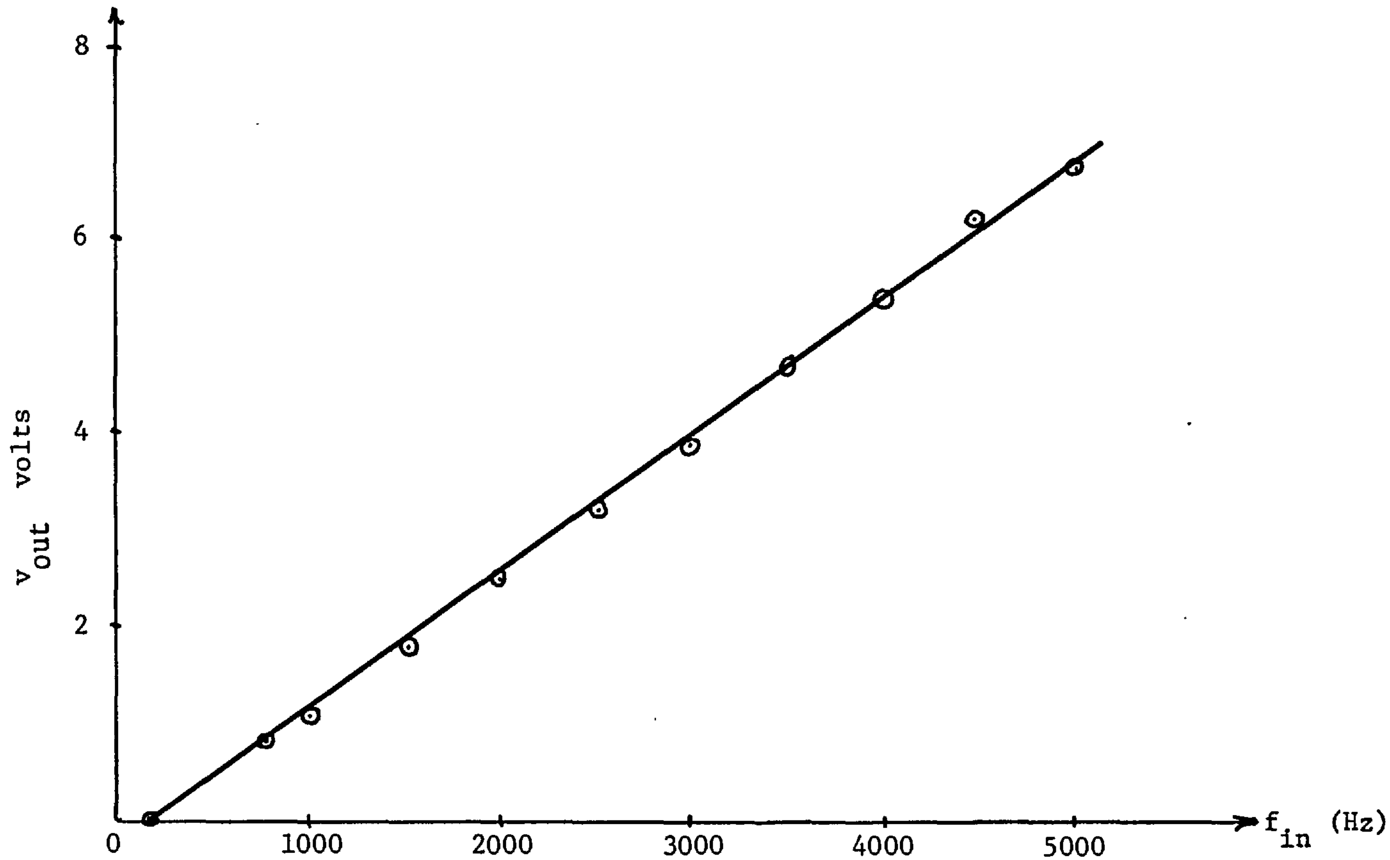


Fig. 7.2 Sine wave transfer characteristic of Zero Crossing Counter.

7.5.7 The Adaptive Recognition Algorithm

The algorithm used for adaptive recognition was that due to Braverman [B-17], [B-19, ch. 3], and is of the Linear, Non-Iterative type. The basis of this algorithm is the following hypothesis:

> Let the observed data be represented in terms of N binary (1 or 0) digits, where N is the number of binary digits necessary to represent each speech (or arbitrary species) sample. (In optical character recognition, the data might be represented by projecting the character on a matrix of N photocells each of which outputs a 1 if more than half of the cell is beneath the projected character and a 0 if not.) Then each set of 1's and 0's corresponding to a sample can be represented by a vector from the origin to a vertex of a hypercube in N-dimensional space. There will be 2^N vertices of this N dimensional unit hypercube.
>
> About *each* vertex belonging to a given category of optical character, or digit (e.g. the set of vertices belonging to the category 'x') we describe a unit *hypersphere*; we then term the vertex "internal" if all vertices lying on the surface of the hypersphere belong to the same category (as the vertex at the centre of the hypersphere.) Otherwise the vertex is termed "boundary".
>
> Then the set of vertices belonging to a given category is *compact* if the ratio of boundary vertices to the number of internal vertices is very small. *This algorithm is designed to operate upon compact sets.*

In the case where each dimension of a sample is an arbitrary number, between 1 and 100 in the present experiment, then the sets of samples belonging to different categories can be said to form "clouds" in hyperspace. A cloud can be said to be *compact* if the number of points lying near the edge of the cloud are much fewer than the number of points within the cloud.

The algorithm proceeds as follows:

(i) Training Phase

The first two known sample points are arbitrarily of different categories. The computer constructs a hyperplane perpendicular to the "line" joining the two points and midway between them. The coordinates of each point are substituted into the equation of the hyperplane. One point will produce a positive output and will be given a "1" output with respect to this plane (plane 1). The other point, being on the other side of the hyperplane, will produce a negative output and be assigned to "0" with respect to this plane.

As each new "known" point is read into the computer, its coordinates are substituted into the equation(s) for the existing hyperplane(s). If the output n-dimensional "binary" vector $\underline{\underline{x}}$ (where n is the number of hyperplanes existing, and x_1 is the output of the point with respect to hyperplane 1, x_2 the output with respect to hyperplane 2, etc.) is different from all previous output vectors *or* if the output vector is the same as that of a previous point belonging to the same category, nothing is done and a new point is read into the computer. If, however, the output vector is the *same* as that of a previous point belonging to a *different* category then a new hyperplane is constructed perpendicular to, and through the midpoint of, the "line" joining the two conflicting points. The output is calculated for all points with respect to this new hyperplane. It is clear that, since the two conflicting points are on opposite sides of the new hyperplane, the outputs of these points will be different with respect to this new hyperplane and hence the output binary vectors of the two points will no longer be the same.

Thus, after all training (known) points have been read into the machine, there will exist an n dimensional vector of 1's and 0's for each point, where n is the number of hyperplanes the machine has found necessary to construct in order to effectively partition the hyperspace into different regions, or "clouds", for the different categories. If the categories do indeed form compact sets, then the n dimensional vectors corresponding to members within a given category should be somewhat similar.

Following the hyperplane construction, the computer methodically tries to eliminate hyperplanes without allowing "conflicting points" to arise. This may be possible since the construction of any given hyperplane during the sequential read-in of points *might have made an earlier hyperplane redundant.*

It is interesting to note that if there are n different categories to which a point may belong, then the *maximum* number of hyperplanes necessary to separate the different categories *if* the categories form compact sets is n(n-1)/2. This is because each of the n categories must be separated from the other n-1 categories; however the hyperplane separating category i from category j can be the same as the hyperplane separating category j from i; hence the factor of 1/2. In practice, this number of hyperplanes is usually not required.

As each point finally produces an n dimensional vector of 1's and 0's, the number of possible vectors is 2^n-1 . Because this number is inevitably much greater than the number of categories, or the number of different *n* vectors which were produced by the training points, the machine must index un-named regions so that they will be identified with the category of an adjacent named region. *When this is done, any input sample will produce an n dimensional vector of 1's and 0's and be classified into some known category.* The accuracy of classification will depend upon the "closeness" of the "unknown" to a particular region.

(ii) Recognition Phase

"Unknown" points are entered into the algorithm by substituting their coordinates into the equations for the existing hyperplanes, as in the training phase. Due to the algorithm construction, the "hypervolume" into which this point falls *must* correspond to a known category.

7.6 Experimental Results

One hundred samples, ten of each of the digits (1-10), [S-14] were prepared on punch cards from the data secured from each of two speakers. Each sample consisted of a category identification number (1-10) and then the 47 samples (range 0 to 9.9 volts in steps of 0.1 volt) punched out by the Zero Crossing Sampler via the digital voltmeter/punch. If a spoken digit provided less than 47 samples (i.e. was less than 47/50 sec. long) the remaining sample positions were termed "0".

The algorithm was programmed in Fortran IV and executed on the IBM 7090 computer at Imperial College.

(i) Recognition of Subject One from Subject One

The machine was given five samples of each of ten digits spoken by subject one (LRM, Canadian) and, after the learning algorithm had been implemented, asked to recognize another 50 unknown digits (five of each).

Results: The machine correctly identified 31 out of 50, i.e., *62 percent*. It constructed 13 hyperplanes.

(ii) Recognition of Subject Two from Subject Two

Same conditions as 1. (Speaker RLW, British)

Results: Correct recognition of 36 out of 50---*72 per cent*. Constructed 9 hyperplanes.

(iii) Recognition of Subject Two from Subject One

The machine was given 100 samples (ten per digit) of digits spoken by subject one and asked, after the learning process, to identify 100 samples (ten per digit) spoken by subject two.

Results: Correct recognition of 51 out of 100---*51 per cent*. Constructed 15 hyperplanes.

(iv) Recognition of Subject One from Subject Two

Reverse of (iii).

Results: Correct recognition of 45 out of 100---*45 per cent*. Constructed 12 hyperplanes.

(v) Mixed Recognition

The machine was given both groups of 50 samples used for learning in (i) and (ii). The machine implemented the learning algorithm without knowing which samples were from which speaker.

The machine was then asked to identify 100 samples (five of each digit from each of the two speakers) without knowing which speaker had spoken the digit.

Results: Correct recognition of 65 out of 100-*65 per cent.* Constructed 16 hyperplanes.

Individual Results: Speaker 1: 29/50 = 58 per cent
Speaker 2: 36/50 = 72 per cent

7.6.1 Remarks and Analysis

(i) and (ii). The machine found less "variance" within categories of the spoken digits of speaker 2 than speaker 1 since it constructed fewer hyperplanes and recognized a larger percentage of unknown samples. If the Confusion Matrices are examined it will be noted that the percentage correct recognition was not evenly distributed over the field of digits. The machine was very accurate in recognizing the digits 1, 2, 6 and 8 for both speakers (and 10 for speaker 1), less accurate on 7 and 9, and inaccurate in recognizing 3, 4, and 5. Examination of the patterns for the digits 3, 4 and 5 shows very little basis for separation in any case. (See Fig. 7.3).

(iii) and (iv). In accepting 50 additional samples from speakers 1 and 2 the machine constructed 15% and 33% more hyperplanes, respectively. This indicates that the machine was adjusting its boundaries to the further refined positions dictated by the additional information. Although the percentage accurate recognition dropped to about 50, it is still high enough to state that the categorical distributions encountered when learning on the samples from one speaker were sufficiently invariant to recognize unknown samples from another speaker. In fact, the confusion matrix shows 80% accuracy in recognizing the digits 1, 6, 9, 10 as spoken by subject 2 after having heard only 10 samples of each digit as spoken by subject 1. It should be noted that the

ACTUAL DIGIT

DIGIT CHOSEN BY MACHINE	1	2	3	4	5	6	7	8	9	10
1	3									
2		5								
3	2									1
4			1	2			1			
5					1					
6			1			5	1			
7				1	2		3			
8			1					4	1	
9					2			1	4	
10			2	2						4

Part i

Speaker:

Training digits: LRM (50)
Unknown digits: LRM (50)

% Correct: 62

ACTUAL DIGIT

DIGIT CHOSEN BY MACHINE	1	2	3	4	5	6	7	8	9	10
1	5									
2		4	1	1						1
3			3	2					3	2
4		1		2				1		
5					3					
6						5				
7							5			
8								4		
9			1		2				2	
10										2

Part ii

Speaker:

Training digits: RLW (50)
Unknown digits: RLW (50)

% Correct: 72

Fig. 7.3 Confusion matrices for Experiment I

	ACTUAL DIGIT									
	1	2	3	4	5	6	7	8	9	10
1	8		1						2	
2		6		2						
3			4	2			1	3		
4	1	4	1	1			3	4		
5				1		1				
6						7	6	2		
7						2				
8								1		
9	1		1	1	10				8	
10			3	3						10

Part iii

Speaker:

Training digits: LRM (100)
Unknown digits: RLW (100)

% Correct: 51

	ACTUAL DIGIT									
	1	2	3	4	5	6	7	8	9	10
1	10									
2		5		1						
3		2	4	1						8
4		3	2	5	3			1	5	1
5				1				2	3	
6						7				
7			2	2	2		10			
8			1		3	3		3		
9			1		1			4	2	1
10					1					

Part iv

Speaker:

Training digits:RLW (100)
Unknown digits: LRM (100)

% Correct: 45

Fig. 7.3 Confusion matrices for Experiment I

ACTUAL DIGIT

DIGIT CHOSEN BY MACHINE	1	2	3	4	5	6	7	8	9	10
1	10								2	
2		10		1						
3			4	1	1					
4			1	6	1		1	2	2	
5			2		2	2		2	1	
6						8	2			
7				2	1		6			
8					5		1	5		
9			3					1	5	1
10										9

Part v

Speaker:

Training digits: } LRM (50), RLW (50)
Unknown digits:

% Correct: 65

Figure 7.3 Confusion matrices for Experiment I machine mistook all of speaker 2's *fives* for *nines*. Speaker 2's *fives* do look like Speaker 1's *nines* when the source patterns are examined, due to a suspected fault in the recording/punchout wherein the initial fricative was lost. Also, it is interesting that "hearing" speaker 1 saying *ten* enabled the machine to accurately identify all of speaker 2's *tens* but the reverse was not the case. This might be expected since, if the region (or volume in hyperspace) containing speaker 2's *tens* is within a larger region containing speaker 1's *tens*, then being trained on speaker 2 will not allow recognition of speaker 1 even though the reverse will be true.

(v). When the machine was trained with 50 samples from *each* of the two speakers, it recognized about the same number of unknown samples from each speaker as it did when trained by the 50 samples only from one speaker as in experiments (i) and (ii). It did not, and this is most important, achieve this proficiency by constructing twice as many hyperplanes as it had required, on the average, for each of the speakers individually.

In fact, the machine operated as follows:

In constructing the hyperplanes for the 50 samples from speaker 1 only, (part i) the machine erected 16 hyperplanes and later eliminated 3 as being redundant. Since the same 50 samples (of part i) from speaker 1 were "learned" first in part v, *initially* the same 16 hyperplanes were constructed. After the next 50 points (from speaker 2) had been examined, 10 more hyperplanes were found necessary. However, the machine later eliminated 11 of the 27 total hyperplanes to leave 16, only 3 more than were needed for speaker 1 alone. Thus we may conclude that, although the machine was roughly as efficient in identifying the unknown samples of both speakers, the memory required was only slightly larger than that needed for identifying one speaker only.

7.6.2 Conclusions

The correct recognition rate for parts i, ii, and v of experiment I (62%, 72% and 65%, respectively) exceeded the chance rate (10%) by at least a factor of 5. Nonetheless, because of the limited amount of experimentation done, no statistical significance can really be attached to the results. The remarks in sec. 7.4.9 concerning the significance of variations in the number of hyperplanes constructed for different teaching sets are basically an interpretation of the algorithm behaviour. The drop in correct recognition rate when the machine was trained using the samples of one speaker and asked to identify those of another speaker is similar to that observed in other speech recognition experiments (see chapter 4).

It was decided that, despite the existence of the data gathering limitations, an improvement should result if zero crossing interval *lengths* could be "sampled" and encoded within the basic punch machine structure. The scheme described in the next section successfully accomplished this goal.

7.7 Experiment II: Motivation

The technique used in experiment I preserved information concerning only number of zero crossings per 20 msec. time interval. In sec. 6.6 we discussed the use of the intervalgram, or histogram, of zero crossing interval length distributions for automatic speech recognition. Figures 6.9 and 6.10 illustrate the fact that displays somewhat analogous to the time-frequency intensity display of a short-term speech spectrogram can be derived from zero crossing interval lengths by linear (or exponential) ramp generators. This was first shown by Sakai and Inoue [S-1]. Figure 6.11 shows the "peaked" structure of the first-order density distribution of zero crossing interval lengths.

That these results obtain for English vowels were confirmed by Bezdel and Chandler ([B-6], sec. 6.7.2), who showed experimentally that such information is sufficient for a high degree of success in sustained vowel classification. Our own experiments (Figs. 6.9 and 6.10, and Figs. 7.3, 7.4, and 7.5 below) further demonstrated that zero crossing interval histograms are highly structured.

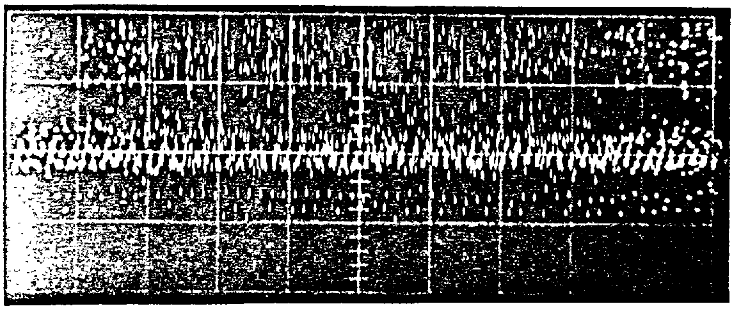


Fig. 7.3 Zero crossing intervalgram, /ɔ/.
Sweep = 50 msec/cm.

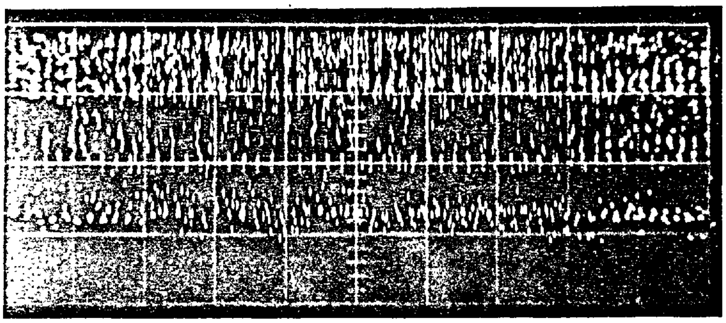


Fig. 7.4 Zero crossing intervalgram, /e/.
Sweep = 50 msec/cm.

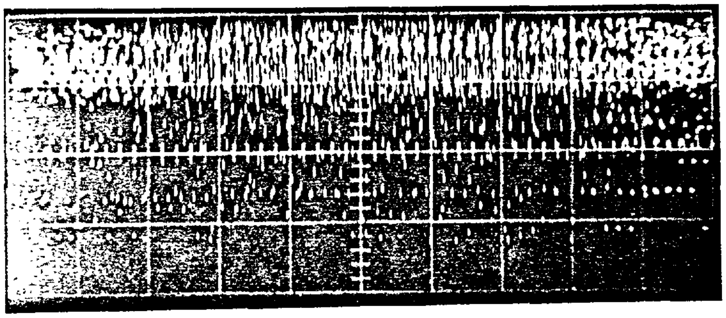


Fig. 7.5 Zero crossing intervalgram,/i/.
Sweep = 50 msec/cm.

The aim of the system described in the next section--based upon zero crossing interval histograms--was twofold:

First, the peaked structure of the histograms suggested that amplitude quantization could be employed to reduce the bit rate required to describe them. An analogous technique had been successfully employed by King and Tunis [K-6] in respect to classification of short-term speech spectrograms.

Secondly, it was decided to make use of the *total sequence* of "short-term" zero crossing histograms which constitutes a spoken digit. The *order* of the sequence members as well as the *constitution* of each member was to be taken into account in the training and recognition process.

In the next section the equipment constructed to produce quantized zero crossing histogram sequences in the form of paper tape output will be described. Then, in sec. 7.5.3, we will briefly outline the algorithm used; this algorithm incorporates the sequential aspects noted as being desirable.

7.8 Experiment II: System Description

The basic limitation on the rate of data flow was still the paper tape punch output of 8 binary digits per 1/100 second; the voltmeter reduced this rate by 50%. Thus it was decided to bypass the voltmeter and output 32 bits of information every 40 msec. or 1/25 second. We recall that the 33.3 msec. averaging time used for the lowpass filter in Peterson's work [P-9] was based on the desirability of averaging over a time interval less than the phonemic rate--10 per second-- and greater than the pitch period--1/100 second.

A block diagram of the system is shown in Fig. 7.5. The system is composed of three sections:

7.8.1 Pulse Production and Gating

Spoken digits were recorded using the setup described in Chapter 9 (Soundproof booth, AKG dynamic microphone and Tandberg 62 tape recorder at 7½ ips). The speech was bandlimited to 4600 Hz by a Mullard switched filter (60 db per octave attenuation out of passband) and then clipped by a cascade of three balanced (long-tail pair) limiting amplifiers. The output of the final amplifier is transmitted by a balanced gate through to a Schmidt trigger which, in turn, drives a monostable multivibrator which thus produces short pulses at each zero crossing of the band-limited signal.

The gate is controlled by an envelope detector which consists of an a.c. signal amplifier, a diode detector, a d.c. amplifier and a three-stage RC lowpass filter.[2] The gate serves two purposes: First, pulses due to clipped system noise are completely eliminated. Second, an internal clock which controls the operation of the following stages is turned on at the start of each spoken digit and remains on for a set period of time after speech ceases to be detected by the envelope detector. This "turn-off delay" is necessary to inhibit system turn-off during intra-word energy gaps. We recall (sec. 3.4.6) that stop consonants, for example, are often preceded by periods of near

---

2 The detector-gate was developed by R. L. Wiley of Imperial College for speech-noise switching purposes. It was capable of distinguishing between speech (even weak, unvoiced fricatives) and system noise with minimal onset delay.

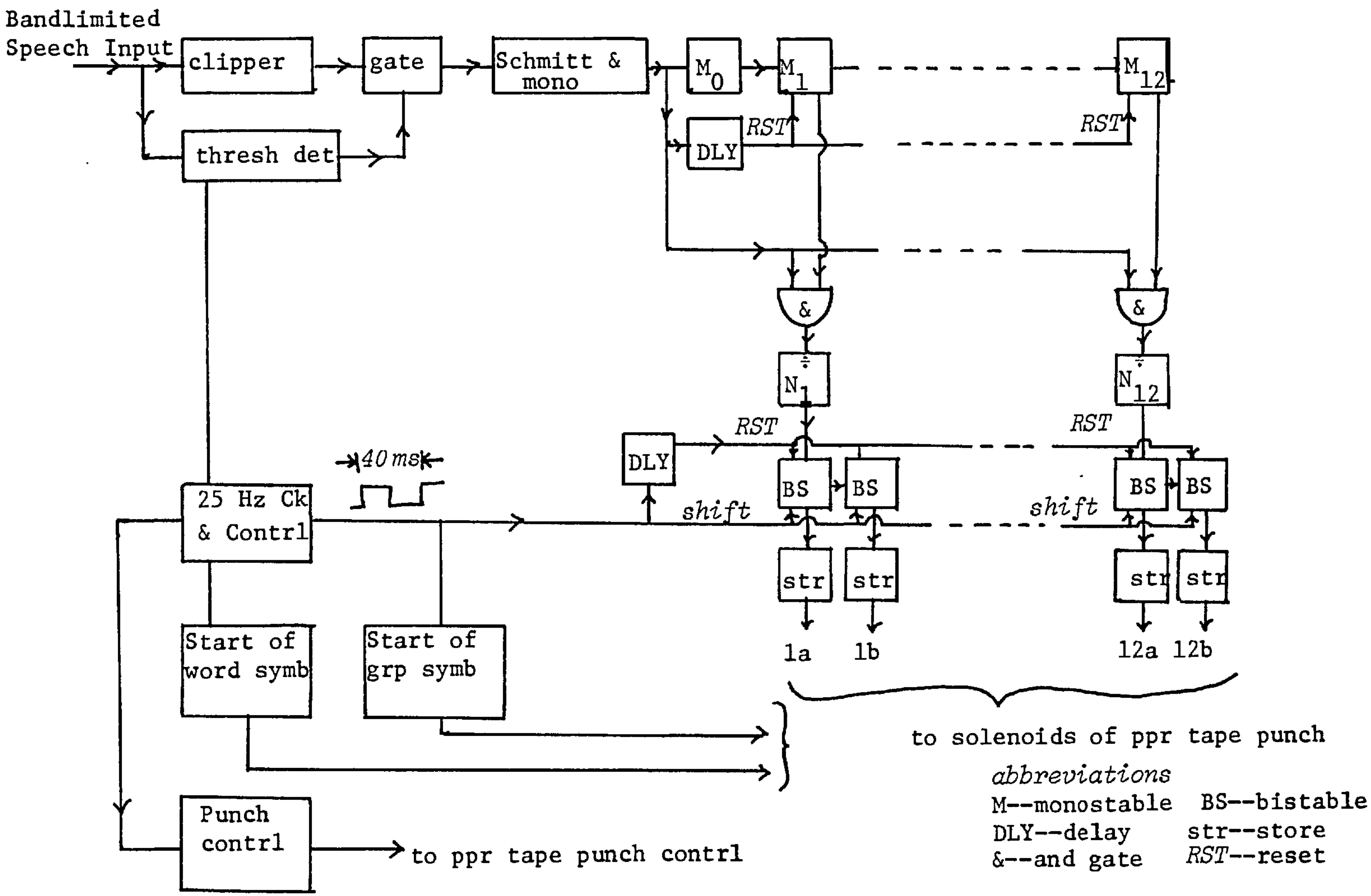


Fig. 7.5 Simplified block diagram of data gathering system of experiment II.

silence. During intra-word energy gaps, the system noise zero crossings are inhibited but the internal clock remains running since the "silent" interval is necessary to the word structure. The maximum delay time needed to allow for intra-word energy gaps was experimentally determined to be about 1/25 second.

7.8.2 Zero Crossing Interval Sorting

The zero crossing pulses enter the first of a chain of 13 monostable multivibrators, M_0 to M_{12}, having "on times" T_0 to T_{13}. Each monostable in the chain is triggered by the "off" edge of the preceding monostable. Thus monostable M_i turns on Δt_i milliseconds after the first monostable is triggered, where

$$\Delta t_i = \sum_{p=0}^{i} T_p \ . \tag{7-3}$$

T_0 is 0.1 msec. and Δt_{13} is 3.33 msec. (1/300 sec.) If a zero crossing occurs at a time t,

$$3.33 \text{ msec.} < t < 0.1 \text{ msec,} \tag{7-4}$$

after a previous zero crossing, *one* of the monstables M_1-M_{12} will be on. The output of this monostable is ANDed to the input of one of 12 divide-by-N_i digital circuits. All of M_1-M_{12} are then rapidly set to "off." Monostable M_0, 0.1 msec. after the zero crossing which initiated the ANDing operation, initiates the start of a new pulse chain down M_1-M_{12}. Thus each zero crossing interval is classified into one of 12 channels, according to the interval length.

Each divide-by-N_i circuit emits an output pulse after every N_i^{th} input pulse. These pulses, in turn, enter a set of two binary counters (12 sets, one per channel). These counters are inhibited from returning to the (0,0) state after the (1,1) state has been reached. Every 40 msec., all 12 sets of counters are parallel shifted to 12 corresponding sets of storage bistables and then returned to the (0,0) state.

Hence, during a given 40 msec. period, the two storage counters for channel p, p=1,..12, will contain information as to whether there have been less than 1 (0,0), between 1 and 2 (0,1), between 2 and 3 (1,0) or more than 3 (1,1) *groups of* N_i *zero crossing interval lengths* between

$$\sum_{i=0}^{p-1} T_i < \Delta\tau_p < \sum_{i=0}^{p} T_i \qquad (7\text{-}5)$$

milliseconds. This yields a weighted and quantized (by the N_i counters), twelve channel zero crossing interval histogram consisting of 24 bits every 40 milliseconds. The "histogram" is punched out onto 4 rows of eight-hole paper tape. Actually, only 7 holes of each of the 2^{nd}, 3^{rd}, and 4^{th} rows may be used for the histogram. The first 4 holes of row 1 are used to indicate start of word and/or start of 4 row sequence. The first hole of rows 2, 3, and 4 is always "blank."

The divide-by-N_i circuits in each channel consist of binary counters which may be adjusted to zero after any count up to 64. We recall that, from sec. 6.6.1, $f_t(\tau_{mi})$, the first order density distribution associated with measurement of zero crossing interval distributions, is actually a weighted histogram. The divide-by-N_i counters are adjusted to approximate this function. The channel boundaries themselves can be adjusted to simulate the various ramp functions as used by Chang *et al.*

(sec. 6.6.1).

7.8.3 The Adaptive Recognition Algorithm

The adaptive recognition algorithm used was actually chosen in conjunction with the design of the data collection system. The paramount requirements for the recognition algorithm were that

(i) the algorithm should cater to data in *binary form*

(ii) the *sequential* aspect of the short-term speech histograms be taken into account in the training and recognition phases.

In Braverman's algorithm (experiment I), each 20 msec. estimate of the zero crossing count was assigned to one dimension of a multidimensional space. It can be shown (see [H-11], for example) that the performance of a linear decision function is unaffected by a non-singular linear transformation, followed by a translation. Therefore, the sequential aspects of the patterns are not really utilized in this class of algorithm.

The algorithm chosen was devised by R. E. Bonner [B-14]. Besides satisfying conditions (i) and (ii), Bonner's algorithm possesses the following desirable attributes:

(i) If a new category or class is added after initial training occurs, excessive revision of the original structure is not required. This was not the case in Braverman's algorithm.

(ii) The algorithm is capable, during recognition phases, of *prediction*. That is, at a certain point during the read-in of the sequence of sub-patterns constituting the spoken word, the machine should be capable of predicting the rest of the sub-patterns which will follow.

(iii) The algorithm provides for the existence of "local stability" in the input sequence of binary sub-patterns. This means that the closer the sub-patterns occur in time, the more correlation there is apt to be between them.

Bonner emphasized that his implied allusion to human performance characteristics (i.e., prediction, correlation of spoken sub-patterns) "has been used only as a source of requirements in an interesting problem situation; there is absolutely no reason to believe that the scheme to be described in any way explains actual human functioning."

The algorithm is described below, as in [B-14], with reference to Fig. 7.6.

At the left is a shift register consisting of M connected segments, each n bits long; these are labelled as "present," "past I", etc. At the start of the test-forming procedure, the first sub-pattern (n bits) of the sequential pattern is introduced into the "present" portion of the register. The ORing procedure is then followed. Here, when bit i in the "present" register is one, the test T_i for output bit i is updated. When updating is necessary, the contents of the *entire* shift register are used to OR to a test. This means that nxM bits of storage are required per test.

After updating the tests, the first sub-pattern is shifted to "past I" and the next subpattern is entered into the "present" segment. Updating again takes place, as before. The process of shifting and updating continues until all the sub-patterns in the sequential pattern have been exhausted. Sub-patterns shifted past "past M-1" are lost. The shift register is then cleared and the procedure repeated for the next sub-pattern. An example of this process is given in Fig. 7.6a.

At completion of formation, test i contains the information on which bit positions in all sequences of sub-patterns of length M were ever one when bit i of the "present segment" was one. The test is therefore designed to reproduce at the output the sub-pattern contained in the "present" portion of the shift register. The tests following training, for this example, are those in Fig. 7.6b, and consist of an $(n.M)$x(n) matrix of binary digits. To use the tests for recognition, the input is introduced into the shift register one sub-pattern at a time, exactly as during test formation.

Sequential pattern used for test formation in Figure 7.6 ;

# 1 Subpattern	0	0	1	1	0	↓ Time
# 2 Subpattern	0	1	0	1	0	
# 3 Subpattern	1	0	1	0	1	

EXAMPLE TRACING TEST FORMATION FOR A SEQUENTIAL PATTERN ($M=2$, $n=5$)

Subpattern number in present segment of register	Shift register condition		ORs representing tests: Test No. 1	2	3	4	5
1	Present	0	0	0	0	0	0
		0	0	0	0	0	0
		1	0	0	1	1	0
		1	0	0	1	1	0
		0	0	0	0	0	0
	Past I	0	0	0	0	0	0
		0	0	0	0	0	0
		0	0	0	0	0	0
		0	0	0	0	0	0
		0	0	0	0	0	0
2	Present	0	0	0	0	0	0
		1	0	1	0	1	0
		0	0	0	1	1	0
		1	0	1	1	1	0
		0	0	0	0	0	0
	Past I	0	0	0	0	0	0
		0	0	0	0	0	0
		1	0	1	0	1	0
		1	0	1	0	1	0
		0	0	0	0	0	0
3	Present	1	1	0	1	0	1
		0	0	1	0	1	0
		1	1	0	1	1	1
		0	0	1	1	1	0
		1	1	0	1	0	1
	Past I	0	0	0	0	0	0
		1	1	0	1	0	1
		0	0	1	0	1	0
		1	1	1	1	1	1
		0	0	0	0	0	0

a) Training Procedure

		ORs representing tests: Test No. 1	2	3	4	5
Present	1	1	0	1	0	1
	0	0	1	0	1	0
	1	1	0	1	1	1
	0	0	1	1	1	0
	1	1	0	1	0	1
Past I	0	0	0	0	0	0
	1	1	0	1	0	1
	0	0	1	0	1	0
	1	1	1	1	1	1
	0	0	0	0	0	0

b) Tests following training

Fig. 7.6 Illustration of Bonner's algorithm. (From [B-14].)

The sequential pattern used to demonstrate the test or recognition procedure is

Subpattern #1	1	1	0	0	1	↓ Time.
#2	0	0	1	0	1	
#3	1	1	1	0	0	

To commence the test, subpattern #1 is introduced into shift register 1, shift register 2 being empty (see Fig. 7.7). The M shift registers--M = 2 here-- are ANDed to Test No. 1 of Fig. 7.6b. The number of one's in the result, divided by the number of one's in the M shift registers gives the "match number" for the test. n match numbers are calculated and an n bit output results with one's wherever the match number exceeds some threshold (0.6 in the example) and zero's otherwise. This n bit number is then ANDed to the "present" segment of the input register and gives the output in column 5 of Fig. 7.7.

Fig. 7.7
Bonner's algorithm, recognition procedure. (From [B-14].)

Subpattern number in present segment of register	Shift register condition	Test No.	Match number for test	Output formed by using threshold = .6 and then ANDing to "present" segment of input reg.	Match number for subpattern
1	1 1 0 0 1 / 0 0 0 0 0	1	2/3	1	
		2	1/3	0	
		3	2/3	0	0.667
		4	1/3	0	
		5	2/3	1	
2	0 0 1 0 1 / 1 1 0 0 1	1	3/5	0	
		2	0/5	0	
		3	3/5	1	1.000
		4	1/5	0	
		5	3/5	1	
3	1 1 1 0 0 / 0 0 1 0 1	1	2/5	0	
		2	2/5	0	
		3	2/5	0	0.000
		4	3/5	0	
		5	2/5	0	
					Average = 0.556

Finally, the number of one's in this output divided by the number of one's in the "present" segment of the input register gives the "match number" for the *subpattern*. This procedure is repeated as each subpattern of the word enters the "present" register forcing previous subpatterns into past I, past II, ... past M-1. The average "match number" for the sequence of subpatterns constituting the "word" gives an indication of the overall match of the unknown word to the $(n.M)\text{x}(n)$ matrix of one's and zero's which is the result of "learning" a particular category.

The key to the *practical* implementation of this algorithm turned out to be the use of the machine dependent (i.e., non-Fortran) AND and OR operations on the 7094 computer. If each spoken digit constitutes p subpatterns of n binary digits each ($n \leqslant$ word length of machine, 32 in this case), then only $(M.n)$ words per storage table (one storage table per category) are needed. In our case, M was varied from 3 to 9, n = 24, and the number of categories was 5. The algorithm description has been necessarily brief and more of the philosophy behind the algorithm development is described in [B-14].

7.8.4 Experimental Procedure

In order to conserve computer time--both in the paper-tape to punched card transcription phase and in the learning-recognition phase--the vocabulary in the tests reported was limited to the spoken digits one, two, three, four, five. Three hundred and five samples (61 of each digit, the author speaking) were recorded using the same equipment setup as in Experiment I. Following tape editing to eliminate inter-word "extraneous noise", the quantized histograms were punched out using the apparatus described earlier. The time and difficulty in tape editing and paper-tape to card transcription (carried out by the University of London Atlas computer) proved to be one of the factors which caused the project to be abandoned when direct input to the 7094 became feasible.

7.9 Experimental Results

The boundaries for the twelve channels in these experiments corresponded to sine wave frequencies of 150, 295, 400, 540, 630, 770, 920, 1130, 1450, 1700, 2380, and 3400 Hz. As mentioned earlier, the shortest zero crossing interval length which could be counted corresponded to a sine wave frequency of 5000 Hz. The divide-by-N_i counters were set so as to produce an approximation to $f_t(\tau_{mi})$. The variable in the learning-recognition phases was the memory length, M.

The results of the limited tests carried out are shown in the confusion matrices of Fig. 7.8. The percentage correct recognition varied from 77%, for $M = 3$, to 88% for $M = 7$ or 9. The recognition reached maximum at this point for the noted conditions. For $M = 9$, the recognition of digits 1, 2, 4, 5, reached over 95%. The digit 3 was mistaken for 2 nearly 35% of the time.

7.9.1 Conclusions

At the conclusion of these initial tests we were faced with a difficult decision. The results were very promising (very comparable to those reported in the literature for other preprocessing methods but with similar or higher bit rate) and other variables which could possibly increase the accuracy were still available for manipulation. Histogram weighting (divide-by-N_i

7.8 *ACTUAL DIGIT*

	1	2	3	4	5
1	24		2	1	7
2		27	7		
3		2	19		1
4	5		1	26	5
5				2	16

Training digits: 160
Unknown digits: 145
% Correct: 77 *M = 3*

7.9 *ACTUAL DIGIT*

	1	2	3	4	5
1	27		1	1	3
2		27	10		
3		2	18		1
4	2			26	6
5				2	19

Training digits: 160
Unknown digits: 145
% Correct: 81 *M = 5*

7.10 *ACTUAL DIGIT*

	1	2	3	4	5
1	29	1			
2		27	9		
3		1	20		1
4				28	1
5				1	24

Training digits: 160
Unknown digits: 145
% Correct: 88 *M = 7*

7.11 *ACTUAL DIGIT*

	1	2	3	4	5
1	28				2
2		28	10		
3		1	18		1
4	1		1	28	1
5				1	25

Training digits: 160
Unknown digits: 145
% Correct: 88 *M = 9*

Figs. 7.9–7.11 Confusion matrices for Experiment II. Speaker: LRM.

circuits) could be varied and signal preprocessing (particularly differentiation) was intended to be applied.

However, three factors suggested that this course of action might not be the most fruitful. First, the paper-tape to card transcription was proving difficult because of erratic Atlas computer service. The paper tape input promised for the 7094 did not become available. Second, an FM tape recorder was acquired so that the Direct Data Channel, with its limited sampling rate, might be employed for *effective* high speed speech input and magnetic tape input. Finally, it was suggested--both by the detailed review of the literature which now constitutes Chapters 5 and 6, and by conversations with Professor H. B. Voelcker, visiting Imperial College at the time--that the role and significance of zero crossings in automatic speech recognition and in clipped speech perception was not clearly understood. Although Voelcker's papers on zeros as informational attributes of signals had contributed significantly to the understanding of links among various modulation schemes, the zero-based signal theory developed therein had not been extended and applied to what were obviously zero-related speech phenomena--*clipping* and *objective estimates of speech spectral parameters using zero crossings*.

We therefore decided that the short-term goal of realizing a reliable, limited vocabulary speech recognition machine--partially accomplished--should be discarded in favour of the theoretical and experimental studies which constitute the latter and more significant portion of this thesis. These studies are intended to clarify the significance and role of zero crossings as parameters for use in automatic speech recognition machines and to provide links between zero crossings and more conceptually meaningful attributes of speech signals.

8 ZERO-BASED SIGNAL MODELS

In chapter 1 we stated that this thesis is concerned with the interpretation of two intimately related themes: that clipped speech is intelligible, and that the same zero crossing interval sequence which defines the clipped speech waveform may be used to obtain objective estimates of certain speech spectral features. In preceding chapters we have described in detail some phenomena associated with clipped speech audition together with some theories proferred as explanations for them. We have also reviewed the use of zero crossing interval sequences in objective speech spectral feature estimation, and as patterns representative of the original speech signal.

Yet, profound doubts linger concerning the conventional methods of dealing with these phenomena. Furthermore, there still exist many unexplained observations associated with clipped speech audition. For example, the power spectrum of Gaussian signals is, in some cases, *roughly* preserved after clipping. Of what relevance is this to voiced speech sounds--specifically vowels-- whose waveforms are not random but quasi-periodic? *Experimentally,* the power spectrum of vowels *is* often preserved after clipping insofar as the formant structure may still be observed in the post-clipping speech spectrograms. But not *all* speech sounds are *equally* intelligible after clipping. Does this imply that the degree of post-clipping power spectrum preservation is somehow related to the time-frequency characteristics of the original signal?

Pre-clipping highpass filtering and/or differentiation enhances post-clipping intelligibility. Why? Is there a process which will ensure almost *complete* intelligibility of the clipped signal? For example, the single sideband clipped speech signal-- cos $\phi(t)$-- is perceptually the same as the original speech waveform, $s(t)=|m(t)|\cos\phi(t)$. What is the relationship between clipped speech--sgn[s(t)]-- and SSB clipped speech--cos $\phi(t)$?

Zero crossing interval sequences can be processed so as to yield estimates of speech spectral features. Can these estimates be made exact? In other words, can the original signal spectrum be recreated *exactly* using only zero crossing information? If so, how? If not, then exactly what measure of information concerning the original signal do zero crossings constitute?

Thus three basic questions remain:

1: Why does clipping-- a process which ostensibly destroys all signal amplitude information except for polarity-- apparently preserve power spectrum features in some speech signals?

2: What measure of information--concerning the original signal-- do zero crossings constitute?

3: Are there signal transformations which will ensure that almost all information contained in the signal is available in its zero crossings *and* -- if the signal is speech -- will its intelligibility be effectively undiminished by clipping?

To answer these questions we must adopt a method of signal analysis which treats zero crossings as *informational attributes*. Such a technique was formalized by H.B. Voelcker in 1966 [V-6]. He applied these ideas to achieve a unified description of modulation processes.

In this chapter we *review* and *expand upon* this technique and demonstrate that it has important applications in speech signal analysis. In section 8.1 we will outline the transition from *Fourier series sum* to *zero-based product* representation of periodic signals and discuss basic relationships between the spectra of the two, fundamental zero-based signal components. This section is based primarily on Voelcker's published [V-6] and unpublished [V-7, 9, 10] work. We then show (sec. 8.2, sec. 8.3) that extension of these models to analytic signals provides a link between some concepts of zero-based and conventional signal theory and--in certain cases--permits conclusions to be made about the zero crossings of the original signal. This work was accomplished primarily by S. Haavik [H-1].

We then apply these concepts to propose a zero-based interpretation of the unexplained observations in the psychoacoustic experiments reviewed in chapter 5. We argue that, *apparently*, the zero crossings of a speech signal generally constitute only a *partial description* of that signal. At the same time, the zero crossings *completely* specify one of the components of the zero-based model. We show that the spectrum of this signal can be explicitly (though not simply) expressed in terms of the zero crossings and review the method devised by Voelcker for generating this signal.

The latter part of the chapter is devoted to a theoretical discussion and experimental demonstration of the significance of zero-based signal theory and complex time domain concepts to practical analysis of simple signals. A large number of graphic examples are included to familiarize the reader with zero-based signal ideas and to prepare for the exploitation of these concepts in the analysis of speech signals and clipping phenomena. Then, in chapters 9 and 10, we will approach the specific problem of explaining *speech clipping phenomena* and *zero crossing signal* *parameter estimation*, respectively, using the tools of zero-based signal analysis.

8.1 Product Representation of Bandlimited Signals

8.1.1 Periodic Signals

In chapter 3 we noted that a signal s(t) periodic in T can be expressed as

$$s(t) = \sum_{k=-\infty}^{\infty} c_k \, e^{jk\Omega t} \, , \tag{8-1}$$

where the $\{c_k\}$ are complex Fourier series coefficients and, because s(t) is real,

$$c_k = c^*_{-k} \quad . \tag{8-2}$$

If s(t) is bandlimited to ±W Hz, where W = nΩ/2π, then the finite Fourier series for a bandlimited periodic signal results:

$$s(t) = \sum_{k=-n}^{n} c_k \, e^{jk\Omega t} \, . \tag{8-3}$$

Letting $w = e^{j\Omega t}$, (8-4)

$$s(t) = \sum_{k=-n}^{n} c_k \, w^k \, . \tag{8-5a}$$

We can write (8-5a) as

$$s(t) = c_{-n} w^{-n} \sum_{k=0}^{2n} c'_k \, w^k \, , \tag{8-5b}$$

where $c'_k = c_{n-k}/c_{-n}$. This is a polynomial in w of degree 2n and therefore, by the fundamental theorem of algebra, has 2n roots.[1] Thus (8-5b) can be written as

$$s(t) = c_{-n} w^{-n} \prod_{i=1}^{2n} (1 - \gamma_i w) , \qquad (8\text{-}5c)$$

where the roots $\gamma_i = |\gamma_i| e^{j\mu_i}$ are, in general, complex. This approach, suggested and developed by Voelcker [V-6], [V-7], is the key to zero-based signal theory. The following exposition (8.1.1, 8.1.2) is based largely upon his unpublished notes. Note that

$$\begin{aligned} c'_{2n} &= c_n / c_{-n} \\ &= c_n / c_n^* \\ &= e^{j2\theta_n} , \end{aligned} \qquad (8\text{-}6)$$

where

$$\theta_n = \arg[c_n] \ . \qquad (8\text{-}7)$$

Equating the summation in (8-5b) with the product in (8-5c) we find that

$$\begin{aligned} (1 + \ . \ . \ . \ . \ + e^{j2\theta_n} w^{2n}) &= \prod_{i=1}^{2n} (1 - \gamma_i w) \\ &= 1 - w \sum_{i=1}^{2n} \gamma_i \ . \ . \ . + w^{2n} \prod_{i=1}^{2n} \gamma_i . \end{aligned} \qquad (8\text{-}8)$$

Then, by equating coefficients of w^{2n} in (8-8), we find that

---

1 To prove the fundamental theorem of algebra it is sufficient to show that every polynomial has a root. This root can then be factored out leaving a polynomial of one less degree. This polynomial also has a root, etc. [L-20].

$$\prod_{i=1}^{2n} |\gamma_i| = |e^{j2\theta_n}| = 1, \tag{8-9a}$$

and

$$\sum_{i=1}^{2n} \mu_i = 2\theta_n \ . \tag{8-9b}$$

For this to obtain, some even number, $2n_R$, of the roots γ_i must have $|\gamma_i| = 1$ *and/or* there must exist n_C *pairs* of roots such that $|\gamma_j| = 1/|\gamma_k|$. Here $2n_R + 2n_C = 2n$. The notion that $|\gamma_j| = 1/|\gamma_k| \cdot |\gamma_\ell|$ is not admissible since the expansion must still obtain if n is reduced by unity. Therefore, (8-5c) can be rewritten as

$$s(t)=|c_n|e^{-j\theta_n}(w^{-\frac{1}{2}})^{2n}\prod_{i=1}^{2n_R} (1-e^{j\mu_i}w) \prod_{\ell=1}^{n_C} (1-|\gamma_\ell|e^{j\mu_\ell}w)\cdot(1-e^{j\mu_\ell}w/|\gamma_\ell|). \tag{8-10a}$$

Using, from (8-9b), $\theta_n = \sum_{i=1}^{2n_R} \mu_i/2 + \sum_{\ell=1}^{n_C} \mu_\ell/2$,

(8-10a) can be further simplified to

$$s(t) = |c_n| \prod_{i=1}^{2n_R} (w^{-\frac{1}{2}}e^{-j\mu_i/2} - w^{\frac{1}{2}}e^{j\mu_i/2}) \ .$$
$$\prod_{\ell=1}^{n_C} (w^{-\frac{1}{2}}e^{-j\mu_\ell/2}-|\gamma_\ell|w^{\frac{1}{2}}e^{j\mu_\ell/2})\cdot(w^{-\frac{1}{2}}e^{-j\mu_\ell/2}$$
$$-w^{\frac{1}{2}}e^{j\mu_\ell/2}/|\gamma_\ell|) \ . \tag{8-10b}$$

Finally, noting that $w = e^{j\Omega t}$, (8-4), we can write -- after some manipulation --

$$s(t) = (-1)^n|c_n| \prod_{i=1}^{2n_R} 2 \sin \frac{\Omega}{2}(t-\tau_i) \prod_{\ell=1}^{n_C} 2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)]. \tag{8-11}$$

Here
$$\tau_i = -\mu_i/\Omega$$
$$\tau_\ell = -\mu_\ell/\Omega$$

and $e^{\Omega\sigma_\ell} = |\gamma_\ell|$ or, $\sigma_\ell = \ln|\gamma_\ell|/\Omega$. s(t) is identically equal to zero for real values of time, $t = \tau_i$, or for complex values of time, $z = \tau_\ell + j\sigma_\ell$. Here we define z to be the complex time variable.

From (8-4) it follows that

$$\mathrm{Re}(z) = \{\tan^{-1}[\mathrm{Im}(w)/\mathrm{Re}(w)]\}/\Omega \tag{8-12a}$$

and
$$\mathrm{Im}(z) = -\ln\{[\mathrm{Re}(w)]^2 + [\mathrm{Im}(w)]^2\}/2\Omega \, . \tag{8-12b}$$

Hence, roots whose location on the w plane satisfy

$$[\mathrm{Re}(w)]^2 + [\mathrm{Im}(w)]^2 = 1,$$

i.e. roots on the unit circle of the w plane, lie on the real time axis of the z plane.

Thus s(t) is described completely, *except* for a multiplicative constant $|c_n|$, by the location of its real and complex zeros. We note that the ambiquity as to multiplicative constant arises because we require only 2n zeros in (8-11) whereas, from (8-3), s(t) requires 2n complex Fourier coefficients (half of which are the complex conjugates of the other half) *plus* one real Fourier coefficient--the D.C. component--for complete description. Therefore

$$s(t) \propto s_{RZ}(t) \cdot s_{CZ}(t) \tag{8-13}$$

where
$$s_{RZ}(t) = \prod_{i=1}^{2n_R} 2 \sin \frac{\Omega}{2} (t-\tau_i) \tag{8-14}$$

is a wholly real zero signal ,

and $$s_{CZ}(t) = \prod_{\ell=1}^{n_C} 2[\cos\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)] \qquad (8\text{-}15)$$
$$\geqslant 0$$

is a wholly complex zero signal [V-6].

Here $2n_R$ = number of *real zeros* or zero crossings per period T

n_C = number of *complex zero pairs* per period T

$2n_R + 2n_C$ = total number of zeros per period T

τ_i = location in time of i^{th} real zero $[0\leqslant\tau\leqslant T]$

and

$\tau_\ell \pm j|\sigma_\ell|$ = location in (complex) time of ℓ^{th} complex zero pair $[0\leqslant\tau\leqslant T]$.

8.1.2 Limiting Forms: Extensions to Aperiodic Signals

Although aperiodic signals can be treated by considering them as periodic signals with infinite period, the periodic signal model, (8-11), should be expected to approach the aperiodic Hadamard form [X-5, p. 246]

$$s(t) = s(0) \prod_{i=1}^{2n_R=\infty} (1-t/\tau_i) \prod_{\ell=1}^{n_C=\infty} (1-t/z_\ell)\cdot(1-t/z_\ell^*) \text{ , } \qquad (8\text{-}15)$$

(where $z_\ell = \tau_\ell + j|\sigma_\ell|$ and $s(0) \neq 0$) as T becomes very large.

In (8-11), as $T\to\infty$, $\Omega=2\pi/T\to 0$. It seems reasonable then to replace the trigonometric products with the first terms of their series expansions [V-9]:

i.e. $$s(t) = \lim_{\substack{T\to\infty \\ n\to\infty}} (-1)^n|c_n| \prod_{i=1}^{2n_R} 2\sin\frac{\Omega}{2}(t-\tau_i)\prod_{\ell=1}^{n_C}2[\cos\Omega\sigma_\ell-\cos\Omega(t-\tau_\ell)] \qquad (8\text{-}16a)$$

$$= \lim_{\substack{T\to\infty \\ n\to\infty}} (-1)^n |c_n| \prod_{i=1}^{2n_R} 2[\tfrac{\Omega}{2}(t-\tau_i) -. . + \ldots] \cdot$$

$$\prod_{\ell=1}^{n_C} 2[1 + \tfrac{1}{2}(\Omega\sigma_\ell)^2 + . . . -\{1-\tfrac{1}{2}\Omega^2(t-\tau_\ell) + . . .\}] \tag{8-16b}$$

$$= \lim_{\substack{T\to\infty \\ n\to\infty}} (-1)^n |c_n| \prod_{i=1}^{2n_R} \Omega(t-\tau_i) \prod_{\ell=1}^{n_C} \Omega^2[\sigma_\ell^2 + (t-\tau_\ell)^2] \text{ ,} \tag{8-16c}$$

if $(t-\tau_i) \lessgtr T$ for all i and $(t-\tau_\ell) \leq T$ for all ℓ.

After some rearrangement, we obtain

$$s(t) = (-1)^n |c_n| \Omega^n \prod_{i=1}^{2n_R} \Omega(t-\tau_i) \prod_{\ell=1}^{n_C} |z_\ell|^2 \prod_{i=1}^{2n_R} (1-t/\tau_i) \prod_{\ell=1}^{n_C} (1-t/z)(1-t/z^*), \tag{8-17a}$$

$$\text{with } s(0) = (-1)^n |c_n| \Omega^n \prod_{i=1}^{2n_R} \tau_i \prod_{\ell-1}^{n_C} |z_\ell|^2 \quad . \tag{8-17b}$$

Therefore (8-15) follows from (8-17a) and (8-17b). The "if" conditions which permit (8-16c) to replace (8-16b) may not always be satisfied. For this reason, the above approach is--at best--a plausibility argument for extending periodic model zero-based theory to the aperiodic case. Product formulations for aperiodic signals are discussed and examined in some detail by Requicha [R-7, part I], [L-10].

However, for the remainder of this thesis, we will be concerned only with *periodic* signals and signal models.

8.1.3 Basic Spectral Relationships

The factorization of s(t) into real zero (RZ) and complex zero (CZ) components immediately suggests certain spectral relationships. For example,

$$s(t) = \sum_{k=-(n_R+n_C)}^{n_R+n_C} c_k \cdot e^{jk\Omega t} , \tag{8-18a}$$

$$s_{RZ}(t) = \sum_{k=-n_R}^{n_R} Rz_k \cdot e^{jk\Omega t} , \tag{8-18b}$$

and

$$s_{CZ}(t) = \sum_{k=-n_C}^{n_C} Cz_k \cdot e^{jk\Omega t} , \tag{8-18c}$$

where $n = n_R + n_C$. Therefore, because

$$s(t) \propto s_{RZ}(t) \cdot s_{CZ}(t) \tag{8-19a}$$

it follows from (8-18) that

$$c_k \propto \sum_{n=\max\{-n_R,k-n_C\}}^{\min\{n_R,k+n_C\}} Rz_n \cdot Cz_{k-n} \tag{8-19b}$$

for $-n_C - n_R \le k \le n_C + n_R$.

This result may be obtained directly from the convolution theorem:

$$\text{i.e.} \quad s_{RZ}(t) \cdot s_{CZ}(t) \leftrightarrow \{Rz_k\} * \{Cz_k\} . \tag{8-20}$$

From (8-18) and (8-19) it is clear that s(t), $s_{RZ}(t)$ and $s_{CZ}(t)$ are bandlimited to $\pm n\Omega$, $\pm n_R\Omega$, and $\pm n_C\Omega$ radians/sec., respectively, so that

$$n\Omega = n_R\Omega + n_C\Omega \ . \tag{8-21}$$

We emphasize that $\{c_k\}$, $\{Rz_k\}$ and $\{Cz_k\}$, the Fourier coefficients of $s(t)$, $s_{RZ}(t)$ and $s_{CZ}(t)$, respectively, are *complex*.

From the relationships established in sections 8.1.1, 2, 3 the first principle regarding the significance of zero crossings as signal descriptors becomes obvious:

P1: Zero crossings (real zeros) *apparently* constitute only a partial description of bandlimited signals. Specifically, for a periodic signal bandlimited to $\pm W = n\Omega/2\pi$ Hz, the "percentage information" available in the form of zero crossings is

$$I[s(t)] = 100\ n_R/(n_R + n_C) \tag{8-22}$$

where $2n_R$ = the number of *zero crossings*,

n_C = the number of *complex zero pairs*,

and $2n = 2n_R+2n_C$ = the *total number of zeros*,

per period T.

Only when $n_C = 0$ is a signal *completely* determined (except for a multiplicative constant) by its zero crossings. The reasoning behind the qualification on P1: will become clear in Chapter 10.

8.2 Analytic Signal Formulation

The relationships developed in the previous sections were based upon Fourier series factorization. In order to exploit the powerful tools of complex variable theory it is necessary to derive certain zero-based relationships using Analytic signal theory.

If we again let $t \rightarrow z = t + j\sigma$, the complex time variable, then the properties of m(z) [defined by equation (2-25)] on the complex z plane may be studied with z = t constituting the special but familiar case of m(t).

H.B. Voelcker showed that the following properties obtain for m(z) [V-6]:

1: m(z) is analytic, i.e., free of singularities, in the closed [$\sigma>0$] upper half of the z plane (the UHP). Voelcker defined this type of analyticity as Analyticity.

2: m*(z) is analytic in the lower half [$\sigma<0$] of the z plane (LHP).

3: For m(t) bandlimited (i.e. $W<\infty$) m(z) is an entire finction of exponential order unity [R-7] and has singularities only for $|z|\rightarrow\infty$ in the LHP. As $|z|\rightarrow\infty$ in the UHP, $m(z)\rightarrow$ a finite constant.

$$\text{That Im } [m(t)] = \mathcal{H}\{\text{Re } [m(t)]\} \tag{8-23}$$

is a consequence of property 1: [V-6, p. 343].

8.2.1 Product Representations

The periodic analytic signal

$$m(t) = \sum_{k=0}^{n} c'_k \, e^{jk\Omega t} \tag{8-24a}$$

can be factored in a manner similar to that applied to s(t) to yield

$$m(t) = K \prod_{i=1}^{n} [1 - a_i e^{j\Omega(t-\tau_i)}] \,, \tag{8-24b}$$

where $|a_i| = e^{\Omega\sigma_i}$. $z_i = \tau_i + j\sigma_i$ is the location of the i^{th} zero and

$$r_i(t) = 1 - a_i e^{j\Omega(t-\tau_i)} \tag{8-25}$$

is termed the "elementary analytic signal." [H-1], [V-6]

It is instructive to ask, at this point, whether knowledge concerning the zero locations of m(t) allows one to make any statements about the zeros of s(t). Furthermore--and perhaps more importantly-- the question arises as to the existence of relationships between zero locations and other, more conventional, signal attributes. We consider these problems in sections 8.2.3 and 8.2.2 respectively.

8.2.2 Phase-Envelope Relationships

The envelope, $|m(t)|$, and phase, $\phi(t)$, of an Analytic signal may be related by studying the behaviour of the logarithm of $m(z) = |m(z)|e^{j\phi(z)}$ on the z plane [V-6]:

$$\text{i.e., } \ln m(z) = \ln|m(z)| + j\,\phi(z)\,. \tag{8-26}$$

Since $m(z) \to$ a finite constant as $|z| \to \infty$ in the UHP, ln m(z) may have--if the constant is zero-- a singularity at this point. The derivative of ln m(z),

$$\ln' m(z) = m'(z)/m(z) \tag{8-27}$$

has no *singularities* for finite UHP z (σ>0) provided that m(z) is free of UHP *zeros*. Under these conditions, it can be shown that [V-6]

$$\phi'(t) = H[\ln'|m(t)|] \tag{8-28}$$

and

$$\ln' |m(t)| = -H[\phi'(t)]. \tag{8-29}$$

Analytic signals with no UHP zeros are termed *minimum phase* (MP) in analogy with network theory. Signals with zeros in *both* half-planes are termed *non-minimum phase* (NMP) while signals with zeros only in the closed UHP (σ>0) are *maximum phase* (MaxP) signals.

In the general case (NMP), the *instantaneous frequency*, $\phi'(t)$, and the derivative of the log envelope, $\ln'|m(t)|$, are related by

$$\phi'(t) = \phi'(0) - \sum_n \frac{r_n \sigma_n}{|z_n|^2} + \sum_n \frac{r_n \sigma_n}{(t-\tau_n)^2 + \sigma_n^2} \tag{8-30a}$$

and $$\ln'|m(t)| = \ln'|m(0)| + \sum_n \frac{r_n \tau_n}{|z_n|^2} + \sum_n \frac{r_n (t-\tau_n)}{(t-\tau_n)^2 + \sigma_n^2} \tag{8-30b}$$

where the zeros of $m(z)$ are located at $z_n = \tau_n + j\sigma_n$ and r_n is the order of the zero at z_n [V-6, p. 345].

Therefore "phase and envelope fluctuations are wholly describable in terms of zeros, and thus the *zeros of a bandlimited wave can be viewed as its informational attributes.*" [V-6].

8.2.3 Relationship Between the Zeros of s(t) and those of m(t)

Using $r_i(t)$, equation (8-25), we may define an elementary real signal [V-6], [H-1]

$$s_i(t) = \text{Re}\,[r_i(t)] = 1 - a_i \cos \Omega t. \tag{8-31}$$

The zeros of $s_i(t)$ occur at

$$z_n = [2\pi n \pm j \cosh^{-1}(1/a_i)]/\Omega, \quad 0 < a_i \leqslant 1 \tag{8-32a}$$

or $$z_n = \tau_n = [2\pi n \pm \cos^{-1}(1/a_i)]/\Omega, \quad a_i > 1 \tag{8-32b}$$

whereas the zeros of the elementary analytic signal, $r_i(t)$ occur at

$$z_n = \tau_n + j\sigma_n = [2\pi n + j \ln(a_i)]/\Omega, \quad 0 < a_i < \infty . \tag{8-33}$$

If $0 < a_i \leq 1$, then the zeros of $s_i(t)$ occur in complex conjugate pairs, one per period $T = 2\pi/\Omega$. As $a_i \to 1$, these zeros approach the real axis (σ=0) and when $a_i = 1$, become second order real zeros. Thus $r_i(t)$, for $0 < a_i \leq 1$, represents a MP signal with one zero per period occuring in the LHP ($\sigma < 0$). To illustrate (8-29), we note that

$$\ln m_i(t)_{MP} = \ln [1-a_i e^{j\Omega t}],\ 0 < a_i \leq 1$$

$$= - \sum_{k=1}^{\infty} a_i^k \cdot e^{jk\Omega t}/k. \qquad (8\text{-}34)$$

Thus $\mathrm{Re}[\ln m_i(t)_{MP}] = \ln |m_i(t)|_{MP} = - \sum_{k=1}^{\infty} a_i^k \cdot \cos k\Omega t/k$ (8-35a)

and $\mathrm{Im}[\ln m_i(t)_{MP}] = \phi_i(t)_{MP} = - \sum_{k=1}^{\infty} a_i^k \cdot \sin k\Omega t/k.$ (8-35b)

The derivatives of (8-35a) and (8-35b) are indeed a Hilbert pair.

It follows that, although the zeros of the *elementary* real and analytic signals are "related", in the general case-- involving products of real or analytic signals-- knowledge of the zeros of m(t) certainly does *not* imply that the zeros of the real part of m(t)--s(t)-- are in any way simpler to locate. However, as Haavik has shown [H-1], knowledge of the *gross* nature of m(t) sometimes enables statements to be made about the overall distribution (i.e., whether complex or real) of the zeros of s(t).

8.2.4 The Properties of MaxP Signals

Using the elementary analytic signal--$r_i(t) = [1 - a_i e^{j\Omega(t-\tau_i)}]$-- we can represent a general MaxP Analytic signal as [H-1]

$$R_n(t) = |R_n(t)| e^{j\phi_{MaxP}(t)}, \qquad (8\text{-}36)$$

where $$|R_n(t)| = \prod_{i=1}^{n} |1 - a_i e^{j\Omega(t-\tau_i)}|, \quad 1 < a_i < \infty \tag{8-37a}$$

and $$\phi_{MaxP}(t) = \sum_{i=1}^{n} \phi_i(t) \quad [H-1] \, . \tag{8-37b}$$

$|r_i(t)|$ and $\phi_i(t)$ may be found by using logarithmic expansions:

i.e., $$\ln r_i(t) = \ln|r_i(t)| + j\,\phi_i(t) \tag{8-38a}$$

$$= [\ln a_i - \sum_{k=1}^{\infty} a_i^{-k} \cos k\Omega(t-\tau_i)/k]$$

$$+ j\,[\Omega(t-\tau_i) + \pi + \sum_{k=1}^{\infty} a_i^{-k} \sin k\Omega(t-\tau_i)/k],$$

$$1 < a_i < \infty \, . \tag{8-38b}$$

Alternatively, $|r_i(t)| = [1 + a_i^2 - 2a_i \cos\Omega(t-\tau_i)]^{\frac{1}{2}}$ so that $|r_i(t)| > 0$ for $a_i > 1$.[2]

It is therefore evident that

$$s_n(t) = \mathrm{Re}[R_n(t)] = |R_n(t)| \cos \phi_{MaxP}(t) \tag{8-39}$$

has real zeros only when

$$\cos \phi_{MaxP}(t) = 0. \tag{8-40}$$

We shall now show that these RZ's are the *only* zeros of s(t). The derivative of the elementary phase contributions to $\phi_{MaxP}(t)$ is found from (8-38b):

$$\phi_i'(t) = \Omega\,[1 + \sum_{k=1}^{\infty} a_i^{-k} \cos k\Omega(t-\tau_i)], \quad 1 < a_i < \infty. \tag{8-41}$$

---

2 Let $a_i = 1 + x$, $x > 0$. Then $|r_i(t)|_{min} = [1+a_i^2 - 2a_i]^{\frac{1}{2}}$
$= x > 0.$

Haavik showed [H-11] that $\phi_i'(t)$ is non-negative so that $\phi'_{MaxP}(t) = \sum_{i=1}^{n} \phi'_i(t) > 0$. This implies that $\phi_{MaxP}(t)$ is a monotone increasing function of time. From (8-38b), $\phi_i(t+T) - \phi_i(t) = \Omega T$ so that

$$\begin{aligned}\phi_{MaxP}(t+T) - \phi_{MaxP}(t) &= n\Omega T \\ &= 2\pi n \ .\end{aligned} \tag{8-42}$$

Thus $\cos\phi_{MaxP}(t)$, and hence $s_n(t)$, passes through odd multiples of $\pi/2$ and $3\pi/2$ n times per period and therefore exhibits 2n zero crossings per period. Since $s_n(t)$ contains only 2n zeros, all of its zeros are real or zero crossings.

$$\text{i.e.,}\quad s_n(t) = \prod_{i=1}^{2n} 2 \sin \frac{\Omega}{2} (t-\tilde{\tau}_i). \tag{8-43}$$

As before, the zeros of $s_n(t)$--$\{\tilde{\tau}_i\}$-- are not *simply* related to the zeros of $R_n(t)$--$\{\tau_i + j\sigma_i\}$.

Since an RZ signal, by definition, is completely determined (except for a multiplicative constant) by its zero crossings, it is interesting to ask whether operations exist such that a general RZ-CZ signal may be transformed into a wholly RZ signal. Given s(t) (and hence m(t)) then a process which could convert m(t) to a MaxP signal would simultaneously transform s(t) into a wholly RZ signal.

Haavik showed that at least two such processes exist [H-1]. We examine these in the next section.

8.3 Zero Conversion (CZ to RZ) Processes

8.3.1 Differentiation and Sinewave Addition

$$\textit{If}\qquad R_n(t) = \sum_{k=0}^{n} c'_k\, e^{jk\Omega t},\quad c'_k = \begin{cases} 2c_k, & k > 0 \\ c_k, & k = 0 \end{cases} \tag{8-44a}$$

happens to be MaxP then

$$F_n(w) = \sum_{k=0}^{n} c'_k\, w^k, \qquad w = e^{j\Omega t} \tag{8-44b}$$

has roots only on and inside the unit circle in the w plane.

S. Haavik showed that repeated differentiation of $s_n(t) = \mathrm{Re}[R_n(t)]$ converts the signal (asymptotically) into a real zero signal *by forcing $R_n(t)$ to become MaxP*. The following alternative proof, suggested by A. Requicha, also encompasses another zero conversion method:

$$\text{We write} \quad F_{n-1}(w) = \sum_{k=0}^{n-1} c'_k\, w^k = F_n(w) - c'_n\, w^n, \tag{8-45}$$

$$\text{and note that} \quad |F_{n-1}(w)| \leqslant \sum_{k=0}^{n-1} |c'_k| \cdot |w^k|. \tag{8-46a}$$

On the unit circle, $|w| = 1$, so that

$$|F_{n-1}(e^{j\theta})| \leqslant \sum_{k=0}^{n-1} |c'_k| \; . \tag{8-46b}$$

Then Rouchés theorem [M-6, p. 2] implies that if

$$|F_{n-1}(w)| < |c'_n\, w^n|, \qquad |w| < 1 \tag{8-47a}$$

then $c'_n\, w^n$ and

$$F_n(w) = F_{n-1}(w) + c'_n\, w^n \tag{8-47b}$$

have the same number of zeros inside the unit circle. But $c'_n\, w^n$ has n zeros, all at the origin. Therefore, from (8-46b), a sufficient condition for $F_n(w)$ to have n zeros within the unit circle, and therefore be MaxP, is

$$\sum_{k=0}^{n-1} |c'_k| < |c'_n| \; . \tag{8-48}$$

It is clear then that if the highest frequency component is "sufficiently large", $s_n(t)$ will be wholly RZ. It is also evident that repeated differentiation will ultimately satisfy this criterion. This suggests a second principle concerned with zeros as informational attributes of signals:

P2: Repeated differentiation of a bandlimited signal asymptotically converts the signal into a real zero signal. That is, differentiation tends to convert CZ's into RZ's--zero crossings.

Combining P1: and P2: we find that

$$I[s'(t)] \geqslant I[s(t)] \tag{8-49}$$

and $$I[s^n(t)] \rightarrow 1 \text{ as } n \text{ increases.}$$

That differentiation cannot decrease the number of zero crossings follows directly from Rolle's theorem. Equation (8-48) also implies that--as first suggested by Haavik--simply increasing $|c_n'|$ so that

$$|c_n'| > \sum_{k=0}^{n-1} |c_k'| \tag{8-50}$$

will ensure that $s_n(t)$ has only real zeros. Therefore, since a real signal bandlimited to $\pm W$ Hz must exhibit precisely 2W zeros per second, the addition of a sine wave of frequency W Hz and "sufficient amplitude" will convert all CZ's to RZ's. Thus:

P3: Addition of a sinewave of frequency W Hz to a bandlimited ($\pm W$ Hz), periodic signal s(t) will--if the sinewave amplitude is sufficient--convert all CZ's to RZ's.

Extension of P2: and P3: to random signals is intuitively straightforward. For example, the mean zero crossing rate of the m^{th} derivative of bandpass white Gaussian noise is [H-1],[R-10]

$$\rho_{o,m} = 2\left[\frac{2m+1}{2m+3}\right]^{\frac{1}{2}} \cdot \left[\frac{f_h^{\,2m+3} - f_\ell^{\,2m+3}}{f_h^{\,2m+1} - f_\ell^{\,2m+1}}\right]^{\frac{1}{2}} \quad , \tag{8-51}$$

where the noise is bandlimited to $[f_\ell, f_h]$ Hz. For $f_\ell = 0$, (8-51) reduces to (6-4). Note especially that $\rho_{o,\,m+1} > \rho_{o,m}$

8.3.2 Bandpass Filtering

Differentiation and sinewave addition convert CZ's to RZ's; highpass filtering sets a *lower bound* on the number of RZ's per period. This can be demonstrated by writing [V-11]

$$s(t) = \sum_{k=-n}^{-n_1} c_k \cdot e^{jk\Omega t} + \sum_{k=n_1}^{n} c_k \cdot e^{jk\Omega t} \; , \; 0 \leqslant n_1 \leqslant n \tag{8-52}$$

and noting that s(t) has 2n zeros per period.

Now $$m(t) = 2 \sum_{k=n_1}^{n} c_k \cdot e^{jk\Omega t} = |m(t)| \; e^{j\phi(t)}. \tag{8-53a}$$

Rearranging, we find that

$$m(t) = 2\; e^{jn_1\Omega t} \sum_{k=0}^{n-n_1} c_{k+n_1} \; e^{jk\Omega t} \tag{8-53b}$$

$$= e^{jn_1\Omega t} \; |m_{LP}(t)| \; e^{j\phi_{LP}(t)} \tag{8-53c}$$

where $$m_{LP}(t) = K \prod_{i=1}^{n-n_1} [1 - e^{j\Omega(t-z_i)}] \tag{8-54}$$

is an $(n-n_1)$ zero lowpass analytic signal.

Because $$\phi(t) = n_1\Omega t + \phi_{LP}(t) \; , \tag{8-55}$$

$[\phi(t+T) - \phi(t)]_{min} = 2\, n_1$, when $m_{LP}(t)$ is MP. Conversely, if $m_{LP}(t)$ is MaxP, then $\phi_{LP}(t+T) - \phi_{LP}(t) = 2\pi(n-n_1)$

and

$$[\phi(t+T) - \phi(t)] = 2\pi n_1 + 2\pi(n-n_1) = 2\pi n.$$

Therefore $$2\pi n_1 \leq \Delta\phi(t) \leq 2\pi n \qquad (8\text{-}56)$$

and s(t) will exhibit not *less* than $2n_1$ and not *more* than 2n real zeros--zero crossings--per period. Hence

P4: A periodic signal bandlimited to $[n_1\Omega/2\pi,\ n\Omega/2\pi]$ Hz exhibits not less than $2n_1$ zero crossings per period.

8.3.3 Application to Clipped Speech Psychoacoustic Phenomena

At this point we reiterate three of Licklider's "unexplained phenomena":

> L5. Pre-Clipping Differentiation: Pre-clipping speech differentiation results in higher word articulation scores (>90%) even for unpracticed listeners.
>
> L7. Ultra-Sonic Bias: Unless the level of an ultrasonic bias--applied to the speech waveform before clipping --is small compared to the speech signal level, the resultant clipped speech signal will be more intelligible than it would be *per se*.
>
> L3. Highpass Filtering: Severe (e.g., infinite) peak clipping is less deleterious to intelligibility if the original speech is filtered so as to remove the low frequency components.

Equating L5 *and* P2, L7 *and* P3, and L3 *and* P4 we find that operations which condition s(t) by *increasing* the number of zero crossings per period (by effectively converting CZ's to RZ's) produce a more intelligible clipped speech waveform. We suspect, therefore, that *the greater the percentage of zeros available as zero crossings then the greater amount of information preserved by clipping* since clipping *apparently* affects only CZ's and leaves RZ's unaffected. (We shall assume, for the remainder of this thesis, that in speech clipping systems the clipped signal is re-bandlimited to the bandwidth of the original signal so that the total number of zeros per period is unchanged. *In practice, this re-bandlimiting is often effectively accomplished by the electrical-to-audio transducer, i.e., the headphones or loudspeaker.)*

However, before these ideas can be consolidated, the effect of clipping on complex zero configuration--and the links between zeros and spectral parameters--must be clarified. For example, the unrestricted manipulation of only one complex zero pair could significantly alter the spectral characteristics of the Fourier series polynomial. If clipping--which can be considered to be a member of a class of operations affecting only the complex zero component of a signal--can be shown to be somehow *restricted* in its freedom to manipulate complex zeros then arguments for gross preservation of the complex zero signal spectrum could be put forward. Explanation of these phenomena requires an investigation into the geometry of the zeros of polynomials.

Our first priority, however, is to review and establish some physical characteristics of RZ and CZ signals and to thus provide a more meaningful link between zeros and signal spectral characteristics.

8.4 Real Zero Signals

Real zero signals possess the minimum bandwidth possible for any signal having the specified set of zero crossings; in this sense they are unique. A real periodic signal having $2n_R$ zero crossings per period and no complex zeros has bandwidth $n_R\Omega/2\pi$ Hz, where $\Omega = 2\pi/T$ and $s(t)$ is periodic in T. It follows directly from Rolle's Theorem that *all* derivatives of real zero signals are real zero.

8.4.1 The Spectrum of RZ Signals

Since $$s_{RZ}(t) = \prod_{i=1}^{2n_R} 2 \sin \frac{\Omega}{2}(t-\tau_i) \qquad (8\text{-}57a)$$

$$= 2 \sum_{k=0}^{n_R} |Rz_k| \cos(k\Omega t + \theta_k), \qquad (8\text{-}57b)$$

then the $\{Rz_k\}$'s and $\{\theta_k\}$'s can be derived explicitly in terms of the $\{\tau_i\}$'s. The following results are primarily of academic interest. In practice, the Fourier coefficients of $s_{RZ}(t)$ are calculated by expanding (8-57a) to yield $s_{RZ}(t)$ at $2n_R$ equispaced time intervals and then employing the discrete Fourier transform.

Expanding (8-57) we find, after much manipulation, that the spectral components of $s_{RZ}(t)$ can be calculated as follows:

i) $k = n_R$

$$|Rz_{n_R}| \cos(n_R\Omega t + \theta_{n_R}) = \cos \frac{\Omega}{2}(2n_R t - \sum_{i=1}^{2n_R} \tau_i), \qquad (8\text{-}58)$$

so that $|Rz_{n_R}| = 1$.

ii) $0 < k < n_R$

$$|Rz_k| \cos(k\Omega t + \theta_k) = (-1)^k \sum_{j=1}^{\phi_{n_R}} \cos \frac{\Omega}{2}[2(n_R-k)t + \tau_{\phi_j}] \qquad (8\text{-}59)$$

where $\phi_{n_R} = \binom{2n_R}{k}$ and τ_{ϕ_j} is the sum of the elements in the j^{th} row of a $\{2n_R x \phi_{n_R}\}$ matrix

$$\tau_\phi = [M] \cdot \begin{bmatrix} \tau_1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \tau_2 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \tau_3 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \cdot & & & & & & & & \cdot \\ \cdot & & & & & & & & \cdot \\ \cdot & 0 & 0 & 0 & & & & & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \tau_{2n_R} \end{bmatrix}$$

Here [M] is a $\{2n_R \times \phi_{n_R}\}$ matrix of signed 1's where j of the $2n_R$ 1's in each row are given plus (+) signs in each of the $\binom{2n_R}{j}$ possible ways and the rest of the 1's are given minus (-) signs.

iii) $k = 0$

$$|Rz_o| \cos\theta_o = (-1)^{n_R} \sum_{j=1}^{\phi_{n_R}} \cos\frac{\Omega}{2}(\tau_1 + \tau_{\phi_j}) \qquad (8\text{-}60)$$

where $\phi_{n_R} = \binom{2n_R-1}{n_R-1} = \frac{1}{2}\binom{2n_R}{n_R}$ and τ_{ϕ_j} is the sum of the elements in the j^{th} row of a $\{(2n_R-1) \times \phi_{n_R}\}$ matrix

$$\tau_\phi = [N] \cdot \begin{bmatrix} \tau_2 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \tau_3 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \tau_4 & & & & & & \cdot \\ \cdot & \cdot & \cdot & & & & & & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \tau_{2n_R} \end{bmatrix}$$

Here [N] is a $\{(2n_R-1) \times \phi_{n_R}\}$ matrix of signed 1's where j of the $(2n_R-1)$ 1's in each row are given plus (+) sign in each of the $\binom{2n_R-1}{n_R-1}$ possible ways and the rest of the 1's are given minus (-) signs.

The details of the preceding computational algorithm make it quite clear that *the spectral nature of $s_{RZ}(t)$ is a very complicated function of the zero crossing positions.* However, from equations (8-58), (8-59), and (8-60), it is evident that, for a given number of RZ's, $2n_R$,

$$|Rz_k|_{max} = \begin{cases} \binom{2n_R}{k} , & 0 < k < n_R \\ \frac{1}{2}\binom{2n_R}{n_R} , & k = 0 \end{cases} \quad ; \qquad (8\text{-}61)$$

When $k = n_R$, $|Rz|$ is, as per (8-58), unity.

8.4.2 Real Zero Interpolation

A real zero signal is specified entirely by its zero crossing (RZ) positions. Thus, clipping-- which preserves RZ positions -- is a lossless process for real zero signals. Given the set of zero crossing positions, $\{\tau_i\}$, then $s_{RZ}(t)$ can be generated using equation (8-14),

$$\text{i.e.,} \qquad s_{RZ}(t) = \prod_{i=1}^{2n_R} 2 \sin \frac{\Omega}{2}(t-\tau_i) \; .$$

However, this requires knowledge of both Ω and all $2n_R$ zero crossings before $s_{RZ}(t)$ can be calculated.

Voelcker showed that $s_{RZ}(t)$ for arbitrary (i.e. aperiodic) signals can be approximated closely, or with arbitrary accuracy, by invoking the phase-envelope relationships noted in 8.2.2([V-6, pt. II):

$$\text{i.e., writing } s_{RZ}(t) = |s_{RZ}(t)| \cos \phi_{s_{RZ}}(t) \quad , \tag{8-62}$$

it can be shown that

$$\phi'_{s_{RZ}}(t) = \sum_i \pi \cdot \delta\ (t-\tau_i) \tag{8-63}$$

$$\text{and that } \ln'|s_{RZ}(t)| = \sum_i 1/(t-\tau_i) \tag{8-64a}$$

$$= H\{\phi'_{s_{RZ}}(t)\} \ . \tag{8-64b}$$

Then, from (8-62), (8-63), and (8-64)

$$s_{RZ}(t) = \mathrm{sgn}[s(t)] \cdot \exp\ \{ \int_{-\infty}^{t} H[\sum_i \delta(t-\tau_i)]\ dt\}. \tag{8-65}$$

This method, defined by Voelcker as *Real Zero Interpolation*, ostensibly removes the periodicity criterion implicit in (8-14). The requirement that all zero crossing positions be known is not relaxed since implementation of (8-65) requires a real, non-ideal Hilbert transformer which is characterized by *finite* memory. However, (8-65) makes is possible to generate an approximation to $s_{RZ}(t)$; because the impulse response of a Hilbert transformer falls off as 1/t (sec. 2.3.2), the influence of zero crossings remote from t = 0 becomes negligible if the non-ideal Hilbert transformer is "sufficiently long." Fig. 8.1, from [V-6, pt. II], illustrates the operation of the Real Zero Interpolator.

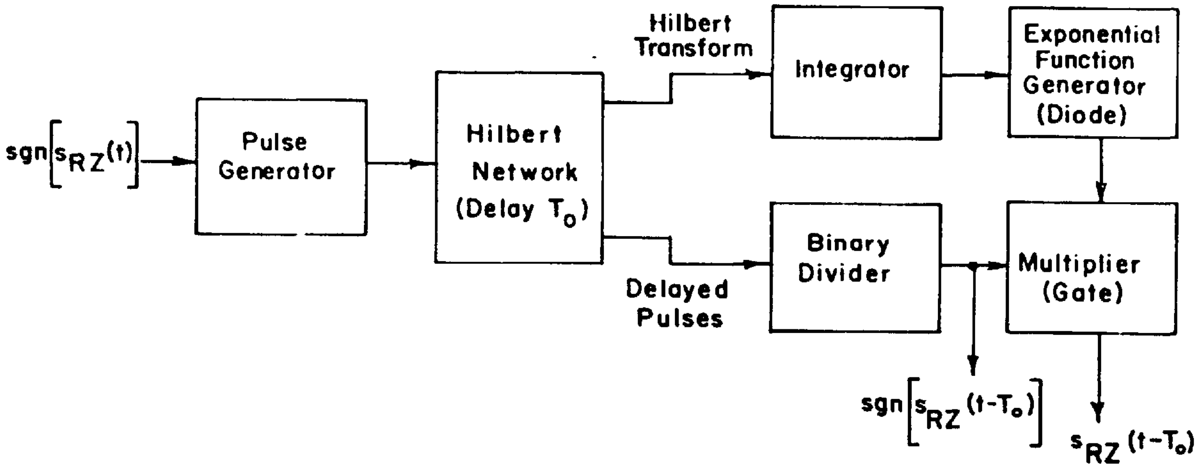


Fig. 8.1 The Real Zero Interpolator, block diagram. (From [V-6].)

The significance of $s_{RZ}(t)$ to clipped speech studies is that all information needed to construct the clipped speech waveform is carried by $s_{RZ}(t)$ in a signal of minimum bandwidth,

i.e., $$\mathrm{sgn}[s(t)] = \mathrm{sgn}\,[s_{RZ}(t)] \quad . \tag{8-66}$$

We note here, for future reference, that the output of the Real Zero Interpolator, for speech input, is almost completely *unintelligible*. Thus, *the intelligibility of clipped speech depends upon more than preservation of zero crossing locations*. Specifically, the nature of the interpolating waveform (i.e., clipped speech results from zero crossing interpolation with a rectangular waveform) is of great importance. V. Sobolev and V. Telepnev have shown, for example [S-17], that zero crossing interpolation with waveforms of the form

$$s_i(t) = (-1)^i \sin[\pi(t-\tau_i)/\Delta\tau_i], \tag{8-67}$$

(a single sine wave half-cycle interpolated between zero crossings) or $s_i(t) = (-1)^i\{\sin[\pi(t-\tau_i)/\Delta\tau_i] + k(\tau_i)\ \sin[3\pi(t-\tau_i)/\Delta\tau_i]\}$, (8-68)

(a two-term square wave approximation half-cycle interpolated between adjacent zero crossings) --where $\Delta\tau_i = \tau_{i+1} - \tau_i$ and $k(\tau_i) \propto \Delta\tau_i$ -- produces speech which is subjectively more pleasant than rectangular interpolated (clipped) speech.

8.5 Complex Zero Signals

A wholly complex zero, bandlimited periodic signal may be defined as

$$s_{CZ}(t) = \prod_{\ell=1}^{n_C} 2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)] \tag{8-69}$$

$$\geqslant 0$$

where the n_C complex zero *pairs* occur at complex times $z_\ell = \tau_\ell \pm j\sigma_\ell$.

Observe that the elementary complex zero signal, in (8-69), $2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)]$, is periodic in $T = 2\pi/\Omega$ whereas the elementary real zero signal, $2 \sin \frac{\Omega}{2}(t-\tau_i)$, is periodic in 2T. Thus--by the product-convolution relationship--addition of each complex zero pair to a signal increases the signal bandwidth by $\Omega/2\pi$ Hz. In contrast, each additional real zero increments the bandwidth by $\frac{1}{2}(\Omega/2\pi)$Hz. However, a *periodic* signal must have an even number of real zeros and real zeros must therefore be added in pairs. For example, in a bandwidth preserving CZ conversion process--e.g., differentiation-- each converted complex zero pair becomes two real zeros.

Real zeros--zero crossings-- are *overt* signal attributes. Complex zeros are ostensibly *covert*, or hidden. Their presence, however, may often be inferred from other signal attributes. In the following two sections we will discuss the nature of complex zero signals and methods of determining the complex zero positions.

8.5.1 Determination of $s_{CZ}(t)$

i) Division

Since
$$s(t) \propto s_{RZ}(t) \cdot s_{CZ}(t) \,, \tag{8-70}$$

$s_{CZ}(t)$ may be extracted from $s(t)$ by noting that

$$s_{CZ}(t) \propto s(t) / \left[\prod_{i=1}^{2n_R} 2 \sin \frac{\Omega}{2}(t-\tau_i) \right] \,. \tag{8-71}$$

$s_{RZ(t)}$ is synthesized using the product formulation, eq. (8-14), which is expanded in terms of the zero crossings of the original signal. As will be noted in sec. 8.6, the positions of zero crossings may be defined to an arbitrary degree of precision by bandlimited interpolation of the Nyquist samples using the FFT implementation of the discrete Fourier transform.

For example, a bandlimited periodic square wave of unity amplitude has the Fourier series representation

$$s(t) = \frac{4\Omega}{\pi} \sum_{\substack{k=1 \\ (k\ \mathrm{odd})}}^{L} \frac{\sin k\Omega t}{k} \,, \tag{8-72}$$

where $\Omega = 2\pi/T$ and $s(t)$ is bandlimited to $\pm L\Omega/2\pi$ Hz. But

$$\sin n\Omega t = \binom{n}{1} \cos^{n-1}\Omega t \cdot \sin\Omega t - \binom{n}{3} \cos^{n-3}\Omega t \cdot \sin^3 t + \ldots - . \tag{8-73a}$$

$$= \sin\Omega t[\binom{n}{1}\cos^{n-1}\Omega t\cdot x^0 - \binom{n}{3}\cos^{n-3}\Omega t\cdot x^1 + \binom{n}{5}\cos^{n-5}\Omega t\cdot x^2 - +] \tag{8-73b}$$

where $x = (1 - \cos^2\Omega t)$. Using the Binomial expansion,

$$(1 - \cos^2\Omega t)^n = \sum_{i=o}^{n} (-1)^i\binom{n}{i}\cos^{2i}\Omega t \quad . \tag{8-74}$$

Therefore, from (8-72) and (8-73),

$$s(t) = \frac{4\Omega}{\pi} \sum_{\substack{k=1 \\ (k\ odd)}}^{L} \{\frac{1}{k} \sum_{j=0}^{(k+1)/2} (-1)^{j+1}\binom{k}{2j-1}\cos^{k-2j+1}\Omega t \quad .$$

$$[\sum_{i=0}^{j-1} (-1)^i\binom{j-1}{i}\cos^{2i}\Omega t]\}\cdot \sin\Omega t \quad . \tag{8-75}$$

It is clear that, for the square wave,

$$s_{RZ}(t) = 4 \sin\frac{\Omega}{2} t \cdot \sin\frac{\Omega}{2}(t-T/2)$$

$$= 4 \sin\frac{\Omega}{2} t \cdot \cos\frac{\Omega}{2} t$$

$$= 2 \sin\Omega t \quad , \tag{8-76}$$

and $s_{CZ}(t)$ is simply (8-75) with the factor 2 sin Ωt removed.

Figures 8.2 and 8.3 illustrate the following features of the bandlimited square wave, (8-72), with L = 15 and 31, respectively, and Ω = 1.

a) $s(t) = s_{RZ}(t) \cdot s_{CZ}(t)$

b) $s_{RZ}(t) = \prod_{i=1}^{2n_R} 2 \sin \frac{\Omega}{2}(t-\tau_i)$

c) $s_{CZ}(t) = \prod_{\ell=1}^{n_C} 2[\cosh\Omega\sigma_\ell - \cos\Omega(t - \tau_\ell)]$

d) $|s_{CZ}(f)|$

e) Root map: the RZ-CZ positions on the complex time plane.

(The method of complex zero location will be discussed in sec. 8.6). Note that the proportionality constant which makes the rms value of s(t) equal to unity has been omitted from the diagrams. Multiplication of all s(t) values by $|c_n|$ $(2/\pi)$ will accomplish this (see (8-11)); Here, $|c_n| = 1/L$.

In practice, dividing s(t) by $s_{RZ}(t)$ to obtain $s_{CZ}(t)$ is complicated by the fact that $s(t) = s_{RZ}(t) = 0$ at all real zeros or zero crossings. This problem is solved using l' Hôpital's Rule:

$$\lim_{t\to\tau_i} \frac{s(t)}{s_{RZ}(t)} = \lim_{t\to\tau_i} \frac{s'(t)}{s'_{RZ}(t)} = \frac{s(\tau_i+\Delta t)}{s_{RZ}(\tau_i+\Delta t)} \quad . \tag{8-77}$$

ii) <u>Deconvolution</u>

Equations (8-19) and (8-20) express the basic convolution relationship which yields the Fourier series coefficient of s(t), $\{c_k\}$, given those of $s_{RZ}(t)$, $\{Rz_k\}$ and $s_{CZ}(t)$, $\{Cz_k\}$:

$$c_k \propto \sum_{n = \max\{-n_R,\ k-n_C\}}^{\min\{n_R,\ k+n_C\}} Rz_n \cdot Cz_{k-n} , \tag{8-78a}$$

$$-n_C-n_R \le k \le n_C+n_R$$

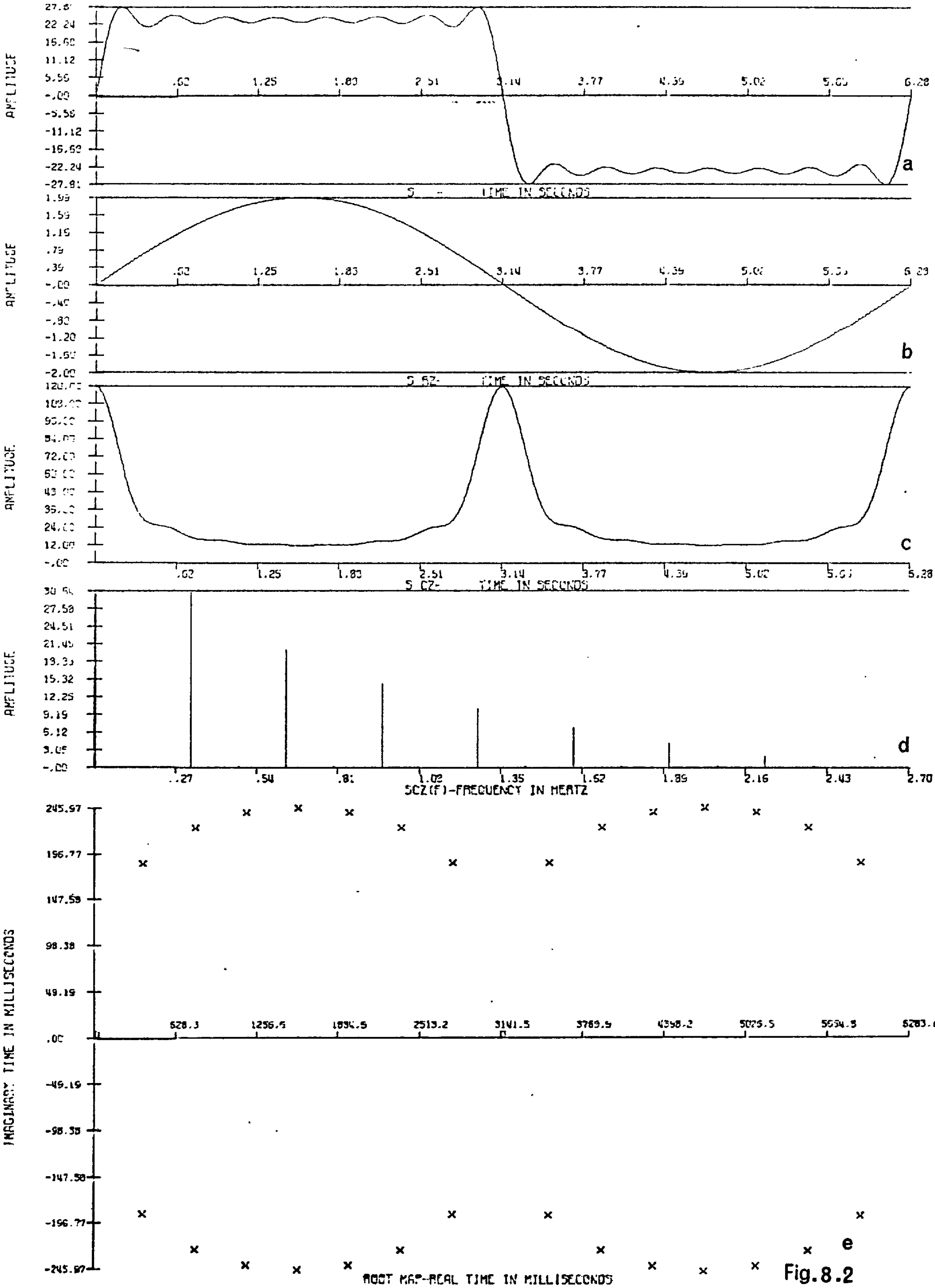
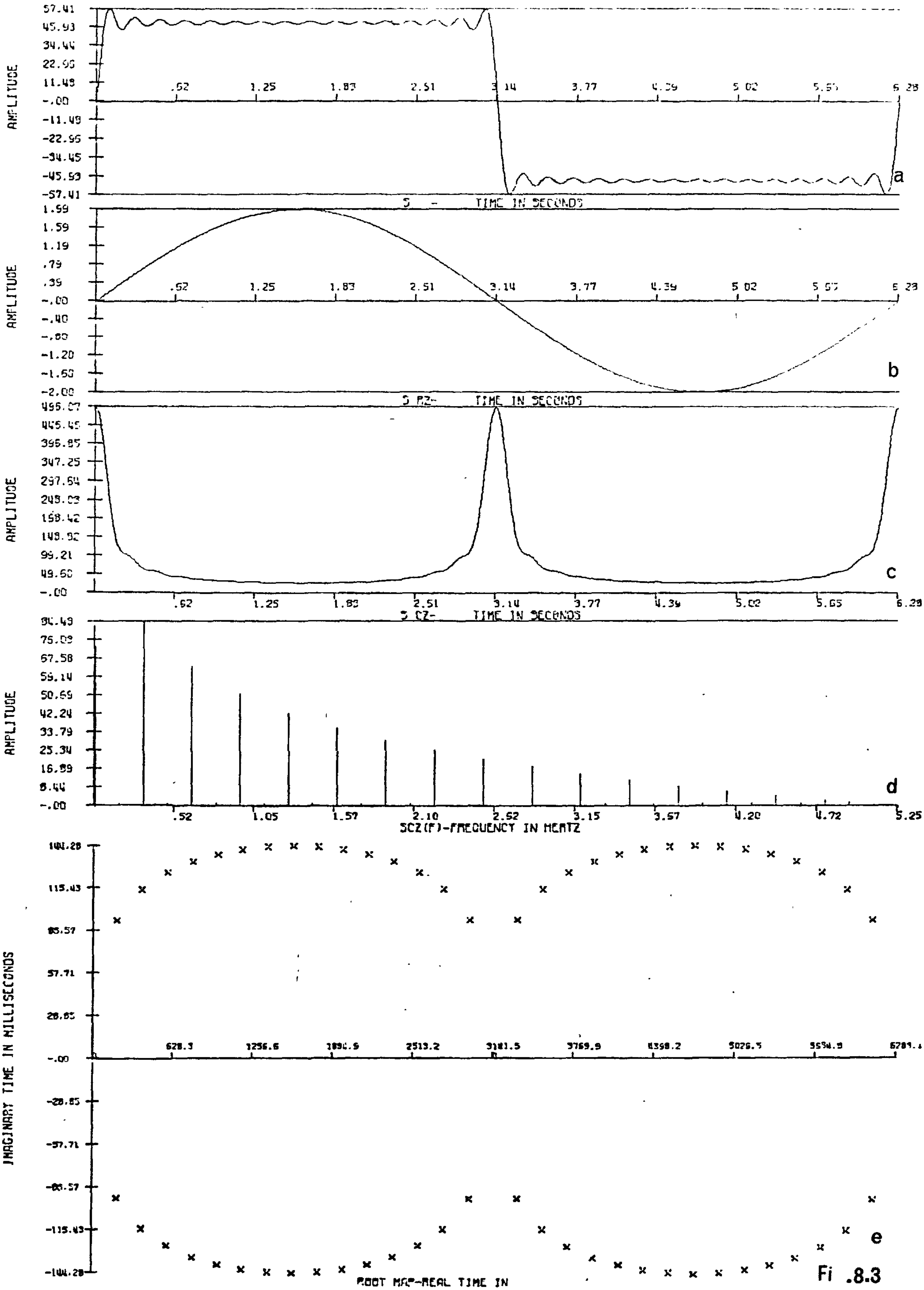


Fig. 8.2


AMPLITUDE
TIME IN SECONDS
a
b
c
d
SCZ(F)-FREQUENCY IN HERTZ
e
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN

Fi .8.3

or $\{c\} \propto \{Rz\}*\{Cz\}$. (8-78b)

{c} is often called the serial product of {Rz} and {Cz} because the sequence {c} consists of the coefficients of the polynomial which is the product of the polynomials represented by {Rz} and {Cz} [B-16; p. 35].

When {c} and {Rz} are known, {Cz} may be found by long division of polynomials. Polynomial division is equivalent to calculating {Cz} using the following relationships [B-16, pp. 35-36].

$$Cz_k = \begin{cases} c_{-(n_R+n_C)} \, / \, Rz_{-n_R} \, , & k = -n_C \\[2ex] \dfrac{c_{-n_R+k} - \sum\limits_{j=-n_C}^{k-1} Cz_j \cdot Rz_{k-j-n_R}}{Rz_{-n_R}} \, , & -n_C < k \leqslant 0 \\[2ex] Cz^*_{-k} \, , & 0 < k \leqslant n_C \, . \end{cases} \qquad (8\text{-}79)$$

Note that each subsequent value of Cz_k depends upon all previous values of Cz_k calculated; thus roundoff errors may accumulate rapidly if the values of {c} and {Rz} are not accurate.

iii) Analytic Factorization

Analytic factorization of polynomials higher than the 2^{nd} degree is cumbersome and only very specific solutions exist. However, certain waveforms possess symmetries which enable conditions --similar to those used in evaluated Fourier integrals (e.g., [P-2, pp. 10-12]) --to be formulated and used to effect factorization of higher degree polynomials.

Generally, $f_n(w) = K \cdot f_{np}(w) = K \prod_{i=1}^{n} (w - w_i)$, so that $f_{np}(w) = w^n - w^{n-1}\left\{\binom{n}{1} w_j\right\} + w^{n-2}\left\{\binom{n}{2} w_j\right\} - + \,..(-1)^n\left\{\binom{n}{n} w_j\right\}$,

(8-80)

where $\left\{\binom{n}{i} w_j\right\}$ consists of $\binom{n}{i}$ terms, each involving all the possible selections of the n roots taken i at a time. For instance,

$$f_{4p}(w) = \prod_{i=1}^{4} (w - w_i) \qquad (8\text{-}81a)$$

$$\begin{aligned} &= w^4 - w^3 \cdot (w_1+w_2+w_3+w_4) \\ &\quad +w^2 \cdot (w_1w_2+w_1w_3+w_1w_4+w_2w_3+w_2w_4+w_3w_4) \\ &\quad -w \cdot (w_1w_2w_3+w_1w_2w_4+w_1w_3w_4+w_2w_3w_4) \\ &\quad +w_1w_2w_3w_4 \text{ , unless } a_1 \text{ and/or } a_2 = 1. \end{aligned} \qquad (8\text{-}81b)$$

Because our polynomials are actually Fourier series representing real signals, we can make the following statements concerning the roots of (8-81):

i) If $w_1 = a_1e^{j\theta_1}$, $w_2 = a_2e^{j\theta_2}$ then $w_3 = e^{j\theta_1}/a_1$ and $w_4 = e^{j\theta_2}/a_2$, unless a_1 and/or $a_2 = 1$.

ii) $|w_1w_2w_3w_4| = 1$

iii) $|w_1+w_2+w_3+w_4| = |w_1w_2w_3+w_1w_2w_4+w_1w_3w_4+w_2w_3w_4|$.

iv) $\mathrm{Im}[\, w_1w_2+w_1w_3+w_1w_4+w_2w_3+w_2w_4+w_3w_4] = 0$.

Two examples follow in which 6 and 10 degree polynomials representing bandlimited square waves (BW = 3 and 5 Ω, respectively) are factored analytically by invoking waveform symmetry conditions and equation (8-81):

Example 1: A 6^{th} degree square wave, $s(t) = \sin\Omega t + \sin 3\Omega t/3$. (8-82)

On the w plane, the roots must lie at

i) 1, -1 -- the real zeros

and ii) jx, j/x, $-jx$, and $-j/x$, (8-83)

by virtue of symmetry conditions. Also,

$$\begin{aligned} &w_1w_2+w_1w_3+w_1w_4+w_1w_5+w_1w_6 \\ &\quad +w_2w_3+w_2w_4+w_2w_5+w_2w_6 \\ &\qquad +w_3w_4+w_3w_5+w_3w_6 \\ &\qquad\quad +w_4w_5+w_4w_6 \\ &\qquad\qquad +w_5w_6 = 0 , \end{aligned} \tag{8-84}$$

because the coefficient of $\sin 2\Omega t$ is zero. Inserting the roots of (8-83) into (8-84) we find that $x^2 = -2 \pm\sqrt{3}$ so that $x = j\ 1.93$ or $j\ 0.52$. We shall later confirm these results with computer factorization of (8-82).

Example 2: A 10^{th} degree square wave,

$$s(t) = \sin\Omega t + \sin 3\Omega t/3 + \sin 5\Omega/5 \quad . \tag{8-85}$$

On the w plane, the roots must lie at

i) 1, -1 -- the real zeros

and ii) $re^{j\theta}$, $re^{j(\pi-\theta)}$, $re^{j(\pi+\theta)}$, $re^{-j\theta}$

$e^{j\theta}/r$, $e^{j(\pi-\theta)}/r$, $e^{j(\pi+\theta)}/r$ and $e^{j\theta}/r$, (8-86)

again, by virtue of waveform symmetries.

We find that $\left\{\binom{10}{2} w_j\right\}$ consists of 45 terms of the form $w_m w_n$, $m\neq n$, which sum to 5/3 and that $\left\{\binom{10}{4} w_j\right\}$ consists of 210 terms of the form $w_m w_m w_o w_p$, m, n, o, p all different integers, which sum to -4/3. After very much manipulation we derive two equations:

$$\cos 2\theta \cdot [r^2 + r^{-2}] = -4/3 \tag{8-87a}$$

and $$[r^4 + r^{-4}] + [r^2 + r^{-2}] \cdot 2\cos 2\theta + 2\cos 4\theta = 3. \tag{8-87b}$$

Letting $r = e^{\phi}$, we obtain

$$\cos 2\theta \cdot \cosh 2\phi = -2/3 \tag{8-87c}$$

and $$2\cosh 4\phi + 4\cosh 2\phi \cdot \cos 2\theta + 2\cos 4\theta = 3\ . \tag{8-87d}$$

Solving, $r = e^{0.476} = 1.61$ and $\theta = 58.45°$.

This type of factorization procedure is generally supplanted by iterative computer based methods when dealing with realistic speech signal models (or actual waveforms) which involve equations of at least the 50th degree.

8.5.2 Inference of CZ Positions in Real Time

Examination of Figs. 8.2 and 8.3 reveals that the complex zero pairs have positions in real time which are associated with the square wave ripple. The following theorem shows that this should be true:

Theorem: Between two successive maxima of a bandlimited periodic signal, there must be a complex zero pair--provided that there is not a minimum of $s_{RZ}(t)$ between these points.

We say that "successive maxima" occur at t_1 and t_2 if

$$\text{i)}\quad s'(t_1) = s'(t_2) = 0 \tag{8-88}$$

and $$\text{ii)}\quad |s(t_o)| < \min\{|s(t_1)|, |s(t_2)|\},\ \text{where } t_1 < t_o < t_2.$$

Condition ii) demands that $s'(t) = 0$ for some t such that $t_1<t<t_2$.

Proof: As usual, we may write

$$s(t) \propto s_{RZ}(t) \cdot \prod_{\ell=1}^{n_C} 2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)] \qquad (8\text{-}89)$$

with (for example) positive maxima of $s(t)$ occurring at $t = t_1, t_2$ and $|s_{RZ}(t)| \geqslant \min\{|s_{RZ}(t_1)|, |s_{RZ}(t_2)|\}$. That is, $s_{RZ}(t)$ monotonically increases, or decreases, between t_1 and t_2.

Assume that there is *no* complex zero pair between t_1 and t_2. Then, for any $\ell = 1, 2, \ldots n_C$ and $t_1 < t < t_2$,

$$2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)] \geqslant K_\ell \qquad (8\text{-}90)$$

where $K_\ell = \min\{2[\cosh\Omega\sigma_\ell - \cos\Omega(t_1-\tau_\ell)],\ 2[\cosh\Omega\sigma_\ell - \cos\Omega(t_2-\tau_\ell)]\}$.

Calculate the following sequence:

$$s_\ell(t) = s_{\ell-1}(t) \cdot \{2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)]\ ,\ \ell = 1, \ldots n_C \qquad (8\text{-}92)$$

where

$$s_0(t) = s_{RZ}(t) \qquad (8\text{-}93)$$

and

$$s_{n_C}(t) = s(t) \quad . \qquad (8\text{-}94)$$

Then, for $t_1 < t < t_2$ and $1 \leqslant \ell \leqslant n_C$,

$$s_\ell(t) = s_{\ell-1}(t) \cdot 2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)] \qquad (8\text{-}95a)$$

$$\geqslant s_{\ell-1}(t) \cdot K_\ell \ . \qquad (8\text{-}95b)$$

That is to say, all points between t_1 and t_2 are multiplied by a value which is greater than or equal to K_ℓ. Thus, if there is no CZ between t_1 and t_2 there can be no $s(t_o)$, $t_1 < t_o < t_2$, such that $|s(t_o)| < \min\{|s(t_1)|, |s(t_2)|\}$. Hence $s(t_1)$ and $s(t_2)$ cannot, by condition ii) of our definition, be successive maxima. Q.E.D.

We emphasize here that the converse does not obtain; the presence of a CZ pair is not always signalled by the presence of "adjacent maxima." Thus, there must be a complex zero pair *between* adjacent ripple maxima on a square wave. Close examination of Figs. 8.2 and 8.3 reveals that the CZ's doe not occur exactly at the minimum between the adjacent maxima. We note that it can be shown that for

$$s(t) = \frac{2}{\pi} \sum_{\substack{k=1 \\ (k \text{ odd})}}^{2n-1} \frac{\sin kt}{k} ,$$

$$s'(t) = \frac{2}{\pi} \sin(2nt)/\sin(t)$$

so that the ripple peaks of $s(t)$ occur at $t = m\pi/2n$, $m = 1, 3, 5\ldots$ and ripple minima of $s(t)$ occur at $t = m\pi/2n$, $m = 2, 4, 6, \ldots$. It follows that ripple on a square wave bandlimited to $\pm W = \pm n\Omega/2\pi$ Hz (n odd) occurs at a frequency W Hz. There are $(n-2)$ ripple minima, each associated with a CZ pair, and 2 real zeros so that, as per (8-21b), $n = n_R + n_C = (2) + (n-2)$. In sec. 8.7 we shall examine further the determination of the positions of CZ's in the complex time domain.

8.6 Computer Factorization of Complex Polynomials

8.6.1 Difficulties in Root Finding

Location of the roots of polynomials with complex coefficients may be carried out in a number of ways. Among the more well known techniques are the secant method, the Newton-Raphson method and the methods of Muller and Laguerre [R-3, ch. 10]. The condition of the polynomial is important in this respect. A polynomial is *ill-condititioned* if very small changes in its coefficients result in large changes in its zero locations. For example, the polynomial

$$f(w) = \prod_{i=1}^{20} (w-w_i), \text{ where } w_i = i$$

$$= w^{20} - 210\, w^{19} + 20,615\, w^{18} - \ldots +20!$$

is highly ill-conditioned. Replacement of -210 by $-(210+2^{-23})$ [-210.000000119] results in

$$\left.\begin{matrix} w_{14} \\ \\ w_{15} \end{matrix}\right\} = 13.992 \pm j\, 2.519$$

and

$$\left.\begin{matrix} w_{16} \\ w_{17} \end{matrix}\right\} = 16.731 \pm j\, 2.813 \text{ [R-3, p. 186].}$$

The viability of the various factorization methods and accuracy considerations are discussed by E. Bareiss in [R-3] and by Delves and Lyness [D-9].

Fortunately, as we shall see, the roots of the polynomials we wish to factorize are such that the polynomials are well-conditioned. This is the case because the roots lie either on the unit circle--and therefore have magnitude unity-- or occur in reflected pairs at $re^{j\theta}$ and $e^{j\theta}/r$. Hence the coefficient of w^{2n} and w^{o} is 1 [L-21, 22, 23].

8.6.2 The Factorization Algorithm

The technique used for polynomial factorization was chosen primarily because i) the algorithm was reasonably efficient and ii) a proven subroutine using the technique was available. The subroutine--NEWRA (now listed in [X-6])--combines the Newton-Raphson technique with polynomial deflation, implicit removal of roots as they are located.

Given a polynomial

$$f(w) = a_{2n}w^{2n}+a_{2n-1}w^{2n-1}+....+a_1w+a_0 \tag{8-96a}$$

$$= K\prod_{i=1}^{2n} (w-w_i) \ , \tag{8-96b}$$

then a root of f(w) may be found by making an estimate of the root, w_k, and using the Newton-Raphson technique [R-2, p. 332] to yield a better estimate, w_{k+1}, of the true root.

i.e., $$w_{k+1} = w_k - f(w_k)/f'(w_k). \tag{8-97}$$

If the iterated estimate diverges--$|w_{k+1}-w_k|$ grows larger--then the polynomial

$$g(w) = w^{2n}\cdot f(1/w) = a_0w^{2n}+a_1w^{2n-1}+...a_{2n-1}w+a_{2n} \tag{8-98}$$

whose roots are the reciprocals of those of f(w) is considered. Iteration of either f(w) or g(w) will therefore, effectively, yield a root of f(w).

Note that

$$f'(w) = 2n\cdot a_{2n}w^{2n-1}+(2n-1)\cdot a_{2n-1}w^{2n-2}+....+a_1 \tag{8-99a}$$

$$= f(w) \cdot \sum_{k=1}^{2n} (w-w_k)^{-1} \ . \tag{8-99b}$$

When m of the 2n roots [$m \geq 1$] have been found, the polynomial of which we desire a root is

$$h(w) = f(w) / \prod_{i=1}^{m} (w-w_i) \quad . \tag{8-100}$$

But $$h'(w)/h(w) = f'(w)/f(w) - \sum_{i=1}^{m} (w-w_i)^{-1}. \tag{8-101}$$

That is, (8-97) may be replaced by

$$w_{k+1} = w_k - [f'(w_k)/f(w_k) - \sum_{i=1}^{m} (w_k - w_i)^{-1}]^{-1} \tag{8-102}$$

during the iteration sequence whose purpose is to find the $m+1^{st}$ root of f(w). For m=0, (8-102) reduces to (8-97).

8.6.3 Accuracy Tests

Subroutine NEWRA was tested by factorizing polynomials representing square waves of various degrees. For

$$s(t) = \frac{4\Omega}{\pi} \sum_{\substack{k=1 \\ (k \text{ odd})}}^{L} \sin k\Omega t/k \quad , \tag{8-103}$$

$$f(w) = K \, [-jw^{2L}/L - jw^{2L-2}/(L-2) - \ldots\ldots$$
$$+ \ldots\ldots + j\, W^2/(L-2) + j/L]. \tag{8-104}$$

For example, when $L = 7$ and $\Omega = 1$:

$$f(w) = K \, [-jw^{14}/7 - jw^{12}/5 - jw^{10}/3 - jw^8$$
$$+ jw^6 + jw^4/3 + jw^2/5 + j/7]. \tag{8-105}$$

The theorem derived in sec. 8.5.2 implies that--due to the ripple associated with a bandlimited square wave--the polynomial is well-conditioned; that is, the zeros are "uniformly" distributed in angle about the origin in the w plane because they are associated with ripple in the z plane.

The roots were located iteratively and are shown in Fig. 8.2e and Fig. 8.3e for L = 15 (30 degree polynomial) and L = 31 (62 degree polynomial), respectively. The transformation

$$w = e^{j\Omega z} = e^{j\Omega t} \cdot e^{-\Omega\sigma} \qquad (8\text{-}106)$$

has been used to map the roots of f(w) from the w plane to the complex time [z] domain. The accuracy of the factorization was checked by substituting the derived roots into the original equation. In all cases the result (theoretically zero) was less than 10^{-3}. In addition, the original waveforms (s(t), $s_{RZ}(t)$ and $s_{CZ}(t)$) were *synthesized* using the derived roots in the product formulation. In fact, all waveforms in Figs. 8.2 and 8.3 were *synthesized* by expanding $s_{RZ}(t)$ and $s_{CZ}(t)$ in terms of the derived real zeros, eq. (8-14), and complex zero pairs, eq. (8-15), respectively, and then forming the product of $s_{RZ}(t)$ and $s_{CZ}(t)$ to yield s(t). Multiplication of all s(t) values by $2/\pi|c_n|$ --$2/15\pi$ for L = 15, $2/31\pi$ for L = 31 -- results in the expected rms value of unity for s(t).

Despite the accuracy of the factorization subroutine, tests on actual speech sounds revealed that reduction of the degree of the polynomial to be factorized was in the best interests of improved accuracy. For this reason, a method of hybrid factorization was developed and used for complex zero location of speech signals (sec. 9.4.1).

8.6.4 Complex Zero Configurations: Some Experimental Observations

In order to provide some familiarity with complex zero concepts and configurations, we have factorized the polynomial representing

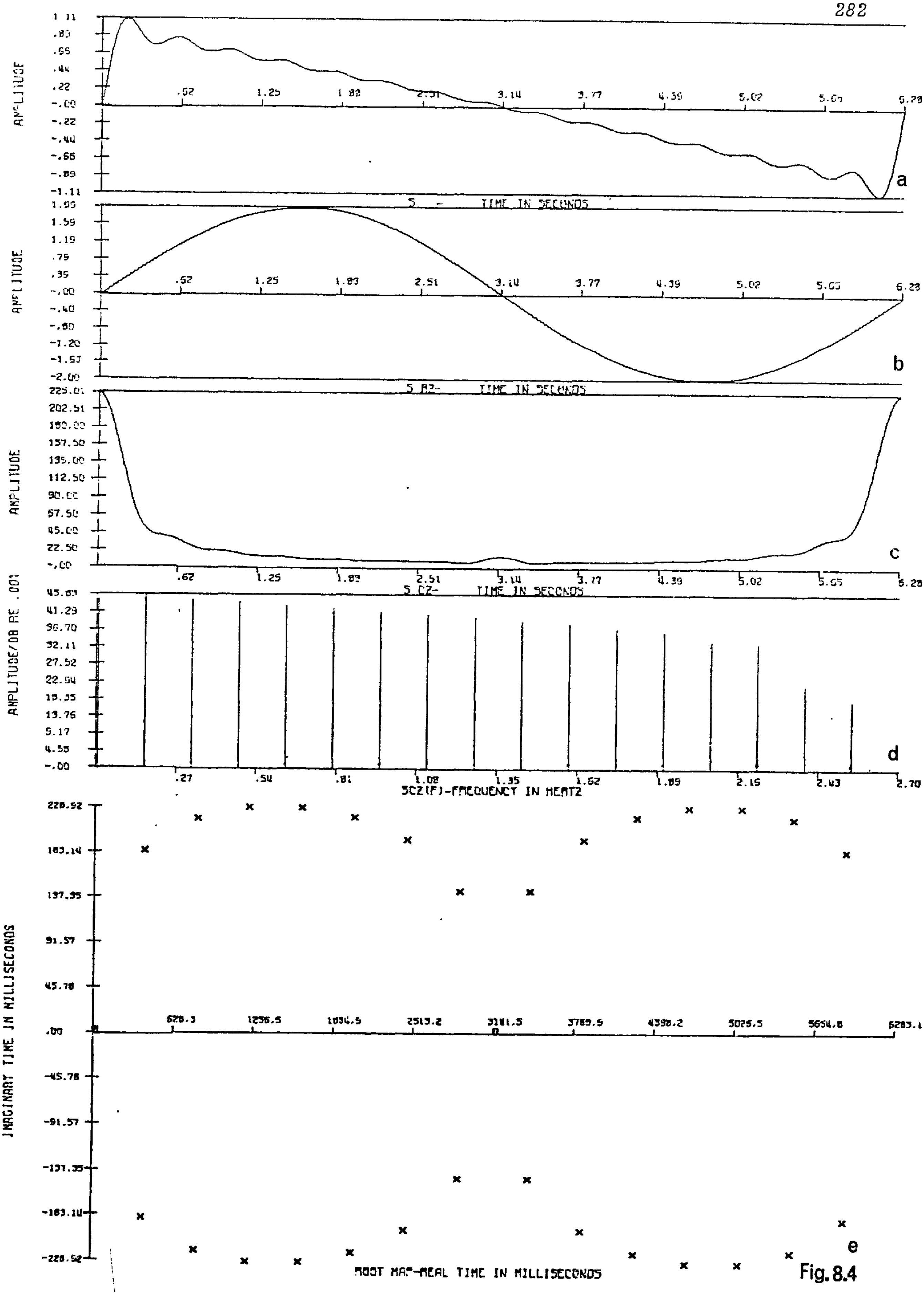
$$s_1(t) = \frac{2}{\alpha(\pi-\alpha)} \sum_{k=1}^{15} \frac{\sin(k\alpha)\cdot\sin(kt)}{k^2}, \qquad (8\text{-}107)$$

for various values of α. This Fourier series represents a "triangular" wave of period $T = 2\pi$ seconds with peaks occuring at $t = \alpha$, $\alpha+T/2$. (Fig. 8.18a.) For $\alpha = 0$, the "triangular" wave becomes a "sawtooth" while when $\alpha = \pi/2$, the "triangular" wave becomes symmetrical about $t = 0, \pm\pi/2, \pm\pi, \cdots$. Figures 8.4a, 8.5a, 8.6a, 8.7a, 8.8a, 8.9a, 8.10a, and 8.11a show $s_1(t)$ for $\alpha = 0, \pi/14, \pi/7, \ldots \pi/2$. The "b" and "c" diagrams of the figures show the respective RZ signals, $s_{RZ}(t)$, (all equal to sin t) and the CZ signals, $s_{CZ}(t)$. $s_{RZ}(t)$, $s_{CZ}(t)$, and $s_1(t)$ are all *synthesized* using the RZ's and CZ's depicted in Figs. 8.4e–8.11e, respectively. The "d" diagrams of each figure show the logarithm of the amplitude spectrum of $s_{CZ}(t)$ re 0.001.

In Figs. 8.11a–8.17a, the signal

$$s_2(t) = \frac{4}{\pi\cdot\alpha} \sum_{k=1}^{15,\ k\text{ odd}} \frac{\sin(k\alpha)\cdot\sin(kt)}{k^2} , \qquad (8\text{-}108)$$

a progressively clipped symmetrical triangular wave of period $T = 2\pi$ seconds. (Fig. 8.18b.) When $\alpha = \pi/2$, $s_1(t) = s_2(t)$; as $\alpha\to0$, $s_2(t)$ becomes progressively clipped. As before, the "b" and "c" diagrams of the figures show the respective *synthesized* RZ and CZ signals while the "d" and "e" diagrams show the log $|S_{CZ}(f)|$



AMPLITUDE
a
5 - TIME IN SECONDS
AMPLITUDE
b
5 HZ- TIME IN SECONDS
AMPLITUDE
c
5 DZ- TIME IN SECONDS
AMPLITUDE/DB RE .001
d
5CZ(F)-FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
e
ROOT MAP-REAL TIME IN MILLISECONDS

Fig. 8.4

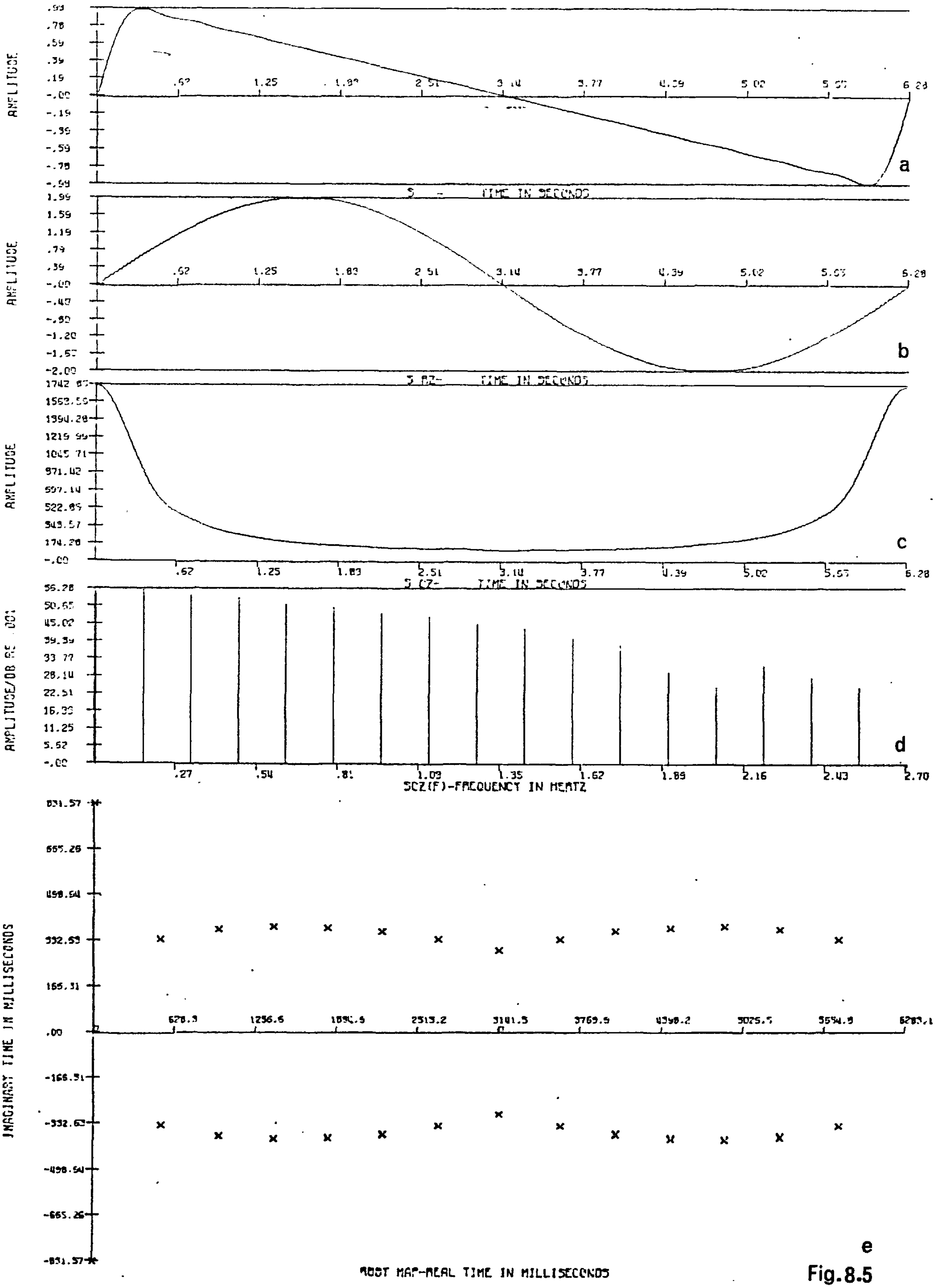

AMPLITUDE
TIME IN SECONDS
a
b
c
d
e
AMPLITUDE/DB RE .001
SCZ(F)-FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS

Fig.8.5

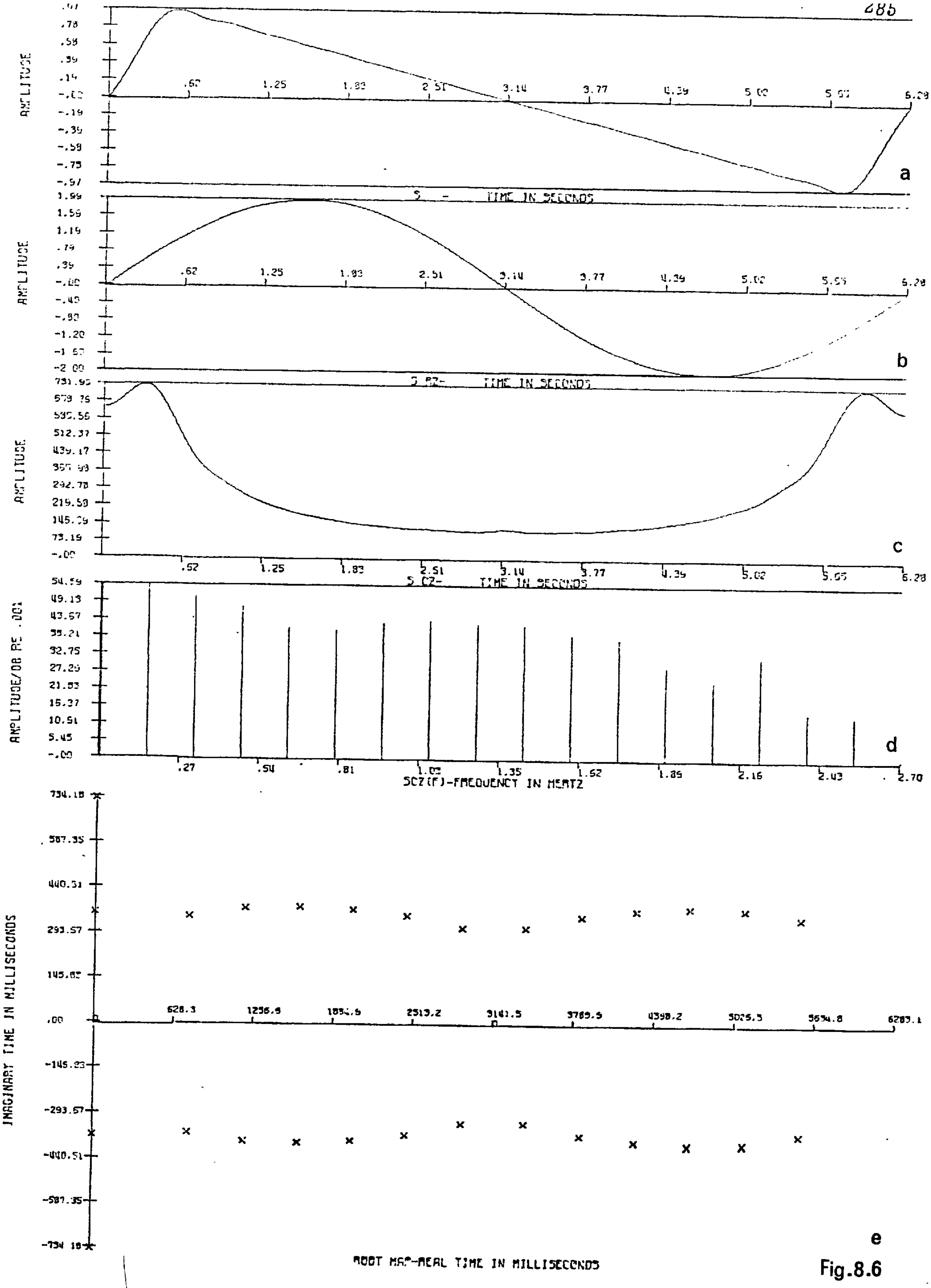
SAWTOOTH

$$S(T)=(2/(ALPHA(PI-ALPHA)))\sum_{N=1}^{15} SIN(N.ALPHA).SIN(N.T)/N**2$$

COMPLEX ZEROS// TIME IN MILLISECONDS

ALPHA=0				ALPHA=PI/14			
390.9281	+/-	J	185.5483	0.0	+/-	J	831.5760
805.3395	P/-	J	217.6496	520.4454	+/-	J	338.2498
1216.1935	+/-	J	228.9254	970.1765	+/-	J	374.8076
1626.1606	+/-	J	228.9232	1397.8863	+/-	J	384.0616
2036.4177	+/-	J	219.1010	1819.4108	+/-	J	381.0325
2448.5605	+/-	J	196.5045	2240.4705	+/-	J	367.2570
2869.3292	+/-	J	144.7215	2668.1998	+/-	J	338.8469
3413.8508	+/-	J	144.7215	3141.5873	+/-	J	296.7100
3834.6194	+/-	J	196.5044	3614.9802	+/-	J	338.8469
4246.7623	+/-	J	219.1010	4042.7095	+/-	J	367.2570
4657.0193	+/-	J	228.9232	4463.7691	+/-	J	381.0325
5066.9864	+/-	J	228.9254	4885.2937	+/-	J	384.0616
5477.8404	+/-	J	217.6496	5313.0035	+/-	J	374.8076
5892.2519	+/-	J	185.5483	5762.7346	+/-	J	338.2497

Ref. FIG. 8.4 Ref. FIG. 8.5



AMPLITUDE
TIME IN SECONDS
a
b
c
AMPLITUDE/DB RE .001
SCZ(F)-FREQUENCY IN HERTZ
d
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
e

Fig.8.6

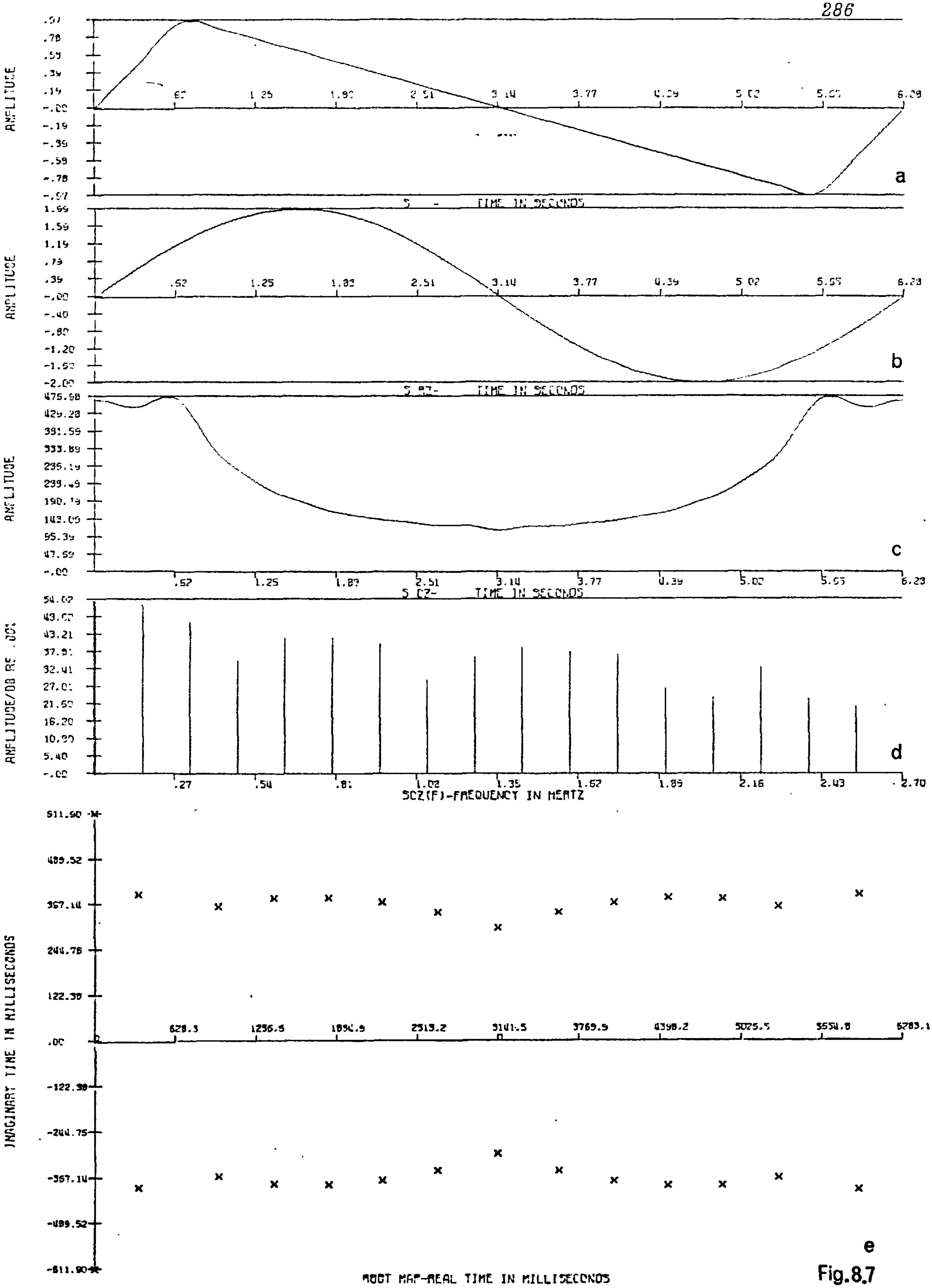

AMPLITUDE
TIME IN SECONDS
AMPLITUDE
TIME IN SECONDS
AMPLITUDE
TIME IN SECONDS
AMPLITUDE/DB RE .001
FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
a
b
c
d
e

Fig.8.7

$$S(T)=(2/(ALPHA(PI-ALPHA)))\sum_{N=1}^{15} SIN(N.ALPHA).SIN(N.T)/N**2$$

ALPHA=PI/7			ALPHA=3.PI/14		
COMPLEX ZEROS// TIME IN MILLISECONDS					
.0	+/- J	734.1893	0.0	+/- J	611.9049
.0	+/- J	364.3000	342.4563	+/- J	394.1668
744.3586	+/- J	353.5619	964.4964	+/- J	361.3032
1185.8023	+/- J	379.6491	1400.4975	+/- J	382.5490
1610.1615	+/- J	384.4113	1822.6156	+/- J	383.8903
2030.7522	+/- J	377.1398	2243.1125	+/- J	372.1940
2453.4827	+/- J	357.4752	2670.1306	+/- J	345.0571
2892.9940	+/- J	318.6825	3141.5873	+/- J	304.8662
3390.1859	+/- J	318.6825	3613.0494	+/- J	345.0571
3829.6973	+/- J	357.4752	4040.0674	+/- J	372.1941
4252.4277	+/- J	377.1398	4460.5693	+/- J	383.8903
4673.0185	+/- J	384.4113	4882.6825	+/- J	382.5490
5097.3776	+/- J	379.6491	5318.6835	+/- J	361.3032
5538.8213	+/- J	353.5619	5940.7237	+/- J	394.1668
Ref. FIG. 8.6			Ref. FIG. 8.7		

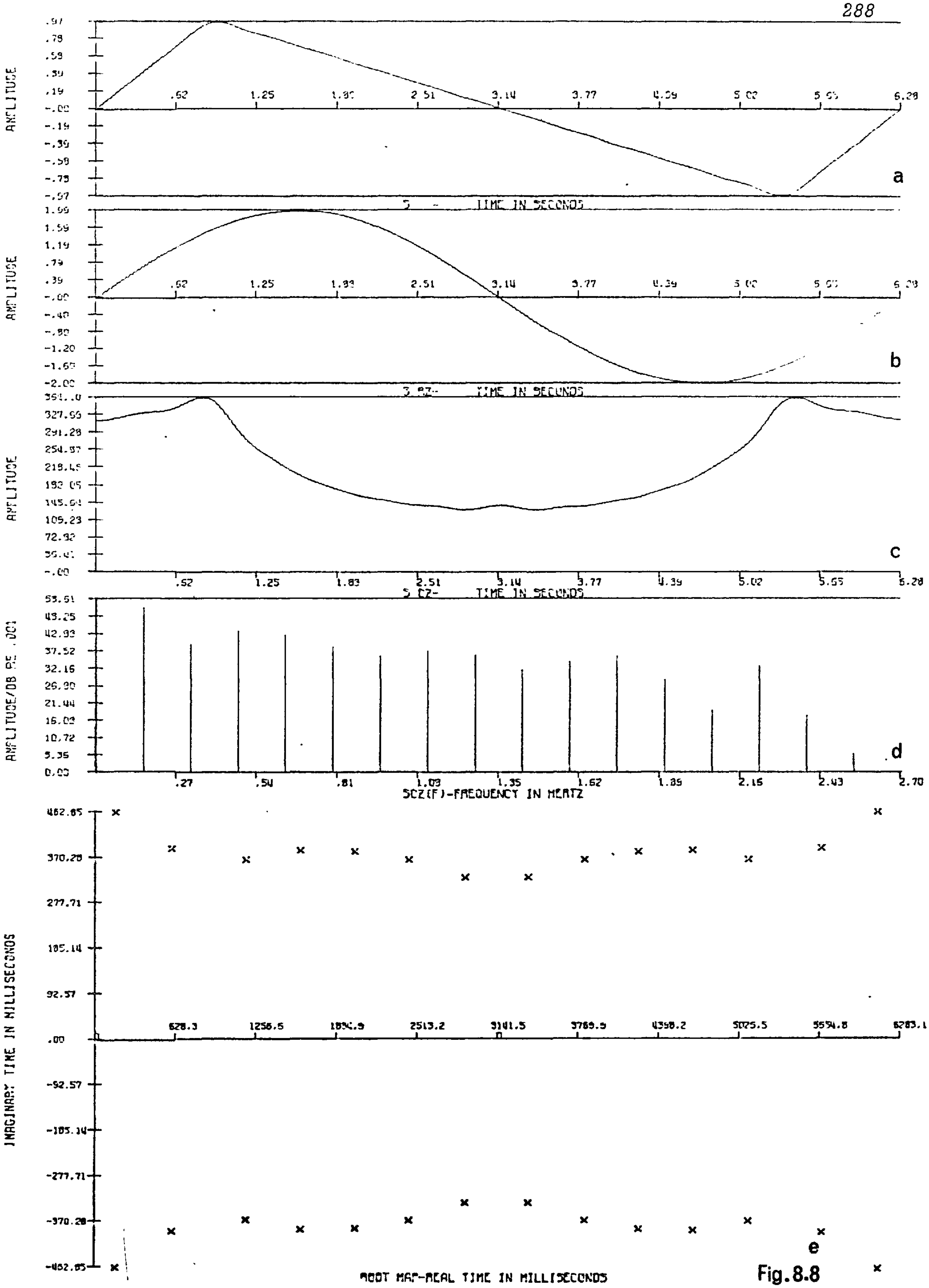

AMPLITUDE
TIME IN SECONDS
a
b
c
d
e
AMPLITUDE/DB RE .001
SCZ(F)-FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS

Fig.8.8

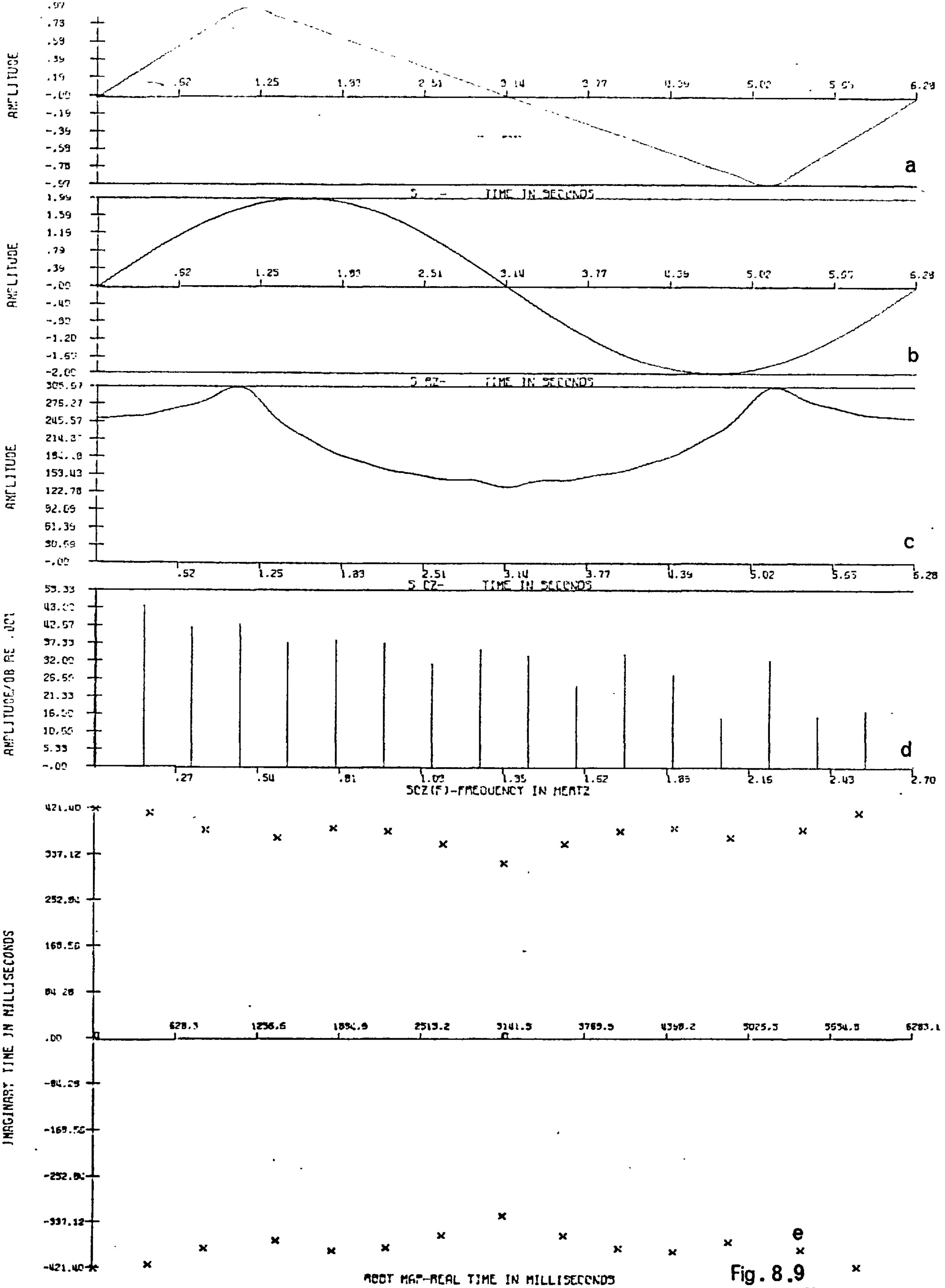

AMPLITUDE
TIME IN SECONDS
a
b
c
AMPLITUDE/DB RE .001
SCZ(F)-FREQUENCY IN HERTZ
d
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
e

Fig. 8.9

$$S(T)=(2/(ALPHA(PI-ALPHA)))\sum_{N=1}^{15} SIN(N \cdot ALPHA) \cdot SIN(N \cdot T)/N**2$$

ALPHA=2.PI/7 ALPHA=5.PI/14

COMPLEX ZEROS// TIME IN MILLISECONDS

169.3257	+/- J	462.8560	0.0	+/- J	421.4052
609.7062	+/- J	389.7043	419.5884	+/- J	414.3765
1183.1398	+/- J	366.1033	849.1005	+/- J	384.4769
1615.1604	+/- J	384.7700	1401.4751	+/- J	369.7138
2035.7023	+/- J	382.9798	1830.4115	+/- J	386.9803
2457.2795	+/- J	366.0712	2249.9749	+/- J	381.9554
2895.4966	+/- J	329.5891	2675.0214	+/- J	358.5050
3387.6833	+/- J	329.5891	3141.5873	+/- J	322.7324
3825.9004	+/- J	366.0711	3608.1586	+/- J	358.5050
4247.4777	+/- J	382.9798	4033.2050	+/- J	381.9554
4668.0196	+/- J	384.7700	4452.7684	+/- J	386.9803
5100.0401	+/- J	366.1033	4881.7048	+/- J	369.7138
5673.4738	+/- J	389.7043	5434.0795	+/- J	384.4769
6113.8542	+/- J	462.8560	5863.5916	+/- J	414.3766

Ref. FIG. 8.8 Ref. FIG. 8.9

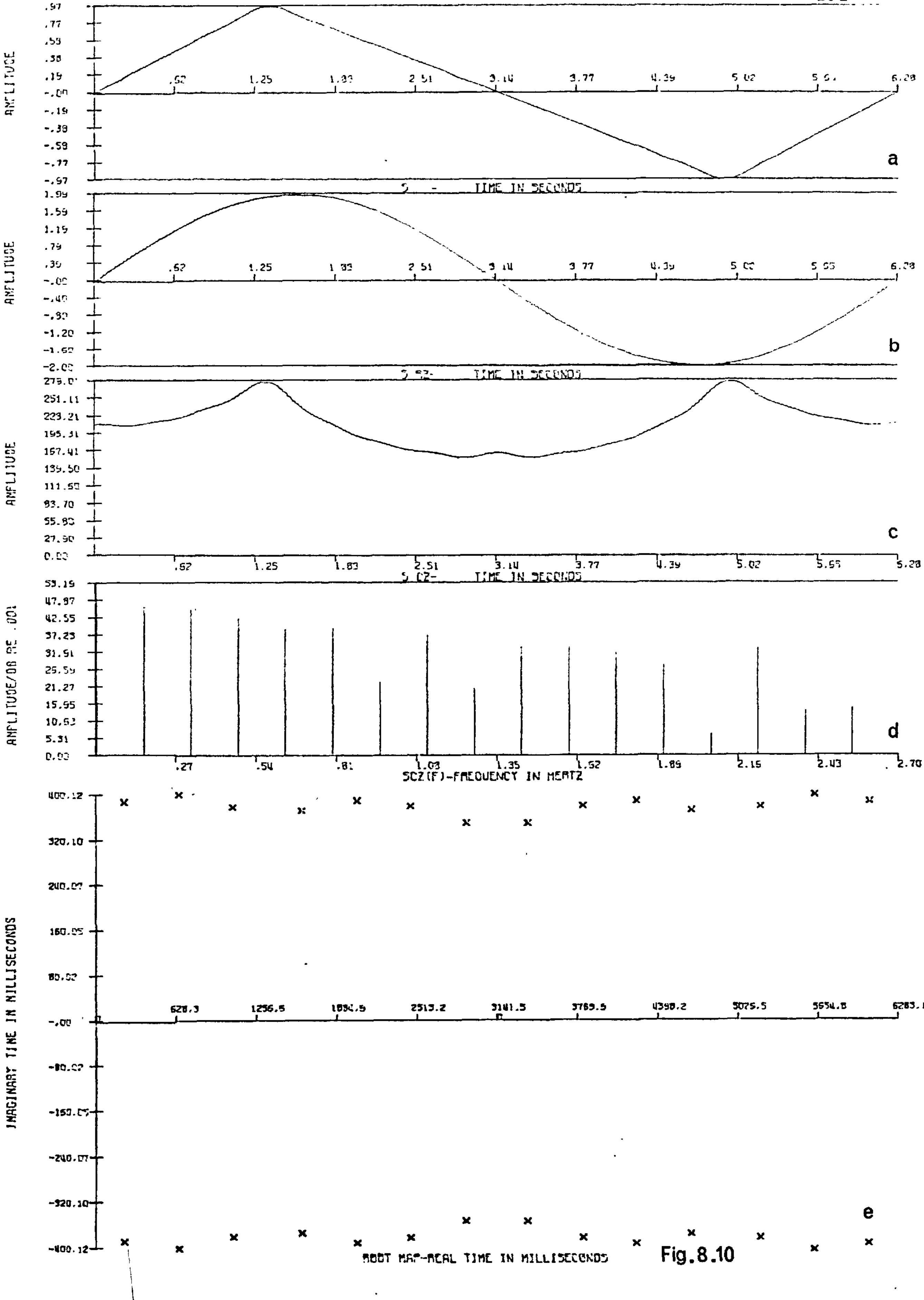

AMPLITUDE
TIME IN SECONDS
a
b
c
AMPLITUDE/DB RE .001
SCZ(F)-FREQUENCY IN HERTZ
d
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
e

Fig.8.10

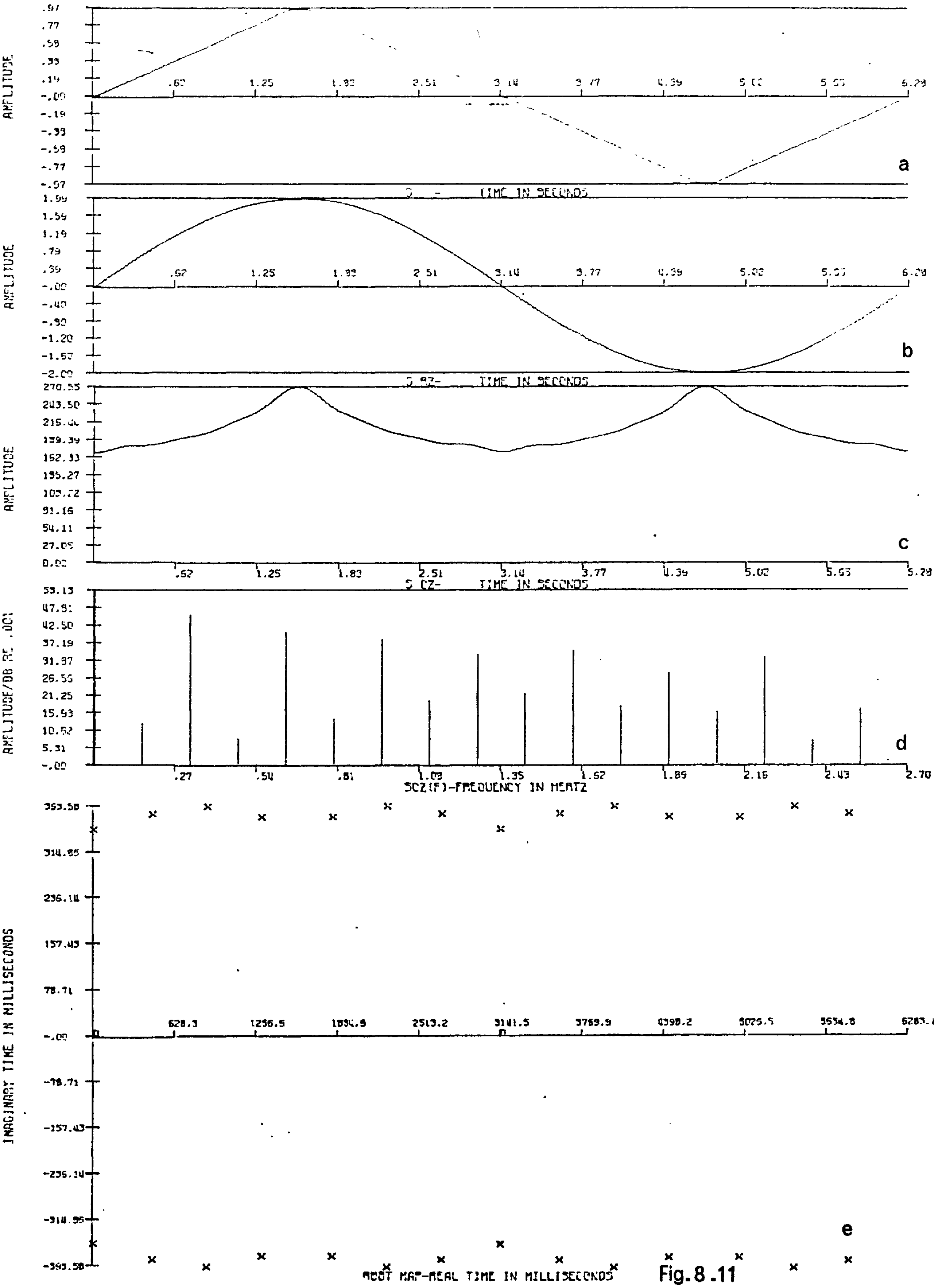

a
b
c
d
e
AMPLITUDE
TIME IN SECONDS
AMPLITUDE/DB RE .001
SCZ(F)-FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS

Fig. 8 .11

$$S(T)=(2/(ALPHA(PI-ALPHA)))\sum_{N=1}^{15} SIN(N.ALPHA).SIN(N.T)/N**2$$

ALPHA=3.PI/7 ALPHA=PI/2

COMPLEX ZEROS// TIME IN MILLISECONDS

226.8819	+/- J	388.7121	0.0	+/- J	355.3194
652.3656	+/- J	400.1266	454.7640	+/- J	381.6530
1077.6081	+/- J	380.0545	875.7541	+/- J	393.5808
1620.3742	+/- J	372.9579	1300.8278	+/- J	376.2999
2046.9420	+/- J	389.6929	1840.7629	+/- J	376.2993
2466.3798	+/- J	381.1516	2265.8369	+/- J	393.5795
2900.2744	+/- J	350.4972	2686.8275	+/- J	381.6511
3382.2056	+/- J	350.4972	3141.5873	+/- J	355.3168
3816.8001	+/- J	381.1516	3596.3524	+/- J	381.6511
4236.2379	+/- J	389.6929	4017.3430	+/- J	393.5795
4662.8057	+/- J	372.9579	4442.4171	+/- J	376.2993
5205.5718	+/- J	380.0545	4982.3522	+/- J	376.2999
5630.8143	+/- J	400.1266	5407.4258	+/- J	393.5808
6056.2980	+/- J	388.7121	5828.4160	+/- J	381.6530

Ref. FIG. 8.10 Ref. FIG. 8.11

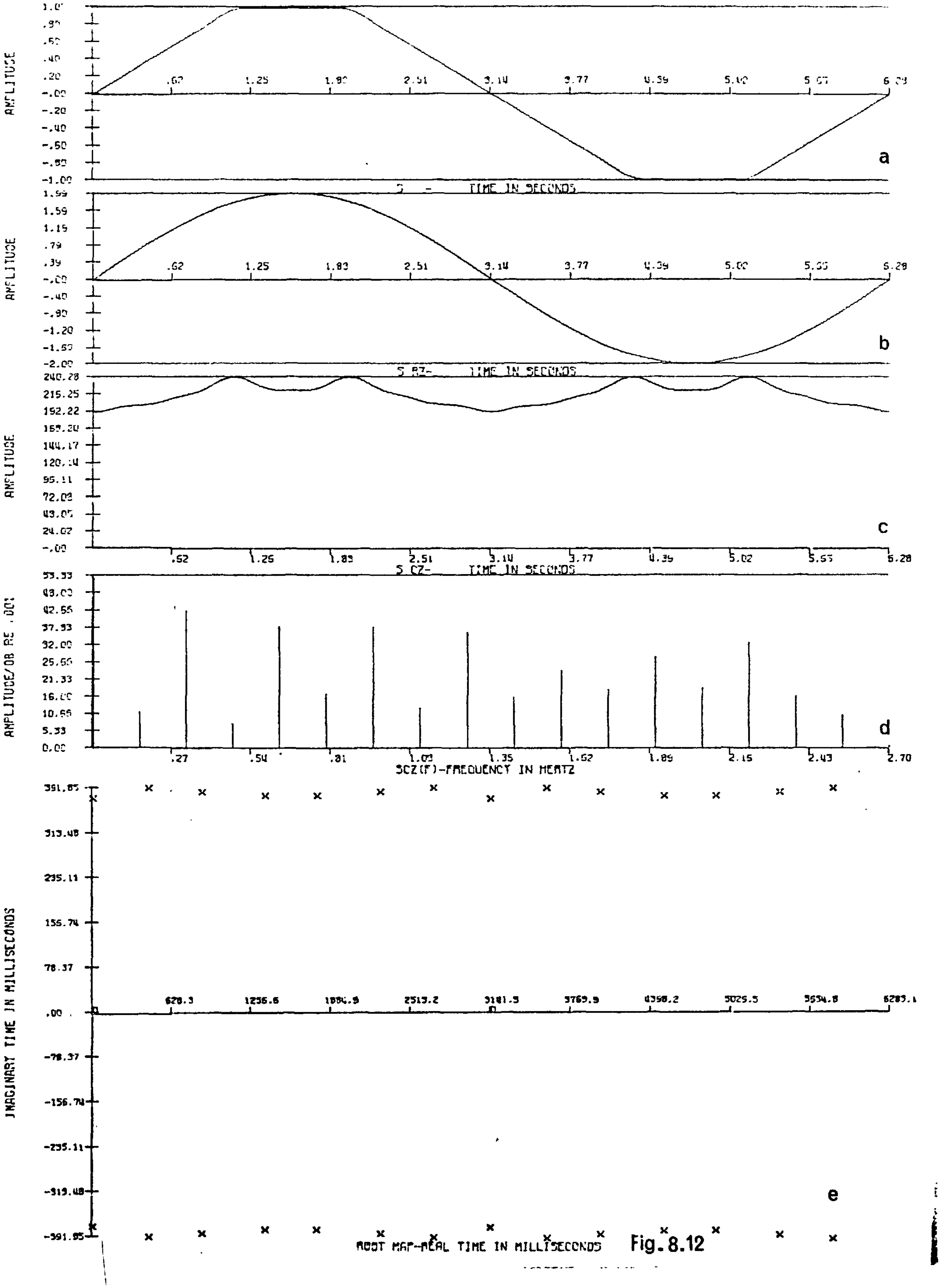

AMPLITUDE
TIME IN SECONDS
AMPLITUDE/DB RE .001
SCZ(F)-FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
a
b
c
d
e

Fig. 8.12

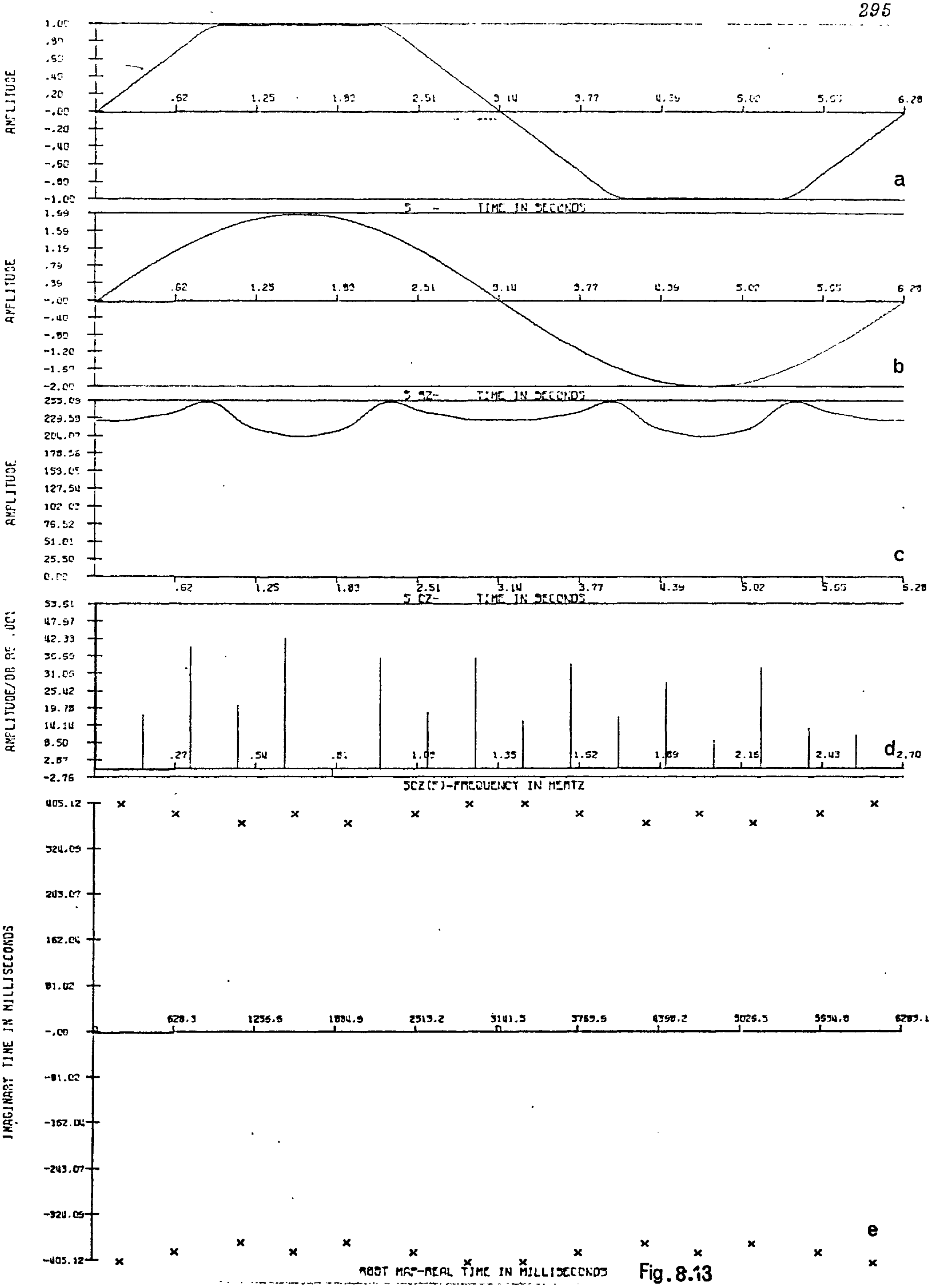

AMPLITUDE
TIME IN SECONDS
AMPLITUDE/DB RE .001
SCZ(F)-FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
a
b
c
d
e

Fig. 8.13

'CLIPPED' TRIANGULAR WAVE

$$S(T)=(4/(PI \cdot ALPHA)) \cdot \sum_{N=1}^{15,\ N\ ODD} SIN(N \cdot ALPHA) \cdot SIN(N \cdot T)/N^{**}2$$

ALPHA=2.PI/7			ALPHA=5.PI/14		
213.6435	+/- J	405.1230	0.0	+/- J	374.3107
636.6764	+/- J	387.5850	443.6772	+/- J	391.8559
1155.6155	+/- J	371.9856	865.1148	+/- J	385.2004
1570.7963	+/- J	387.9891	1364.2341	+/- J	378.5919
1985.9771	+/- J	371.9856	1777.3586	+/- J	378.5919
2504.9162	+/- J	387.5850	2276.4778	+/- J	385.2004
2927.9491	+/- J	405.1230	2697.9154	+/- J	391.8559
3355.2309	+/- J	405.1230	3141.5873	+/- J	374.3107
3778.2638	+/- J	387.5850	3585.2645	+/- J	391.8559
4297.2029	+/- J	371.9856	4006.7021	+/- J	385.2009
4712.3836	+/- J	387.9891	4505.8214	+/- J	378.5919
5127.5644	+/- J	371.9856	4918.9459	+/- J	378.5919
5646.5035	+/- J	387.5850	5418.0652	+/- J	385.2004
6069.5364	+/- J	405.1230	5839.5028	+/- J	391.8559

Ref. FIG. 8.13 Ref. FIG. 8.12

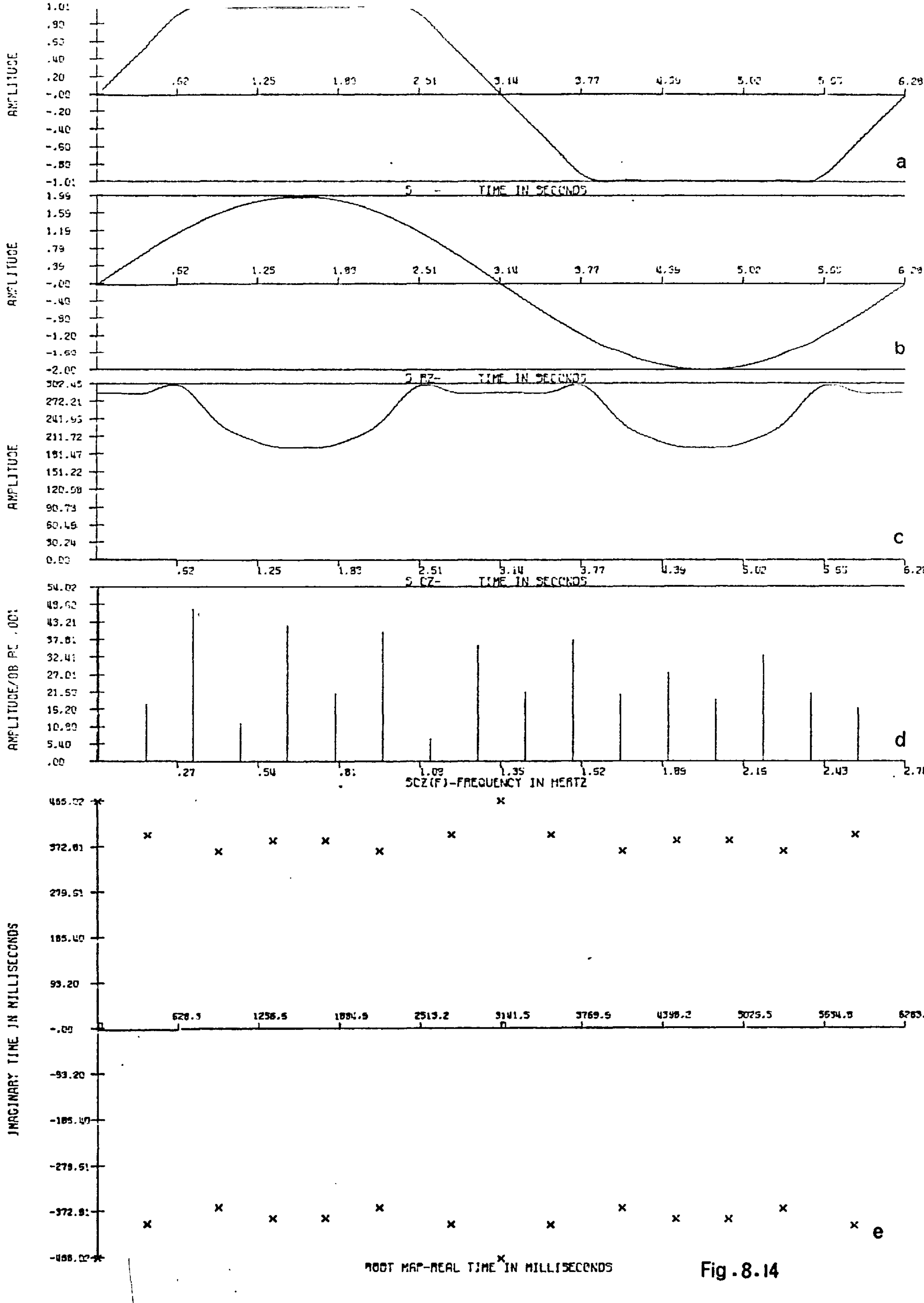


Fig.8.14

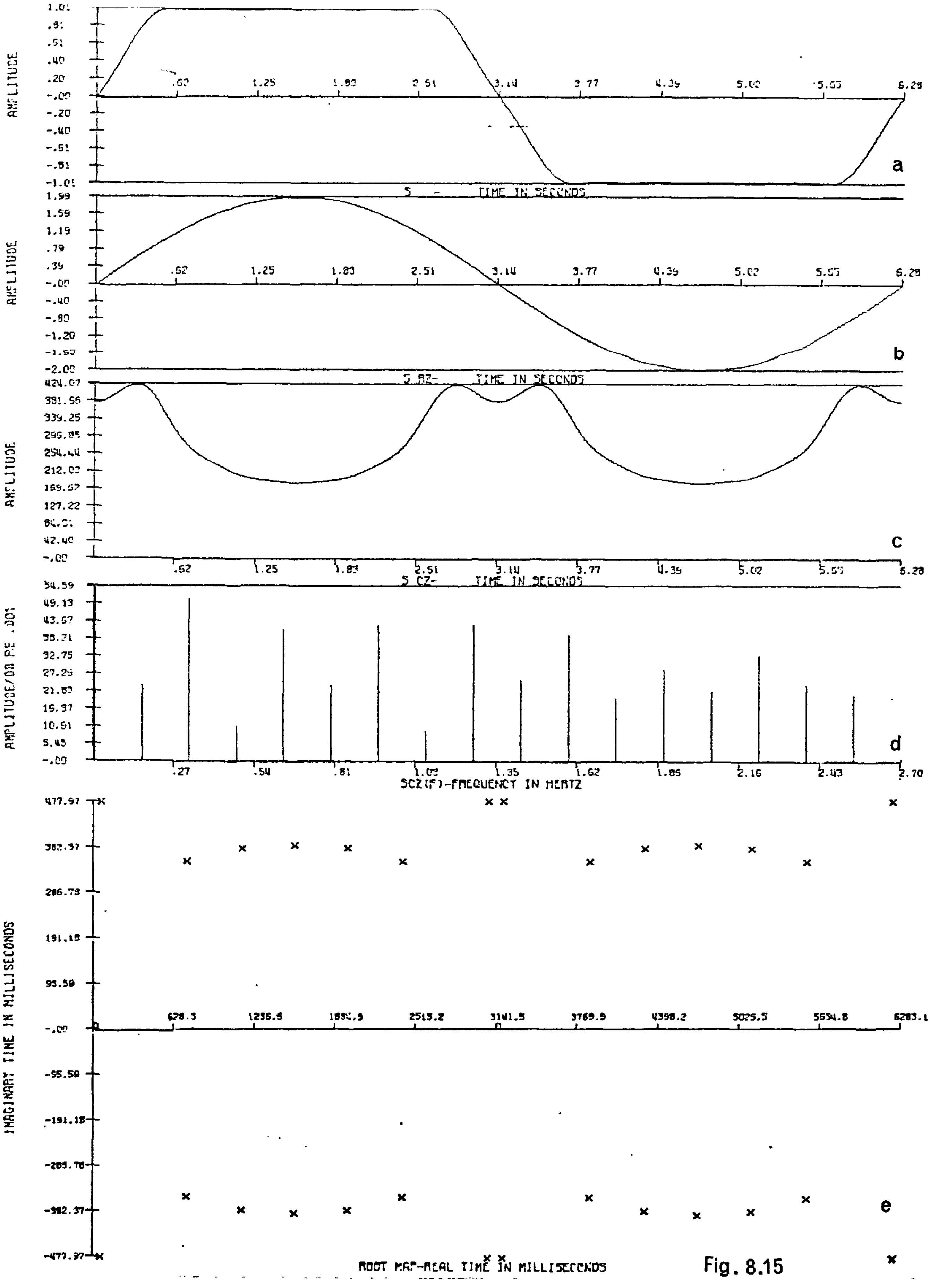

a
b
c
d
e
AMPLITUDE
TIME IN SECONDS
AMPLITUDE/DB RE .001
SCZ(F)-FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS

Fig. 8.15

$$S(T)=(4/(PI\cdot ALPHA))\cdot\sum_{N=1}^{15,\ N\ ODD} SIN(N\cdot ALPHA)\cdot SIN(N\cdot T)/N^{**}2$$

COMPLEX ZEROS// TIME IN MILLISECONDS

ALPHA=PI/7		ALPHA=3.PI/14	
57.2989 +/- J	477.9712	0.0 +/- J	466.0206
730.0919 +/- J	351.9913	388.9414 +/- J	397.9154
1157.4008 +/- J	379.5836	944.1641 +/- J	363.0827
1570.7963 +/- J	386.6301	1365.1062 +/- J	385.3008
1984.1918 +/- J	379.5836	1776.4864 +/- J	385.3008
2411.5007 +/- J	351.9913	2197.4285 +/- J	363.0827
3084.2937 +/- J	477.9712	2752.6512 +/- J	397.9154
3198.8862 +/- J	477.9712	3141.5873 +/- J	466.0206
3871.6793 +/- J	351.9913	3530.5287 +/- J	397.9154
4298.9881 +/- J	379.5836	4085.7515 +/- J	363.0827
4712.3836 +/- J	386.6301	4506.6935 +/- J	385.3008
5125.7791 +/- J	379.5836	4918.0737 +/- J	385.3008
5553.0880 +/- J	351.9913	5339.0158 +/- J	363.0827
6225.8810 +/- J	477.9712	5894.2386 +/- J	397.9154

Ref. FIG. 8.15   Ref. FIG. 8.14

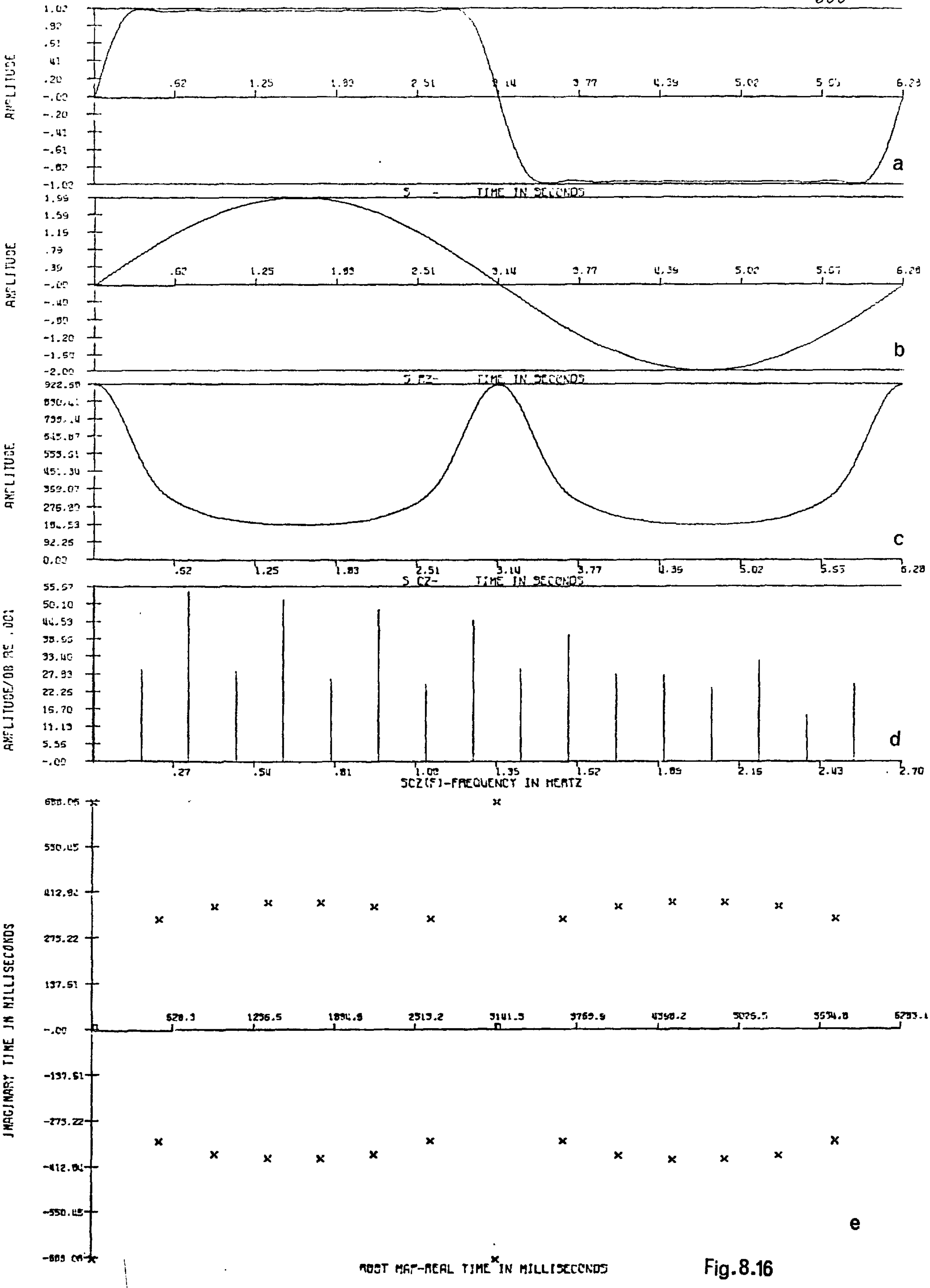

AMPLITUDE
TIME IN SECONDS
a
AMPLITUDE
TIME IN SECONDS
b
AMPLITUDE
TIME IN SECONDS
c
FREQUENCY IN HERTZ
d
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
e

Fig.8.16

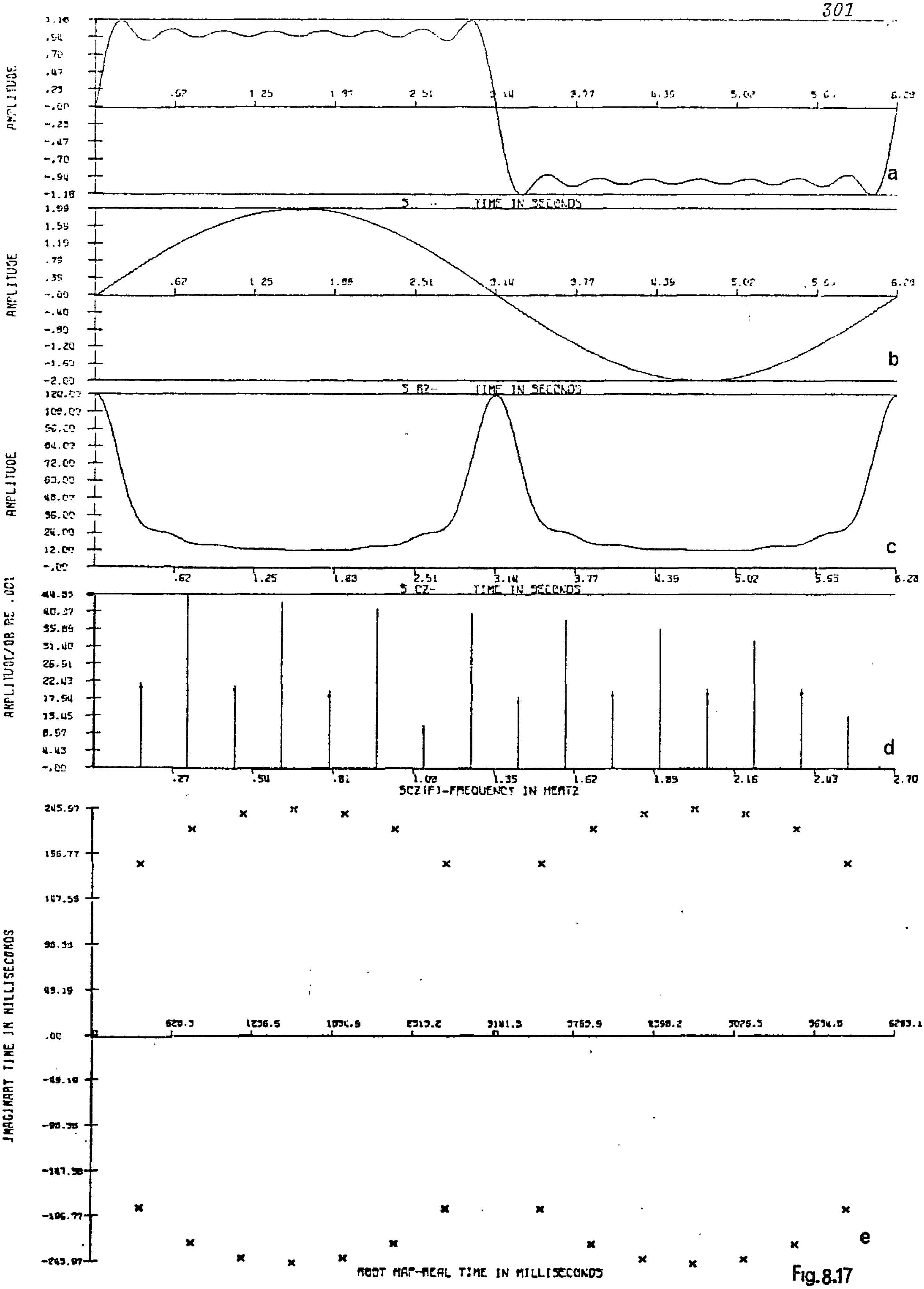

AMPLITUDE
TIME IN SECONDS
a
b
c
d
AMPLITUDE/DB RE .001
FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
e

Fig.8.17

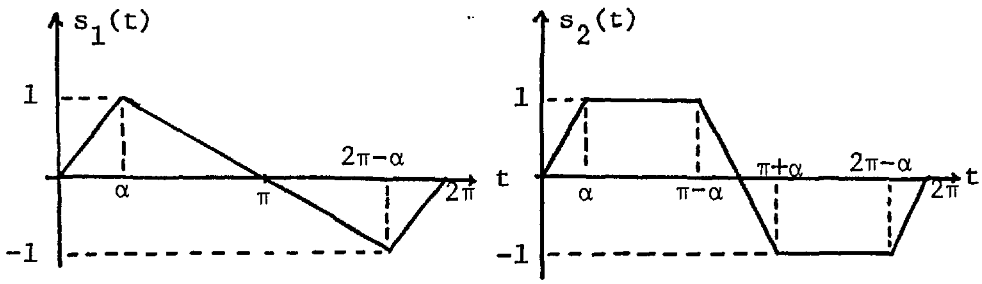
$$S(T)=(4/(PI\cdot ALPHA))\cdot\sum_{N=1}^{15,\ N\ ODD} SIN(N\cdot ALPHA)\cdot SIN(N\cdot T)/N^{**}2$$

COMPLEX ZEROS// TIME IN MILLISECONDS

ALPHA=0				ALPHA=PI/14			
374.0449	+/-	J	186.6256	0.0	+/-	J	688.0681
775.3279	+/-	J	224.4835	511.6899	+/-	J	335.1230
1173.4956	+/-	J	241.0565	947.8636	+/-	J	371.9515
1570.7963	+/-	J	245.9723	1364.3655	+/-	J	384.4134
1968.0970	+/-	J	241.0565	1777.2272	+/-	J	384.4134
2366.2647	+/-	J	224.4835	2193.7290	+/-	J	371.9515
2767.5477	+/-	J	186.6256	2629.9027	+/-	J	335.1230
3515.6322	+/-	J	186.6256	3141.5927	+/-	J	688.0681
3916.9152	+/-	J	224.4835	3653.7273	+/-	J	335.1230
4315.0829	+/-	J	241.0565	4089.4509	+/-	J	371.9515
4712.3836	+/-	J	245.9723	4505.9528	+/-	J	384.4134
5109.6843	+/-	J	241.0565	4918.8145	+/-	J	384.4134
5507.8521	+/-	J	224.4835	5335.3163	+/-	J	371.9515
5909.1351	+/-	J	186.6256	5771.4900	+/-	J	335.1230

Ref. FIG. 8.17 Ref. FIG. 8.16

re 0.001 and the RZ-CZ arrays.

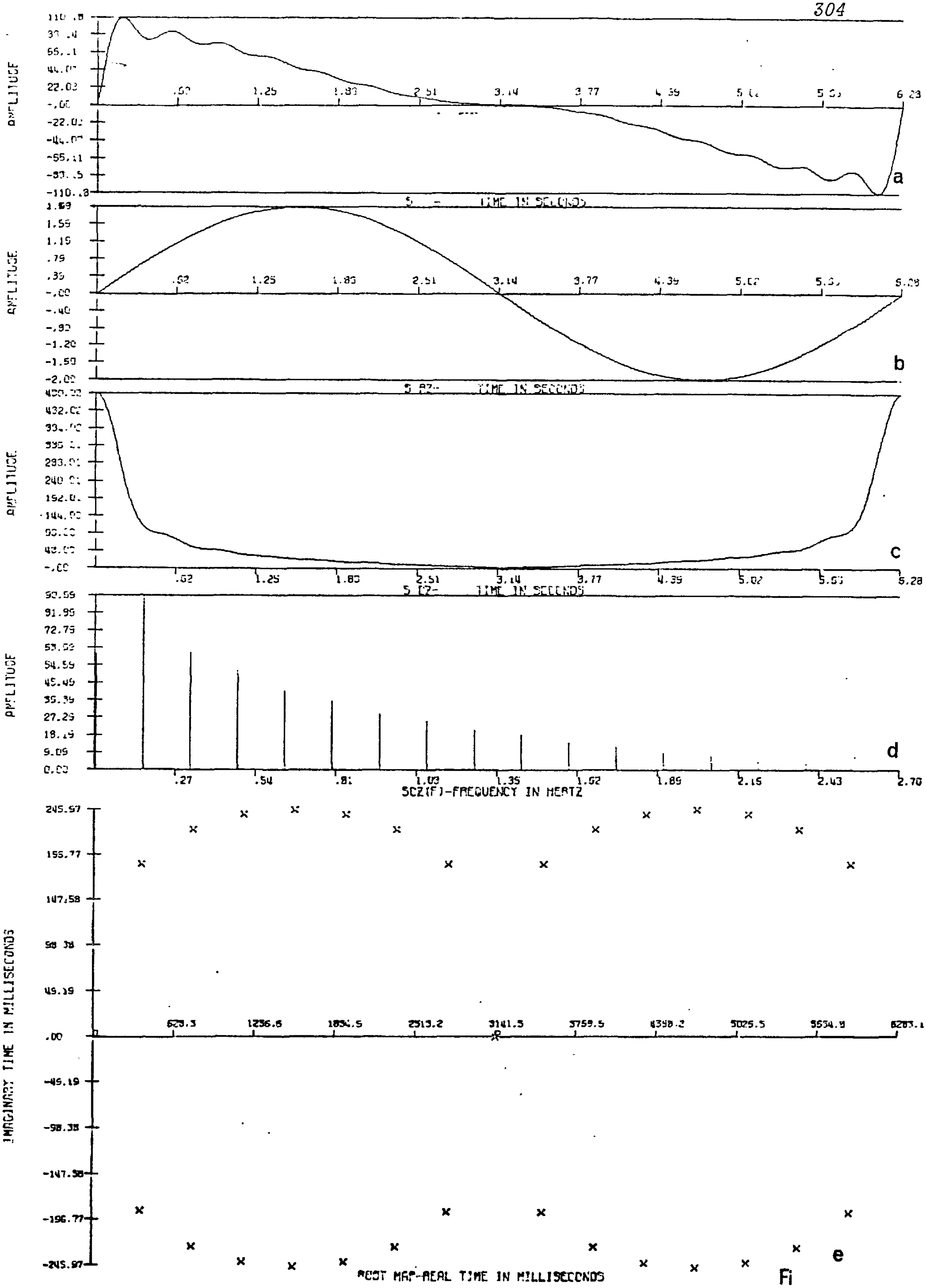


a) $s_1(t)$ b) $s_2(t)$

Fig. 8.18 Triangular and progressively clipped triangular waves.

Observe that, in Figs. 8.12e-8.16e the CZ configuration corresponding to the clipped portion of the original waveform assumes the "arced" configuration typical of a square wave (Fig. 8.17e). We note that the same symmetries observed in the waveforms are seen in the zero arrays. For example, for $\alpha = \pi/2$, (Fig. 8.11a), $s_1(t)$ is symmetrical about 0, $\pi/2$, π, $3\pi/2$, 2π . . while $s_2(t)$ is symmetrical about the same points for all α (Figs. 8.11a-8.17a). In these cases, the zero arrays are symmetrical about the same points.

Note the apparent regularity, in real time, of zeros generally. That is, a real zero or a complex zero pair occurs about once every $T/(n_R+n_C)$ seconds. In some cases the regularity is "forced" by ripple; the square wave and sawtooth, for example. Other waveform characteristics which lead us to expect real time regularity of zeros will be examined in sec. 9.4


AMPLITUDE
TIME IN SECONDS
a
b
c
d
e
SCZ(F)-FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS

Fi

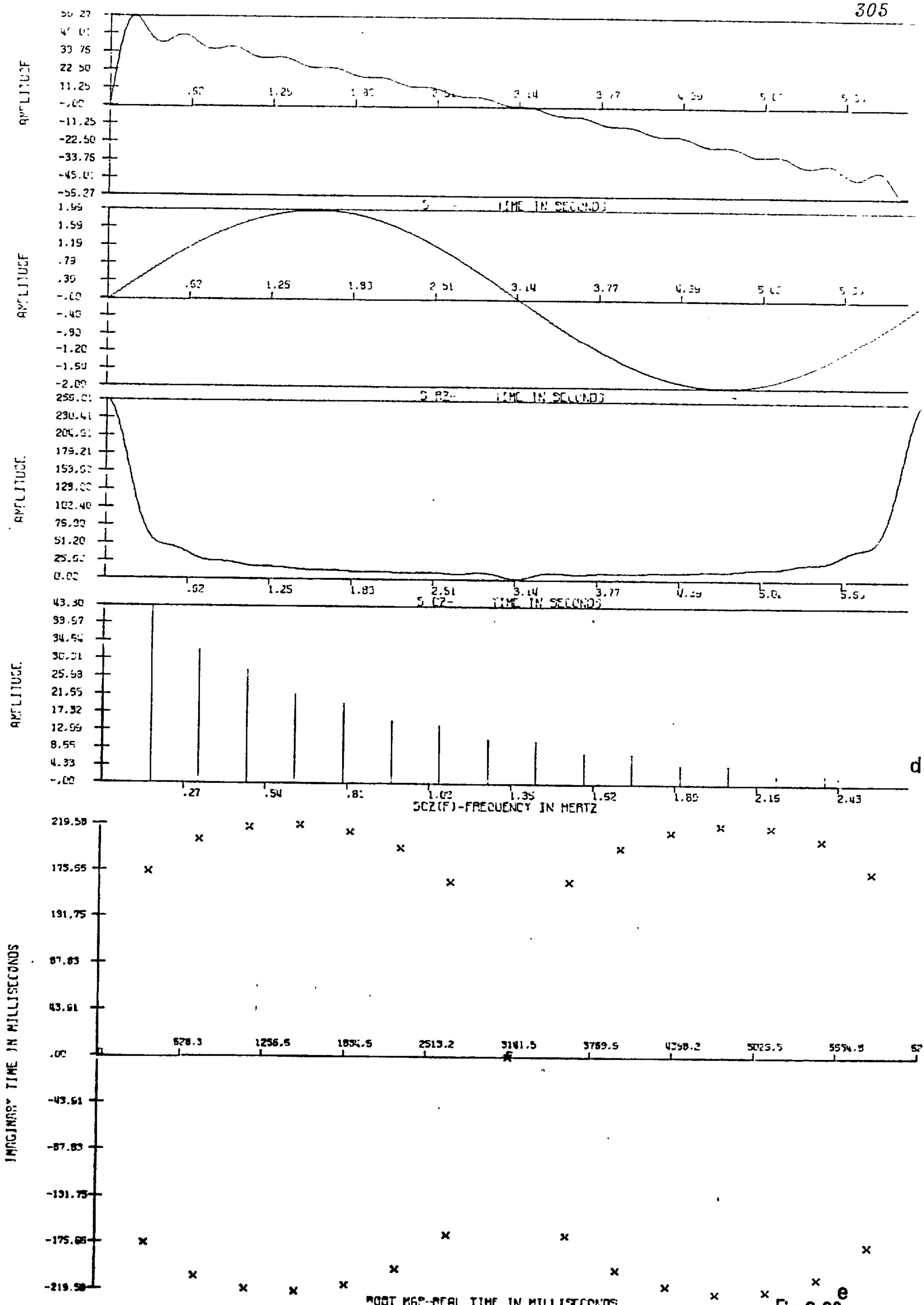
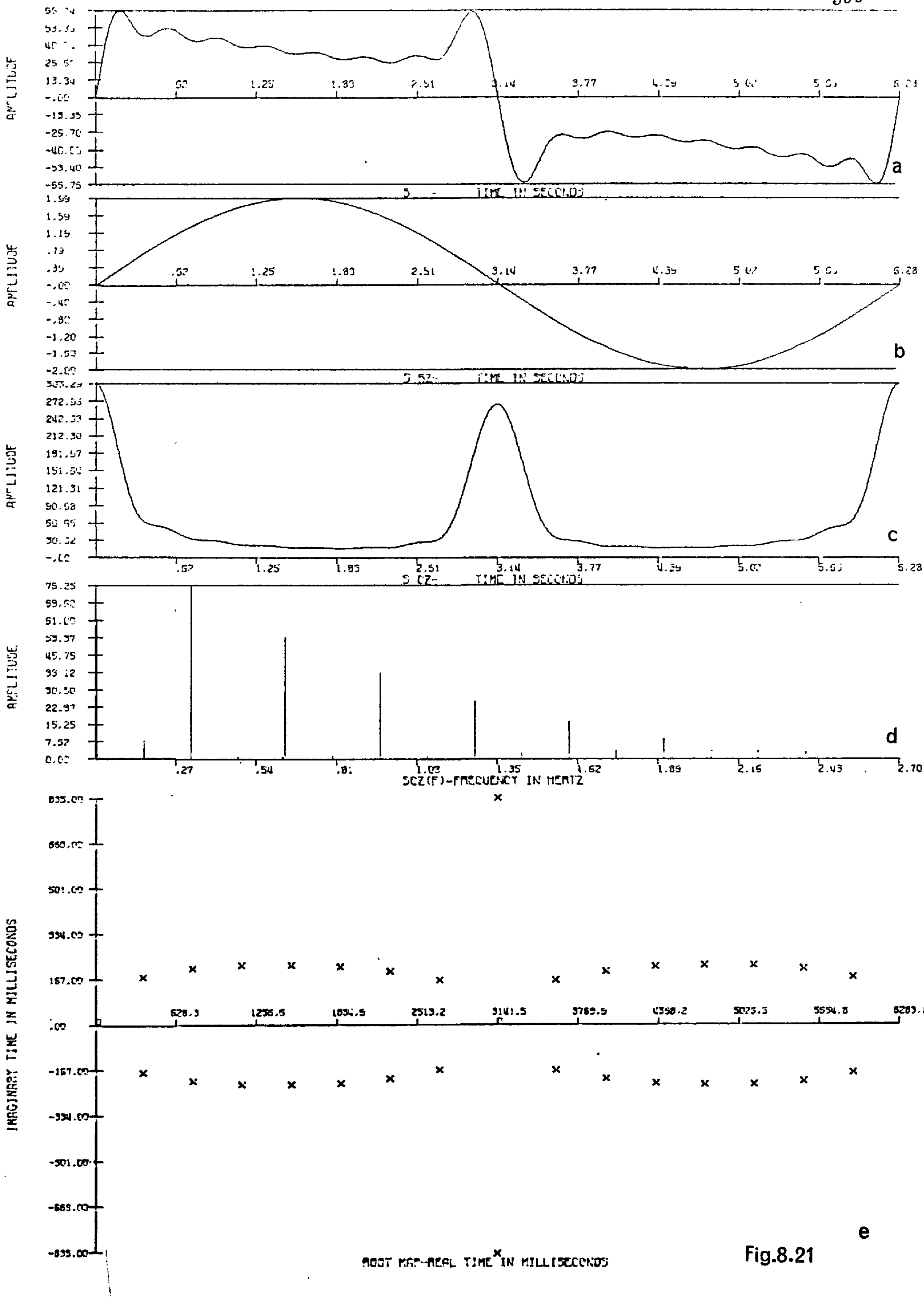

AMPLITUDE
TIME IN SECONDS
SCZ(F)-FREQUENCY IN HERTZ
d
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
e

Fig.8.20



a
b
c
d
e
AMPLITUDE
TIME IN SECONDS
SCZ(F)-FREQUENCY IN HERTZ
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS

Fig.8.21

and 9.5.

8.6.5 Complex Zero Manipulation

Figure 8.2a, the square wave, differs from Fig. 8.4a, the sawtooth, only in their respective complex zero waveforms. Fig. 8.19 demonstrates the effect of adding a CZ pair with coordinates $z = \pi \pm j\ 0.00001$ to the square wave CZ waveform. The even order harmonics appear in $|S_{CZ}(f)|$ (and hence $|S(f)|$) and the wave shape of $s(t)$ is forced to become roughly triangular. This follows because the CZ pair at $z = \pi \pm j\ 0.00001$ suppresses the central peak of $s_{CZ}(t)$.

Conversely, Fig. 8.20a-e shows $s(t)$, $s_{RZ}(t)$, $|S_{CZ}(f)|$ and the root map for a triangular wave similar to that shown in Fig. 8.4; the difference is that the upper limit in eq. (8-107) has been increased by 1, to 16. This has the effect of adding a CZ pair at $z = \pi \pm j\ 0.5303$ msec. In Fig. 8.21 we have increased the imaginary time coordinate of this CZ pair from $\pm\ 0.5303$ to $\pm\ 835$ milliseconds. This permits an excursion of $s_{CZ}(t)$ at $t = \pi$ seconds so that $s(t)$ very roughly approximates the square wave of Fig. 8.2a. Note specifically that the odd harmonics of $S_{CZ}(f)$ have been greatly suppressed.

8.7 The Complex Time Domain

The product representation for a bandlimited periodic signal,[1]

$$s(t) = \prod_{i=1}^{2n_R} 2\sin\frac{\Omega}{2}(t-\tau_i) \cdot \prod_{\ell=1}^{n_C} 2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)],$$

---

[1] We shall ignore the real, multiplicative constant.

specifies s(t) as a function of t and its real zeros (zero crossings) and complex zeros. If the behaviour on the complex time plane is to be investigated, we let $t \to z$, the complex time variable.

Then

$$s(z=t+j\sigma) = \prod_{i=1}^{2n_R} 2[\sin\frac{\Omega}{2}(t-\tau_i)\cdot\cosh\frac{\Omega}{2}\sigma - j\cos\frac{\Omega}{2}(t-\tau_i)\cdot\sinh\frac{\Omega}{2}\sigma]\cdot$$

$$\prod_{\ell=1}^{n_C} 2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)\cdot\cosh\Omega\sigma - j\sin\Omega(t-\tau_\ell)\cdot\sinh\Omega\sigma] \quad (8\text{-}109a)$$

$$= \prod_{i=1}^{2n_R} A_i(z).e^{ja_i(z)} \prod_{\ell=1}^{n_C} B_\ell(z).e^{jb_\ell(z)}. \quad (8\text{-}109b)$$

At a real zero, s(z) is zero because $\sin\frac{\Omega}{2}(t-\tau_i)$ and $\sinh\frac{\Omega}{2}\sigma$ are zero; at a complex zero, s(z) is zero because $\cos\frac{\Omega}{2}(t-\tau_\ell)\cdot\cosh\Omega\sigma$ is equal to $\cosh\Omega\sigma_\ell$, and $\sin\Omega(t-\tau_\ell)$ is zero.

The phase function is the sum of the contributions to the phase from all the RZ's $[a_i]$ and the CZ's $[b_\ell]$;

i.e., $$\Psi(z) = \Psi_{RZ}(z) + \Psi_{CZ}(z) \quad (8\text{-}110a)$$

$$= \sum_{i=1}^{2n_R} \tan^{-1}[\frac{-\cos\frac{\Omega}{2}(t-\tau_i)\cdot\sinh\frac{\Omega}{2}}{\sin\frac{\Omega}{2}(t-\tau_i)\cdot\cosh\frac{\Omega}{2}\sigma}]$$

$$+ \sum_{\ell=1}^{n_C} \tan^{-1}[\frac{-\sin\Omega(t-\tau_\ell)\cdot\sinh\Omega\sigma}{\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)\cdot\cosh\Omega\sigma}]\cdot \quad (8\text{-}110b)$$

If $\cosh\Omega\sigma >> \cosh\Omega\sigma_\ell$, or equivalently, $\mathrm{Im}\,[z] >> \max\{\sigma_1, \sigma_2, .. \sigma_{n_C}\}$, then (8-110b) reduces to

$$\Psi(z) \simeq \sum_{i=1}^{2n_R} \tan^{-1}[-\cot\tfrac{\Omega}{2}(t-\tau_i)] + \sum_{\ell=1}^{n_C} \tan^{-1}[\tan\Omega(t-\tau_\ell)], \tag{8-111a}$$

which, after some manipulation, becomes

$$\Psi(z) \simeq \sum_{i=1}^{2n_R} \frac{\Omega}{2}\,(t-\tau_i-T/2) + \sum_{\ell=1}^{n_C} \Omega(t-\tau_\ell) \tag{8-111b}$$

$$\simeq \Omega t(n_R+n_C) - \Omega T n_R/2 - \frac{\Omega}{2}\cdot\sum_{i=1}^{2n_R}\tau_i - \Omega\cdot\sum_{\ell=1}^{n_C}\tau_\ell . \tag{8-111c}$$

But $n_R+n_C = n$ and $\Omega = 2\pi/T$. Therefore,

$$\Psi(z) \simeq n\Omega t - \pi n_R - \Omega[\sum_{i=1}^{2n_R}\tau_i/2 + \sum_{\ell=1}^{n_C}\tau_\ell\,]. \tag{8.111d}$$

We emphasize that, apparently, the reduction of (8-111b) to (8-111d) depends on the following:

i) σ is large enough so that

$$\tanh\frac{\Omega}{2}\sigma \simeq 1 \tag{8-112a}$$

$$\text{ii)}\quad \cosh\Omega\sigma >> \cosh\Omega\sigma_m\,,\ \sigma_m = \max\{\sigma_1, \sigma_2, \,..\, \sigma_{n_C}\}. \tag{8-112b}$$

Under these conditions, by (8-111d), $\Psi(z)$ is a linear function of t and is independent of σ. It is dependent upon the number of real zeros and the sum of the RZ and CZ positions. The slope of $\Psi(z)$, $d\Psi(t+j\sigma)/dt$, is equal to the bandwidth of the signal, $n\Omega$.

In Fig. 8.23 a and b we have plotted constant $\Psi(z)$ contours for the square wave of Fig. 8.1 and the sawtooth of Fig. 8.4, respectively. The principal values of $\Psi(z)$ $[0 \leq \Psi(z) \leq 2\pi]$ were calculated for $0<\mathrm{Re}[z]<2\pi$ seconds and $-0.5 \leq \mathrm{Im}[z] \leq 0.5$ seconds at the 8328 intersections of a 128 point (t) by 65 point (σ) grid. The contour plotting algorithm [M-13] searches for pairs of adjacent grid points between which the calculated phase function assumes the desired level, based upon a linear interpolation. Unfortunately, when the phase moves from 2π to 0, the contour plotter thinks that all phases Ψ, such that $0<\Psi<2\pi$, may be found between these two points. This results in the thick "bars" of Fig. 8.22 (whose thickness is that of the actual grid spacing) which, in reality, mark the transition from 2π to 0 radians. Fig. 8.22 exhibits a detailed section of Fig. 8.23a, showing the contour levels.

The significance of the constant phase contours becomes evident if we define

$$s_r(z) = [s(z^*)+s(z)]/2 = \int_{-W}^{W} S(f).\cosh 2\pi f\sigma.e^{j2\pi ft}df \; , \qquad (8\text{-}113a)$$

and

$$j.s_i(z) = [s(z^*)-s(z)]/2 = \int_{-W}^{W} S(f).\sinh 2\pi f\sigma.e^{j2\pi ft}df \; . \qquad (8\text{-}113b)$$

$s_r(z)$ is a real signal because $S(f).\cosh 2\pi f\sigma$ has real (even), imaginary (odd) symmetry about f=0; $s_i(z)$ is an imaginary signal because $S(f).\sinh 2\pi f\sigma$ has real (odd), imaginary (even) symmetry about f=0 [P-2, p. 11]. From (8-109b).

$$s_r(z) = |s(z)| \cos\Psi(z) \qquad (8\text{-}114a)$$

and

$$j.s_i(z) = |s(z)| \sin\Psi(z). \qquad (8\text{-}114b)$$

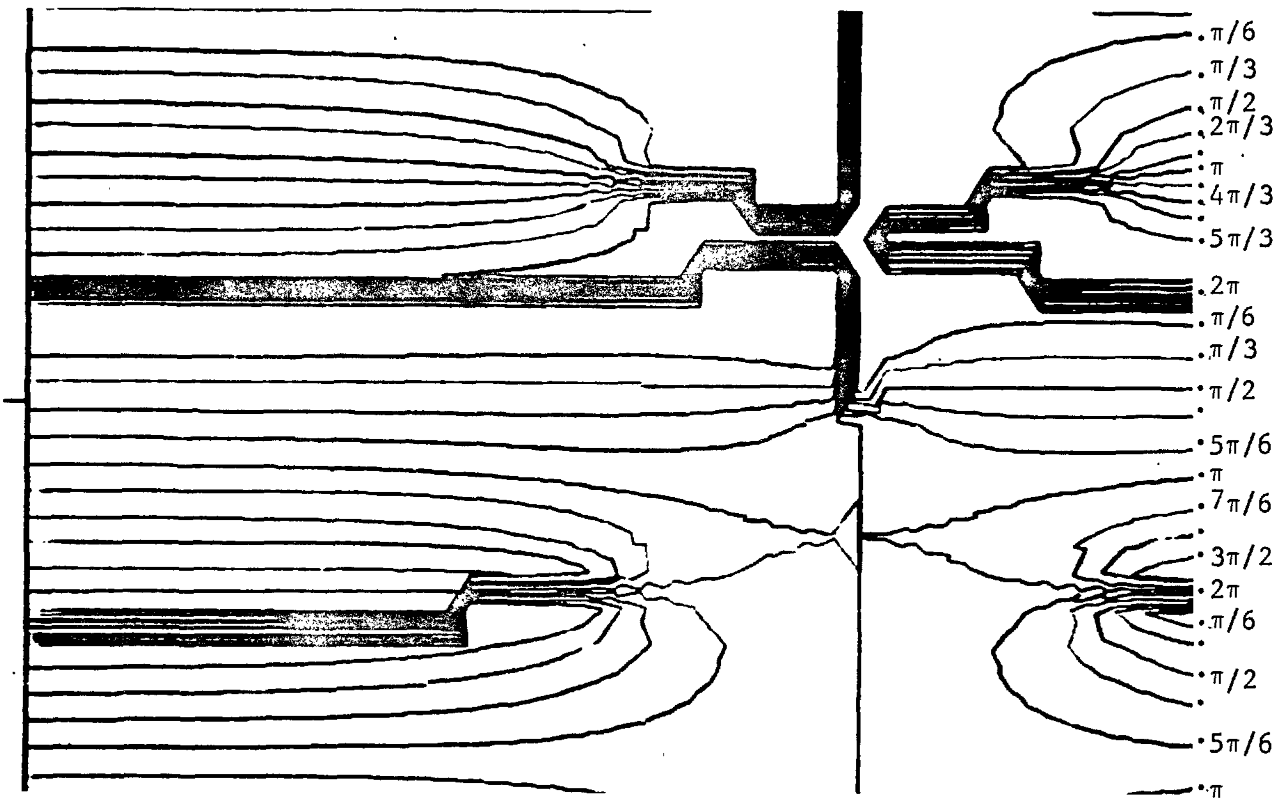


Fig. 8.22 Enlarged section of Fig. 8.23b showing contour levels of $\pi/6$, $\pi/3$, $\pi/2$, $2\pi/3$, $5\pi/6$, $7\pi/6$, $4\pi/3$, $3\pi/2$, $5\pi/3$.

$s_r(z)$ exhibits zero crossings whenever $\Psi(z) = \pm\frac{1}{2}p\pi$, p odd, while $s_i(z)$ exhibits zero crossings whenever $\Psi(z) = \pm p\pi$, p even. In particular,

$$s_r(t) = |s(t)| \cos\Psi(t) = s(t) \quad . \qquad (8\text{-}115)$$

From (8-110b), $\cos\Psi(t) = 1$ or -1 only. Both $s_r(z)$ and $s_i(z)$ are zero at all RZ's and CZ's of $s(z)$.

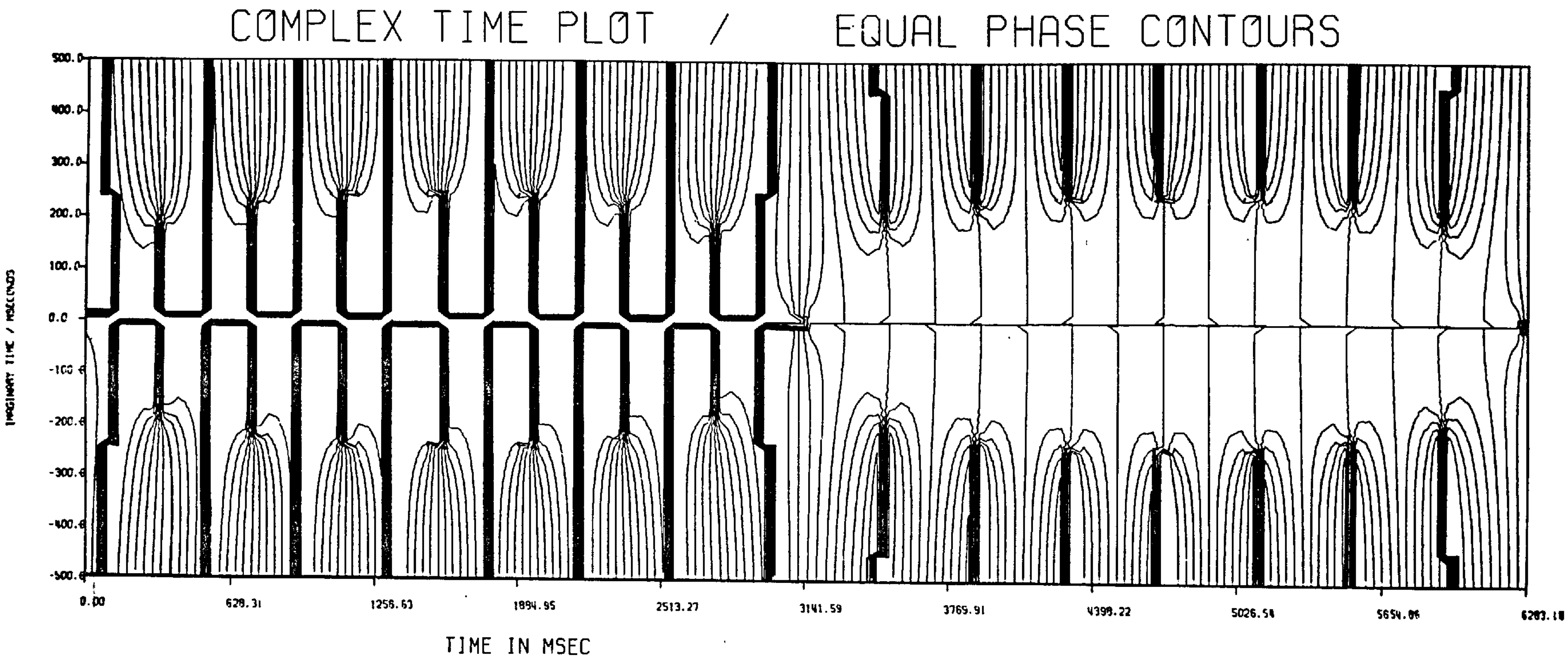


Fig. 8.23a Equal phase [Ψ(z)] contours for square wave of Fig. 8.2 .

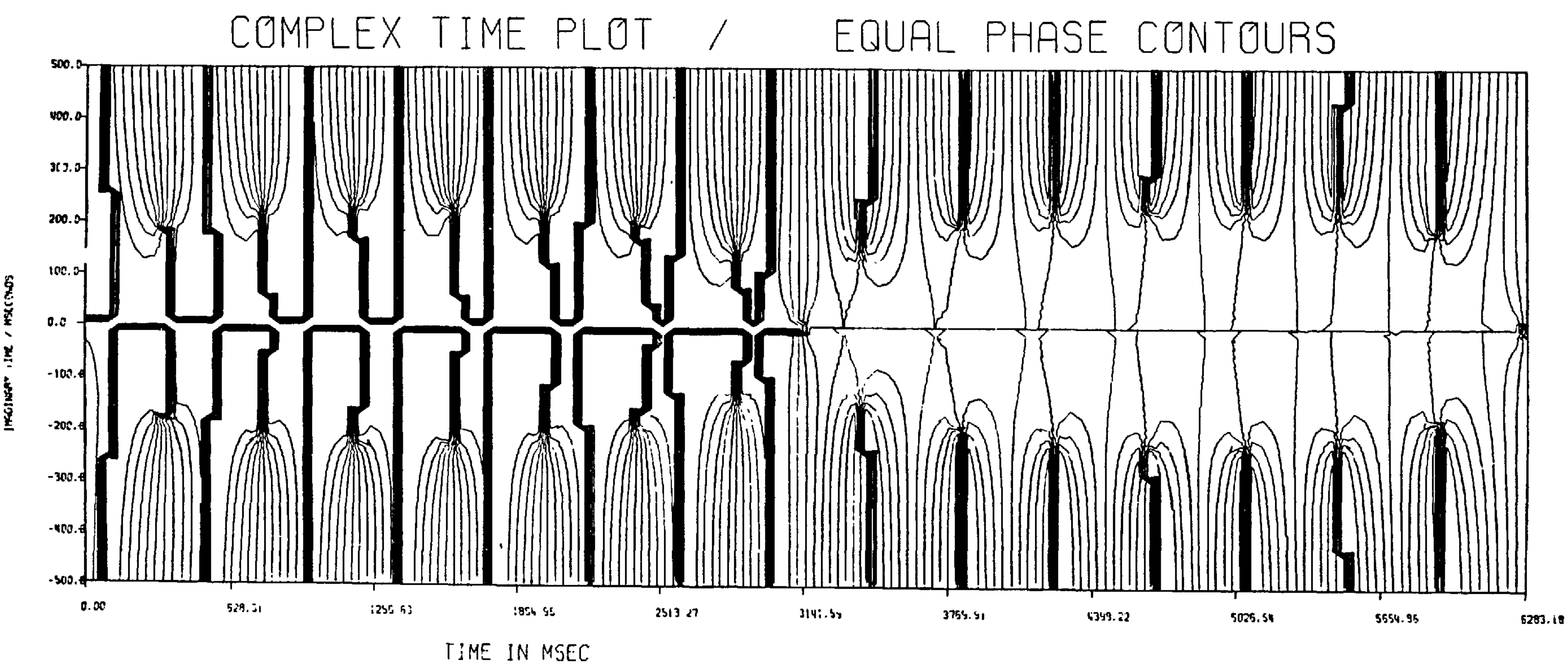


Fig. 8.23b Equal phase [Ψ(z)] contours for sawtooth waveform of Fig. 8.4.

It can easily be shown that, because the mapping $w = e^{j\Omega z}$ is conformal, constant magnitude and phase contours on the z plane are orthogonal. In addition, all constant phase lines may not intersect any others and must terminate at a pole and a zero. (The poles of s(z) occur for Im $[z]\to\infty$.) This behaviour is quite evident in Fig. 8.23. At Im[z] = $\pm$ 0.5 seconds, the constant phase contours are nearly perpendicular to the t axis and nearly uniformly spaced, as predicted by equation (8-111).

For example, at z = 0.62831 = $2\pi/10$ seconds, (8-111) predicts that

$$\Psi(z) = n\Omega t - \pi n_R - \Omega[\sum_{i=1}^{2n_R} \tau_i/2 + \sum_{\ell=1}^{n_C} \tau_\ell] \quad ,$$

if σ is sufficiently great. In this example, $\Omega = 1$; $n = 1+14 = 15$; $\sum\tau_i/2 = \pi/2$; and $\sum\tau_\ell = 43.97$ seconds. This results in $\Psi(z) = -12.5\pi$ or $1.5\pi = 4.71$ rad. The constant phase contour at $z = 2\pi/10 \pm 0.5$ seconds is the 4.71 radian contour. But $\tanh\frac{\Omega}{2}\sigma = \tanh 0.25 \approx 0.24$; $\cosh\Omega\sigma = \cosh 0.5 \approx 1.12$ and $\cosh\Omega\sigma_m = \cosh 0.245 \approx 1.03$ so that $\tanh \frac{\Omega}{2}\sigma \neq 1$ and $\cosh\Omega\sigma \neq \cosh\Omega\sigma_m$.

Thus, the fact that the phase function $\Psi(z)$ seems, for $\mathrm{Im}[z] \geq 0.5$ seconds, to be almost exactly described by (8-111) is probably linked to the regularity of the complex zeros. That is, the conditions, (8-112), required so that (8-110) can be simplified to (8-111) are sufficient but are probably not necessary.

8.8 Significance of Zero Based Signal Characteristics to Clipped Speech Studies

We have seen that it is possible to describe a bandlimited signal *completely* in terms of its real zeros--zero crossings-- *and* complex zeros. Thus, the zero crossing interval sequence constitutes only a *partial* description of the signal; it is only sufficient to construct the real zero component, $s_{RZ}(t)$.

The bandwidth of $s_{RZ}(t)$ is a fraction--$[n_R/(n_R+n_C)]$-- of the bandwidth of the original signal, s(t). Therefore, in a sense, knowledge of the $2n_R$ real zeros (of a bandlimited periodic signal) ostensibly constitutes information concerning the same fraction--$[n_R/(n_R+n_C)]$-- of the total number of parameters necessary to completely describe the signal (to a multiplicative constant).

Nevertheless, since sgn $[s(t)]$ = sgn $[s_{RZ}(t)]$, the RZ signal is sufficient to construct the clipped signal. It is in fact the *minimum* bandwidth signal which carries sufficient information to do so.

We have also noted that operations which tend to convert CZ's to RZ's, viz., differentiation, addition of a sine wave carrier and (indirectly) high pass filtering, are associated with signals which-- when clipped-- are more intelligible and/or pleasant than the original clipped signal. Such signals--by virtue of their higher zero crossing count-- also contain a greater fraction of preserved information concerning the original signal in their clipped version.

These observations seem to suggest that clipping is, among other things, a type of "imperfect" sampling process: the positions of the real zeros (zero crossings) are preserved--or sampled-- by the clipper while the complex zeros are preserved in number (if the clipped signal is re-bandlimited to the original signal bandwidth) but not (apparently) in position.

The imperfections in "sampling by clipping" are two-fold then:

i) Only a fraction, $n_R/(n_R+n_C)$, of the information necessary to completely specify the original signal is exactly[3] preserved by clipping.

ii) The rest of the information, a fraction $n_C/(n_R+n_C)$, is, apparently lost by the clipping process.

The index of efficiency of clipping as an imperfect sampler is, therefore, lower bounded by the percentage of real zeros and this index may be increased by zero count preserving complex zero conversion processes.

What remains to be explained is *whether the complex zero information is truly lost in the clipping process.* The high intelligibility of clipped speech suggests that it is not.

---

3 We shall see in sec. 9.5 that the bandlimiting operation following clipping does not usually *significantly* alter the RZ positions.

9 CLIPPED SPEECH II: CLIPPING AS A ZERO CROSSING SAMPLER AND A SPECTRAL OPERATOR ON THE COMPLEX ZERO SIGNAL-- A NEW APPROACH TO THE PSYCHOACOUSTIC PROBLEM

9.1 Review of the Product Formulation for Bandlimited Periodic Signals

We have seen that factorization of the Fourier series polynomial enables a periodic bandlimited signal to be expressed-- except for a multiplicative constant--*completely* in terms of its real zeros (zero crossings) and complex zeros; i.e.,

$$s(t) = (-1)^n |c_n| \prod_{i=1}^{2n_R} 2 \sin\frac{\Omega}{2}(t-\tau_i) \prod_{\ell=1}^{n_C} 2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)].$$

In chapter 8 we outlined some of the basic relationships and ideas concerning zero-based signal models and argued that clipping, followed by re-bandlimiting, could be regarded as an operation which may significantly alter only the complex zero signal.

In addition we showed that those pre-clipping signal processing operations which have been observed to enhance the intelligibility of the clipped signal are those which tend to convert complex zeros into real zeros without altering the total zero count.

In this chapter we will consolidate these ideas and develop a zero-based rationale for the intelligibility of clipped speech in terms of overall power spectrum feature preservation.

9.2 Signal Spectra as a Function of Zero Positions

9.2.1 A Product Expansion for Sgn[s(t)]

We noted in sec. 8.8 that $s_{RZ}(t)$ contains sufficient information to create sgn[s(t)]. That is,

$$\mathrm{sgn}[s(t)] = \mathrm{sgn}[s_{RZ}(t)] \tag{9-1a}$$

$$= \mathrm{sgn}\Big[\prod_{i=1}^{2n_R} 2\sin\frac{\Omega}{2}(t-\tau_i)\Big] \tag{9-1b}$$

$$= \prod_{i=1}^{2n_R} \mathrm{sgn}\Big[2\sin\frac{\Omega}{2}(t-\tau_i)\Big] \tag{9-1c}$$

$$= \prod_{i=1}^{2n_R} \mathrm{sgn}[t-\tau_i], \quad |t|\leq T/2. \quad [\text{V-11}] \tag{9-1d}$$

9.2.2 The Fourier Series Coefficients of Sgn[s(t)] in Terms of its Zero Crossing Positions

The Fourier series pair for sgn[s(t)] is

$$\mathrm{sgn}[s(t)] = \sum_{k=-\infty}^{\infty} c_k \cdot e^{jk\Omega}, \tag{9-2a}$$

where

$$c_k = \frac{1}{T}\int_{-T/2}^{T/2} \mathrm{sgn}[s(t)]\cdot e^{-jk\Omega t}\,dt. \tag{9-2b}$$

Substituting (9-1d) in (9-2b),

$$c_k = \frac{1}{T}\int_{-T/2}^{T/2} e^{-jk\Omega t}\,\Big\{\prod_{i=1}^{2n_R} \mathrm{sgn}[t-\tau_i]\Big\}dt, \quad k\neq 0. \tag{9-3}$$

Integrating by parts, we let $dv = e^{-jk\Omega t}dt$ (so that $v = -e^{-jk\Omega t}/jk$) and $u = \prod_{i=1}^{2n_R} \mathrm{sgn}\,[t-\tau_i]$. Then $du/dt = 2 \sum_{i=1}^{2n_R} a_i \cdot \delta(t-\tau_i)$, where a_i is a polarity switching function and is equal to $(-1)^{i-1}$.

Thus

$$c_k = \frac{1}{T}\left[uv \Big|_{-T/2}^{T/2} - \int_{-T/2}^{T/2} v\, du \right] \tag{9-4a}$$

$$= \frac{2}{jk\Omega T} \int_{-T/2}^{T/2} e^{-jk\Omega t}\{\sum_{i=1}^{2n_R} a_i \cdot \delta(t-\tau_i)\}dt, \quad k\neq 0 \tag{9-4b}$$

$$= \frac{2}{jk\Omega T} \sum_{i=1}^{2n_R} (-1)^{i-1} \cdot e^{-jk\Omega \tau_i}, \quad k\neq 0. \tag{9-4c}$$

Note that $uv \Big|_{-T/2}^{T/2}$ is zero because of the periodicity in T of u and v.

$$\text{For } k=0, \quad c_0 = \frac{1}{T} \int_{-T/2}^{T/2} \{\prod_{i=1}^{2n_R} \mathrm{sgn}[t-\tau_i]\}dt$$

$$= -2 \sum_{i=1}^{2n_R} (-1)^{i-1} \cdot \tau_i/T - 1, \tag{9-4d}$$

= net area under the square wave.

In summary,

$$\mathrm{sgn}[s(t)] = \begin{cases} \sum_{k=-\infty}^{\infty} c_k \cdot e^{jk\Omega t} \\ \\ c_o + 2 \sum_{k=1}^{\infty} [a_k \cos k\Omega t + b_k \sin k\Omega t] \end{cases}$$

where $\Omega = 2\pi/T$ and $\tau_1, \tau_2, \tau_3, \ldots\ldots \tau_{2n_R}$ are the RZ's of s(t) in $-T/2 \leq t \leq T/2$. Then

$$c_0 = -1 - 2 \cdot \sum_{i=1}^{2n_R} (-1)^{i-1}\, \hat{\tau}_i,\ \hat{\tau}_i = \tau_i/T \tag{9-5a}$$

$$c_k = \frac{1}{j\pi k} \sum_{i=1}^{2n_R} (-1)^{i-1} \cdot e^{-j2\pi k\hat{\tau}_i} \tag{9-5b}$$

$$a_k = \frac{-1}{2\pi k} \sum_{i=1}^{2n_R} (-1)^{i-1} \cdot \sin 2\pi k\hat{\tau}_i \tag{9-6a}$$

$$b_k = \frac{1}{2\pi k} \sum_{i=1}^{2n_R} (-1)^{i-1} \cdot \cos 2\pi k\hat{\tau}_i \quad . \quad [V-11] \tag{9-6b}$$

Previous work has emphasized zero crossing intervals rather than distance from reference point (e.g., t = 0). Letting

$$\tau_i = \sum_{q=1}^{i} \Delta_q, \text{ where } \Delta_q = \tau_q - \tau_{q-1}, \tag{9-7}$$

then

$$c_0 = -1 - 2 \cdot \sum_{i=1}^{2n_R} (-1)^{i-1} \cdot \sum_{q=1}^{i} \hat{\Delta}_q,\ \hat{\Delta}_q = \Delta_q/T \tag{9-8a}$$

and

$$c_k = \frac{1}{j2\pi k} \sum_{i=1}^{2n_R} (-1)^{i-1} \cdot \prod_{q=1}^{i} e^{-j2\pi k\hat{\Delta}q} \quad . \tag{9-8b}$$

These results are analogous to those of sec. 8.4.1. In each case, the spectrum of the signal in question ($s_{RZ}(t)$ or sgn[s(t)]) -- *implicitly* determined by the product expansion, (8-57a) and (9-1b), respectively--are *explicitly* expressed as a function of the real zero (zero crossing) positions. Again, the results are qualitatively uninforming.

Thus our method of attack will be to examine the effect of the clipping-bandlimiting operation on the *positions* of the zeros of the speech signals, particularly the complex zeros.

9.3 The Zeros of Speech Signals

In this section, we will describe experiments and observations regarding the zeros of voiced speech signals. These experiments are the first step in applying the elements of zero-based signal analysis to the speech clipping problem.

9.3.1 Hybrid Factorization

The speech signals to be factorized were single pitch periods of sustained vowels spoken by the author, a Canadian. Five second segments of sustained vowels were recorded in the Imperial College silent room, constructed by Sound Control, Limited.

The silent room--of dimensions 6' x 7' x 8'--was isolated from the main building structure by nine rubber feet and was of doublewalled wooden construction lined with Bondacoust. The Bondacoust lining reduced reverberation time to less than 0.05 seconds and therefore allowed "dead" recordings to be made.

The speech was bandlimited to ± 3 KHz, with a Krohn-Hite filter, model 310-AB, having fully variable upper cut-off frequency and attennation of 24 db octave above cutoff frequency, and then recorded on a Tandberg model 62 tape recorder (#684744) operating at $7\frac{1}{2}$ ips via an AKG type D19C dynamic microphone (#19765). The tape was then played back at 1 7/8 ips into a B & K type 7001 FM tape recorder (#204688) operating at 60 ips. Finally, the output of the B & K recorder, operating at 1.5 ips, was sampled 1250 times per second by the Direct Data Channel of the Imperial College IBM 7094 computer. The most significant fidelity limitation of this arrangement was the low frequency response of the Tandberg tape recorder. This response dropped off at about 30 Hz thus effectively highpass limiting the 4x slowed down speech to 4x30 or 120 Hz. Comparison of this waveform with speech slowed down by recording at 60 ips on the FM tape recorder (response 0-20 KHz at 60 ips) and playing back at 15 ips (also a 4:1 speed reduction) revealed no significant changes in overt signal structure.

The effective speech sampling rate was therefore

$$1250 \times \frac{7.500}{1.875} \times \frac{60}{1.5} = 200{,}000 \text{ samples per second.}$$

Since only 6000 independent (Nyquist) samples per second were required, the effective sampling rate was 200,000/6000 or 33.33 times the Nyquist rate. A detailed Calcomp signal location index was prepared, and selected "typical" pitch periods--each showing no evidence of FM dropout--were located and read into storage arrays. The author's normal pitch period varies from approximately 8.5 to 10.0 milliseconds giving about 1800 speech samples per pitch period. A linear interpolation was performed to increase the number of samples in a selected pitch period to 2048. A discrete Fourier transform was then implemented via the FFT to yield the complex Fourier series coefficients for the range 0 to 100 KHz. The original signal had been "non-ideally" bandlimited to 3 KHz. In addition, noise in the 3 KHz–100 KHz region was assumed to have been introduced by the tape recording-sampling process. This spectral noise--observed to be very small compared to the passband coefficients--, along with the speech signal spectral components above 3 KHz, were eliminated and an inverse discrete Fourier transform (IDFT) resulted in a smooth, virtually noise free 2048 point signal waveform--"ideally" bandlimited to ±3 KHz--with zero crossings defined to within 0.0048 milliseconds. (Note that Licklider's results (observation L8) showed that zero crossing position specification to 0.1 milliseconds is sufficient for high intelligibility.) It should be emphasized that the above procedure is equivalent to perfect sampling of a truly bandlimited signal, at the Nyquist rate, and then carrying out a bandlimited interpolation [G-4, pp. 199-200] of the Nyquist samples. However, this method guarantees that aliasing errors due to insufficient sampling rate after (necessarily) imperfect bandlimiting to 3 KHz, and high frequency noise, are eliminated.

The positions of the zero crossings were further refined by a linear interpolation between samples at which a signal polarity change occurs. $s_{RZ}(t)$ was then synthesized from this positional information using (8-14):

$$s_{RZ}(t) = \prod_{i=1}^{2n_R} 2 \sin \frac{\Omega}{2}(t-\tau_i) \quad . \tag{9-9}$$

$s_{CZ}(t)$ was then derived by division of $s(t)$ by $s_{RZ}(t)$, each signal being defined at 2048 points. Since both $s(t)$ and $s_{RZ}(t)$ have the same zero crossings, L' Hôpital's rule was applied when necessary. The resultant $s_{CZ}(t)$ contains slight high frequency noise at the times corresponding to the zero crossings of $s(t)$ [or $s_{RZ}(t)$]. However, a discrete Fourier transform of $s_{CZ}(t)$--which has been derived from two signals sampled at 33 times the Nyquist rate--allows this noise to be eliminated by lowpass filtering; only the Fourier coefficients which fall within the known bandwidth of $s_{CZ}(t)$, i.e.,

$$n_C = (3000/F_o - n_R),\ \text{where}\ F_o = 1/T, \quad (9\text{-}10)$$

are used to form the polynomial which is factorized to yield the complex zeros of the signal.

We shall examine the experimental findings regarding the complex zeros of vowels in sec. 9.3.3.

9.3.2 Organization of the Experimental Observations

Single pitch periods of the vowels /u/, /o/, /ʌ/, /e/ and /ɛ/ (b*oo*t, *o*bey, b*u*t, h*a*te, b*e*t) were analyzed and graphical results are presented in groups of 6 pages per vowel. The "page" organization for *each* vowel is as follows:

1/ The zero crossings of s(t), s'(t) and s''(t):

Data concerning the real zeros of s(t), where s(t) is a single pitch period of the vowel in question, are given. Both the distance of the zero crossings from t = 0 and the distance between pairs of adjacent zero crossings are tabulated (in milliseconds).

2/ A graphical presentation (2048 pts) of s(t), s'(t), s''(t) and s'''(t):

The signals are periodic and bandlimited so that differentiation is easily carried out in the frequency domain using the FFT implementation of the DFT, i.e.,

$$\frac{d^n s(t)}{d\,t^n} = (jk\Omega)^n\,S(k\Omega) \quad [\text{P-2, p. 16}]\ . \qquad (9\text{-}11)$$

In practice, a 2048 point transform of s(t) was carried out, the complex Fourier coefficients altered as per (9-11) and an IDFT yielded s'(t). Note that only block capital letters are available on the Calcomp machine so that "T" = "t". The vowel /i/ is represented only by the original waveform and its first three derivatives. Factorization problems prevented further studies at the time.

3/ "Page 3" of each vowel group shows

a) s(t)

b) $s_{RZ}(t)$ for s(t)

c) $s_{CZ}(t)$ for s(t)

d) a root map showing the real zeros (zero crossings)--signified by "O"'s on the real time axis--and the complex zero pairs--signified by "X"'s--of s(t).

4/ "Page 4" of each vowel group is identical to page 3 except that s(t) has been replaced by the signal BL{C s(t)} -- the clipped, then bandlimited (3 KHz), signal.

5/ The amplitude spectrum of

a) s(t)

b) $s_{RZ}(t)$ of a)

c) $s_{CZ}(t)$ of a)

d) BL{C s(t)}

e) $s_{RZ}(t)$ of d)

f) $s_{CZ}(t)$ of d)

Here the amplitude spectrum is defined as $[a_k^2 + b_k^2]^{\frac{1}{2}}$, where

$$s(t) = a_o/2 + \sum_{k=1}^{n} [a_k \cos k\Omega t + b_k \sin k\Omega t].$$

The spectral line components have been interpolated with straight line segments in order to emphasize the spectral envelope features.

6/ "Page 6" gives the positional data concerning the real and complex zeros presented on pages 3/ and 4/

Note the following concerning the overall presentation:

i) The imaginary time (σ) scale for the root map of the zeros of the clipped, then bandlimited, signal has been scaled to approximately match that of the original signal. This has been done for comparison purposes. In the case where a CZ of the clipped, then bandlimited, signal has an imaginary coordinate (σ) significantly greater than the maximum σ found in the original signal, arrows having the same real time position as the complex zero pair and labelled with the value of the σ ordinate (in milliseconds) have been used. (e.g., /o/)

ii) In one case, /o/, the bandlimiting operation following clipping has caused a real zero pair (consisting of two zero crossings very close together) to disappear. This phenomenon will be discussed in sec. 9.4.1.

iii) Due to the minute "ripple" error caused (in regions where $s_{CZ}(t)$ is very small) by the filtering process used to remove the high frequency noise in $s_{CZ}(t)$ (sec. 9.3.1), there are instances of complex zeros falling on the real time axis. However, the method of hybrid factorization ensures that $\{Rz\}*\{Cz\}\rightarrow\{c\}$, with small error.

The figure numbers for the vowel diagram sets are as follows: /u/, Figs. 9.1-9.4; /o/, Figs. 9.5-9.8; /ʌ/, Figs. 9.9-9.12; /e/, Figs. 9.13-9.16; /ε/, Figs. 9.17-9.20; /i/, Fig. 9.21.

9.3.3 Experimental Observations: Original Signal

i) Differentiation

Table 9.12 shows the number of zero crossings per period for each of the six vowel pitch periods and their first three derivatives. Also listed are the fraction of the zeros which appear as zero crossings for each signal. Fig. 9.22 summarizes the data.

Note that the vowels /u/, /o/, and /i/ have a significantly smaller percentage of zero crossings than the other vowels. These three vowels are those specifically singled out by Ahmend and Fatechand ([A-2] and sec. 5.1.3) as having the least resistance to post-clipping degradation by time domain truncation.

ii) $s_{RZ}(t)$

The real zero signal, as might be expected from its formulation [eq. (9-21)] is a smoothly varying signal alternately changing polarity between successive zero crossings.

Indeed, the results obtained in sec. 8.5.2 lead us to believe that ripple in $s_{RZ}(t)$ or even points of inflection, would suggest the presence of complex zeros. Points of inflection would, after a finite number of differentiations, give rise to ripple and then real zeros. *Since all derivatives of RZ signals are real zero (sec. 8.4), RZ signals may not exhibit points of inflection.*

Examination of the RZ signals for /o/, Fig. 9.6b, and /e/, Fig. 9.14b, reveals that where there are relatively long periods of time without any zero crossings, $s_{RZ}(t)$ has large excursions; conversely, closly spaced RZ's tend to cause signal amplitude suppression. These two effects are not unrelated. Irregularity of zero crossing spacing is apparently greatly magnified in the effect produced on signal excursions. We shall discuss the relationship between zero spacing and signal growth in sec. 9.3.4.

iii) $s_{CZ}(t)$

Since $s_{CZ}(t) \propto s(t)/s_{RZ}(t)$, then--assuming that s(t) exhibits no significant excursions from its rms value (sec. 9.4.5)-- $s_{CZ}(t)$ will (intuitively) be "large" when $s_{RZ}(t)$ is small and vice versa. Observationally, this is indeed the case. The time segments during which $s_{RZ}(t)$ has amplitudes which are "visually" insignificant (compared to the segments of large excursion) correspond to time segments containing large excursions of $s_{CZ}(t)$. Figs. 9.6c, d, 9.10c,d, 9.14c,d, and 9.18c,d should be examined and the following points noted:

Time segments of $s_{CZ}(t)$ containing *large amplitude excursions* correspond to time periods which have an absence of CZ pairs (note especially Figs. 9.14c,d and 9.18c,d) *and/or* contain CZ pairs having large imaginary components (see Figs. 9.10c,d and 9.6c,d).

In accordance with the theory developed in sec. 8.5.2 (concerning ripple and CZ's), we expect a CZ pair between adjacent maxima (or minima). This is, of course, observed. It is also noted that amplitude suppression effects of CZ pairs on $s_{CZ}(t)$ are, roughly, inversely proportional to the imaginary component of the CZ. This is to be expected from the formulation of $s_{CZ}(t)$,

$$s_{CZ}(t) = \prod_{\ell=1}^{n_C} 2[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)] \ .$$

That is, for $\Omega\sigma_\ell$ small, $[\cosh\Omega\ \sigma \approx 1]$, $s_{CZ}(t)$ *must* become very small in the vicinity of $t = \tau_\ell$; conversely, for $\Omega\sigma_\ell$ very large, the percentage amplitude variation of $[\cosh\Omega\sigma_\ell - \cos\Omega(t-\tau_\ell)]$ for variations in t, i.e.,

$$\Delta\ s_{CZ}(t) = \frac{[(\cosh\Omega\sigma_\ell + 1) - (\cosh\Omega\sigma_\ell - 1)]}{\cosh\Omega\sigma_\ell} \qquad (9\text{-}12a)$$

$$= 2\ /\ \cosh\Omega\sigma_\ell\ , \qquad (9\text{-}12b)$$

approaches zero rapidly.

In Fig. 9.10c,d note the small ripple effect on $s_{CZ}(t)$ of the CZ pair at $z = 2.2622 \pm j\ 0.7646$ milliseconds compared with the ripple effect caused by the CZ pair at $z = 2.6883 \pm j\ 0.2516$ milliseconds. Similarly, in Fig. 9.2c,d note the same type of effect of the CZ pairs at $z = 5.3525 \pm j\ 0.0982$ and $z = 7.1571 \pm j\ 0.2429$. The former -- because of its very small imaginary component -- greatly reduces the amplitude of $s_{CZ}(t)$ while the latter causes only a small ripple effect to occur. Another observation to be noted (Figs. 9,10c,d, 9.14c,d and 9.18c, d) is that a succession of CZ pairs of "small" imaginary component reduces $s_{CZ}(t)$ to a very small value. This effect occurs in those time segments when $s_{RZ}(t)$ is large and is quite analogous to the dynamic suppression caused in $s_{RZ}(t)$ by a succession of closely spaced RZ's.

iv) s(t)

The original signal (s), s(t), possesses no apparent time segments, containing large excursions or of virtually zero amplitude, similar to those observed in $s_{RZ}(t)$ and $s_{CZ}(t)$. The reasons for this will be discussed in sec. 9.3.5.

Since -- observationally -- significant gaps without RZ's produce huge amplitude excursions in $s_{RZ}(t)$ and significant gaps without CZ's produce huge amplitude excursions in $s_{CZ}(t)$, then signal segments with no RZ's or CZ's must produce huge excursions in s(t). Thus, we would expect to find no significant time segments without either RZ's or CZ's since we observe no huge amplitude excursions in s(t). *Observationally*, this seems to be the case; where RZ's are sparse, CZ's are plentiful and vice versa. There are no "significant" gaps without either an RZ or a CZ pair. "Significant" means much greater than

$$T/(n_R+n_C) = T/n = 1/W \quad . \qquad (9\text{-}13)$$

That is, *the zeros of vowel waveforms seem to occur regularly in real time*.

/u/ - BOOT PERIOD=8.72 MILLISECONDS

REAL ZEROS- TIME/MILLISECONDS DELTA(N)=TAU(N)-TAU(N-1)

	S(T)		S'(T)		S''(T)	
N	TAU(N)	DELTA(N)	TAU(N)	DELTA(N)	TAU(N)	DELTA(N)
1	2.2766	2.2816	0.2851	0.5499	0.0212	0.1285
2	3.7277	1.4511	0.7276	0.4425	0.1574	0.1362
3	5.1405	1.4128	1.3531	0.6255	0.3957	0.2383
4	6.9873	1.8468	1.9999	0.6468	0.5616	0.1659
5	7.4937	0.5064	2.1616	0.1617	0.6553	0.0937
6	8.7150	1.2213	2.8340	0.6724	0.8765	0.2212
7			3.1191	0.2851	1.0255	0.1490
8			3.3233	0.2042	1.2255	0.2000
9			4.0553	0.7320	1.4723	0.2468
10			4.2510	0.1957	1.6723	0.2000
11			4.6383	0.3873	1.8510	0.1787
12			5.2425	0.6042	2.0850	0.2340
13			5.3574	0.1149	2.3702	0.2852
14			5.9957	0.6383	2.6084	0.2382
15			6.5999	0.6042	2.7319	0.1235
16			6.7872	0.1873	2.9701	0.2382
17			7.2808	0.4936	3.2127	0.2426
18			7.7999	0.5191	3.5999	0.3872
19			7.9403	0.1404	3.7148	0.1149
20			8.4552	0.5149	3.8595	0.1447
21					4.1446	0.2851
22					3.3659	0.7870
23					5.0255	1.6596
24					5.2935	0.2680
25					5.4638	0.1703
26					5.5999	0.1361
27					5.8127	0.2128
28					6.2169	0.4042
29					6.7106	0.4937
30					6.9574	0.2468
31					7.4552	0.4978
32					7.9148	0.4596
33					8.1148	0.2000
34					8.2467	0.1319
35					8.3743	0.1276
36					8.6127	0.2384

REF. FIG. 9.1

A B C

NOTE DELTA(1)=(PERIOD+TAU(1))-TAU(LAST)

Table 9.1

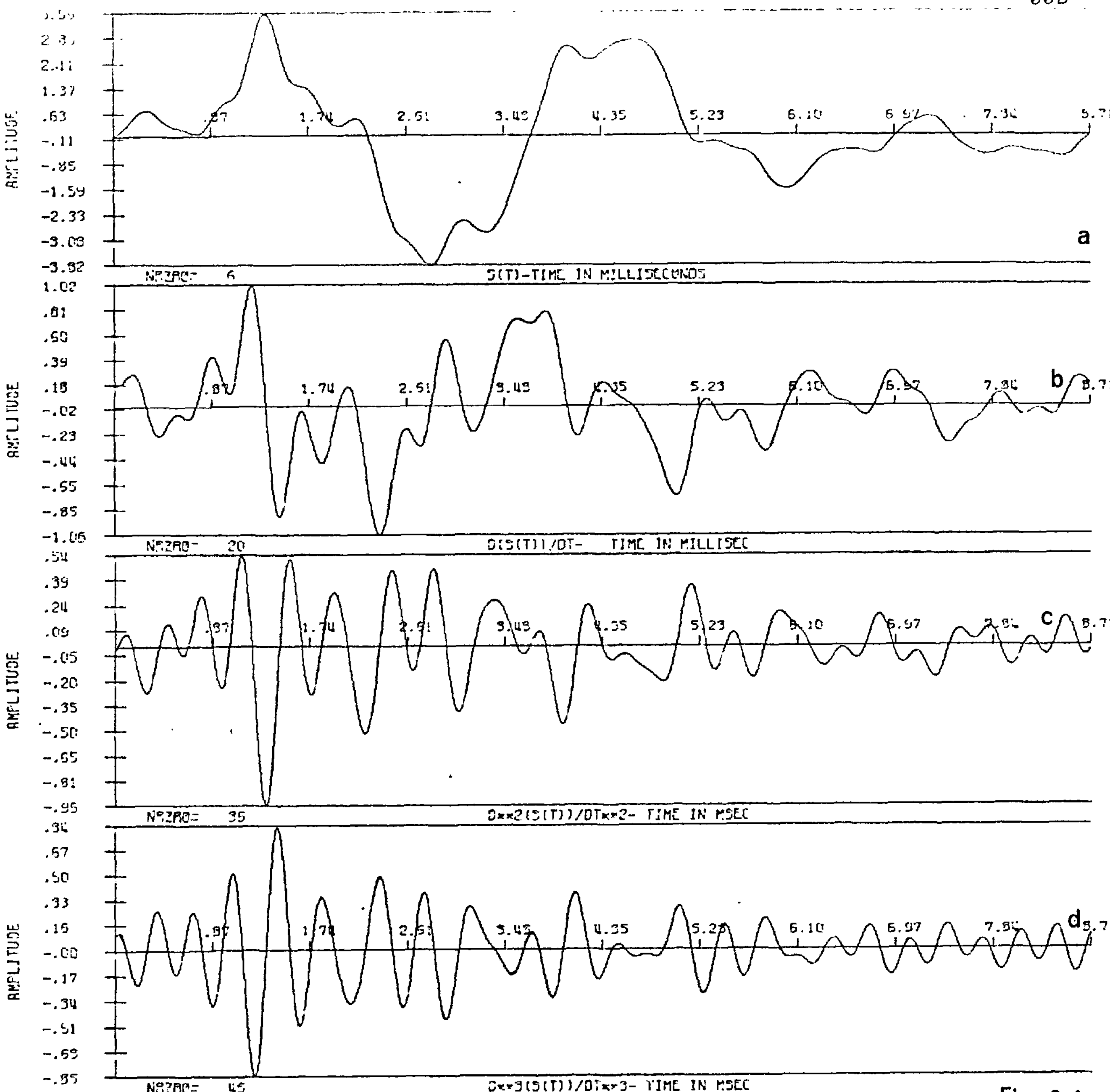

AMPLITUDE
S(T)-TIME IN MILLISECONDS
D(S(T))/DT- TIME IN MILLISEC
D**2(S(T))/DT**2- TIME IN MSEC
D**3(S(T))/DT**3- TIME IN MSEC
a
b
c
d

Fig. 9.1

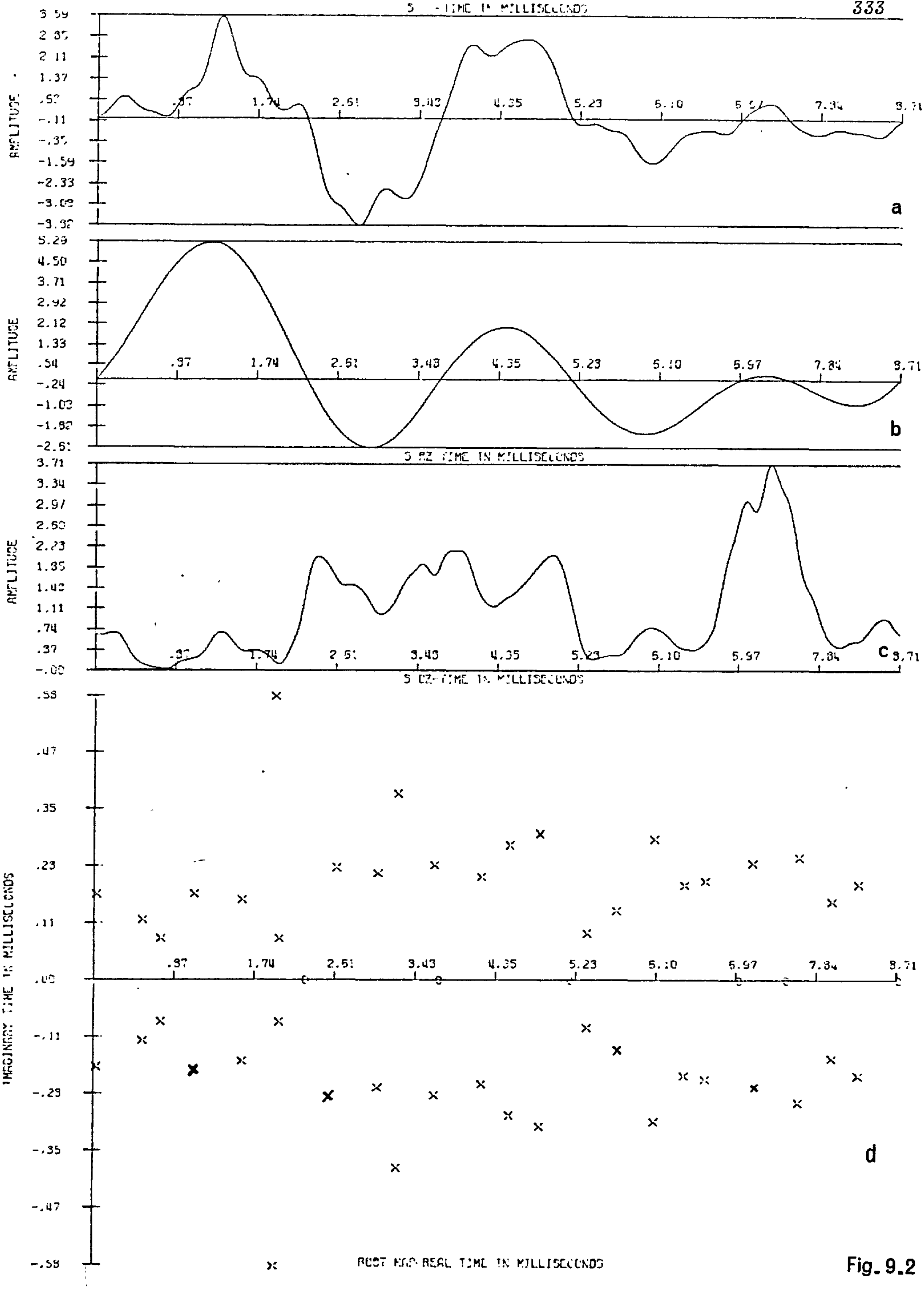
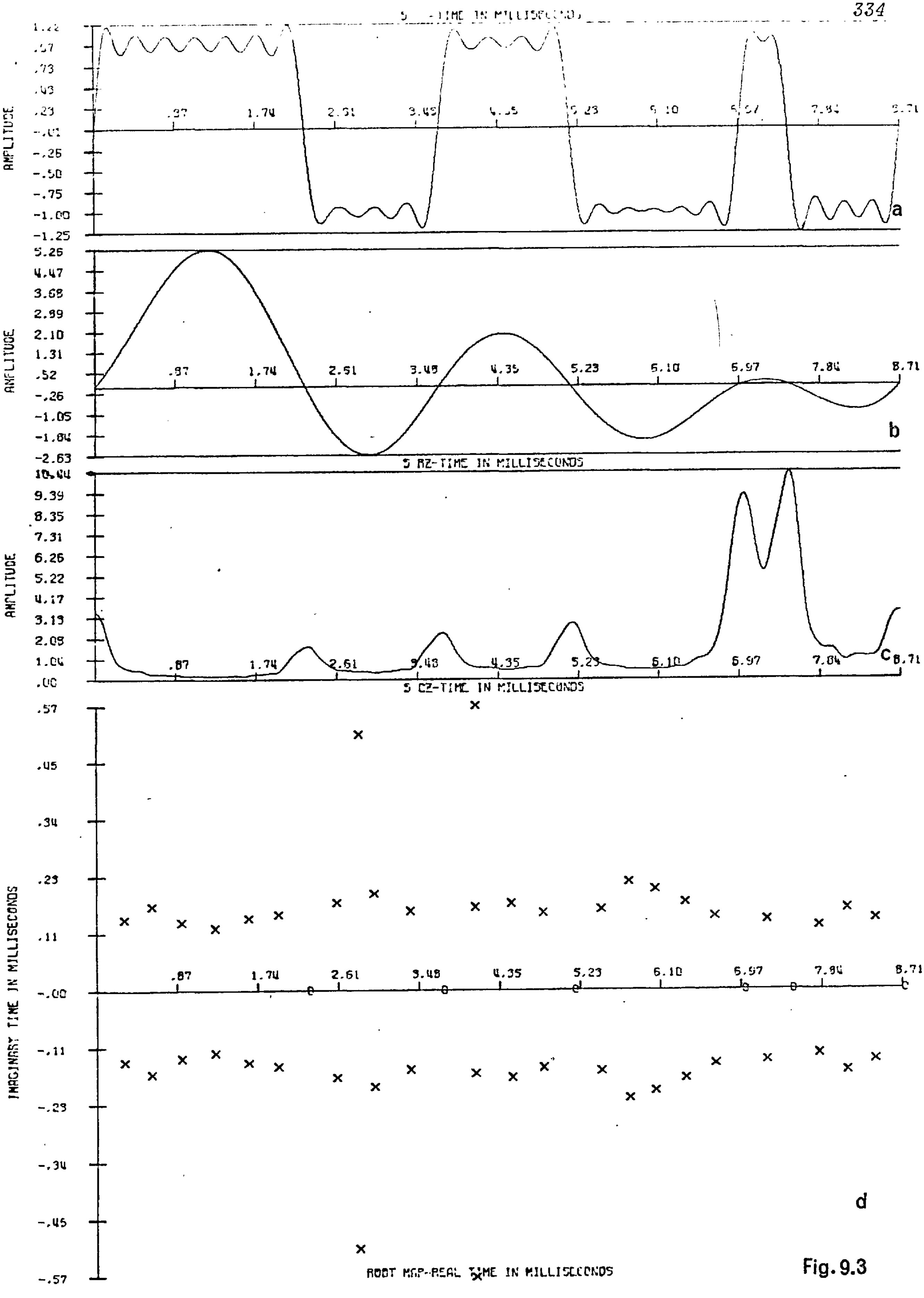

AMPLITUDE
TIME IN MILLISECONDS
a
b
c
d
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS

Fig. 9.2



5 -TIME IN MILLISECONDS
AMPLITUDE
a
AMPLITUDE
b
5 RZ-TIME IN MILLISECONDS
AMPLITUDE
c
5 CZ-TIME IN MILLISECONDS
IMAGINARY TIME IN MILLISECONDS
d
ROOT MAP-REAL TIME IN MILLISECONDS

Fig. 9.3

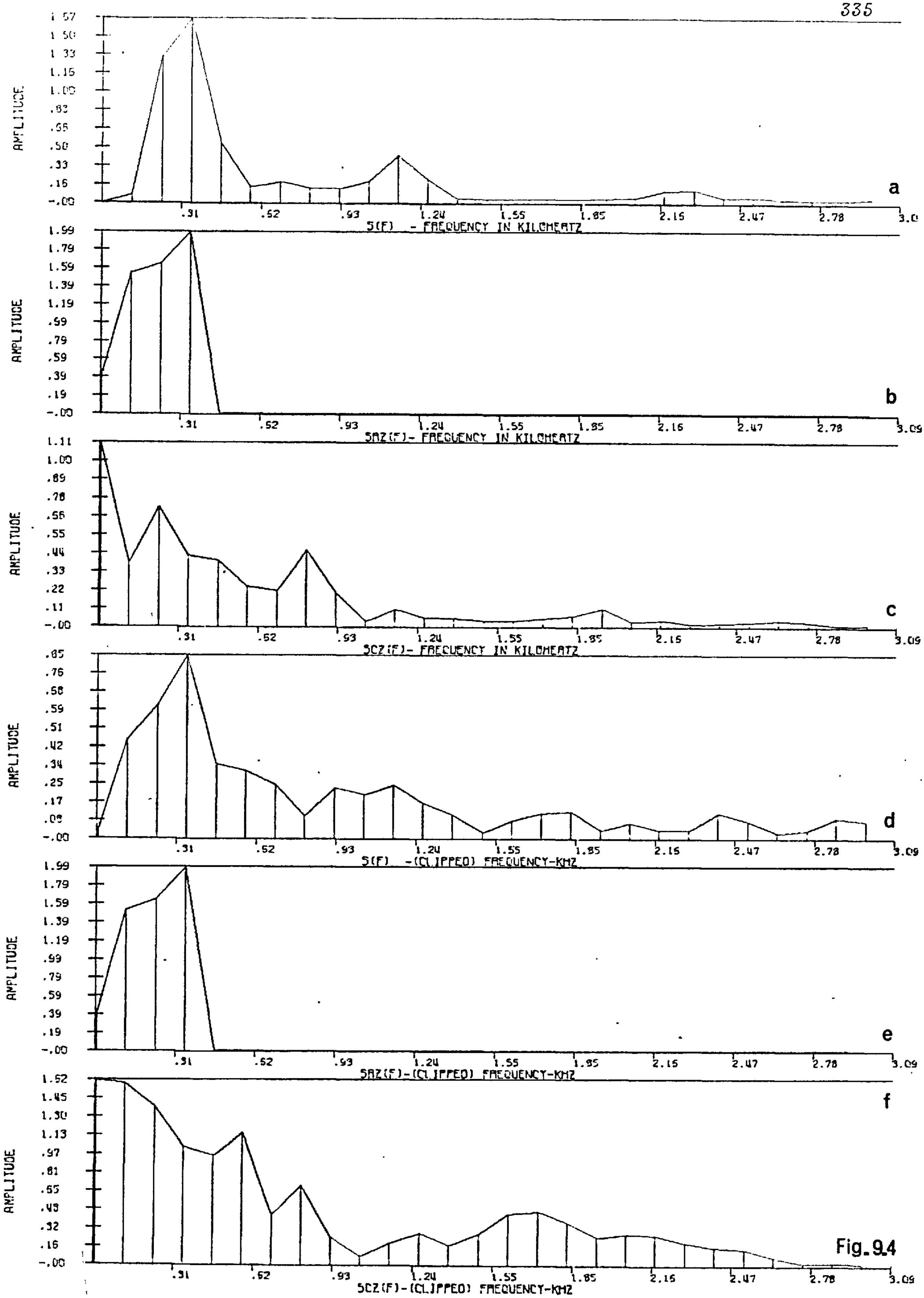

AMPLITUDE
a
S(F) - FREQUENCY IN KILOHERTZ
b
SAZ(F)- FREQUENCY IN KILOHERTZ
c
SCZ(F)- FREQUENCY IN KILOHERTZ
d
S(F) -(CLIPPED) FREQUENCY-KHZ
e
SAZ(F)-(CLIPPED) FREQUENCY-KHZ
f
SCZ(F)-(CLIPPED) FREQUENCY-KHZ
.31
.62
.93
1.24
1.55
1.85
2.16
2.47
2.78
3.09

Fig. 9.4

-/u/ - BOOT PERIOD=8.72 MILLISECONDS

ZEROS- TIME/MILLISECONDS

ORIGINAL SIGNAL	CLIPPED AND B.L. SIGNAL
.0329 +/- J 0.1790	0.2927 +/- J 0.1444
.5345 +/- J 0.1246	0.5899 +/- J 0.1692
.7349 +/- J 0.0861	0.9161 +/- J 0.1376
1.0717 +/- J 0.1769	1.2800 +/- J 0.1263
1.6172 +/- J 0.1672	1.6375 +/- J 0.1461
1.9654 +/- J 0.5896	1.9587 +/- J 0.1534
2.0111 +/- J 0.0863	2.2724
2.2766	2.5911 +/- J 0.1766
2.6391 +/- J 0.2341	2.8317 +/- J 0.5191
3.0821 +/- J 0.2222	2.9962 +/- J 0.1956
3.2890 +/- J 0.3895	3.3835 +/- J 0.1600
3.6963 +/- J 0.2383	3.7277
3.7277	4.0936 +/- J 0.5764
4.2034 +/- J 0.2154	4.1807 +/- J 0.1729
4.5094 +/- J 0.2804	4.4824 +/- J 0.1763
4.8362 +/- J 0.3035	4.8297 +/- J 0.1563
5.1405	5.1405
5.3525 +/- J 0.0982	5.4511 +/- J 0.1626
5.6682 +/- J 0.1449	5.7553 +/- J 0.2194
6.0833 +/- J 0.2929	6.0391 +/- J 0.2042
6.4079 +/- J 0.1968	6.3647 +/- J 0.1775
6.6337 +/- J 0.2056	6.6841 +/- J 0.1489
6.9873	6.9831
7.1571 +/- J 0.2429	7.2491 +/- J 0.1418
7.4937	7.5022
7.6546 +/- J 0.2547	7.8083 +/- J 0.1295
8.0076 +/- J 0.1629	8.1190 +/- J 0.1641
8.2986 +/- J 0.1984	8.4203 +/- J 0.1421
8.7150	8.7150
REF. FIG. 9.2	REF. FIG. 9.3

Table 9.2

/0/- OBEY   PERIOD=8.99 MILLISECONDS

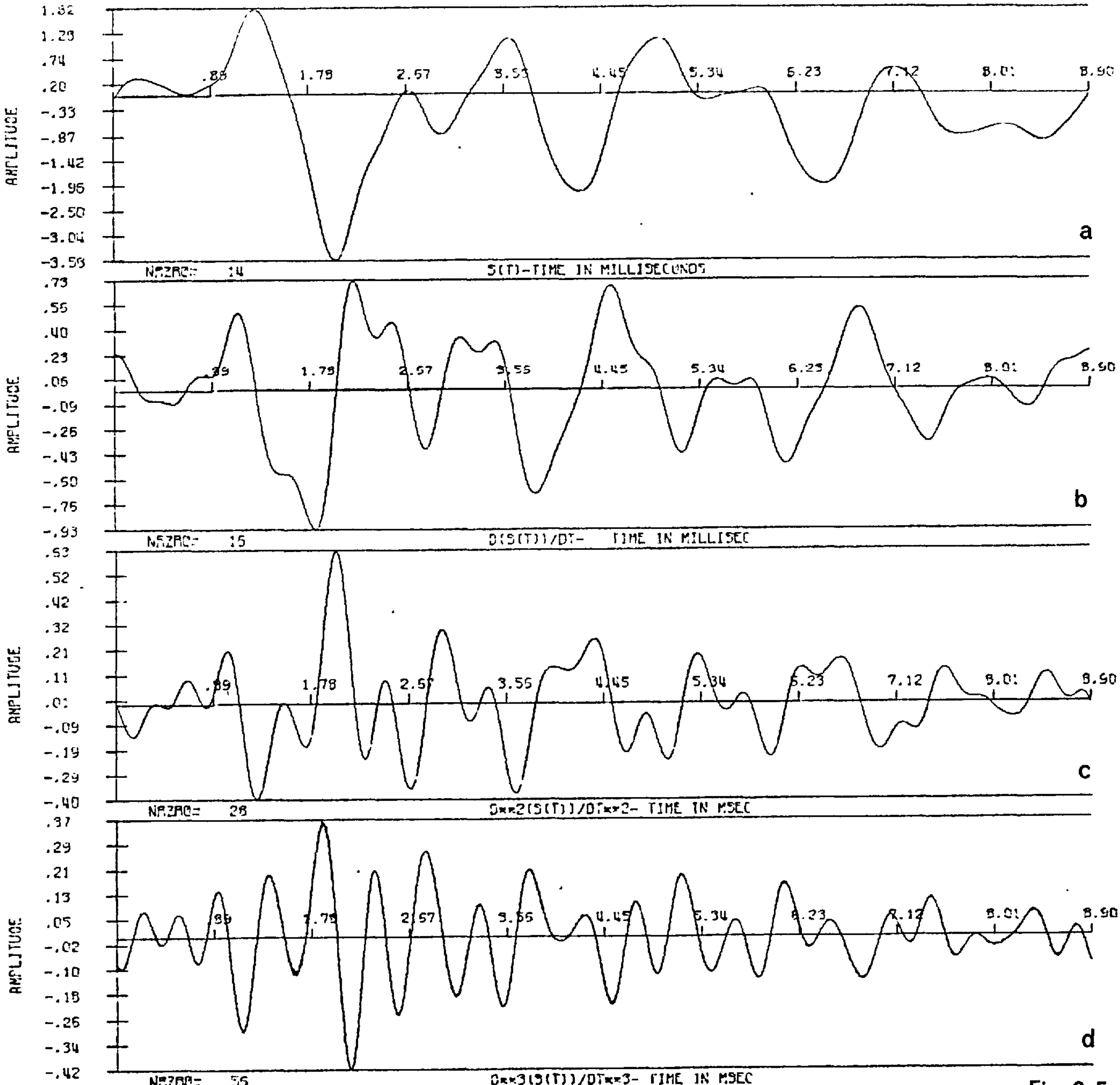
REAL ZEROS-   TIME/MILLISECONDS   DELTA(N)=TAU(N)-TAU(N-1)

	S(T)		S'(T)		S''(T)	
N	TAU(N)	DELTA(N)	TAU(N)	DELTA(N)	TAU(N)	DELTA(N)
1	0.0130	1.6327	0.2085	0.7244	0.3433	0.4376
2	0.6171	0.6041	0.6475	0.4390	0.3694	0.0261
3	0.6779	0.0608	1.2863	0.6388	0.5128	0.1434
4	1.6079	0.9300	2.0208	0.7345	0.7779	0.2651
5	2.6292	1.0213	2.6900	0.6692	0.8561	0.0782
6	2.7465	0.1173	2.9898	0.2998	1.1212	0.2651
7	3.2419	0.4954	3.6069	0.6171	1.5080	0.3868
8	3.8242	0.5823	4.2501	0.6432	1.5514	0.0434
9	4.6108	0.7866	4.9758	0.7257	1.8295	0.2781
10	5.2974	0.6866	5.4017	0.4259	2.1772	0.3477
11	5.6538	0.3564	5.8797	0.4780	2.3858	0.2086
12	5.9710	0.3172	6.4794	0.5997	2.5249	0.1391
13	6.9010	0.9300	7.1052	0.6258	2.8334	0.3085
14	7.3703	0.4693	7.5311	0.4259	3.1506	0.3172
15			8.1221	0.5910	3.3158	0.1652
16			8.4741	0.3520	3.4592	0.1434
14					3.8286	0.3694
18					4.5282	0.6996
19					5.1757	0.6475
20					5.4930	0.3173
21					5.6451	0.1521
22					5.7928	0.1477
23					6.1231	0.3303
24					6.7967	0.6736
25					7.4225	0.6258
26					7.9613	0.5388
27					8.3264	0.3651
28					8.8957	0.5693

REF. FIG. 9.5

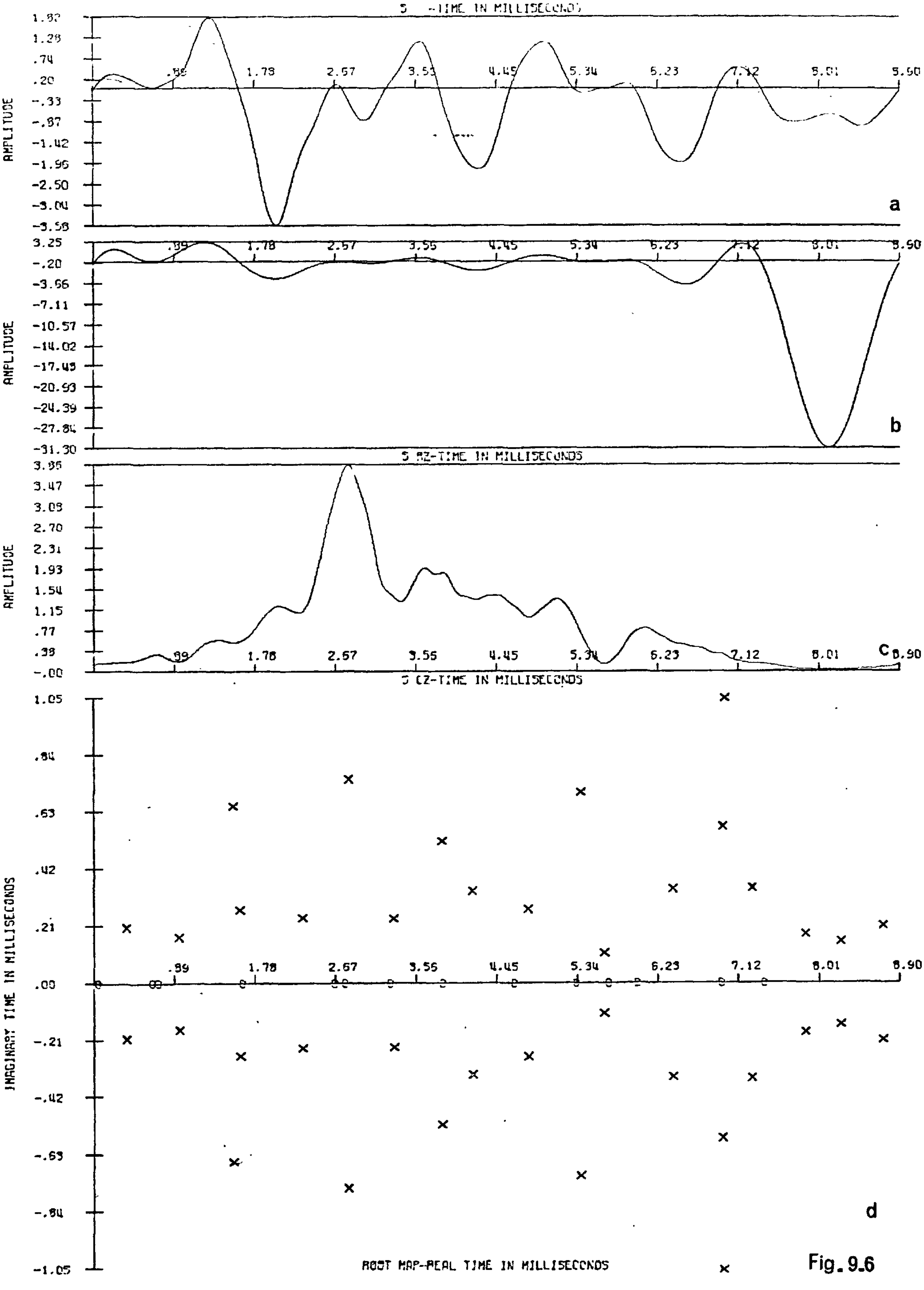
A   B   C

Table 9.3



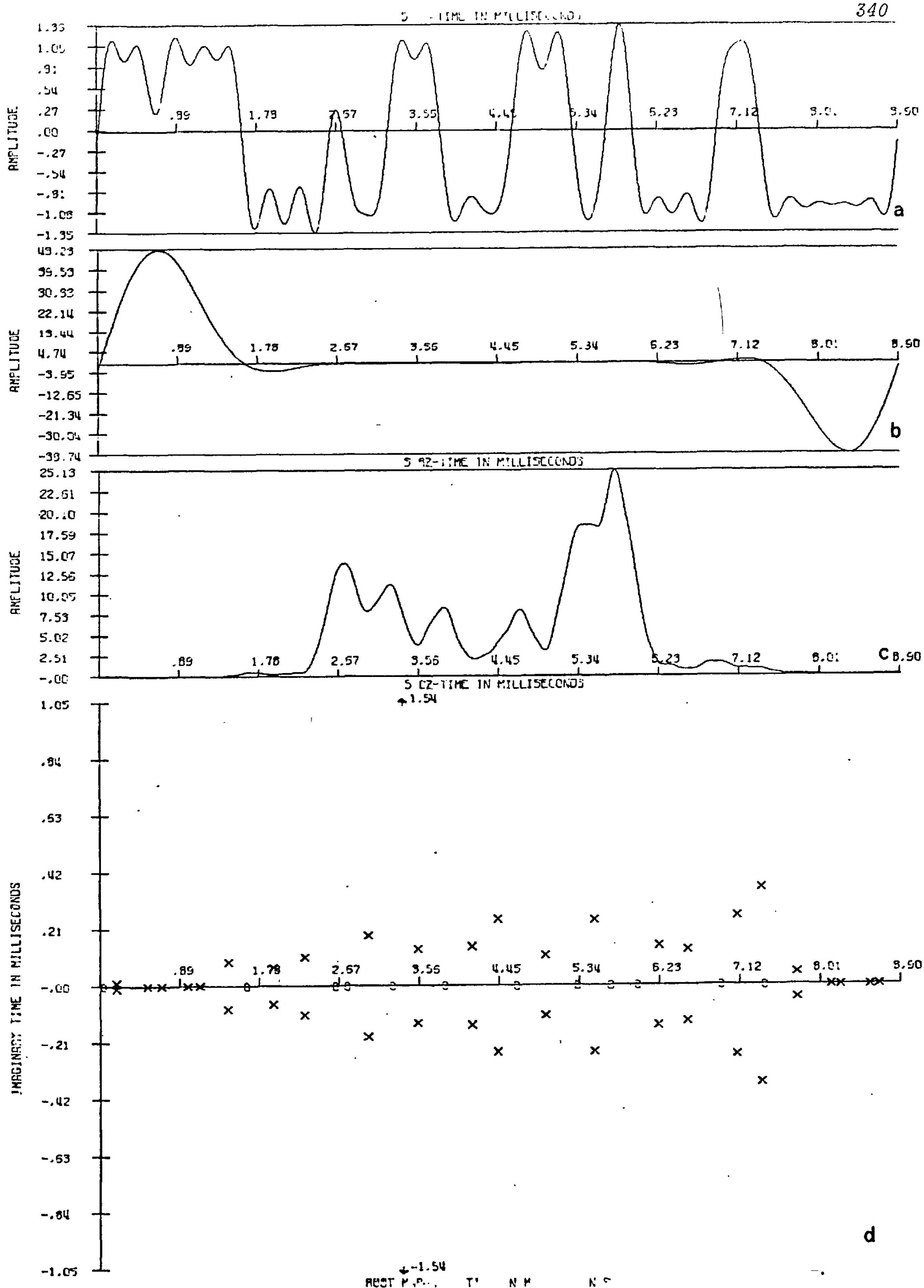
AMPLITUDE
a
NRZRO= 14
S(T)-TIME IN MILLISECONDS
b
NRZRO= 15
D(S(T))/DT- TIME IN MILLISEC
c
NRZRO= 28
D**2(S(T))/DT**2- TIME IN MSEC
d
NRZRO= 56
D**3(S(T))/DT**3- TIME IN MSEC

Fig. 9.5



TIME IN MILLISECONDS
AMPLITUDE
a
b
MZ-TIME IN MILLISECONDS
AMPLITUDE
c
CZ-TIME IN MILLISECONDS
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
d

Fig. 9.6


AMPLITUDE
a
b
5 AZ-TIME IN MILLISECONDS
AMPLITUDE
c
5 CZ-TIME IN MILLISECONDS
IMAGINARY TIME IN MILLISECONDS
d

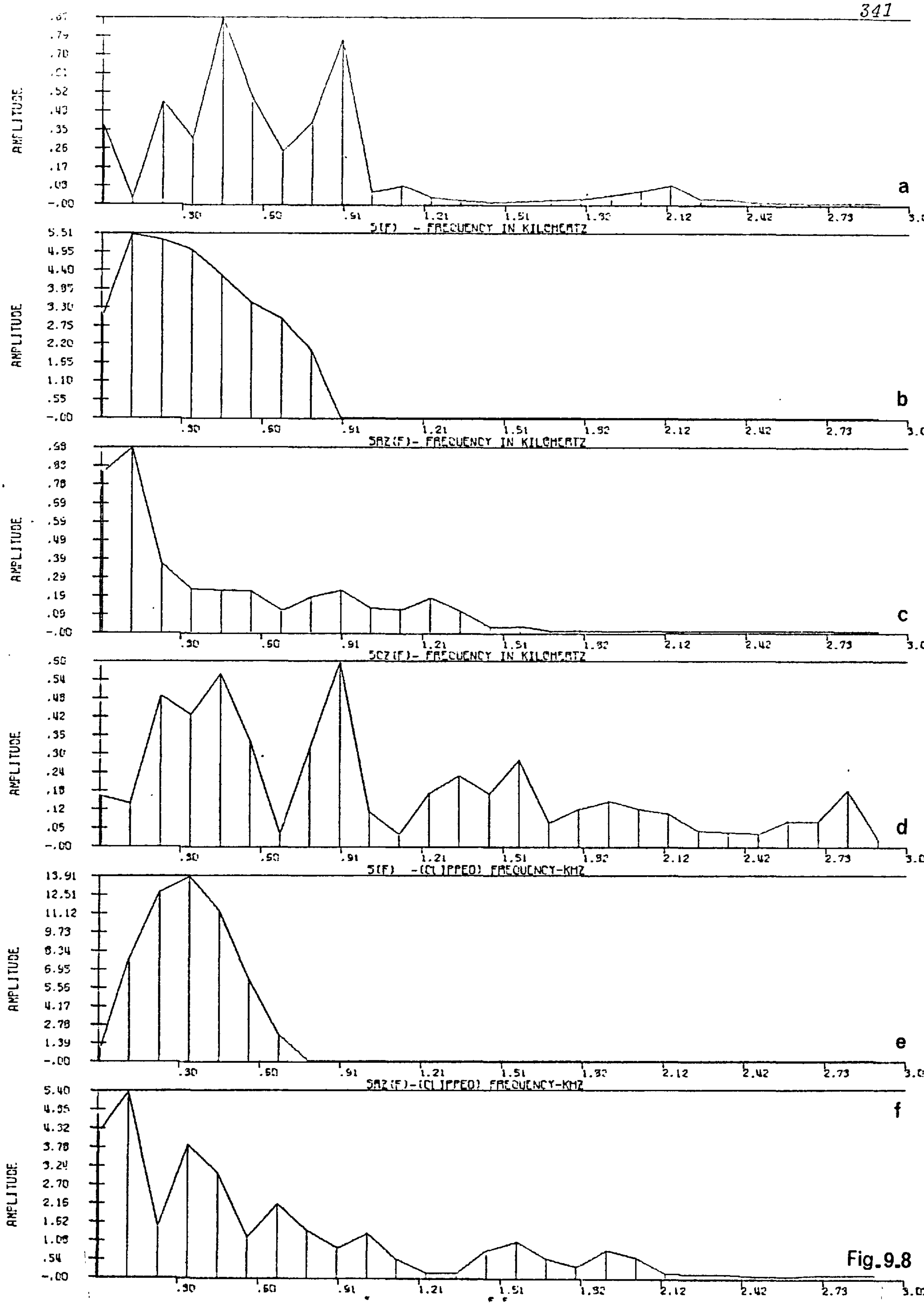

AMPLITUDE
a
.30 .60 .91 1.21 1.51 1.82 2.12 2.42 2.73 3.03
S(F) - FREQUENCY IN KILOHERTZ
AMPLITUDE
b
SAZ(F)- FREQUENCY IN KILOHERTZ
AMPLITUDE
c
SCZ(F)- FREQUENCY IN KILOHERTZ
AMPLITUDE
d
S(F) -(CLIPPED) FREQUENCY-KHZ
AMPLITUDE
e
SAZ(F)-(CLIPPED) FREQUENCY-KHZ
AMPLITUDE
f

Fig.9.8

/o/ - OBEY PERIOD=8.899 MILLISECONDS

ZEROS- TIME/MILLISECONDS

ORIGINAL SIGNAL	CLIPPED AND B.L. SIGNAL
.0130	0.0130
.3553 +/- J 0.2061	0.1886 +/- J 0.0105
.6171	0.5349*
.6779	0.6928*
.9347 +/- J 0.1713	0.9838*
1.5396 +/- J 0.6592	1.1145*
1.6079	1.4397 +/- J 0.0876
1.6130 +/- J 0.2707	1.6166
2.3044 +/- J 0.2411	1.9341 +/- J 0.0680
2.6292	2.2844 +/- J 0.1072
2.7465	2.6074
2.8099 +/- J 0.7571	2.7378
3.2419	2.9944 +/- J 0.1885
3.3041 +/- J 0.2393	3.2419
3.8242	3.3780 +/- J 1.5402
3.8445 +/- J 0.5273	3.5449 +/- J 0.1372
4.1837 +/- J 0.3410	3.8286
4.6108	4.1484 +/- J 0.1450
4.7964 +/- J 0.2721	4.4333 +/- J 0.2458
5.2974	4.6195
5.3728 +/- J 0.7092	4.9669 +/- J 0.1120
5.6324 +/- J 0.1117	5.2931
5.6538	5.5107 +/- J 0.2443
5.9710	5.6625
6.3957 +/- J 0.3487	5.9623
6.9010	6.2209 +/- J 0.1473
6.9592 +/- J 1.0582	6.5414 +/- J 0.1324
6.9592 +/- J 0.5801	6.8923
7.2685 +/- J 0.3531	7.0888 +/- J 0.2569
7.3703	7.3652
7.8529 +/- J 0.1817	7.3653 +/- J 0.3610
8.2461 +/- J 0.1548	7.7532 +/- J 0.0470
8.7147 +/- J 0.2119	8.1305*
	8.2406*
	8.5674*
	8.6669*
REF. FIG. 9.6	REF. FIG. 9.7

* COMPLEX ZEROS WITH '0' IMAGINARY COMPONENT DUE TO IMPERFECT FACTORISATION

NOTE: THIS SIGNAL HAD TWO REAL ZEROS CONVERTED INTO A COMPLEX ZERO BY THE LOW PASS FILTERING FOLLOWING CLIPPING. TOTAL NUMBER OF ZEROS REMAINS CONSTANT.

Table 9.4

/ʌ/ - BUT PERIOD=10.23 MILLISECONDS

REAL ZEROS- TIME/MILLISECONDS DELTA(N)=TAU(N)-TAU(N-1)

	S(T)		S'(T)		S''(T)	
N	TAU(N)	DELTA(N)	TAU(N)	DELTA(N)	TAU(N)	DELTA(N)
1	0.0150	0.5045	0.4945	0.6643	0.1249	0.1549
2	0.7143	0.6993	1.0240	0.5295	0.3547	0.2298
3	1.2887	0.5744	1.4636	0.4396	0.6743	0.3196
4	1.8782	0.5895	1.9581	0.4945	1.2538	0.5795
5	2.0280	0.1498	2.1679	0.2098	1.5685	0.3147
6	2.2728	0.2448	2.4876	0.3197	1.6784	0.1099
7	2.7823	0.5095	3.0220	0.5344	1.8282	0.1498
8	3.2568	0.4745	3.4017	0.3797	2.0680	0.2398
9	3.8163	0.5595	3.9602	0.5585	2.3177	0.2497
10	4.0061	0.1898	4.0860	0.1258	2.8472	0.5295
11	4.1509	0.1448	4.4117	0.3257	3.1869	0.3397
12	4.7054	0.5545	4.9602	0.5485	3.4766	0.2897
13	5.1749	0.4695	5.4896	0.5294	3.5865	0.1099
14	6.1140	0.9391	5.8343	0.3447	3.7763	0.1898
15	6.6585	0.5445	5.9991	0.1648	4.0011	0.2248
16	7.2229	0.5644	6.3438	0.3447	4.2508	0.2497
17	7.8473	0.6244	6.9082	0.5644	4.8053	0.5545
18	8.5816	0.7343	7.5027	0.5945	5.1200	0.3147
19	9.1610	0.5794	8.1870	0.6843	5.6745	0.5545
20	9.7405	0.5795	8.8364	0.6494	5.9192	0.2447
21			9.4308	0.5944	6.1740	0.2548
22			10.0602	0.6294	6.7534	0.5794
23					7.0481	0.2947
24					7.2229	0.1748
25					7.3228	0.0999
26					7.6775	0.3547
27					7.7974	0.1199
28					8.0022	0.2048
29					8.3319	0.3297
30					8.4318	0.0999
31					8.6466	0.2148
32					9.1961	0.5495
33					9.6506	0.4545
34					9.8504	0.1998
35					9.9503	0.0999
36					10.2000	0.2497

REF. FIG. 9.9

A B C

Table 9.5

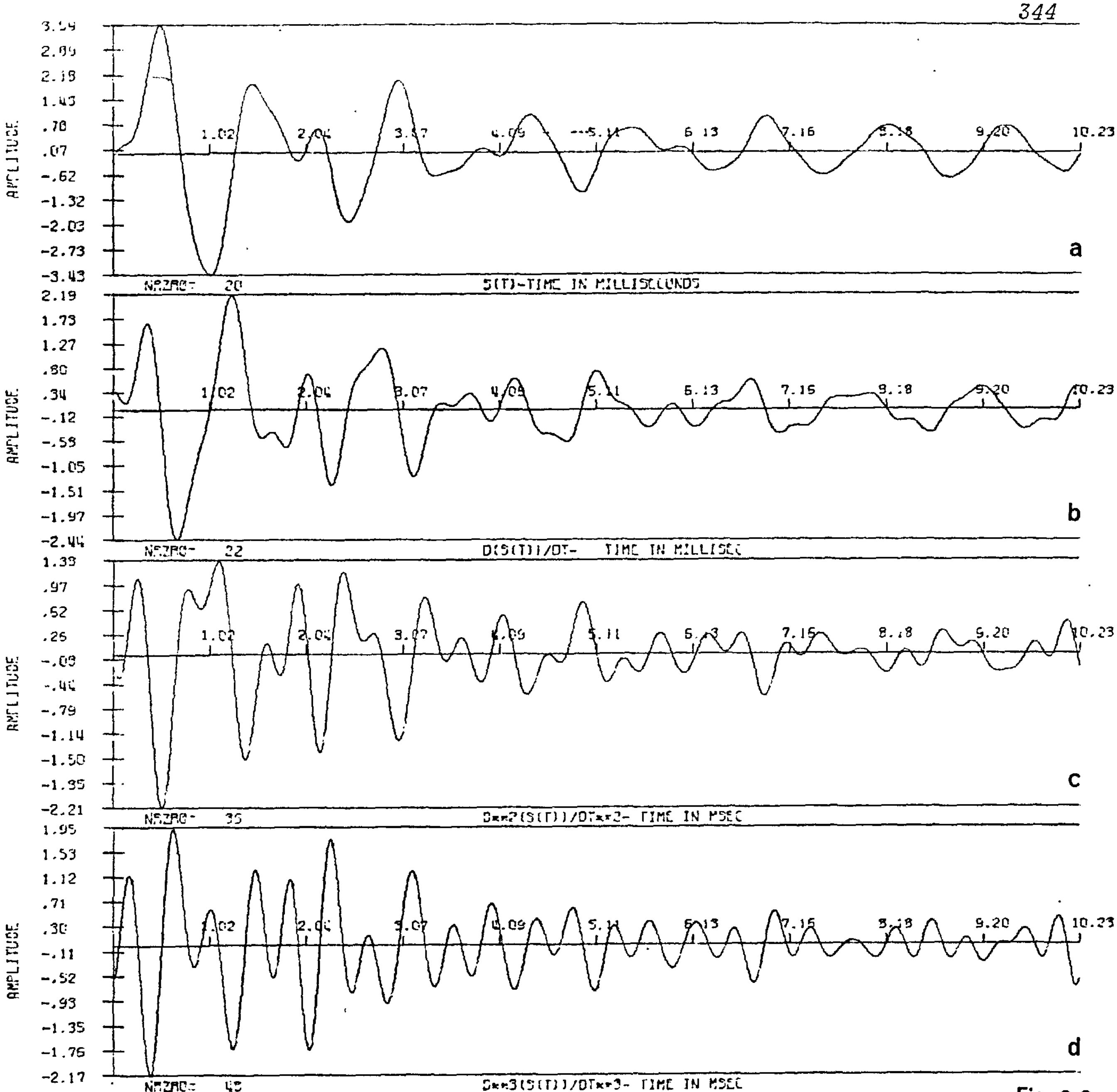

AMPLITUDE
S(T)-TIME IN MILLISECONDS
D(S(T))/DT- TIME IN MILLISEC
D**2(S(T))/DT**2- TIME IN MSEC
D**3(S(T))/DT**3- TIME IN MSEC
a
b
c
d

Fig.9.9

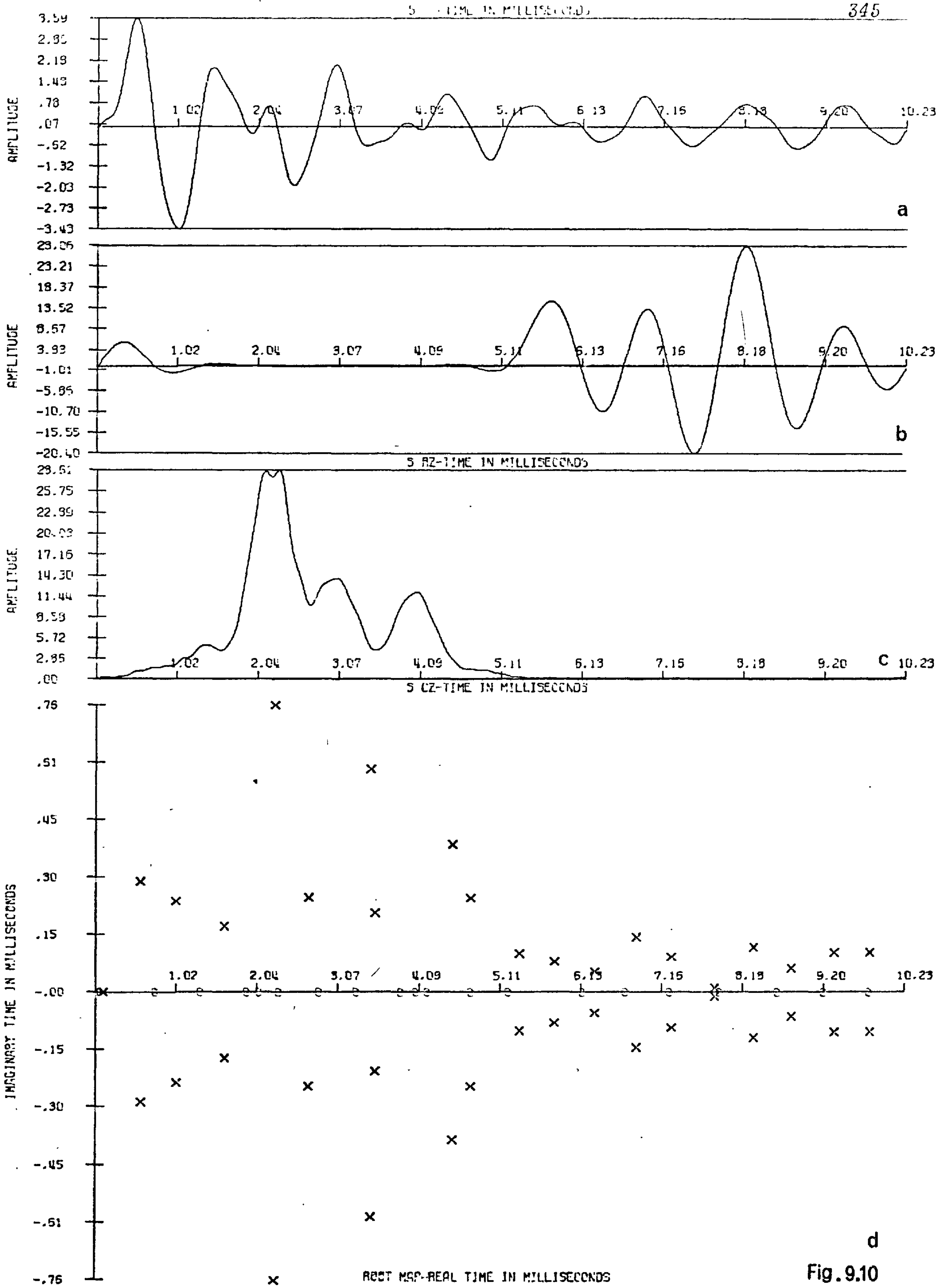

AMPLITUDE
a
AMPLITUDE
b
5 RZ-TIME IN MILLISECONDS
AMPLITUDE
c
5 CZ-TIME IN MILLISECONDS
IMAGINARY TIME IN MILLISECONDS
d
ROOT MAP-REAL TIME IN MILLISECONDS
1.02
2.04
3.07
4.09
5.11
6.13
7.15
8.18
9.20
10.23

Fig. 9.10

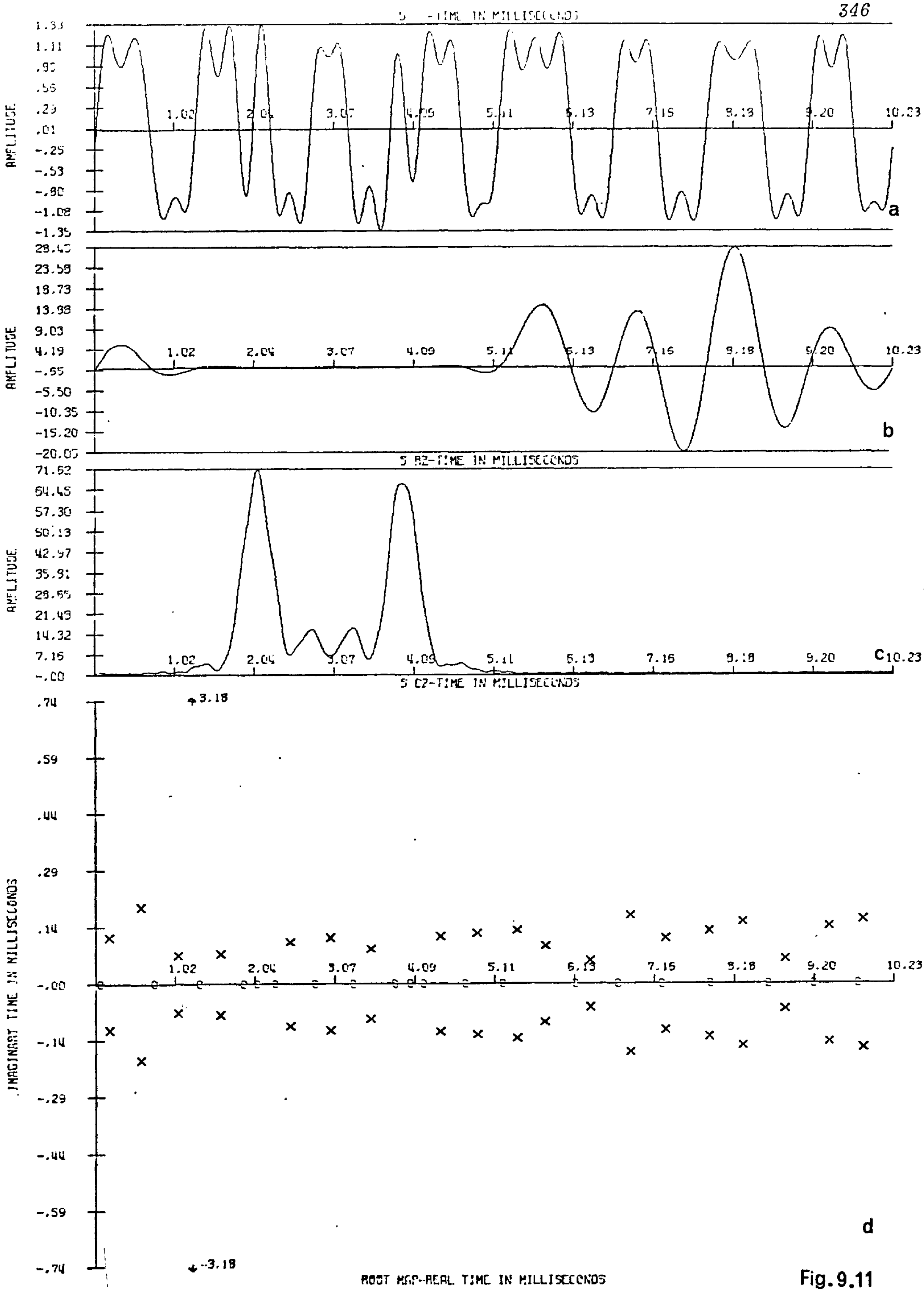

5 -TIME IN MILLISECONDS
AMPLITUDE
a
b
5 RZ-TIME IN MILLISECONDS
c
5 CZ-TIME IN MILLISECONDS
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
d

Fig. 9.11

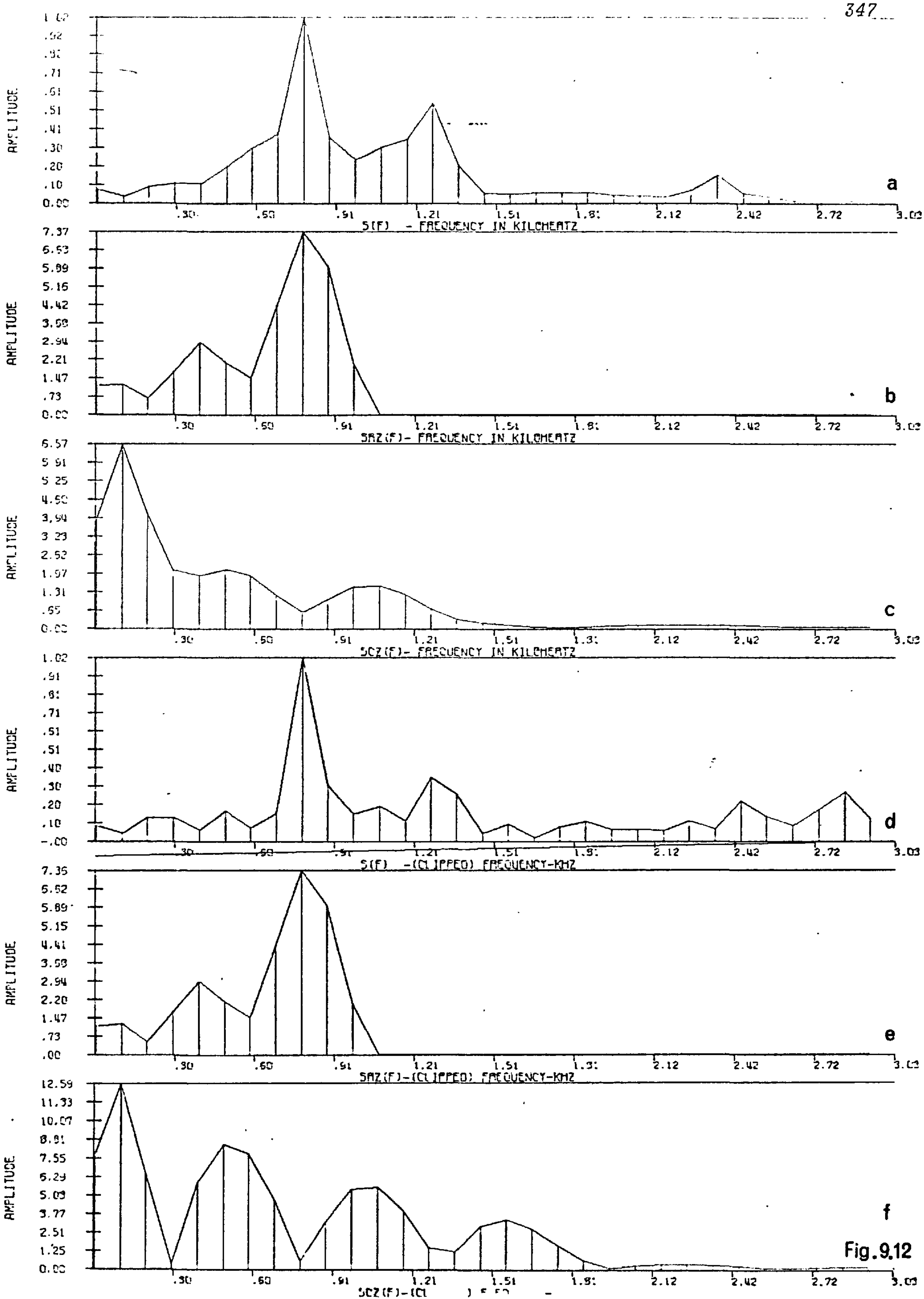

AMPLITUDE
S(F) - FREQUENCY IN KILOHERTZ
SAZ(F)- FREQUENCY IN KILOHERTZ
SCZ(F)- FREQUENCY IN KILOHERTZ
S(F) -(CLIPPED) FREQUENCY-KHZ
SAZ(F)-(CLIPPED) FREQUENCY-KHZ
a
b
c
d
e
f

Fig. 9.12

ZEROS- TIME/MILLISECONDS

ORIGINAL SIGNAL	CLIPPED AND B.L. SIGNAL
.0150	0.0200
.1035 +/- J 0.0001	0.1693 +/- J 0.1217
.5661 +/- J 0.2942	0.5806 +/- J 0.2003
.7143	0.7143
1.0190 +/- J 0.2422	1.0560 +/- J 0.0671
1.2887	1.2375
1.6263 +/- J 0.1751	1.2376 +/- J 3.1827
1.8782	1.5990 +/- J 0.0802
2.0280	1.8632
2.2622 +/- J 0.7646	2.0380
2.2728	2.2678
2.6883 +/- J 0.2516	2.4934 +/- J 0.1100
2.7823	2.7773
3.2568	3.0054 +/- J 0.1212
3.4742 +/- J 0.5952	3.2618
3.4743 +/- J 0.2079	3.5272 +/- J 0.0912
3.8163	3.8013
4.0061	4.0011
4.1509	4.1659
4.5138 +/- J 0.3938	4.4186 +/- J 0.1244
4.7054	4.7054
4.7454 +/- J 0.2506	4.8898 +/- J 0.1326
5.1749	5.1848
5.3596 +/- J 0.1025	5.4002 +/- J 0.1410
5.8015 +/- J 0.0821	5.7576 +/- J 0.0995
6.1140	6.1090
6.3118 +/- J 0.0551	6.3365 +/- J 0.0614
6.6585	6.6585
6.8334 +/- J 0.1467	6.8469 +/- J 0.1791
7.2229	7.2229
7.2653 +/- J 0.0941	7.2977 +/- J 0.1192
7.8273 +/- J 0.0113	7.8608 +/- J 0.1385
7.8473	7.8472
8.3198 +/- J 0.1198	8.2818 +/- J 0.1617
8.5816	8.5816
8.7979 +/- J 0.0647	8.8279 +/- J 0.0656
9.1610	9.1610
9.3388 +/- J 0.1060	9.3933 +/- J 0.1510
9.7405	9.7355
9.7963 +/- J 0.1054	9.8317 +/- J 0.1681
REF. FIG. 9.10	REF. FIG. 9.11

Table 9.6

/e/ - HATE PERIOD=9.43 MILLISECONDS

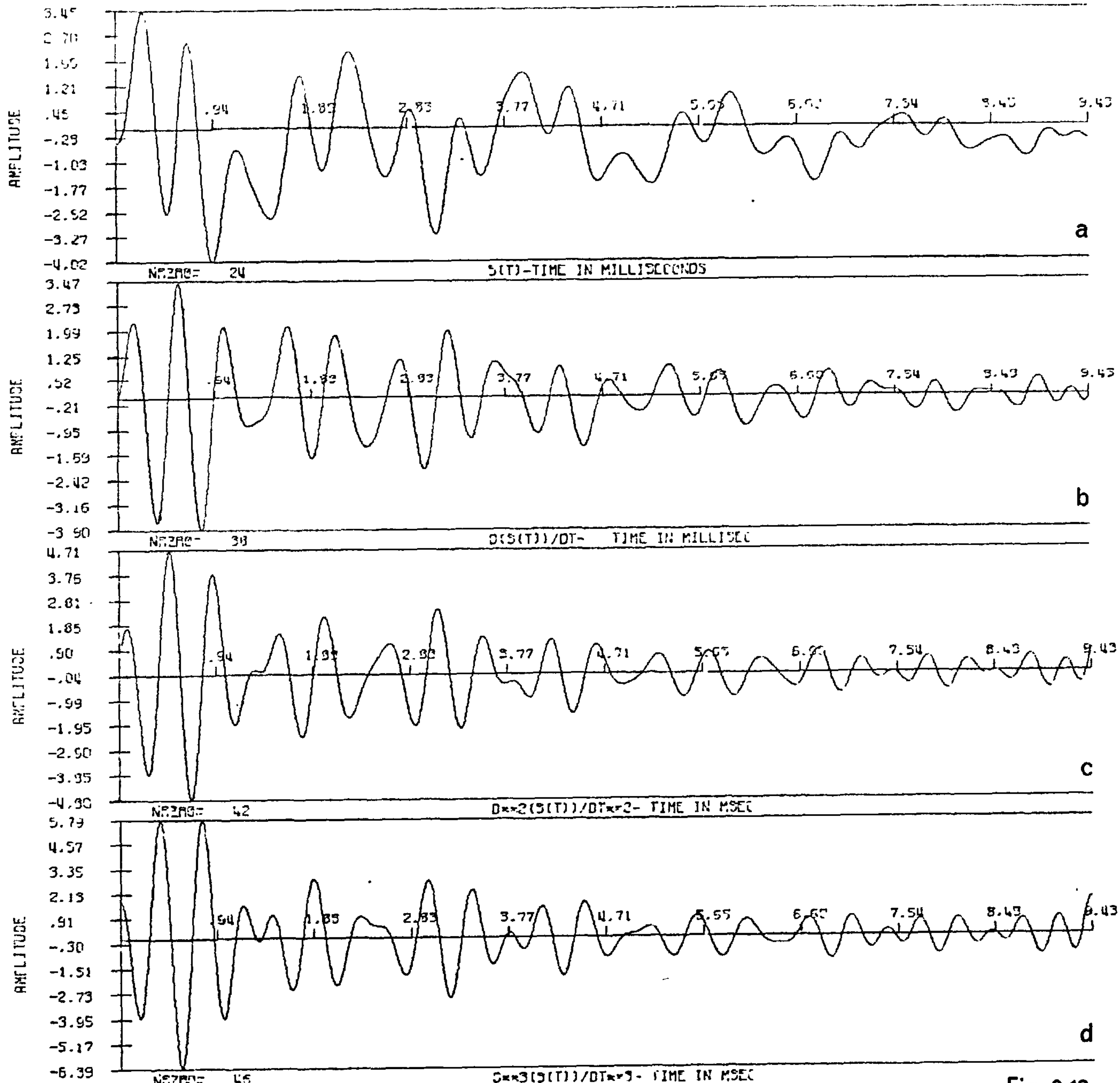
REAL ZEROS- TIME/MILLISECONDS DELTA(N)=TAU(N)-TAU(N-1)

	S(T)		S'(T)		S''(T)	
N	TAU(N)	DELTA(N)	TAU(N)	DELTA(N)	TAU(N)	DELTA(N)
1	0.0737	1.4090	0.2579	0.2580	0.1566	0.2026
2	0.3868	0.3131	0.4881	0.2302	0.3776	0.2210
3	0.5939	0.2071	0.6953	0.2072	0.5940	0.2164
4	0.7919	0.1980	0.9347	0.2394	0.8058	0.2118
5	1.6760	0.8841	1.1603	0.2256	1.0314	0.2256
6	1.8970	0.2210	1.4919	0.3316	1.2616	0.2302
7	2.1043	0.2073	1.7819	0.2900	1.6576	0.3960
8	2.4588	0.3545	2.0030	0.2211	1.8924	0.2348
9	2.7765	0.3177	2.2654	0.2624	2.1227	0.2303
10	2.9192	0.1427	2.6154	0.3500	2.4220	0.2993
11	3.2968	0.3776	2.8502	0.2348	2.7489	0.3269
12	3.3935	0.0967	3.1034	0.2532	2.9837	0.2348
13	3.6974	0.3039	3.3429	0.2395	3.2185	0.2348
14	4.4190	0.7166	3.5459	0.2030	3.4350	0.2165
15	4.2546	0.1594	3.9460	0.4001	3.6652	0.2302
16	4.5170	0.2624	4.1993	0.2533	4.0842	0.4190
17	5.4103	0.8933	4.3973	0.1980	4.3006	0.2164
18	5.5899	0.1796	4.6781	0.2808	4.5262	0.2256
19	5.8063	0.2164	4.9084	0.2303	4.7703	0.2441
20	6.1102	0.3039	5.1893	0.2809	5.0695	0.2992
21	7.4132	1.3030	5.4978	0.3085	5.3504	0.2809
22	7.7355	0.3223	5.7050	0.2072	5.5991	0.2487
23	7.9611	0.2256	5.9582	0.2532	5.8339	0.2348
24	8.0947	0.1336	6.2713	0.3131	6.0963	0.2624
25			6.5015	0.2302	6.3864	0.2901
26			6.7732	0.2717	6.6581	0.2717
27			7.0219	0.2487	6.8975	0.2394
28			7.1968	0.1749	7.1047	0.2072
29			7.6204	0.4236	7.3027	0.1980
30			7.8461	0.2257	7.4685	0.1685
31			8.0302	0.1841	7.5329	0.0644
32			8.2743	0.2441	7.7448	0.2119
33			8.6012	0.3269	7.9381	0.1933
34			8.8268	0.2256	8.1407	0.2026
35			9.0432	0.2164	8.3525	0.2118
36			9.1952	0.1520	8.4769	0.1244
37			9.3241	0.1289	8.5229	0.0460
38			9.4299	0.1058	8.7347	0.2118
39					8.9327	0.1980
40					9.1169	0.1842
41					9.2596	0.1427
42					9.3840	0.1244

REF. FIG. 9.13

A B C

Table 9.7



AMPLITUDE
S(T)-TIME IN MILLISECONDS
NAZMO= 24
a
D(S(T))/DT- TIME IN MILLISEC
NAZMO= 38
b
D**2(S(T))/DT**2- TIME IN MSEC
NAZMO= 42
c
D**3(S(T))/DT**3- TIME IN MSEC
NAZMO= 45
d

Fig. 9.13

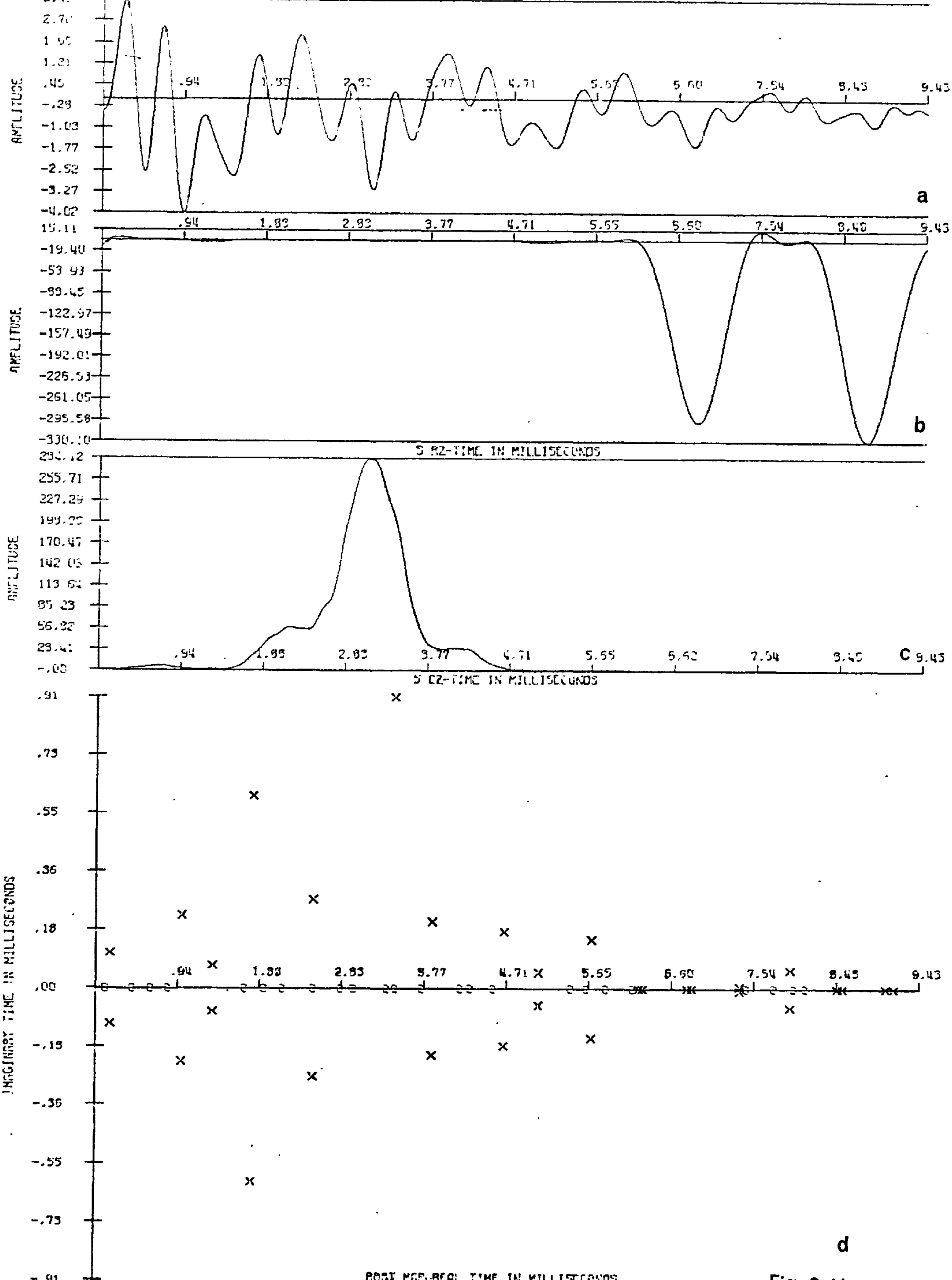

a
b
c
d
AMPLITUDE
S RZ-TIME IN MILLISECONDS
S CZ-TIME IN MILLISECONDS
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS

Fig. 9.14

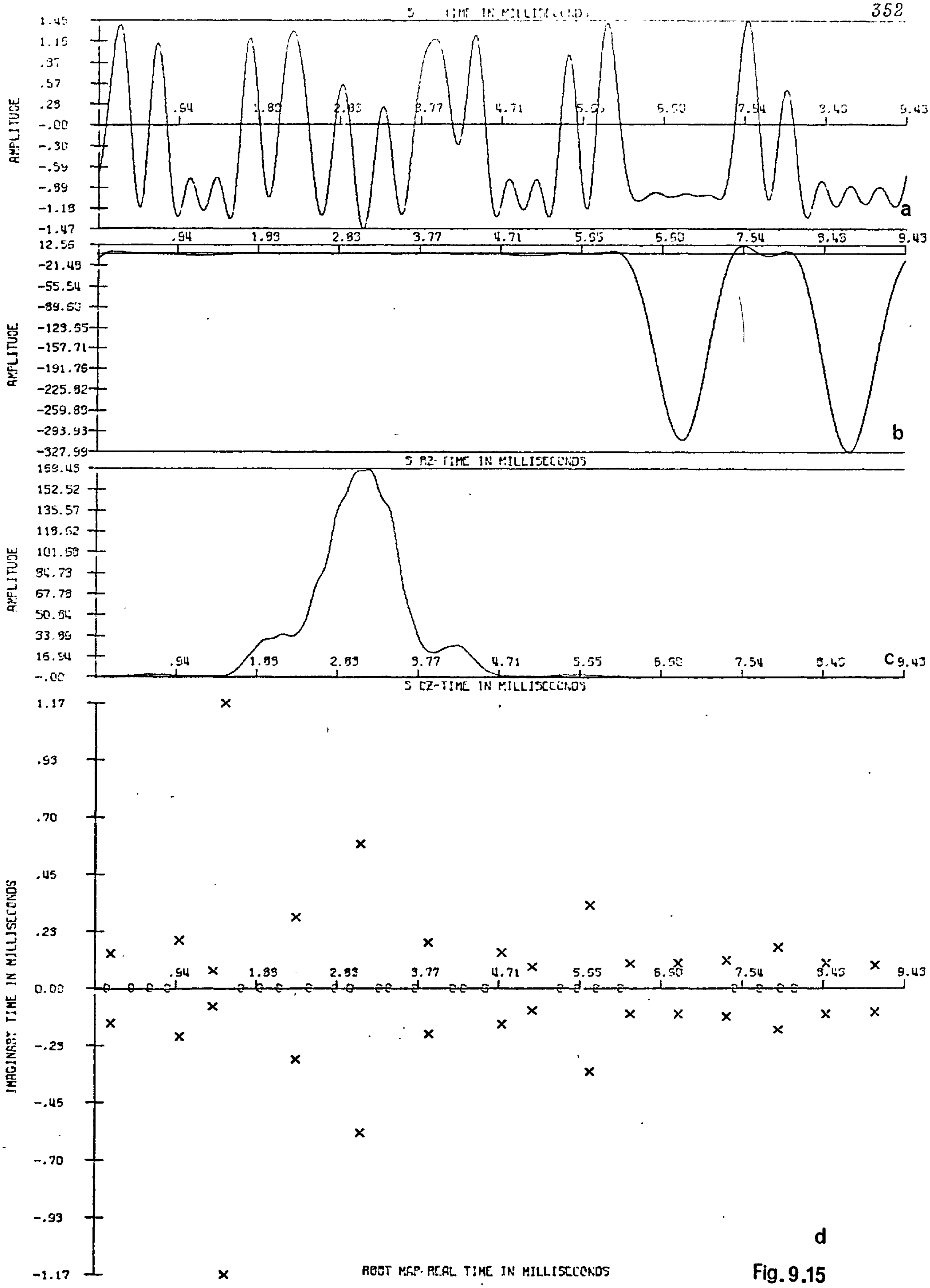

S TIME IN MILLISECONDS
AMPLITUDE
a
b
S RZ-TIME IN MILLISECONDS
c
S CZ-TIME IN MILLISECONDS
IMAGINARY TIME IN MILLISECONDS
d
ROOT MAP-REAL TIME IN MILLISECONDS

Fig. 9.15

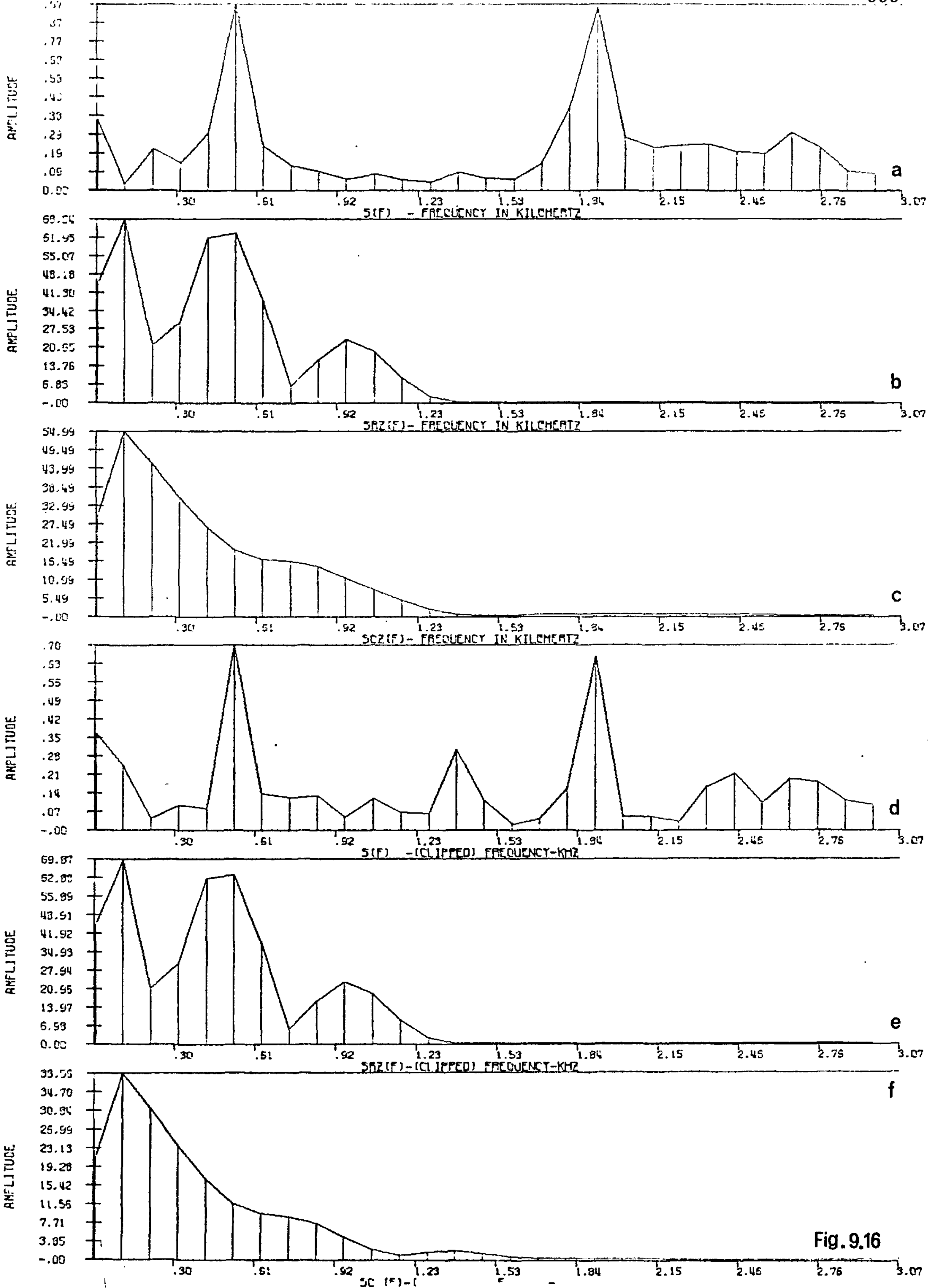

a
b
c
d
e
f
AMPLITUDE
S(F) - FREQUENCY IN KILOHERTZ
SRZ(F)- FREQUENCY IN KILOHERTZ
SCZ(F)- FREQUENCY IN KILOHERTZ
S(F) -(CLIPPED) FREQUENCY-KHZ
SRZ(F)-(CLIPPED) FREQUENCY-KHZ
.30 .61 .92 1.23 1.53 1.84 2.15 2.46 2.76 3.07

Fig. 9.16

ZEROS- TIME/MILLISECONDS

.0737		0.0829	
.1600	+/- J 0.1107	0.1645	+/- J 0.1428
.3868		0.3776	
.5940		0.5940	
.7920		0.8012	
.9825	+/- J 0.2289	0.9698	+/- J 0.1987
1.3323	+/- J 0.0729	1.3634	+/- J 0.0729
1.6760		1.5062	+/- J 1.1748
1.7984	+/- J 0.6161	1.6668	
1.8971		1.8878	
2.1043		2.1135	
2.4588		2.3357	+/- J 0.2924
2.4864	+/- J 0.2786	2.4588	
2.7765		2.7673	
2.9192		2.9423	
3.2968		3.0900	+/- J 0.5957
3.3935		3.2738	
3.4121	+/- J 0.9176	3.3889	
3.6974		3.6928	
3.8456	+/- J 0.2101	3.8837	+/- J 0.1885
4.1440		4.1394	
4.2546		4.2638	
4.5170		4.5216	
4.6746	+/- J 0.1794	4.7369	+/- J 0.1471
5.0672	+/- J 0.0524	5.0942	+/- J 0.0907
5.4103		5.3965	
5.5899		5.5945	
5.6784	+/- J 0.1546	5.7657	+/- J 0.3430
5.8063		5.8155	
6.1102		6.0963	
6.2159*		6.2375	+/- J 0.1040
6.2589*		6.7963	+/- J 0.1050
6.7776*		7.3573	+/- J 0.1167
6.8284*		7.4270	
7.4132		7.7170	
7.3552	+/- J 0.0074	7.9558	
7.7355		7.9599	+/- J 0.1699
7.9611		8.1131	
7.9499	+/- J 0.0600	8.5112	+/- J 0.1055
8.0947		9.0838	+/- J 0.0965
8.4905*			
8.5498*			
9.0454*			
9.1393*			

REF. FIG. 9.14 REF. FIG. 9.15

*COMPLEX ZEROS WITH '0' IMAGINARY COMPONENT.DUE TO IMPERFECT FACTORISATION

Table 9.8

/Ɛ/ - BET PERIOD=9.455 MILLISECONDS

REAL ZEROS- TIME/MILLISECONDS DELTA(N)=TAU(N)-TAU(N-1)

	S(T)		S'(T)		S''(T)	
N	TAU(N)	DELTA(N)	TAU(N)	DELTA(N)	TAU(N)	DELTA(N)
1	0.8219	0.9509	0.1062	0.1198	0.0508	0.1613
2	1.0388	0.2169	0.3278	0.2216	0.2216	0.1708
3	1.1911	0.1523	0.3970	0.0692	0.3647	0.1431
4	1.7082	0.5171	0.6648	0.2678	0.5633	0.1986
5	1.9345	0.2263	0.9237	0.2589	0.7987	0.2354
6	2.1191	0.1846	1.1227	0.1990	1.0157	0.2170
7	2.4700	0.3509	1.3666	0.2439	1.2281	0.2124
8	2.9132	0.4432	1.8236	0.4570	1.4451	0.2170
9	3.0148	0.1016	2.0314	0.2078	1.5328	0.0877
10	3.3518	0.3370	2.2761	0.2447	1.7129	0.1801
11	3.6058	0.2540	2.5577	0.2816	1.9298	0.2169
12	3.7627	0.1569	2.6546	0.0969	2.1515	0.2217
13	4.2798	0.5171	2.7793	0.1247	2.3869	0.2354
14	4.4921	0.2123	2.9686	0.1893	2.5993	0.2124
15	4.6860	0.1939	3.2087	0.2401	2.7193	0.1200
16	5.0785	0.3925	3.4673	0.2586	2.8855	0.1622
17	5.9003	0.8218	3.6796	0.2123	3.1025	0.2170
18	6.7590	0.8587	4.1136	0.4340	3.3334	0.2309
19	7.5347	0.7757	4.3814	0.2678	3.5642	0.2308
20	8.0425	0.5078	4.5892	0.2078	3.7766	0.2124
21	8.1896	0.1471	4.8292	0.2400	3.9566	0.1800
22	8.2964	0.1068	5.3186	0.4894	4.0490	0.0924
23	9.1506	0.8542	5.5033	0.1847	4.2660	0.2170
24	9.3260	0.1754	5.7295	0.2262	4.4876	0.2216
25			5.9927	0.2632	4.7046	0.2170
26			6.1681	0.1754	4.9354	0.2308
27			6.3620	0.1939	5.1062	0.1708
28			6.5098	0.1478	5.2170	0.1108
29			6.6298	0.1200	5.4156	0.1986
30			6.9530	0.3232	5.6233	0.2077
31			7.1700	0.2170	5.8496	0.2263
32			7.3593	0.1893	6.0712	0.2216
33			7.6547	0.2954	6.2558	0.1846
34			8.1026	0.4479	6.4312	0.1754
35			8.2411	0.1385	6.5744	0.1432
36			8.4858	0.2447	6.7960	0.2216
37			8.8181	0.3323	7.0546	0.2586
38			9.0398	0.2217	7.2715	0.2169
39			9.2429	0.2031	7.5023	0.2308
40			9.4414	0.1985	7.7332	0.2309
41					7.8717	0.1386
42					8.0102	0.1385
43					8.1764	0.1662
44					8.3704	0.1990
45					8.5873	0.2169
46					8.9705	0.3832
47					9.1552	0.1847
48					9.3445	0.1893

REF. FIG. 9.17

A B C

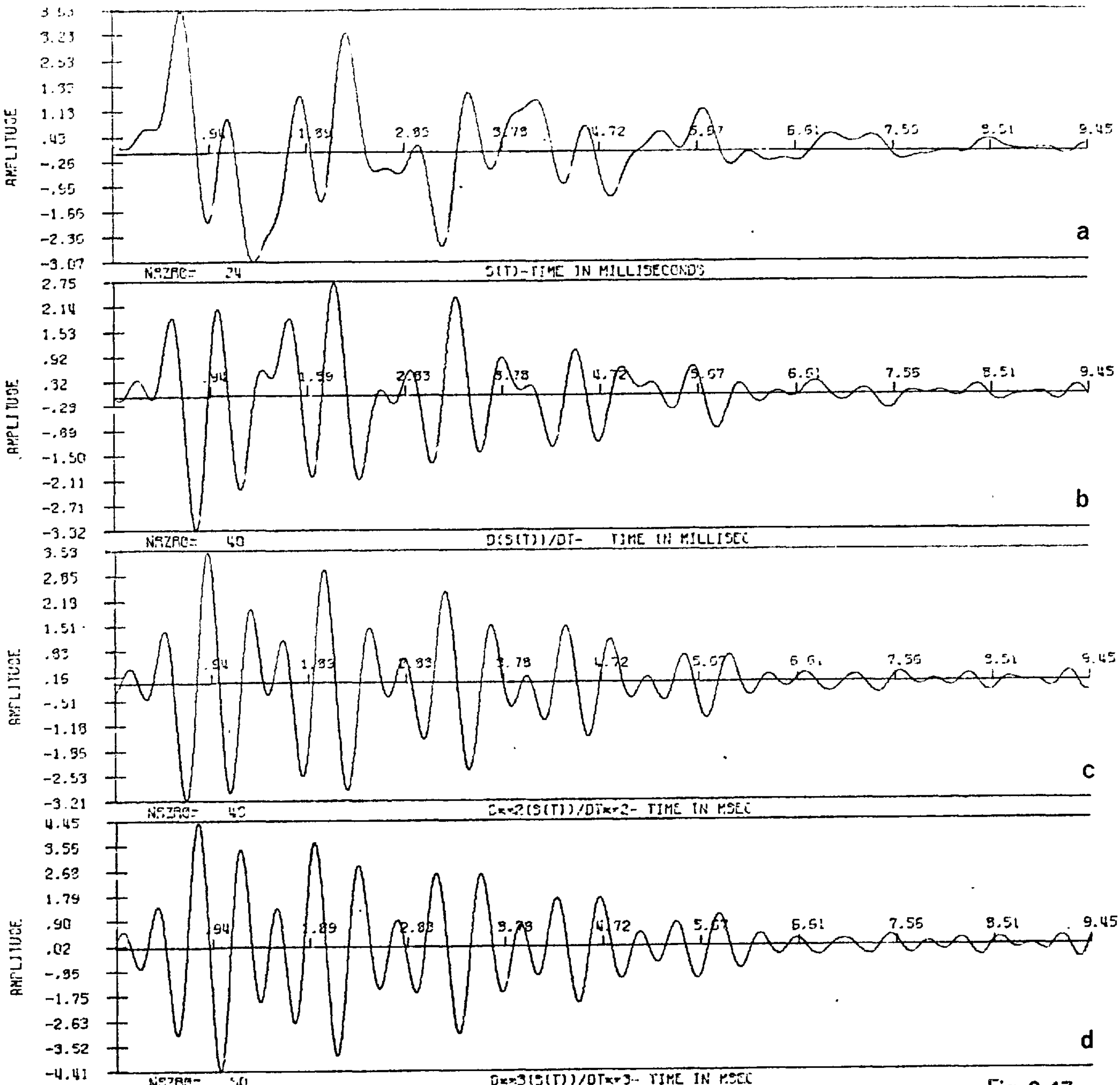
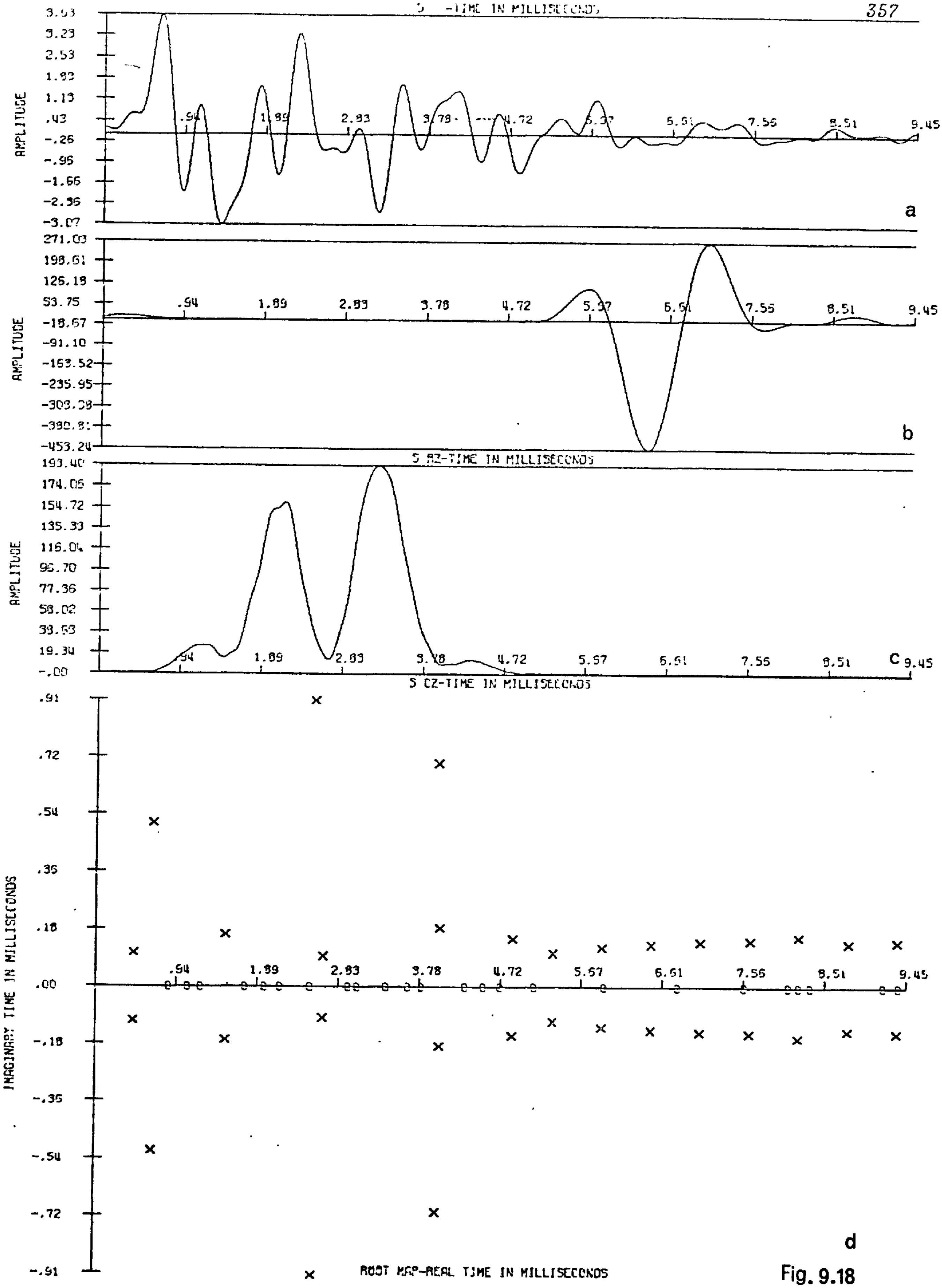

AMPLITUDE
NAZRO= 24
S(T)-TIME IN MILLISECONDS
a
NAZRO= 40
D(S(T))/DT- TIME IN MILLISEC
b
NAZRO= 40
c
NAZRO= 50
d

Fig.9.17



TIME IN MILLISECONDS
AMPLITUDE
.94
1.89
2.83
3.78
4.72
5.67
6.61
7.56
8.51
9.45
a
b
S AZ-TIME IN MILLISECONDS
c
S CZ-TIME IN MILLISECONDS
IMAGINARY TIME IN MILLISECONDS
d
ROOT MAP-REAL TIME IN MILLISECONDS

Fig. 9.18

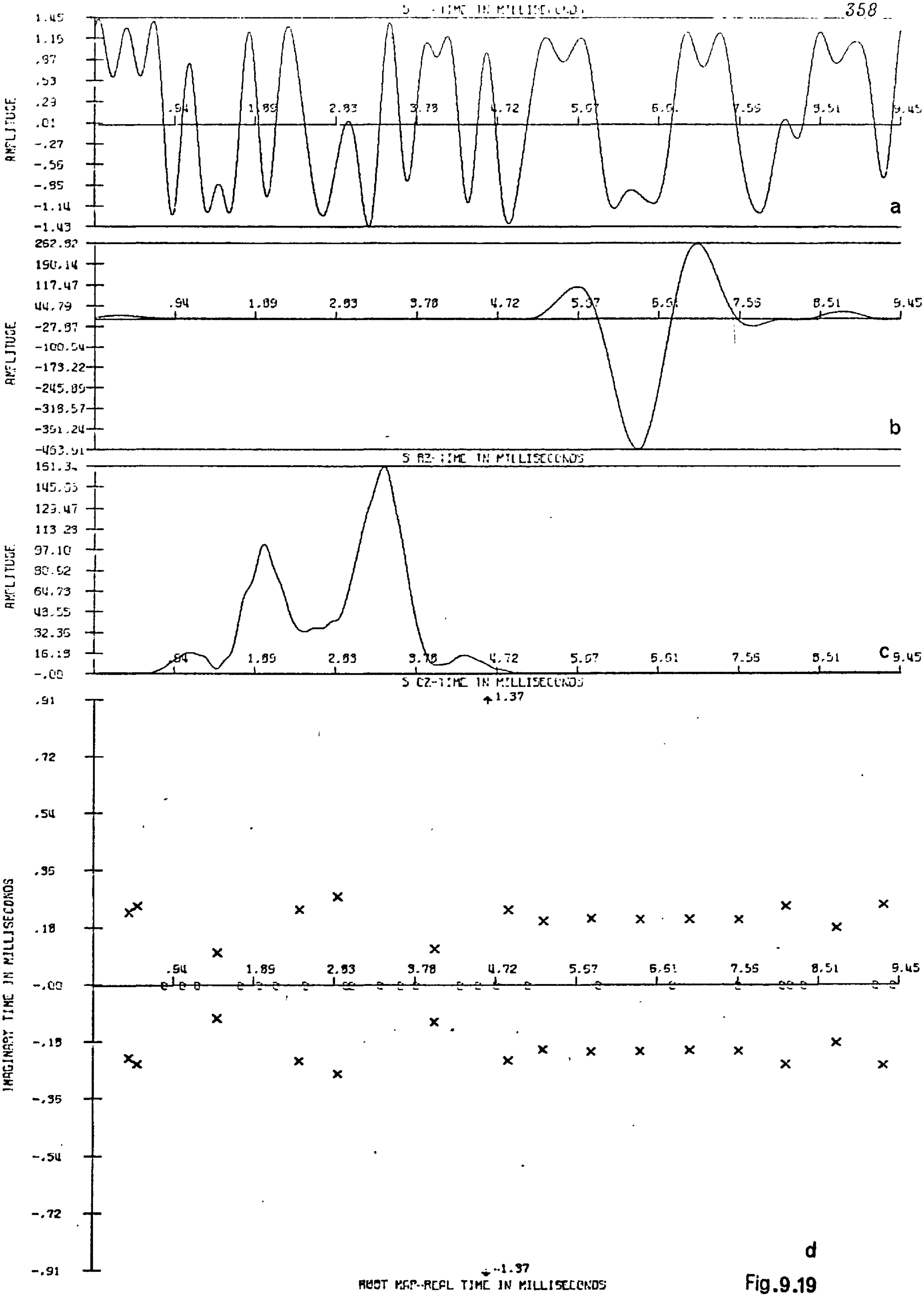

AMPLITUDE
a
b
S RZ-TIME IN MILLISECONDS
c
S CZ-TIME IN MILLISECONDS
IMAGINARY TIME IN MILLISECONDS
ROOT MAP-REAL TIME IN MILLISECONDS
d

Fig.9.19

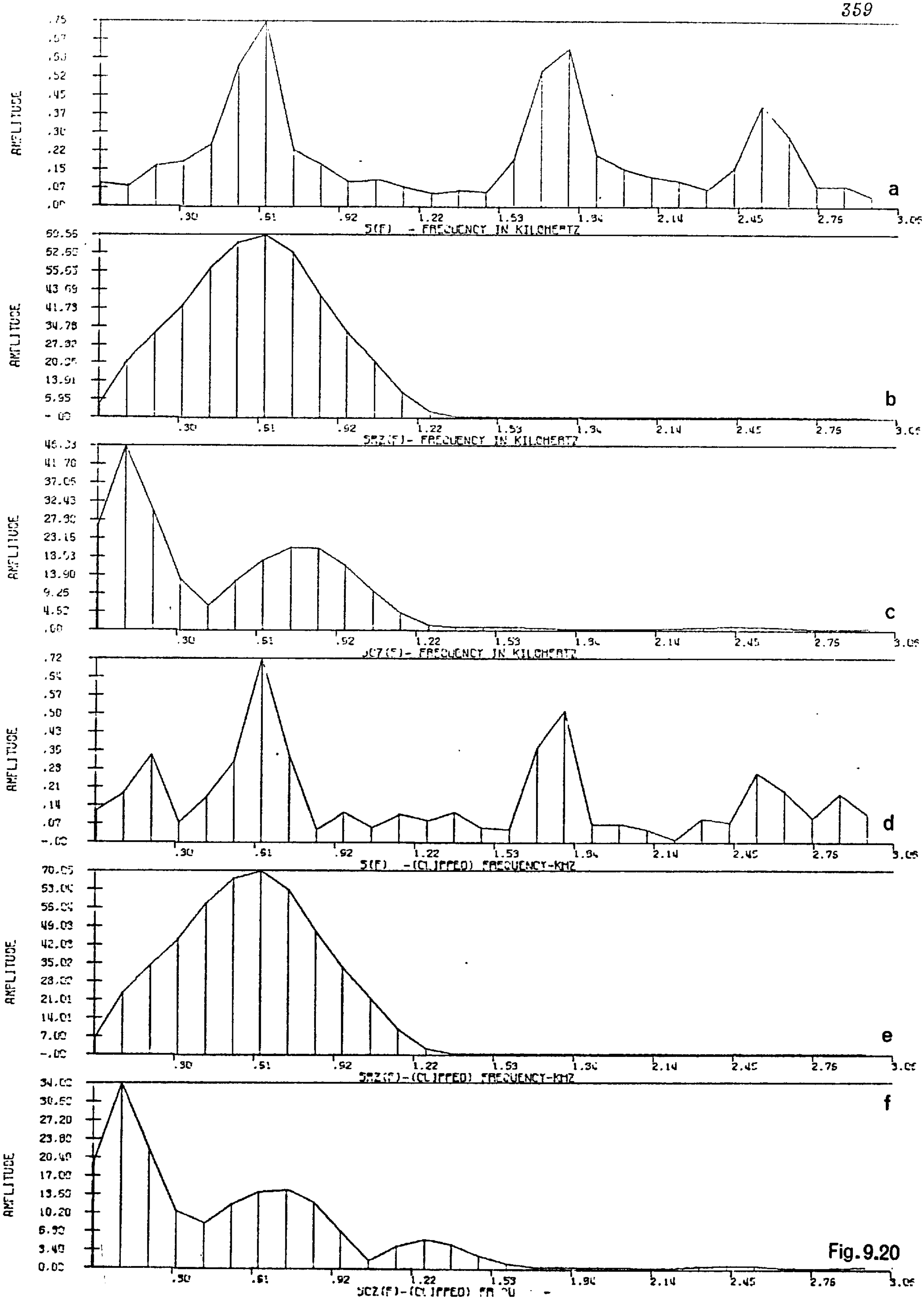

AMPLITUDE
S(F) - FREQUENCY IN KILOHERTZ
a
b
c
d
S(F) -(CLIPPED) FREQUENCY-KHZ
e
f

Fig. 9.20

ZEROS- TIME/MILLISECONDS

ORIGINAL SIGNAL	CLIPPED AND B.L. SIGNAL
.4438 +/- J 0.1085	0.4261 +/- J 0.2322
.6673 +/- J 0.5192	0.5238 +/- J 0.2518
.8218	0.8125
1.0388	1.0249
1.1911	1.2003
1.5090 +/- J 0.1676	1.4598 +/- J 0.1043
1.7082	1.7082
1.9344	1.9279
2.1191	2.1238
2.4699	2.4264
2.5403 +/- J 0.9121	2.4265 +/- J 0.2417
2.6501 +/- J 0.0989	2.8783 +/- J 0.2822
2.9131	2.8784
3.0147	3.0147
3.3517	3.3471
3.6056	3.5826
3.7626	3.7626
3.9835 +/- J 0.7113	4.0107 +/- J 0.1179
3.9836 +/- J 0.1782	4.2751
4.2797	4.4920
4.4920	4.6258 +/- J 1.3757
4.6859	4.7044
4.8656 +/- J 0.1547	4.8716 +/- J 0.2405
5.0784	5.0738
5.3343 +/- J 0.1096	5.2812 +/- J 0.2045
5.9003	5.8462 +/- J 0.2123
5.9040 +/- J 0.1271	5.9001
6.4750 +/- J 0.1373	6.4219 +/- J 0.2109
6.7588	6.7681
7.0481 +/- J 0.1441	6.9990 +/- J 0.2100
7.5345 +/- J 0.1490	7.5713 +/- J 0.2100
8.0423	8.0515
8.1895	8.1884
8.1896 +/- J 0.1607	8.1285 +/- J 0.2528
8.2962	8.3054
8.7656 +/- J 0.1404	8.7172 +/- J 0.1843
9.1503	9.1503
9.3257	9.2708 +/- J 0.2556
9.3423 +/- J 0.1458	9.3534
REF. FIG. 9.18	REF. FIG. 9.19

Table 9.10

REAL ZEROS- TIME/MILLISECONDS DELTA(N)=TAU(N)-TAU(N-1)

N	S(T) TAU(N)	S(T) DELTA(N)	S'(T) TAU(N)	S'(T) DELTA(N)	S''(T) TAU(N)	S''(T) DELTA(N)
1	0.0272	2.2192	0.1223	0.1862	0.0045	0.1358
2	1.5669	1.5397	0.2490	0.1267	0.1902	0.1857
3	1.6666	0.0997	0.4619	0.2129	0.3623	0.1721
4	1.8839	0.2173	0.6204	0.1585	0.5389	0.1766
5	2.0289	0.1450	0.8061	0.1857	0.7155	0.1766
6	2.2779	0.2490	0.9782	0.1721	0.8876	0.1721
7	2.4319	0.1540	1.1367	0.1585	1.0552	0.1676
8	2.6765	0.2446	1.2997	0.1630	1.2182	0.1630
9	2.8214	0.1449	1.4356	0.1359	1.3722	0.1540
10	3.0795	0.2581	1.6167	0.1811	1.5352	0.1630
11	3.2335	0.1540	1.7888	0.1721	1.7073	0.1721
12	3.4283	0.1948	1.9609	0.1721	1.8794	0.1721
13	5.0425	1.6142	2.1602	0.1993	2.0697	0.1903
14	7.0830	2.0405	2.3549	0.1947	2.2644	0.1947
15			2.5542	0.1993	2.4591	0.1947
16			2.7489	0.1947	2.6539	0.1948
17			2.9482	0.1993	2.8486	0.1947
18			3.1520	0.2038	3.0433	0.1947
19			3.3286	0.1776	3.2335	0.1902
20			3.5687	0.2401	3.4283	0.1948
21			3.6456	0.0769	3.6049	0.1760
22			3.8087	0.1631	3.7272	0.1223
23			3.9128	0.1041	3.8630	0.1358
24			4.1257	0.2129	4.0306	0.1676
25			4.3250	0.1993	4.2253	0.1947
26			4.5107	0.1857	4.4220	0.1967
27			4.7371	0.2226	4.6193	0.1973
28			4.9092	0.1721	4.8186	0.1993
29			5.1764	0.2672	5.0269	0.2083
30			5.3621	0.1857	5.2670	0.2401
31			5.6157	0.2536	5.5025	0.2355
32			5.8104	0.1947	5.7153	0.2128
33			6.0142	0.2038	5.9146	0.1993
34			6.2135	0.1993	6.1094	0.1948
35			6.3992	0.1857	6.3041	0.1947
36			7.2098	0.8106	6.4988	0.1947
37			7.3457	0.1359	6.6755	0.1767
38			7.5721	0.2264	6.7977	0.1222
39			7.7352	0.1631	6.9154	0.1177
40			7.9164	0.1812	7.0921	0.1767
41			8.1066	0.1902	7.2777	0.1856
42			8.2741	0.1675	7.4635	0.1858
43			8.7542	0.4801	7.6536	0.1901
44			8.8311	0.0769	7.8257	0.1721
45			9.0757	0.2446	8.0069	0.1812
46			9.2111	0.1354	8.1836	0.1767
47					8.3692	0.1856
48					8.5141	0.1449
49					8.6455	0.1314
50					8.7949	0.1494
51					8.9670	0.1721
52					9.1437	0.1767
REF. FIG. 9.21	A		B		C	

Table 9.11

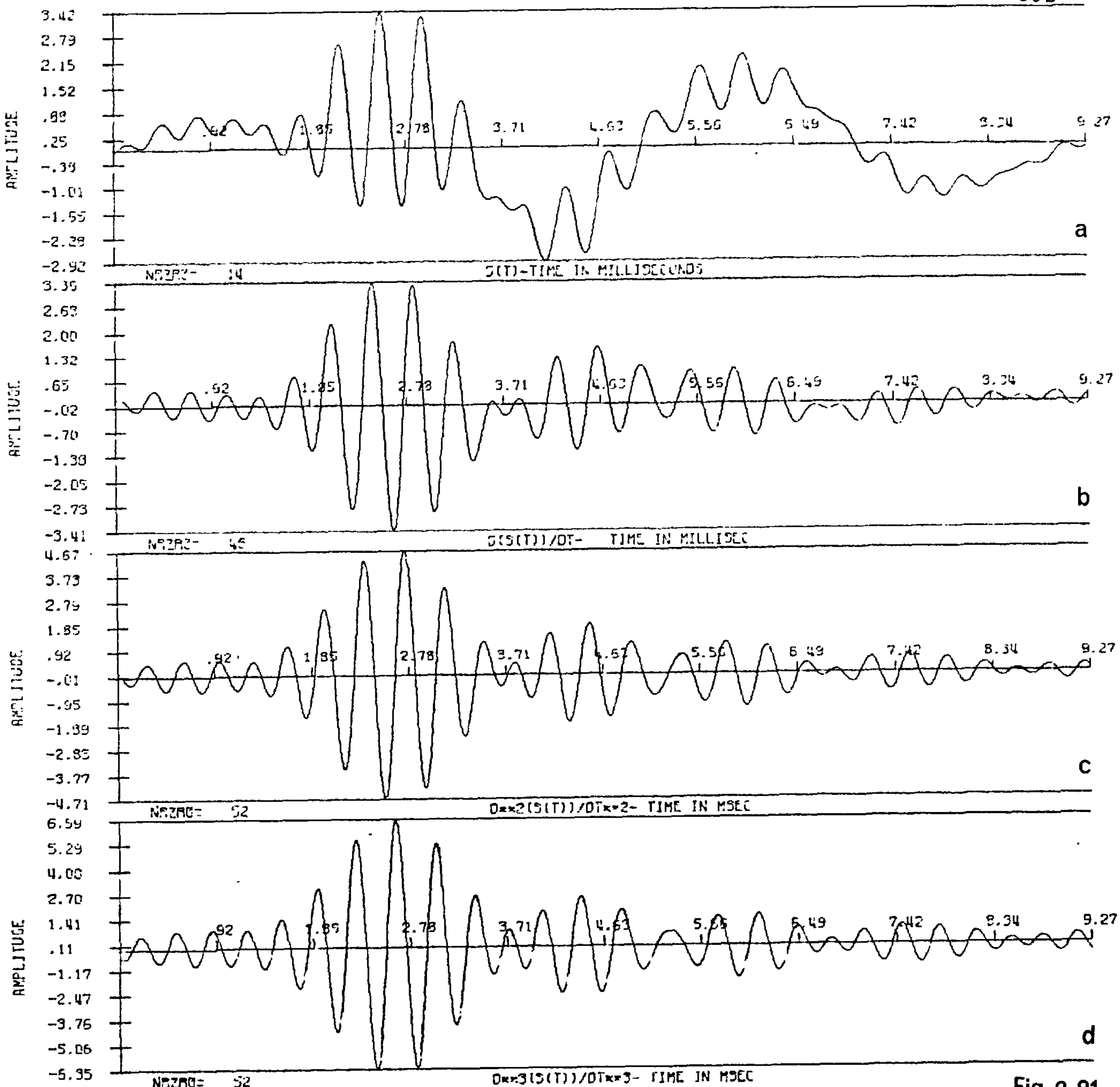

AMPLITUDE
a
S(T)-TIME IN MILLISECONDS
b
D(S(T))/DT- TIME IN MILLISEC
c
D**2(S(T))/DT**2- TIME IN MSEC
d
D**3(S(T))/DT**3- TIME IN MSEC

Fig.9.21

Zero Conversion by Differentiation

Table 9.12

Vowel	s(t)	s'(t)	s''(t)	s'''(t)	Total Zeros/ Period
/u/	6 0.12	20 0.38	36 0.69	46 0.89	52 = $2n_R+2n_C$
/o/	14 0.27	16 0.31	28 0.54	40 0.77	52
/ʌ/	20 0.33	22 0.37	36 0.60	48 0.80	60
/e/	24 0.43	38 0.68	42 0.75	46 0.82	56
/ɛ/	24 0.43	40 0.71	48 0.86	50 0.89	56
/i/	14 0.26	46 0.85	52 0.96	52 0.96	54
	• +	• +	• +	• +	

• $2n_R$ + $n_R/(n_R+n_C)$

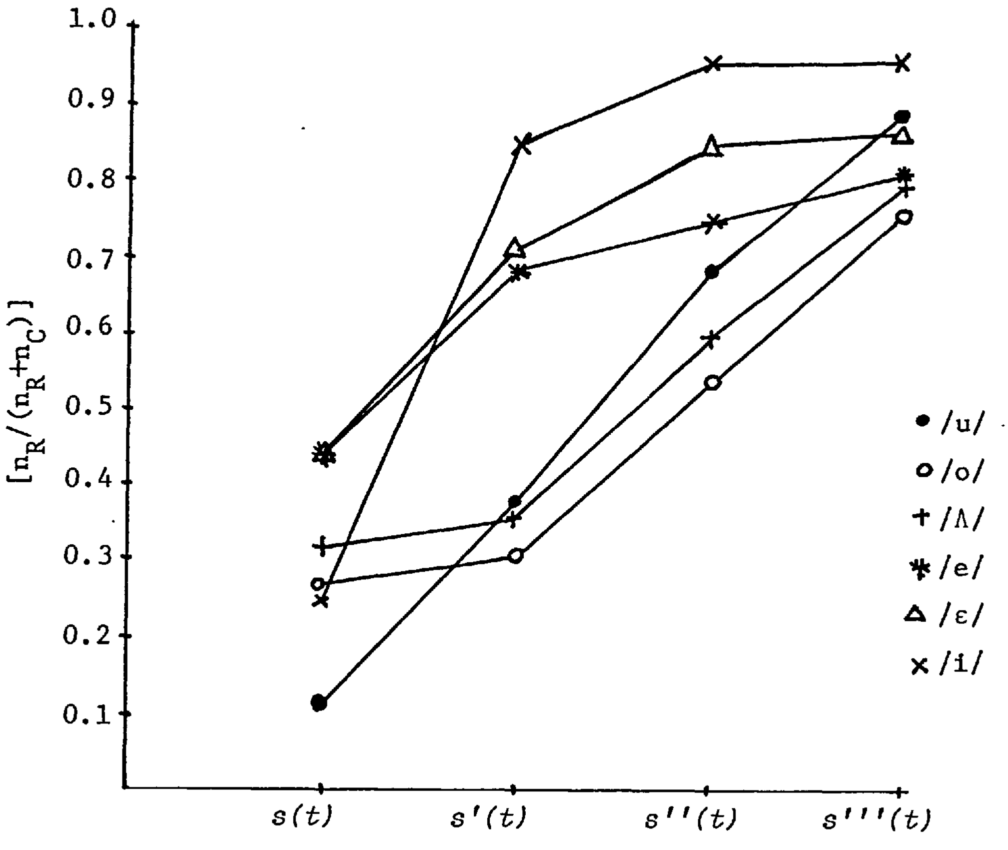


Fig. 9.22 $[n_R/(n_R+n_C)]$ for 6 vowel samples.

Figure 9.23 shows the constant $\Psi(z)$ contours [eq. (8-110)] for the vowel /u/, Fig. 9.2. Note that the linearization of phase with real time which *must* occur for large values of σ [eq.(8-111)] occurs for σ fairly small, as in the square wave and sawtooth examples, Figs. 8.23a and b. Again, we suggest that this is because of zero regularity in real time.

9.3.4 Signal Growth and Zero Distribution

We have seen that, observationally, real time segments without either CZ's or RZ's would correspond to areas of rapid signal growth. Also, time segments adjacent to areas with zero gaps experience a relative signal amplitude suppression. Is it possible to obtain a quantitative measure of signal growth in area of zero voids?

For the simple example of a CZ signal in which *all* CZ's are located at $z = T/2 \pm j\ 0.0$, from (8-15),

$$s(t) = \prod_{\ell=1}^{n} 2\ (1 + \cos\Omega t) \tag{9-14a}$$

$$= 2^{n} \cdot (1 + \cos\Omega t)^{n} \tag{9-14b}$$

$$= 2^{2n} \cdot \cos^{2n}\Omega t/2 \tag{9-14c}$$

$$= 2[\cos n\Omega t + \binom{2n}{1}\cdot\cos(n-1)\Omega t + \binom{2n}{2}\cos(n-2)\Omega t + \ldots$$

$$\ldots. + \binom{2n}{n}\cos\Omega t + \tfrac{1}{2}\binom{2n}{n}] \quad . \tag{9-14b}$$

The DeMoivre-Laplace theorem [P-3, p. 66] states that, for $2npq >> 1$,

$$\binom{2n}{k}\ p^{k}\cdot q^{2n-k} \simeq \frac{1}{\sqrt{4\pi npq}}\ e^{-(k-2np)^{2}/4npq} \quad ; \tag{9-15a}$$

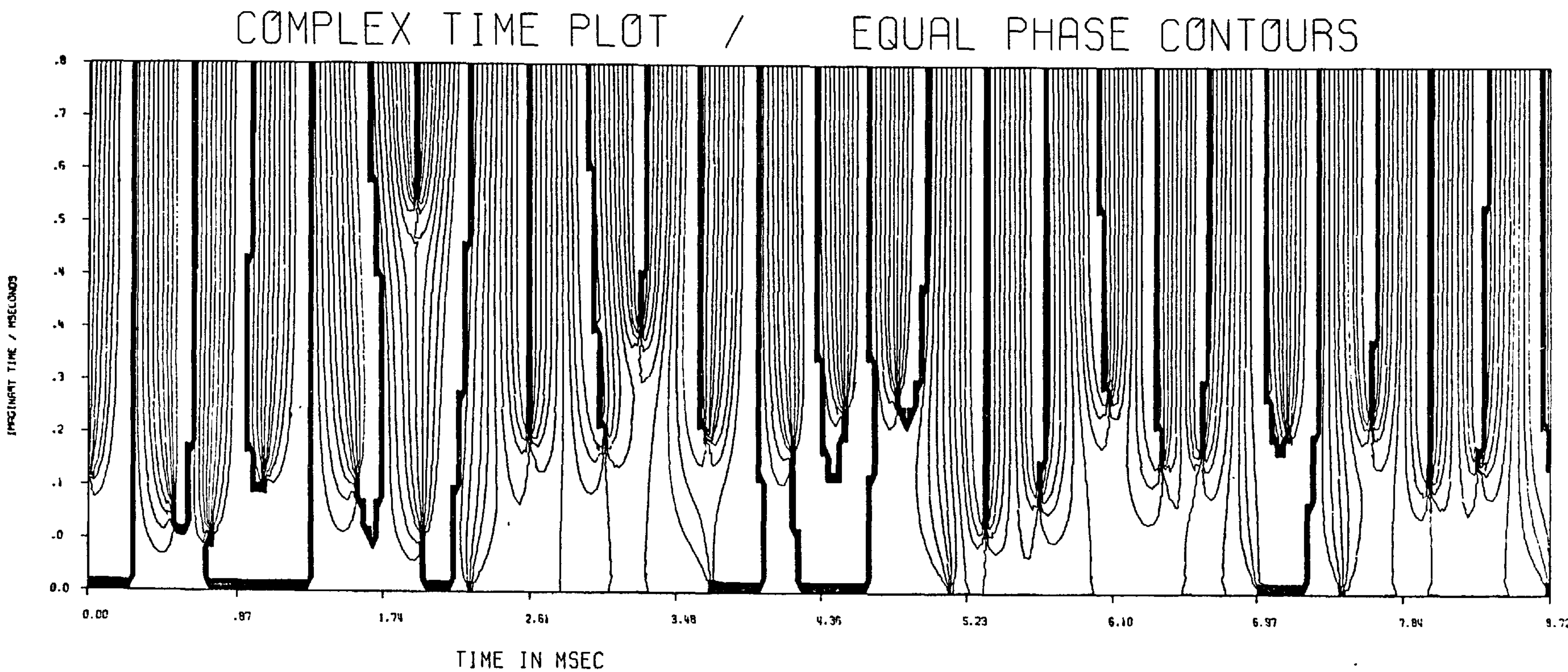


Fig. 9.23 Equal phase contours [$\Psi(z)$] for vowel /u/, upper half z plane.

For p=q= ½, this reduces to

$$(\tfrac{1}{2})^{2n}\binom{2n}{k} \simeq \frac{1}{\sqrt{\pi n}} \; e^{-(k-n)^2/n} \quad , \tag{9-15b}$$

for n >> 2. Thus, for n large, we can write,

$$2^n \cdot (1 + \cos\Omega t)^n \simeq \frac{2^{2n+1}}{\sqrt{\pi n}} \sum_{k=0}^{n} e^{-k^2/n} \cdot \cos k\Omega t \quad . \tag{9-16}$$

If we regard the line spectrum as the sampled spectrum of a signal consisting of one period of the periodic signal, then, because

$$\sqrt{\pi/\alpha} \;\; e^{-\omega^2/4\alpha} \leftrightarrow e^{-\alpha t^2} \quad , \text{ [P-2, p. 25]} \tag{9-17}$$

the time function whose Fourier transform is normal, or Gaussian shaped, *is itself the same shape.* Indeed, the d.c. component of the "pulse" is $2^{2n+1}/\sqrt{\pi n}$ and the signal passes through zero at $t = T/2$ and 2^{2n} at $t = 0$.

The problem of relating signal dynamic range to regularity of zeros is quite involved for more complicated signals. As noted by Requicha [R-7, p. 121] only qualitative observations -- those in the previous section, for example -- are thus far available. Nevertheless, it is experimentally true that abrupt changes in the "short-term zero density" -- i.e., zero gaps -- are associated with large excursions in signal amplitude.

9.3.5 The Dynamic Range of Vowel Waveforms

We observed in sec. 9.3.3 that, experimentally, vowel waveforms do not possess time segments containing either "huge" amplitude excursions or, conversely, prolonged time segments of "negligible" amplitude. The descriptions "huge" and "negligible" must be considered as relative terms. In the RZ and CZ plots, for example, we might say that if linear scaling of waveform values with reference to the peaks results in significant time segments of signal which are essentially indistinguishable from the zero amplitude axis then these segments are of "negligible" amplitude.

More formally, we could state that these conditions are present if Fourier series expansion of such waveforms -- from an amplitude spectrum viewpoint -- is heavily dependent upon the segments of large excursion and is not significantly affected by setting the waveform values in the segments of "negligible" amplitude to zero.

Intuitively, the spectrum of $s_{RZ}(t)$ in Figs. 9.10b, 9.14b and 9.18b would not be significantly affected if the signal values in the relevant time segments (1.5-4.5, 0.5-5.7, and 0.9-4.8 msec., respectively) were set to zero. That is, the contributions to the spectrum of $s_{RZ}(t)$ may vary *considerably* over different portions of the signal. On an energy basis, for this type of signal,

$$\int_0^T |s_{RZ}(t)|^2 dt \approx \int_0^a |s_{RZ}(t)|^2 dt + \int_b^T |s_{RZ}(t)|^2 \, dt, \qquad (9\text{-}18)$$

where a and b are the beginning and end of the segment of "negligible" amplitude. We wish to show that this type of behaviour is not characteristic of speech waveforms generally, specifically vowels.

Fant has shown [F-2] that the Laplace transform relating volume velocity at the glottis to sound pressure at a distance ℓ

from the lips for a 3 formant vowel is approximately

$$P(s) = \left[\left\{\frac{1}{1-e^{-sT}}\right\}\left\{\frac{U_{qo}}{\prod_{r=1}^{4}(1-s/s_r)}\right\}\right]\left[\frac{K(s)}{\prod_{n=1}^{3}(1-s/\hat{s}_n)(1-s/\hat{s}_n^*)}\right]\left[\frac{\hat{\rho}\, s}{4\pi\ell} K_T(s)\right]$$

where s is the complex frequency variable.

The first factor, $[U_{qo}/\prod_{r=1}^{4}(1-s/s_r)][1/(1-e^{-sT})]$ is the Laplace transform of the glottal volume velocity waveform (see sec. 3.3.1) for a given voice effort and constant fundamental period T. According to Fant [F-2, p. 52) s_1, s_2, s_3, and s_4 are real poles having typical values $-2\pi.100$, $-2\pi.2000$, $-2\pi.4000$ and $-2\pi.5000$ rad/sec. respectively.

The second factor, $[K(s)/\prod_{n=1}^{3}(1-s/\hat{s}_n)(1-s/\hat{s}_n^*)]$, is the vocal tract transfer function relating volume velocity through the lips to volume velocity at the glottis. The effects of the three primary complex conjugate pole-pairs whose presence is revealed as F1, F2, and F3 is directly included as $\{\hat{s}_n, \hat{s}_n^*\}$ while K(s) is a factor to correct for the presence of higher order poles [F-2, p. 42].

Finally, the third factor $\hat{\rho}s \cdot K_T(s)/4\pi\ell$ -- is the approximate transfer function from volume velocity through the lips to pressure in the sound field at a distance ℓ from the lips [F-2, p. 44]. $\hat{\rho}$ is the ambient air density.

The inverse transform of P(s) represents the sound pressure vs. time waveform of a sustained (ideally) vowel. For stationary (sustained) conditions, a single period can be expressed as

$$p(t) = \sum_{r=1}^{4} A_r' \cdot e^{s_r t} + \sum_{n=1}^{3} (-1)^n \cdot A_n' \cdot e^{\sigma_n t} \cdot \cos[2\pi(F_n t + \Phi_n)],$$

$$0 \le t \le T \qquad (9\text{-}20)$$

where

$$A_r' = A_r(1-e^{s_r T})^{-1} \qquad (9\text{-}21)$$

$$A_n' = A_n(1-2 \cdot e^{\sigma_n T} \cdot \cos 2\pi F_n T + e^{2\sigma_n T})^{-\frac{1}{2}}$$

and

$$\Phi_n = \frac{1}{2\pi}\tan^{-1} [\sin 2\pi F_n T/(e^{-\sigma_n T} - \cos 2\pi F_n T)] \ . \qquad (9\text{-}23)$$

Note that $\sigma_n = -\pi B_n$, where B_n is the formant bandwidth. Because $B_n < 45$ Hz for vowels, then, for $F_o = 100$ Hz, $|\Phi_n| < 14°$.

If the poles are moved onto the $j\omega$ axis (9-20) reduces to

$$\tilde{p}(t) = \sum_{r=1}^{4} A_r' \cdot e^{s_r t} + \sum_{n=1}^{3} (-1)^n \cdot A_n' \cdot \cos 2\pi(F_n t + \Phi_n) \ . \qquad (9\text{-}24)$$

In reality, the poles are quite near the $j\omega$ axis [D-18], [F-8, p. 152], [F-2, p. 51] so that $\sigma_n << \omega_n = 2\pi F_n$. Reducing the $\{\sigma_n\}$ to zero eliminates the damping on each sinusoidal component of p(t).

Equation (9-20) involves specification of only 3 parameters per formant -- A_n, F_n and σ_n -- and 8 parameters for the voicing source. Since σ_n is highly dependant upon F_n[F-8, p. 152], the number of necessary parameters for complete vowel specification is reduced further. This model, therefore, is a less redundant method of specifying the speech waveform than the general

Fourier series representation. Briefly, this is so because once the formant amplitudes and locations are known, the Fourier series is basically determined because the form of the resonators is known.

More important, (9-20) demonstrates that the dynamic range of p(t) must be small compared to the dynamic range possible for an arbitrary signal with the same bandwidth. The upper bound on the amplitude of p(t) is $\sum A'_r + \sum A'_n$ but because of the phase relationships which give rise to the $(-1)^n$ factors in (9-20), and because the phase perturbations Φ_n are small, this bound is not approached. The upper bound on *any* signal with the same amplitude spectrum as (9-20) and arbitrary phase spectrum is simply the *sum* of the absolute values of *all* the Fourier series coefficients of the signal represented by (9-20). The A'_n factors, as per (9-22), are proportional to the formant amplitudes, A_n, which are simply the value of the amplitude spectrum of p(t) *at the formant frequencies*. It is easily shown that, for the σ_n, T and F_n combinations associated with vowels, $A'_n < 2A_n$.

Therefore, vowels can be modelled realistically as the summation of a very small number of damped sinusoids, each of which has amplitude not much greater than that of the single Fourier coefficient nearest the relevant formant frequency. For this reason, we do not expect, and indeed, do not observe, vowel waveforms with time segments of either "huge" or "negligible" amplitude. It follows that, because gaps devoid of *both* RZ's and CZ's must give rise to "huge" amplitude excursions, we should not (and did not, in our experiemntal work) expect to find such voids in voiced speech sounds. We will discuss this contention further in sec. 9.5.3.

9.4 The Zeros of Bandlimited Clipped Speech Waveforms

9.4.1 The Effects of Bandlimiting on Sgn[s(t)]

In this section we will attempt to show that bandlimiting a clipped, periodic signal should not significantly affect the zero crossing positions provided that certain bandwidth-related zero crossing separations are satisfied in the clipped signal.

A clipped, periodic signal may be expressed in the form (9-1d)

$$C\ s(t) = \mathrm{sgn}[s(t)] = \prod_{i=1}^{2n_R} \mathrm{sgn}[t-\tau_i], \quad |t| \leqslant T/2. \tag{9-25a}$$

$$= 2\sum_{i=1}^{2n_R} (-1)^{i-1}\, U(t-\tau_i) - 1 \tag{9-25b}$$

$$0 \leqslant t \leqslant T,$$

$$= 2\sum_{i=1}^{n_R} [U(t-\tau_{2i-1}) - U(t-\tau_{2i})] - 1\,,$$

$$0 \leqslant t \leqslant T\ . \tag{9-25c}$$

Here U(x) is the unit step,

$$U(x) = \begin{cases} 1, & x \geqslant 0 \\ 0, & x < 0\ . \end{cases}$$

We wish to predict the effects of bandlimiting C s(t). Invoking the Fourier series relationships, equations (2-6) and (2-7), we may write [P-2, p. 46]:

$$BL_n\ \{C\ s(t)\} = \int_{-T/2}^{T/2} C\ s(\tau)\ \frac{\sin[(n+\frac{1}{2})\Omega(t-\tau)]}{T\cdot\sin[\Omega(t-\tau)/2]}\ d\tau, \tag{9-26}$$

where $BL_n\{C\,s(t)\}$ is bandlimited to $\pm W = \pm n\Omega/2\pi$ Hz. Combining (9-25c) and (9-26)

$$BL_n\{C\,s(t)\} = 2\left\{\int_{-T/2}^{T/2} \sum_{i=1}^{n_R} [U(\tau-\tau_{2i-1})-U(\tau-\tau_{2i})]\cdot \frac{\sin[(n+\frac{1}{2})\Omega(t-\tau)]}{T\sin[\Omega(t-\tau)/2]}d\tau\right\}-1,$$

$$0 < t < T. \qquad (9\text{-}27)$$

i) Ripple

For convenience, we write

$$k_n(t) = \frac{\sin[(n+\frac{1}{2})\Omega t]}{T\sin[\Omega t/2]}\quad , \qquad (9\text{-}28)$$

where $k_n(t)$ is the *Fourier series kernel* [P-2, p.44]. Then

$$BL_n\{C\,s(t)\} = 2\left\{\sum_{i=1}^{n_R}\int_0^T [U(\tau-\tau_{2i-1}).k_n(t-\tau)-U(\tau-\tau_{2i}).k_n(t-\tau)]d\tau\right\} -1 \qquad (9\text{-}29a)$$

$$= 2\left\{\sum_{i=1}^{n_R}\int_{\tau_{2i-1}}^{T} k_n(t-\tau)\,d\tau - \sum_{i=1}^{n_R}\int_{\tau_{2i}}^{T} k_n(t-\tau)\,d\tau\right\}-1 \qquad (9\text{-}29b)$$

$$= 2\left\{\sum_{i=1}^{n_R}\int_{\tau_{2i-1}}^{\tau_{2i}} k_n(t-\tau)\,d\tau\right\}- 1,\ \tau_{2i-1} < \tau_{2i}.\ [1] \qquad (9\text{-}29c)$$

1 The $2n_R$ RZ's may arbitrarily be arranged into n_R pairs with one member of each pair greater than the other.

Differentiating both sides of (9-29c) with respect to t,

$$\frac{d\ BL_n\{C\ s(t)\}}{dt} = 2 \sum_{i=1}^{n_R} [k_n(t-\tau_{2i-1})-k_n(t-\tau_{2i})] \qquad (9\text{-}30a)$$

because $\frac{\partial\ k_n(t-\tau)}{\partial t} = -\frac{\partial\ k_n(t-\tau)}{\partial\ \tau}$. For convenience, we will write

$$\frac{d\ BL_n\{C\ s(t)\}}{dt} = 2 \sum_{i=1}^{n_R} \underline{k}_{n,k}(t) \qquad (9\text{-}30b)$$

where $$\underline{k}_{n,i}(t) = \left[\frac{\sin[(n+\frac{1}{2})\Omega(t-\tau_{2i-1})]}{T\cdot\sin[\Omega(t-\tau_{2i-1})/2]} - \frac{\sin[(n+\frac{1}{2})\Omega(t-\tau_{2i})]}{T\cdot\sin[\Omega(t-\tau_{2i})/2]}\right] . \qquad (9\text{-}31)$$

We contend that, if $(n+\frac{1}{2})\Omega >> \Omega/2$, then--because $\sin[(n+\frac{1}{2})\Omega t]$ varies so much more rapidly than $\sin(\Omega t/2)$--during time segments *of approximate duration* $T/(n+\frac{1}{2})$ *seconds* which are located at least $T/(n+\frac{1}{2})$ seconds away from τ_{2i-1} or τ_{2i} we may write

$$\underline{k}_{n,i}(t) = K_i \cos[(n+\tfrac{1}{2})\Omega t + \theta_i] \qquad (9\text{-}32)$$

in the sense that--*over this time interval*--zero crossings are occuring regularly at the rate of one every $T/(2n+1)$ seconds. K_i and θ_i are constants which are calculated by assuming the denominators of (9-31) to be constant over the interval in question. In summary, over short time segments [on the order of the ripple period, $T/(n+\frac{1}{2})$ seconds] $\underline{k}_{n,i}(t)$ is approximately a sinusoid of frequency $(n+\frac{1}{2})\Omega/2\pi$ Hz *provided that the time segment is located "far enough" from the zero crossings which define* $\underline{k}_{n,i}(t)$.

Now, from (9-30),

$$\frac{d\ BL_n\{C\ s(t)\}}{dt} = 2 \sum_{i=1}^{n_R} \underline{k}_{n,i}(t) \qquad (9\text{-}33)$$

so that we can similarly extend the contentions of the previous paragraph and state that, over short time segments [on the order of the ripple period, $T/(n+\frac{1}{2})$ seconds], $d\ BL_n\{C\ s(t)\}/dt$ is approximately a sinusoid of frequency $(n+\frac{1}{2})\Omega/2\pi$ Hz *provided that the time segment is located "far enough" from any of the zero crossing positions of C s(t).*[2]

Thus, since (9-33) is the derivative of $BL_n\{C\ s(t)\}$ we would expect to find ripple of "frequency" $(n+\frac{1}{2})\Omega/2\pi$ Hz in time segments of $BL_n\{C\ s(t)\}$ "far enough" away from zero crossings of $BL_n\{C\ s(t)\}$ *if* the criterion that $(n+\frac{1}{2})\Omega >> \Omega/2$ is satisfied. For voiced speech, Ω is typically $2\pi\cdot 100$ radians/second and a reasonable minimum bandwidth is 3 KHz, i.e., n = 30.

Examination of Figs. 9.3a, 9.7a, 9.11a, 9.15a, and 9.19a shows that, experimentally, this is the case. The ripple is nothing more than a manifestation of Gibb's phenomenon [P-2, pp. 30-31] and by measurement, has a frequency close to $(n+\frac{1}{2})\Omega/2\pi$ Hz, where C s(t) has been bandlimited to $n\Omega/2\pi$ Hz.

---

2 The question of whether bandlimiting causes *significant* changes in the positions of the zero crossings of C s(t) will be examined in subsection ii) of this section.

ii) Migration and Annihilation of Zero Crossings

At the $2j-1^{th}$ zero crossing of s(t), from (9-30a),

$$\left.\frac{\partial\, BL_n\,\{C\ s(t)\}}{\partial\ t}\right|_{t=\tau_{2j-1}} = 2\ k_n(0) - 2\ k_n(\tau_{2j-1}-\tau_{2j})$$

$$+\ 2 \sum_{\substack{i=1\\ i\neq j}}^{n_R} [k_n(\tau_{2j-1}-\tau_{2i-1}) - k_n(\tau_{2j-1}-\tau_{2i})] \tag{9-34a}$$

$$\simeq 2(n+\tfrac{1}{2})/T + \sum_{\substack{i=1\\ i\neq j}}^{2n_R} K_i \cdot \cos[(n+\tfrac{1}{2})\Omega\tau_{2j-1} + \theta_i] \tag{9-34b}$$

$$\simeq 2W\ +\ K \cdot \cos[(n+\tfrac{1}{2})\Omega\tau_{2j-1} + \theta]\ , \tag{9-34c}$$

for $\tau_{2j-1} - T/(2n+1) > \tau_i > \tau_{2j-1} + T/(2n+1)$, $i \neq 2j-1$ (see Fig. 9.24a).

Now, as $t \to T/2$, the value of the envelope of $k_n(t) \to 1/T$. In fact, for $T/6<t<5T/6$ the value of the envelope of $k_n(t)$ lies between 2/T and 1/T. Therefore, the value of K in (9-34c) is *upper* bounded by $2(2/T)\cdot(2n_R-1)$--at $t = \tau_{2j-1}$--if the other zero crossings of s(t) are located more than T/6 seconds away from $t = \tau_{2j-1}$.

Note that for n "large",

$$2(n+\tfrac{1}{2})/T \simeq 2W\ , \tag{9-35}$$

where $BL_n\{C\ s(t)\}$ is bandlimited to $\pm n\Omega/2\pi = \pm W$ Hz. Thus, *for a given T*, the value of the *derivative* of $BL_n\{C\ s(t)\}$ at a zero crossing of s(t) is the sum of a factor, 2W, which is proportional to the highest frequency present in the signal, and a factor whose *upper bound* is proportional to the number of zero crossings.

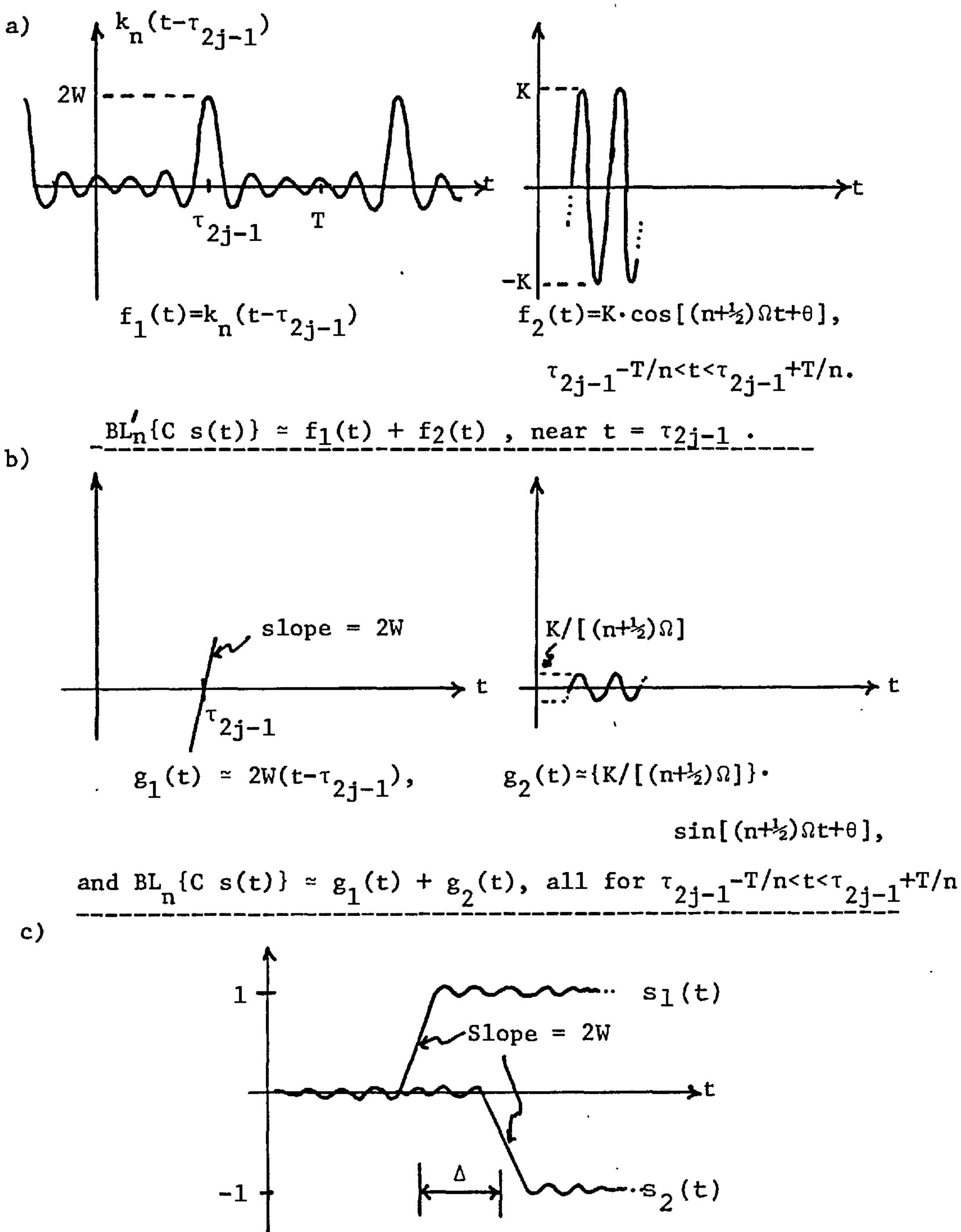


Fig. 9.24 Approximating $BL'_n\{C\ s(t)\}$ (a) and $BL_n\{C\ s(t)\}$ (b) *near* $t = \tau_{2j-1}$, and (c) geometry illustrating zero crossing annihilation by bandlimiting (see text).

Under the above conditions, *near* $t = \tau_{2j-1}$,

$$BL_n\{C\ s(t)\} \approx 2W(t-\tau_{2j-1}) + [K/(n+\tfrac{1}{2})\Omega] \cdot \sin[(n+\tfrac{1}{2})\Omega t + \theta] \quad \text{(see Fig. 9.29b)} . \tag{9-36}$$

Then, for $BL_n\{C\ s(t)\} = 0$ *near* $t = \tau_{2j-1}$,

using $K = K_{max} = 4(2n_R-1)/T$,

$$t \approx \tau_{2j-1} - [K/\{2(n+\tfrac{1}{2})\Omega W\}] \cdot \sin\ [(n+\tfrac{1}{2})\Omega t + \theta] \tag{9-37a}$$

$$\approx \tau_{2j-1} - [(2n_R-1)T/\pi n^2] \cdot \sin[(n+\tfrac{1}{2})\Omega t + \theta] . \tag{9-37b}$$

Experimentally, for speech signals, $n_R < 0.3n$. Thus, for n "large", the *maximum* value of the coefficient of the sine function in (9-37b) is approximately

$$(0.6/\pi) \cdot (T/n) . \tag{9-38}$$

For the speech signals we are concerned with $n \approx 30$ and, experimentally, the average value of the real time interval between the *zeros* of s(t) is T/n (see sec. 9.3.3, eq. (9-13)). Therefore, the factor in (9-38) -- the coefficient of the sine function --represents a *maximum* zero crossing perturbation of less then 20% of the *average inter-zero spacing*.

Thus, to a good approximation, for the signals we are concerned with, the zero crossings of $BL_n\{C\ s(t)\}$ should be relatively undisturbed *if the zero crossings of C s(t) are farther apart than T/n seconds*.

A visual superimposition of the s(t) and $BL_{30}\{C\ s(t)\}$ diagrams and a similar comparison of the $s_{RZ}(t)$ signals corresponding to s(t) and $BL_{30}\{C\ s(t)\}$ shows that experimentally , the effect of the bandlimiting operations on the zero crossings of

C s(t) -- for speech vowels -- is *almost negligible* -- except in one case (Fig. 9.7a), where two zero crossings of the original signal (Fig. 9.6a, t = 0.6177, 0.6799 msec.) have been converted to a complex zero pair by the bandlimiting operation. In this instance, the two relevant zero crossings of s(t) are very close together. Assuming that the arguments of the previous paragraphs can be extended to consider the effects of bandlimiting on two adjacent zero crossings, then -- using geometrical arguments -- there is indeed the possibility that the two RZ's will be converted to a CZ pair by the bandlimiting operation if they are closer than 0.25(T/n) seconds apart.

From Fig. 9.24c, $[s_1(t) - s_2(t)] < \frac{1}{2}$ if $\Delta < 1/4W = 0.25(T/n)$ seconds. Figs. 9.25a, b, c demonstrate zero crossing annihilation by bandlimiting in the practical case.

9.4.2 Experimental Observations: Clipped, then Bandlimited Signal

i) $s_{RZ}(t)$

As per the previous section, $s_{RZ}(t)$ appears to be little changed by the bandlimiting operation except in the one case [Fig. 9.7] where a zero crossing pair ≈ 0.08 milliseconds apart (≈0.17/W) in the original signal have been converted to a CZ pair by bandlimiting.

ii) $s_{CZ}(t)$ and the complex zeros

As shown in the previous section, ripple is associated with the bandlimiting operation following clipping. Since a complex zero pair *must* fall between pairs of successive maxima in the waveform (sec. 8.5.2), the complex zeros occurring between zero crossings in BL{C s(t)} should be "regular" in real time, at

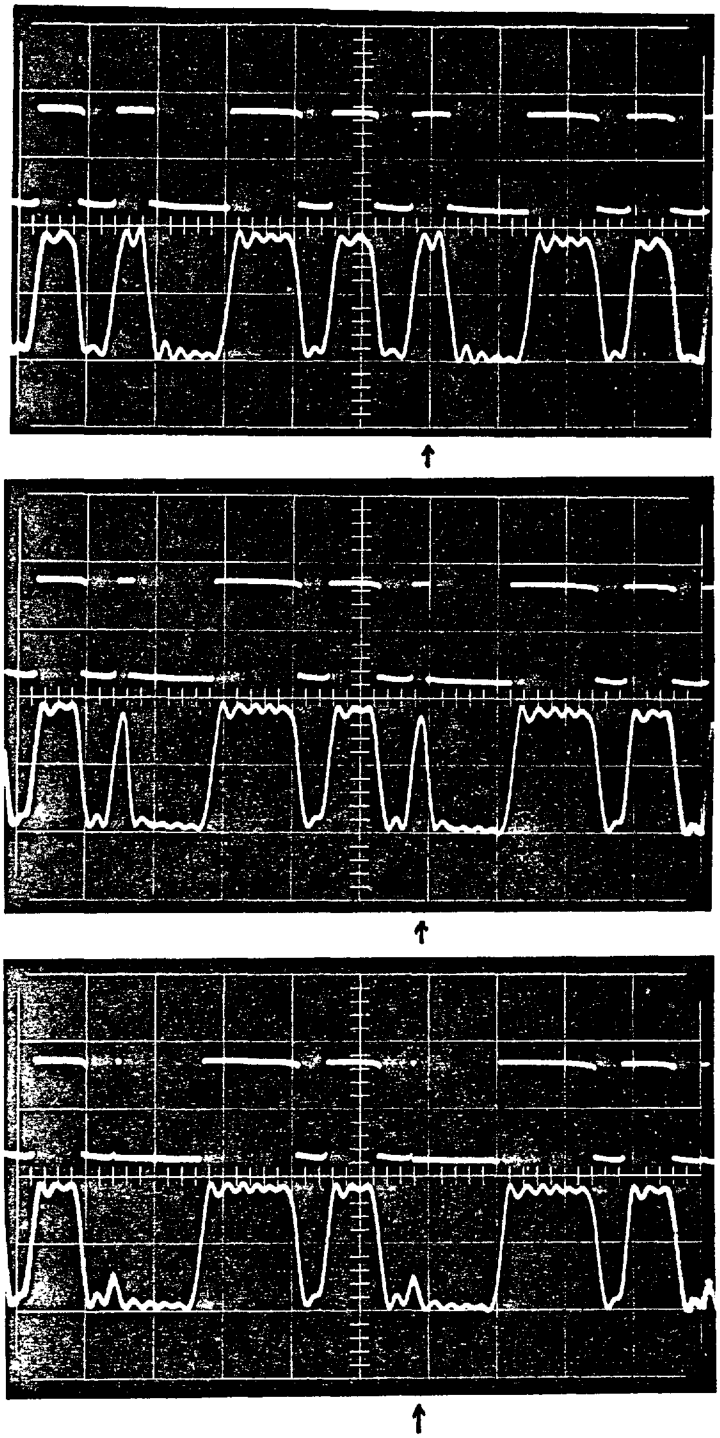


Fig 9.25 Loss of zero crossings by bandlimiting a clipped signal. Zero crossing pair to right of centre line is lost as pre-bandlimiting spacing is decreased.

intervals of approximately T/n. This is observed in Figs. 9.3, 9.7, 9.11, 9.15, and 9.19 except for a few cases corresponding to time segments where s_{CZ} was *extremely* small and hence somewhat inaccurate (sec. 9.3.1). In all (real) time segments where $s_{CZ}(t)$ is not of negligible amplitude -- i.e., visible on the experimental diagrams, resolution 0.001 " -- the CZ-ripple correspondence is exact.

In summary, the following observations were noted:

First, because of the ripple, the CZ pairs are "regular" in real time. Secondly, because smaller imaginary parts of CZ pairs are associated with larger amplitude ripple (sec. 9.3.3 iii), and because the ripple amplitude is largest near the discontinuities of the clipped waveform, the CZ configuration for the bandlimited rectangular waveform exhibits a characteristic "arced" configuration (see Figs. 8.2, 8.3).

Thus, for regular ripple to occur, the CZ's must be regular in real time. The larger the ripple amplitude, the closer the CZ's must be to the real time axis. In effect, the post-clipping and bandlimiting positions of the CZ's are highly restricted by the nature of the bandlimited clipped waveform.

9.5 The Geometry of the Zeros of Polynomials

We have shown that, although the precise positions of the real zeros and complex zero pairs on the complex time (z) plane may be determined by factorization of the Fourier series polynomial representing s(t), certain explicit relationships obtain between *waveform characteristics* and *zero locations*.

The RZ's of s(t) may be located by bandlimited interpolation of s(t) in the time domain (sec. 9.3.1). In addition, the real time positions of the CZ pairs are often "signalled" by overt signal characteristics such as ripple. However, this is not always the case and the imaginary time positions of the CZ's are not at all obvious.

Since the finite Fourier series of $s(t)$ or $BL\{C\ s(t)\}$ -- both periodic signals bandlimited to $\pm\ W = n\Omega/2\pi$ Hz -- can be represented by a polynomial of degree $2n$ in $w = e^{j\Omega t}$, it is of interest to ask whether the significant features of $f(w)$, the Fourier series polynomial, are of value in determining the character of its roots (or zeros) and therefore the nature of the zeros of the signal.

For this reason, this section is devoted to a study of the zeros of $f(w)$ as a function of its coefficients, which are the complex Fourier coefficients of the signal.

9.5.1 Self-Inversive Polynomials

As noted in chapter 8, because

$$s(t) = \sum_{k=-n}^{n} c_k \cdot w^k, \quad w = e^{j\Omega t} \tag{9-39}$$

then

$$g(w) = w^{-n} \sum_{k=0}^{2n} c_{k-n} \cdot w^k \tag{9-40}$$

and the zeros of $g(w)$ are the zeros of

$$f(w) = \sum_{k=0}^{2n} c_{k-n} \cdot w^k \tag{9-41}$$

Due to the complex conjugate symmetry of the Fourier coefficients,

$$\text{i.e., } c_{-k} = c_k^* , \tag{9-42}$$

f(w) possesses zeros which are either on, or reflected in,[1] the unit circle, $|w| = 1$. Polynomials satisfying this criterion have the property that $f(1/w^*)$ has the same zeros as f(w) [B-15]. Such polynomials are called *self-inversive* and we shall deal with them exclusively.

Since *vertical* strips in the z plane map into *sectors* of the w plane (eq. (8-12)) our concern with z plane distribution in real time is transformed into an interest in the angular distribution of zeros about the origin of the w plane. Similarly, a *horizontal* strip in the z plane maps into an *annulus* in the w plane. Due to the self-inversive nature of the Fourier series polynomial, investigation of the *maximum* distance from the origin at which roots may be found on the w plane is equivalent to determining the *minimum* distance. The above relationships are illustrated in Fig. 9.26.

9.5.2 Circle Theorems

Real zero signals have roots only on the unit circle in the w plane. That is,

$$f(w) = 0 \text{ for } w = e^{j\theta}. \tag{9-43}$$

A. Kempner has shown [K-2] that, if f(w) has real coefficients only, then the roots which lie upon the unit circle become real roots of

$$\phi(w) = (w^2+1)^{2n} \cdot f\left(\frac{w+j}{w-j}\right) \cdot f\left(\frac{w-j}{w+j}\right) = 0, \tag{9-44}$$

which contains only even powers of w. Letting

$$\phi(w) = \Psi(w^2) = \Psi(w') , \tag{9-45}$$

---

[1] This means that a root at $r\,e^{j\theta}$ must be accompanied by another at $e^{j\theta}/r$.

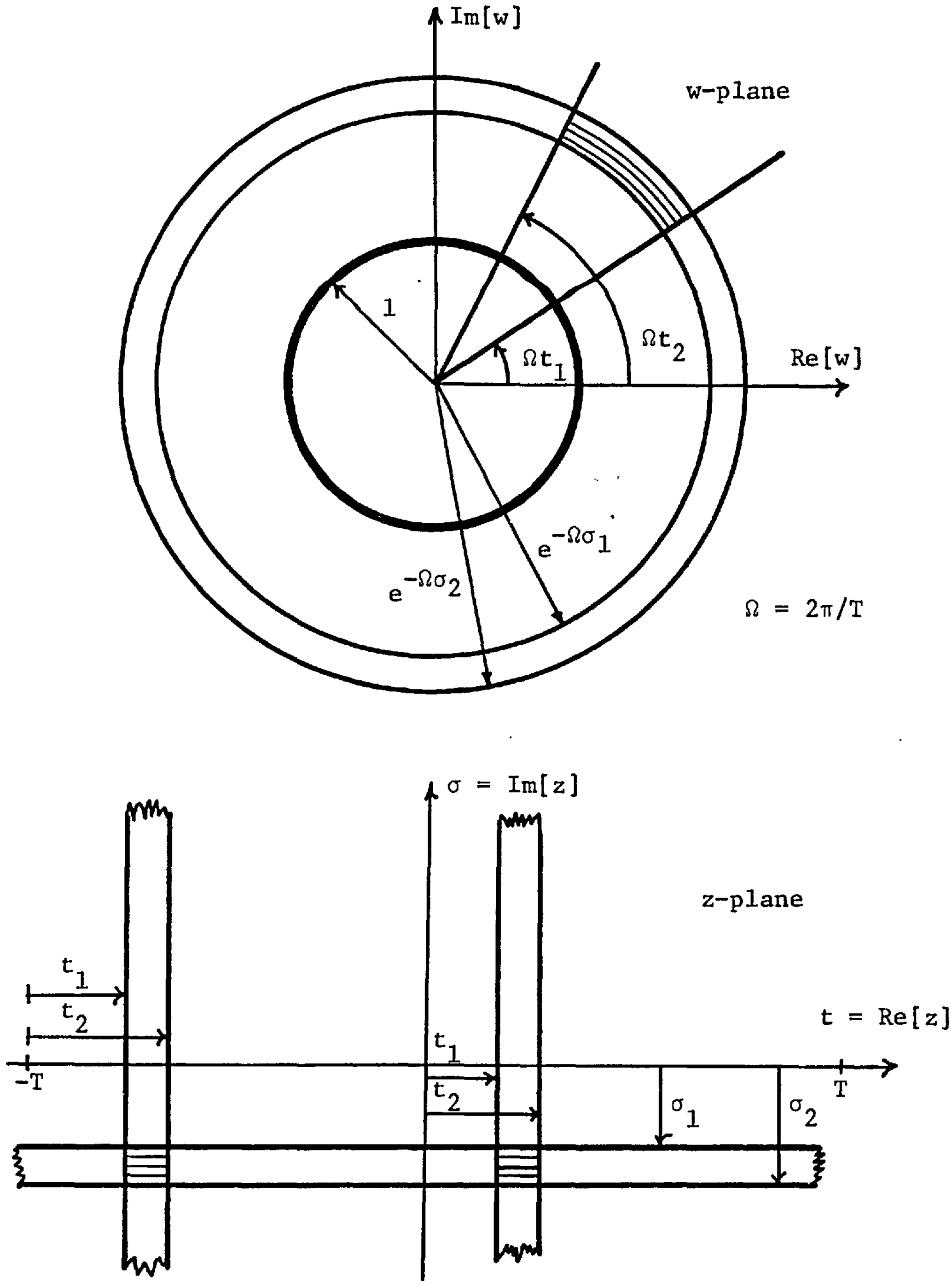


Fig. 9.26 Mapping from w- to z-plane, $w = e^{j\Omega z}$.

then the necessary and sufficient condition that all roots of $f(w) = 0$ are of the form $e^{j\theta}$ is that $\Psi(w')$ has only real, positive roots. Tests for real positive roots are outlined in chapter 9 of Marden [M-6]. However, most of the polynomials we are concerned with have complex coefficients. Although other tests for "real zero" spectra are possible [B-15], [M-6, p. 206, ex. 3] they are qualitatively uninformative.

i) Loose Bounds

We are specifically interested in establishing bounds for the magnitude of the zeros of the polynomial

$$f(w) = a_o + a_1 w + \ldots + a_{2n-1} w + a_{2n} w^{2n} \tag{9-46}$$

as a function of the (2n+1) complex coefficients. Marden deals with this problem at length in his *Geometry of Polynomials* [M-6, pp. 122-165]. For example, it can be shown that all the zeros of $f(w)$ lie on or outside the circle

$$|w| = \min\{|a_0|/(|a_0| + |a_k|)\} \qquad k = 0, 1, \ldots 2n. \tag{9-47}$$

The Fourier series kernel, $k_n(t)$ (eq. (9-28)), has $a_k = 1$ so that

$$|w|_{min} > 1/2 \text{ and } |w|_{max} < 2 \;. \tag{9-48}$$

Note that

$$k_n(t) = \frac{\sin(n+\frac{1}{2})\Omega t}{T\cdot\sin(\Omega t/2)} \tag{9-49a}$$

$$= \frac{2\cdot\sin(2n+1)\Omega' t}{T'\cdot\sin\Omega' t}, \text{ where } \Omega'=\Omega/2 \text{ and } T'=2T \tag{9-49b}$$

$$= \frac{K \prod_{i=0}^{4n+1} 2\cdot\sin\frac{\Omega'}{2}\,[t-iT'/(4n+2)]}{T' \prod_{i=0}^{1} 2\cdot\sin\frac{\Omega'}{2}\,[t-iT'/2]} \qquad (9\text{-}49c)$$

$$= \frac{K}{T'} \prod_{\substack{i=1\\ i\neq 2n+1}}^{4n+1} 2\cdot\sin\frac{\Omega'}{2}\,[t-iT'/(4n+2)] \qquad (9\text{-}49d)$$

so that $k_n(t)$ possesses 4n real zeros per period T' -- or 2n RZ's per period T -- and is bandlimited to $\pm 2n(\Omega/2) = \pm n\Omega$ rad/sec. Thus $k_n(t)$ is an RZ signal and the bounds given by (9-47) are very conservative.

ii) <u>The Lehmur-Schur Algorithm and its Repercussions</u>

The Lehmur-Schur algorithm [L-9], [R-2, pp. 355-359] is used directly to determine whether or not a given polynomial, f(w), has roots within the unit circle. Unfortunately, the basic algorithm breaks down for self-inversive polynomials. However, if the polynomial $f(r\cdot w)$, $r<1$, is substituted for f(w), the algorithm may be used to determine whether f(w) has roots within the circle $|w| = r$.

In appendix A, we show that the Lehmur-Schur algorithm can be modified so as to derive close bounds on the minimum radius (and hence, because of the self-inversive nature of the Fourier series polynomial, the maximum radius) at which roots of the polynomial representing a three-tone vowel model are found. The example used is

$$f(w) = a_3 r^{50+2N} \cdot w^{50+2N} + a_2 r^{35+2N} \cdot w^{35+2N} + a_1 r^{29+2N} \cdot w^{29+2N}$$

$$+ a_1^* r^{21} w^{21} + a_2^* r^{15} w^{15} + a_3^* \quad . \qquad (9\text{-}50)$$

This represents a three-tone vowel model with fundamental or voicing frequency of 100 Hz and formants (tones) located at $(400 + 100 \cdot N)$ Hz, $(1000 + 100 \cdot N)$ Hz and $(2500 + 100 \cdot N)$ Hz, respectively. Thus N represents an SSB modulation (translation) of 100.N Hz, with N=0 corresponding to the original lowpass vowel model. The complex amplitudes of F1, F2, and F3 are a_1, a_2, and a_3, respectively with the usually valid assumption being made that $|a_1|>|a_2|>|a_3|$. As noted in sec. 9.3.5, the three-tone model lacks the damping present in actual vowel sounds.

It is shown in appendix A that the minimum radius r at which roots are found is

$$r \approx [|a_2| \cdot |a_3|]^{-1/15} \quad , \qquad (9\text{-}51)$$

the approximation being more exact as N is increased from zero, i.e., as the signal is SSB translated upwards in frequency. More generally, for $(F_3-F_2) = p(\Omega/2\pi)$ Hz -- $5 < p < 15$, for vowels -- then for

$$|a_1|>|a_2|>|a_3| \quad , \qquad (9\text{-}52)$$

$$r \approx [|a_2| \cdot |a_3|]^{-1/p} \quad . \qquad (9\text{-}53)$$

Again, SSB modulation improves the estimate.

For the two-tone model the results are similar with p being $(F_2-F_1)2\pi/\Omega$ and $|a_3|$ replaced by $|a_1|$ in (9-53). We have tested this extension of the Lehmur-Schur algorithm for the two-tone signal. The results are as follows:

i) Signal: $s(t) = 3\sin\Omega t + \sin 3\Omega t$

r_{min} (predicted) $= (3)^{-1/2} = 0.576$

r_{min} (actual) $= 0.517$

ii) SSB Modulation: translation of i) by 3Ω.

$s(t) = 3 \sin 4\Omega t + \sin 6\Omega t$

r_{min} (predicted) $= 0.576$ (same as i))

r_{min} (actual) $= 0.5807$

iii) SSB Modulation: translation of i) by 5Ω

$s(t) = 3\sin 6\Omega t + \sin 8\Omega t$

r_{min} (predicted) $= 0.576$ (same as i),ii))

r_{min} (actual) $= 0.5777$

iv) Increased Separation of Tones

$s(t) = 3\sin\Omega t + \sin 5\Omega t$

r_{min} (predicted) $= (3)^{-1/4} = 0.769$

r_{min} (actual) $= 0.735$

v) SSB Modulation: translation of iv) by 2Ω

$s(t) = 3\sin 3\Omega t + \sin 7\Omega t$

r_{min} (predicted) $= 0.769$ (same as iv))

r_{min} (actual) $= 0.751$

vi) Increase of "First Formant" Amplitude in iv)

$s(t) = 5\sin\Omega t + \sin 5\Omega t$

r_{min} (predicted) $= (5)^{-1/4} = 0.669$

r_{min} (actual) $= 0.645$

Table 9.13

Roots of experimental two-tone models

i) $s(t) = 3\sin\Omega t + \sin 3\Omega t$

Roots on w plane: ± 1, $0.0 \pm j\,1.9319$, $0.0 \pm j\,0.5176$

ii) $s(t) = 3\sin 4\Omega t + \sin 6\Omega t$

Roots on w plane: ± 1, $\pm j\,1$, $0.0 \pm j\,1.7221$, $0.0 \pm j\,0.5807$
$\pm 0.7587 \pm j\,0.6514$

iii) $s(t) = 3\sin 6\Omega t + \sin 8\Omega t$

Roots on w plane: ± 1, $\pm j\,1$, $0.0 \pm j\,1.7310$, $0.0 \pm j\,0.5777$
$\pm 0.5481 \pm j\,0.8364$, $\pm 0.8844 \pm j\,0.4668$

iv) $s(t) = 3\sin\Omega t + \sin 5\Omega t$

Roots on w plane: ± 1, $\pm 0.4588 \pm j\,0.5661$, $\pm 0.8641 \pm j\,1.0661$

v) $s(t) = 3\sin 3\Omega t + \sin 7\Omega t$

Roots on w plane: ± 1, $\pm 0.9759 \pm j\,0.9047$, $\pm 0.4021 \pm j\,0.9156$,
$\pm 0.5511 \pm j\,0.5109$

vi) $s(t) = 5\sin\Omega t + \sin 5\Omega t$

Roots on w plane: ± 1, $\pm 0.4149 \pm j\,0.4968$, $\pm 0.9904 \pm j\,1.1858$

Thus, as predicted, both SSB modulation and increasing the tone separation increases the accuracy of the predicted r_{min}. Again because of the self-inversive nature of the polynomials, $r_{max} = 1/r_{min}$.

This model predicts that, for three-tone models of vowel sounds, the complex zeros will be located "near" the unit circle in the w plane and hence close to the real time axis in the z plane. This is because $(x)^{1/p} \to 1$ for "large" p, $|x|<1$. In table 9.14, we have used the data of Peterson and Barney [P-11] to calculate the maximum value of $\sigma = |Im[z]|$ for three-tone vowel models. The calculated values of r_{min} range from 0.81 to 0.94 while the corresponding values of σ -- assuming T = 10.0 msec or F_0 = 100 Hz -- are 0.07 and 0.34 milliseconds, respectively. We observed that, in the experimentally factorized vowels, the *majority* of the CZ pairs were located within 0.3 - 0.4 milliseconds of the real time axis.

It should be noted that in the experimental factorizations, the third formant was generally *not* located at the upper bandlimit of the factorized signal. However, assume that

$$s(t) = a_1 \sin n_1\Omega t + a_2 \sin n_2\Omega t + a_3 \sin n_3\Omega t + \varepsilon \sin n_o\Omega t,$$
$$n_o > n_3 > n_2 > n_1 \quad . \qquad (9\text{-}54)$$

If $\varepsilon <<< \min\{a_1, a_2, a_3\}$, then the behaviour of s(t) should be, intuititively, almost unaffected by the presence of the term $\varepsilon \sin n_o \Omega t$.

However, by dimensionality arguments, s(t) *must* possess a number of zeros per period equal to $2n_o$. Thus the effect of the $\varepsilon \sin n_o\Omega t$ term must be to add $(n_o - n_3)$ CZ pairs in such a way so as to leave the signal behaviour essentially unchanged, so that the original $2n_3$ zeros still determine the net signal behaviour. It would be then expected that the "extra" CZ pairs forced into the signal by the $\varepsilon \sin n_o\Omega t$ term would appear relatively far from

(1) Vowel	(2) $(F_3-F_2)/Fo$ $=P$	(3) $\lvert F2\rvert-\lvert F3\rvert$ db	(4) $X=$ (3)/20	(5) 10^x	(6) $\sqrt[P]{\frac{1}{x}}=r_{min}$	(7) $\Omega\sigma_{min}$	σ_{min}, msec
/i/	8	4	0.2	1.6	0.945	0.0565	0.089
/I/	6	4	0.2	1.6	0.926	0.0768	0.123
/ε/	7	7	0.35	2.24	0.89	0.116	0.185
/æ/	16	10	0.50	3.16	0.93	0.0726	0.0726
/a/	15	23	1.15	14.0	0.838	0.177	0.282
/ɔ/	16	27	1.35	22.4	0.823	0.195	0.310
/u/	12	22	1.10	12.6	0.81	0.211	0.339
/u/	14	24	1.20	15.8	0.82	0.199	0.318
/ʌ/	12	17	0.85	7.1	0.85	0.163	0.260
/ɝ/	3	5	0.25	1.78	0.825	0.186	0.297

Table 9.18

Calculation of Radius of Smallest CZ for Three-tone Vowel Model

Data of Peterson and Barney [P-11], Modified

Lehmur-Schur Algorithm

the real time axis so as not to cause any perceptable change in the signal (sec. 9.4.3 iii)).

9.5.3 Angular Distributions

The problem of determining the positions of the real zeros and complex zero pairs in real time requires an investigation into the angular distribution of zeros about the origin in the w plane.

i) Loose Bounds

P. Erdös and P. Turán have shown [E-1, 2, 3] that for the polynomial

$$f(w) = a_0+a_1w+a_2w^2+\ldots\ldots + a_{2n}w^{2n}, \tag{9-55}$$

if $$\max_{|w|=1} |f(w)| = M, \tag{9-56}$$

then for arbitrary fixed $0 \lesssim \alpha < \beta \lesssim 2\pi$, an "Index of Regularity", I_ν can be defined such that

$$I_\nu = \Big| \sum_{\substack{\nu \\ \alpha < \text{arc } w_\nu < \beta}} 1-2n\cdot[(\beta-\alpha)/2\pi] \Big|$$

$$<16[2n\cdot\log(M/\sqrt{|a_0\cdot a_{2n}|})]^{1/2} \tag{9-57}$$

A verbal statement of the Erdös-Turán theorem is that "the absolute value of the difference between the number of roots in a given sector -- $\alpha \lesssim \text{arc } w_\nu \lesssim \beta$ -- and the number of roots that would be found in the same sector if the roots were *uniformly* distributed about the origin (i.e., $2n[(\beta-\alpha)/2\pi]$) is less than $16[2n \log(M/\sqrt{|a_0\cdot a_{2n}|})]^{\frac{1}{2}}$."

Note that

$$f(|w| = 1) = f(e^{j\theta}) = a_0 + a_1 e^{j\theta} + a_2 e^{j2\theta} + \ldots\ldots + a_{2n} e^{j2n\theta} \qquad (9\text{-}58a)$$

$$= e^{jn\theta}(a_0 e^{-jn\theta} + \ldots\ldots\ldots + a_{2n} e^{jn\theta}) \qquad (9\text{-}58b)$$

$$= e^{jn\theta} \cdot s(\theta) \qquad (9\text{-}58c)$$

if f(w) is a Fourier series polynomial. Since $\theta = \Omega t$ in this case,

$$\max_{|w| = 1} |f(w)| = \max |s(t)| . \qquad (9\text{-}59)$$

Finally, M can be replaced by P, where

$$P = |a_0| + |a_1| + \ldots\ldots + |a_{2n}| , \qquad (9\text{-}60)$$

and $$P \geqslant M \cdot \text{[E-3]} \qquad (9\text{-}61)$$

In the Erdös-Turán theorem, I_ν represents an *index of regularity* for the angular distribution of zeros about the origin in the w plane. However, the angular distribution of the zeros of the polynomial representing the Fourier series kernel, $k_n(t)$ -- eq. (9-28) -- is nearly uniform, yet, because $|s(t)|_{max} = (2n+1)/T$, I_ν represents a rather poor bound; i.e.,

$$I_\nu < 16\{2n \log[(2n+1)/T]\}^{\frac{1}{2}} . \qquad (9\text{-}62)$$

In summary, the theorem of Erdös-Turán represents a rather conservative bound on zero regularity about the origin in the w plane.

ii) Kempner's Planetarium Theorems

A.J. Kempner also considered the problem of angular distribution of zeros in the w plane [K-3, -4, -5]. He studied the general polynomial equation, replacing a_k and w in

$$f(w) = a_n w^n + a_{n-1} w^{n-1} + \ldots . + a_1 w + a_0 = 0 \tag{9-63}$$

with

$$R_k \cdot e^{j\Psi k}, \quad 0 \leqslant \Psi_k \leqslant 2\pi, \quad R_k > 0 \tag{9-64}$$

and

$$r \cdot e^{j\theta}, \quad 0 \leqslant \theta \leqslant 2\pi, \quad r > 0 \ , \tag{9-65}$$

respectively.

Kempner described his first theorem as follows [K-5, p. 816]:

"In the [w] plane of complex numbers, mark from the origin the . . . vectors Ψ_k. The vector Ψ_0 is to remain in its original position. The vector Ψ_1 is to rotate in a positive sense with a constant angular velocity Ω, while for k = 1, 2,..n the vector Ψ_k rotates with a uniform angular velocity k times that of Ψ_1. Vectors Ψ_k for which $R_k = 0$, that is, for which the coefficient $a_k=0$, are to be ignored. At any moment the vectors give the directions of the vectors representing the terms

$$a_k \cdot w^k = R_k \cdot r^k \cdot e^{j(\Psi_k + k\theta)}.$$

These directions depend only on θ and the Ψ_k of the coefficients. Most of the theorems [presented] are immediate consequences of the fact that *the sum of vectors from a common point can certainly not vanish when it is possible to draw a line through the point such that all vectors lie on one side of the line.*" (Italics mine.)

Thus, Kempner emphasized, it is impossible to have roots of f(w) = 0 in any sector (vertex at the origin of the w plane) $\alpha \leqslant \theta \leqslant \beta$ for which *all vectors lie on one side of a straight line through the origin.* He noted that such θ intervals -- if they exist -- "are determined by simple inequalities which in many cases enable us to determine sectors [which may have zero width], depending only on the *arguments* of the coefficients, *not* on the *absolute* values, and which are free of roots."[2]

The sectors which interlace the "forbidden sectors" -- the sectors which may not have zeros -- *must* have at least one root per sector [K-3].

Kempner's theorem can be couched in more familiar terms if we immobilize the middle term of the Fourier series polynomial -- i.e., the d.c. component -- instead of the constant term. Then we have the common visualization of sinusoids being composed of pairs of contra-rotating vectors or phasors with angular velocity equal to their radian frequency. It then becomes clear that Fourier series polynomials represent special cases of the "planetarium". Because of the symmetry involved, the only possible "straight line through the origin that all vectors can be on one side of" is effectively in the line $\theta = \pm 90°$. And there is only a choice with respect to the half-plane when the d.c. component is zero. When the d.c. component is positive, then all vectors must lie in the half-plane $-90° \leqslant \theta \leqslant 90°$ for a zero void to occur. Conversely, when the d.c. component is negative, zero voids may only occur for $90° \leqslant \theta \leqslant 270°$. The other possible case for a zero width occurs when all the vectors fall along the 0° phase line, colinear with the d.c. component.

Substituting (9-64) and (9-65) into (9-63) we find that

$$f(w) = u(r,\theta) + j\, v(r,\theta) \qquad (9\text{-}66)$$

where $$u(r,\theta) = \sum_{k=0}^{n} R_k \cdot r^k \cdot \cos(k\theta + \Psi_k) \quad , \qquad (9\text{-}67a)$$

[2] Classically, a planetarium is an instrument with dials rotating at different angular rates.

and $$v(r,\theta) = \sum_{k=0}^{n} R_k \cdot r^k \cdot \sin(k\theta + \Psi_k). \tag{9-67b}$$

For the Fourier series polynomials,

$$f(w) = a_n w^{2n} + a_{n-1} w^{2n-1} + \ldots + a_0 w^n + \ldots . a^*_{n-1} w + a^*_n \tag{9-68a}$$

$$= w^n (a_n w^n + a_{n-1} w^{n-1} + \ldots + a_0 \ldots . a^*_{n-1} w^{-(n-1)} + a^*_n w^{-n}) . \tag{9-68b}$$

After substitution of (9-64) and (9-65) into (9-68),

$$f(|w| = 1) = u(1,\theta) + j\ v(1,\theta)$$

$$= \cos n\theta \cdot s(\theta) + j \sin n\theta \cdot s(\theta) , \tag{9-69}$$

where $\theta = 2\pi(t/T) = \Omega t$. (see (9-58))

Thus,

$$u(1,\theta) = \cos n\Omega t \cdot s(t) \tag{9-70a}$$

and $$v(1,\theta) = \sin n\Omega t \cdot s(t), \ \theta = \Omega t \tag{9-70b}$$

Kempner's second theorem states [K-4, p. 80]: Plot $u(r,\theta)$, (9-67a), and $v(r,\theta)$, (9-67b), against θ, $0 \leq \theta \leq 2\pi$, for a given radius r. Name the points of intersection of the u-curve with the θ-axis α_1, α_2, α_3, .., and the points of intersection of the v-curve with the θ-axis β_1, β_2, ... Consider the combined sequence of the α, β in their natural order of magnitude. Assume the sequence closed cyclically, and let ε_1, ε_2, ε_3, . . . ε_ν, respectively be the number of β between two consecutive α; then the number of roots of $f(w) = 0$ for which $|w| < r$ is given by

$$N = \tfrac{1}{2}\left|\sum_{k=1}^{\nu} (-1)^{(\varepsilon_1+1)+(\varepsilon_2+1)+\ .\ .\ .\ +(\varepsilon_k+1)}\right| \quad . \qquad (9\text{-}71)$$

Kempner's second theorem is simply a recasting of *the Principle of the Argument* [M-6, p. 1]. This principle states that if f(w) is analytic to a simple closed Jordan curve C, and continuous and different from zero on C, *then the net number of times that f(w) encircles the origin of the f(w) plane as w traverses the closed curve C on the w plane equals the number of zeros of f(w) interior to C.*

Figure 9.27illustrates Kempner's second theorem and the Principle of the Argument for $s_{RZ}(t)$ of the vowel /u/, which has 6 RZ's. (see Fig. 9.2). For $r = |w| = 1.05$, the u vs v curve encircles the origin of the $f(w) = u + j\ v$ plane 6 times (Fig. 9.27a). For $r = |w| = 0.95$, there are no net encirclements of the origin; $s_{RZ}(t)$, of course, has all its zeros *on* the unit circle (Fig. 9.27c). When $r = |w| = 1.0$, the curve passes through the origin 6 times (Fig. 9.26b). Application of the second theorem to Figs. 9.27d, e, f) yields the same results (see [B-21]).

Figure 9.28 shows the application of the theorem to the signal s(t) for the vowel /u/. Note that as we traverse the unit circle on the w plane, f(w) passes through the origin on the f(w) plane $2n_R$ times (6 in this case) and encircles the origin n_C times (23 in this case).

Due to the cos $n\Omega t$ and sin $n\Omega t$ factors in (9-69) -- which result from the nature of the Fourier series polynomial -- encirclement of the origin of the f(w) plane by f(w), as w traverses the unit circle in the w plane, will be regular between unit circle zeros of f(w) (i.e., between RZ's of s(t)). This is demonstrated vividly in Fig. 9.28 and obtains *whether or not*

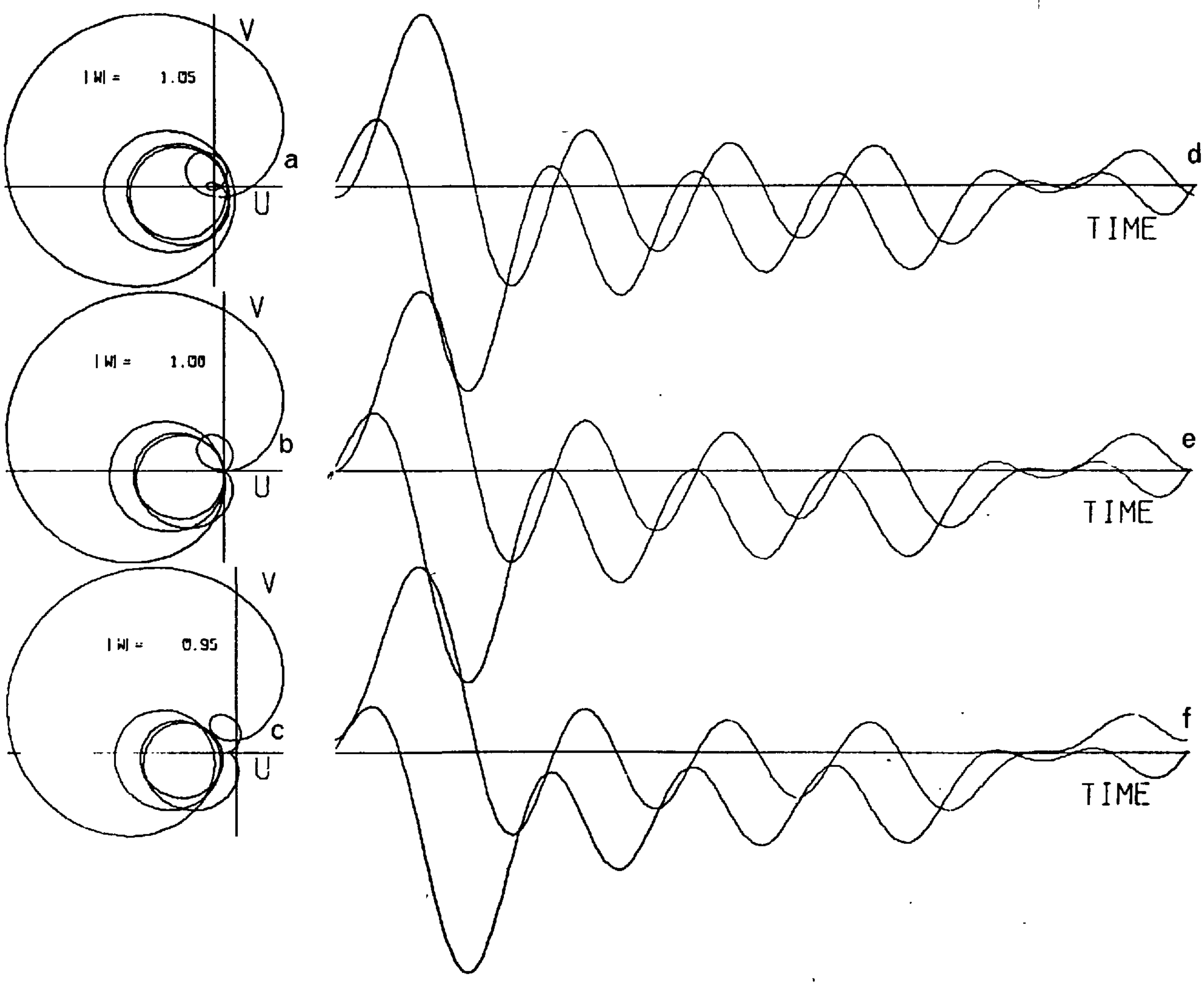


Fig. 9.27 $u(r,\theta)$ vs $v(r,\theta)$ for $r = 1.05$ (a), 1.00 (b) and 0.95 (c) for $s_{RZ}(t)$ of vowel /u/.
$u(r,\theta)$, $v(r,\theta)$ vs $\theta = \Omega t$ for $r = 1.05$ (d), 1.00 (e), and 0.95 (f) for $s_{RZ}(t)$ of vowel /u/.

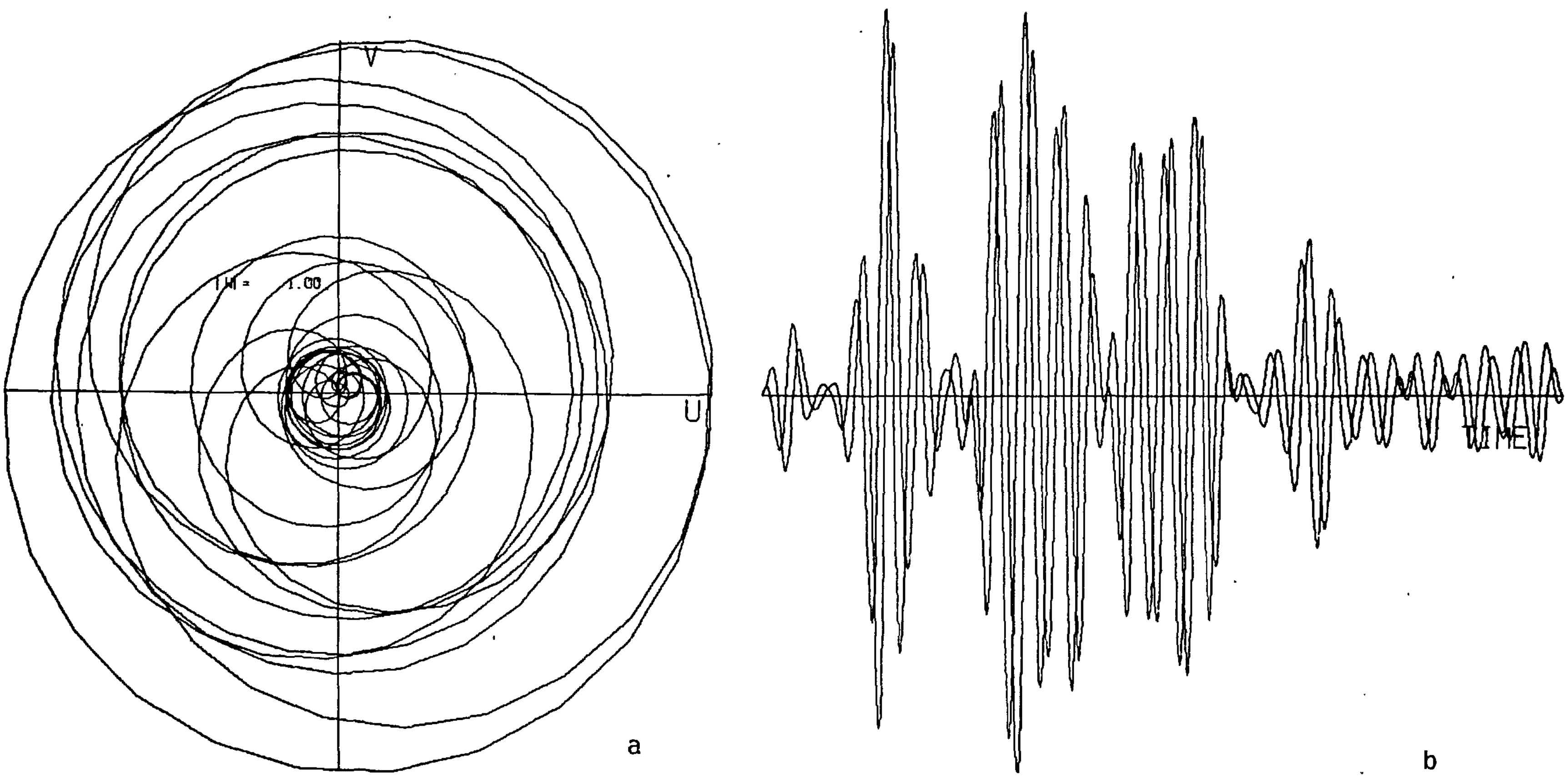


Fig. 9.28 a) u(1,θ) vs v(1,θ) for s(t) of vowel /u/.
b) u(1,θ), v(1,θ) vs θ = Ωt for s(t) of vowel /u/.

the CZ's of s(t) are regular in real time between the RZ's. Note, in Fig. 9.28b, that the RZ's of s(t) occur when $u(1,\theta)$ and $v(1,\theta)$ *simultaneously* pass through the time axis.

Our objective now is to discover whether or not there is reason to believe, as a result of Kempner's theorems (the Principle of the Argument), that the polynomials which interest us have zeros which are regularly distributed about the origin in the w plane. Consider the following path on the w plane:

i) along the arc of unit circle from $\theta+\pi$ to θ.
ii) along the diameter from $w = e^{j\theta}$ to $w = 0$ to $w = e^{j(\theta+\pi)}$

This is a closed path on the w plane. Assume, for convenience that the signal we are interested in is entirely CZ. Then $n = n_C$. Now as w traverses the semi-circle arc from $\theta+\pi$ to θ, $f(w)$ must encircle the origin of the $f(w)$ plane $n/2$ times *whether or not the n_C zeros interior to $|w| = 1$ are distributed uniformly in angle about $w = 0$.* Again, this is due to the cos $n\Omega t$ and sin $n\Omega t$ factors in (9–70).

If the n_C CZ pairs *are* uniformly distributed in θ, then, when the path $w = e^{j\theta} \rightarrow e^{j(\theta+\pi)}$ -- a diameter of the unit circle -- is traversed, the closed path on the w plane is completed and, by the Principle of the Argument, $f(w)$ must have circled the origin of the $f(w)$ plane $n_C/2 = n/2$ times. But this was already accomplished during the semi-circle traversal. Therefore, the trajectory of $f(w)$ as w moves from one end of the diameter to the other must be simply to close the path on the $f(w)$ plane without incurring more encirclements of the origin.

If the n_C CZ pairs *are not* uniformly distributed -- say there are $n_C/2 + p$ zeros in the semi-circle considered and $n_C/2 - p$ in the other semi-circle -- then when the diameter is traversed in the w plane, f(w) must encircle the origin of the f(w) plane p *more* times before the path is closed. Similarly, if there are $n_C/2 - p$ zeros in the semi-circle considered, f(w) must "un-encircle" the origin p times as the diameter is traversed on the w plane so that the net encirclement is $n_C/2 - p$.

In summary, for absolutely regular angular distribution of zeros (about the origin of the w plane) in a Fourier series polynomial, it is *sufficient* that traversal of *any* diameter of the unit circle in the w plane does not cause encirclements of the origin in the f(w) plane. The preceding arguments also apply to RZ-CZ signals if every other RZ is moved slightly outwards from the unit circle and the others are moved slightly inwards. Then the number of zeros within the unit circle is $2n_R/2 + n_C = n$.

We now ask, "What type of signals have this property?" Consider $s(t) = \cos n\Omega t$. Then

$$f(w) = w^n(w^n/2 + w^{-n}/2) \tag{9-72}$$

and
$$f(r,\theta) = r^n \cdot e^{jn\theta}(r^n \cdot e^{jn\theta}/2 + r^{-n} \cdot e^{-jn\theta}/2) \tag{9-73a}$$

$$= r^{2n} \cdot e^{j2n\theta}/2 + 1 \; . \tag{9-73b}$$

Then
$$u(r,\theta) = \tfrac{1}{2}r^{2n}\cos 2n\theta + 1 \tag{9-74a}$$

and
$$v(r,\theta) = \tfrac{1}{2}r^{2n}\sin 2n\theta \quad . \tag{9-74b}$$

Thus as a diameter of the circle $|w| = 1$ is traversed, u(r) and v(r) tend to values unity and zero, respectively, quite rapidly -- as r becomes less than unity -- with the actual rate being dependent on the value of n. Table 9.15 demonstrates that r^n tends to zero quite rapidly with increasing "n" even for "large" values of r, r < 1.

$n \backslash r$	0.95	0.90	0.85	0.80
5	0.767	0.590	0.444	0.328
10	0.599	0.350	0.197	0.107
15	0.463	0.205	0.088	0.035
20	0.358	0.122	0.039	0.011
25	0.277	0.074	0.025	0.0038

Table 9.15

Value of r^n as a function of r and n, r<1.

In this case, $f(r\cdot e^{j\theta})$ does not exhibit origin encircling behaviour as r varies from +1 to -1 along a diameter. Indeed, the zeros of cos $n\Omega t$ are regularly distributed around the unit circle in the w plane at intervals of $2\pi/2n$ radians.

Now consider a three-tone vowel model,

$$s(t) = a_1\cos(n_1\Omega t+\phi_1) + a_2\cos(n_2\Omega t+\phi_2) + a_3\cos(n_3\Omega t+\phi_3) \quad , \quad (9\text{-}75)$$

where the a_i are real and $\phi_i \simeq i.\pi$ (sec. 9.3.5). Then

$$f(r\cdot e^{j\theta}) = a_3 r^{2n_3} e^{j2n_3\theta}/2 - a_2 r^{n_2+n_3} e^{j(n_2+n_3)\theta}/2 + a_1 r^{n_1+n_3} e^{j(n_1+n_3)\theta}/2$$

$$+a_1 r^{n_3-n_1} e^{j(n_3-n_1)\theta}/2 - a_2 r^{n_3-n_2} e^{j(n_3-n_2)\theta}/2 + a_3/2, \quad (9\text{-}76)$$

where $\theta = \Omega t$. For r = 1, (9-76) = $e^{jn_3\theta}\cdot s(\theta)$ as per (9-69). For actual vowels [P-11], $(n_3-n_2)_{min}$ = 3 and $(n_3-n_2)_{max}$ = 15. The minimum values of $2n_3$, n_2+n_3, n_1+n_3, and n_3-n_1 are approximately 33, 30, 21, and 12, respectively. As r becomes less then unity --

i.e., as a diameter of the unit circle is traversed -- the higher powers of r become small and $a_3/2$ quickly becomes the dominant term. The relevant question is whether it can be shown *rigorously* that the three-tone structure is truly sufficient to ensure, via (9-76), that $f(r\cdot e^{j\theta})$ does not make multiple encirclements of the origin as r varies from +1 to -1 for fixed θ; that is, as an arbitrary diameter of the unit circle in the w plane is traversed. This question is left open for future studies.

However, limited experiments on actual vowels have demonstrated that the implied behaviour of $f(r\cdot e^{j\theta})$ for the structure of (9-76) does occur. As r decreases from unity, $f(r\cdot e^{j\theta})$ *rapidly* tends to c_n, the highest frequency Fourier coefficient of the signal. This occurs in a "simple" manner; that is with zero, or perhaps one, encirclement(s) of the origin.

Thus, *experimentally* at least, Kempner's origin circling theorems (really the Principle of the Argument) provide a further plausibility argument for zero regularity (in real time) of speech vowels.

This argument, and the contentions made on the basis of signal growth (sec. 9.3.4) and time-domain vowel structure (sec. 9.3.5) tend to support the assertion that the observations made in the limited experimental studies of actual speech vowels (regarding zero regularity) are typical and can be qualitively predicted.

9.5.4 Summary

In summary, *observationally*, vowels have zeros which occur "regularly" in real time *and* are "near" the real time axis. Theoretically, we have shown, using three-tone models and the theory of the geometry of polynomials, that this behaviour is to be expected because the formant structure (modelled using the three tones) is sufficient to allow this type of behaviour to occur. Specifically, the inter-formant (tone) gaps of low (or zero, in the case of the model) energy seem to be of prime importance in the derivation of both sets of results (sec. 9.5.2, 9.5.3, respectively).

9.6 Single Sideband Clipped Speech

In section 5.1.7 we noted that Marcou and Daguet experimentally determined that the phase function, cos $\phi(t)$, is perceptually the same as s(t). That is

$$\cos\ \phi(t) \overset{P}{=} |m(t)|\ \cos\ \phi(t)\ . \tag{9-77}$$

cos $\phi(t)$ has been defined as "single sideband clipped speech" and is obtained by clipping the SSB translate of s(t), bandpass filtering and then retranslating to the baseband (see sec. 5.1.7).

9.6.1 The Relationship between C s(t) and cos $\phi(t)$

Since

$$s(t) = s_{RZ}(t) \cdot s_{CZ}(t) = |m(t)|\ \cos\ \phi(t)\ , \tag{9-78}$$

ignoring the multiplicative constant, it is clear that $s_{RZ}(t)$ and cos $\phi(t)$ have the same zero crossings. Thus

$$C\ s(t) = \mathrm{sgn}[s(t)] = \mathrm{sgn}[s_{RZ}(t)] = \mathrm{sgn}[\cos\ \phi(t)]\ . \tag{9-79}$$

Expanding (9-76) in a Fourier series gives [S-3, p. 171], [V-11]

$$\mathrm{sgn}[\cos\ \phi(t)] = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{\cos(2k+1)\phi(t)}{2k+k} \cdot (-1)^k\ . \tag{9-80}$$

Therefore, $\mathrm{sgn}[s(t)] \propto \cos\phi(t) - \frac{1}{3}\cos 3\phi(t) + \frac{1}{5}\cos 5\phi(t) - \ldots$ (9-81)

It is clear then that lowpass clipped speech is related to SSB clipped speech; lowpass clipped speech effectively results from the addition to SSB clipped speech of odd order harmonics of $\cos\phi(t)$ -- with the proper polarity and attenuation.

9.6.2 Clipping and Critical Band Theories

So far we have not referred to critical band theories of hearing (sec. 3.2.2) in our discussions of the extent of power spectrum preservation in speech clipping. A. Rimskii-Korsakov, in a paper concerning the audibility of non-linear distortion [R-12], noted that in order to calculate "the probability of distortion being audible in the band $\omega \pm \Delta\omega/2$ [a critical band] . . . we must take into consideration the masking effect created by the fundamental signal . . .[in that band] . . . which tends to mask the distortion . . ." He further noted that "it is evident from the masking curves of pure tones (e.g., [K-7], p. 407) that for masking tone intensities that are not too great [i.e., less than 80 db] the audibility threshold of the masked tone in a band of frequencies surrounding the frequency of the masked tone is approximately 20 db below the level of the masking tone." *In effect, the masked waveform is audible if it is less than 20 db below the level of the masking waveform.*

Rimskii-Korsakov added that two subsidiary effects must be noted. First, the masking tone must be present over "a sufficiently long period of time" before it begins to have an effect. Second, even when no tone is present to be masked, the masking tone itself must exceed some threshold value before it becomes audible [K-7, p. 391].

For these reasons, the subjective effects of clipping are related not only to the extent to which clipping preserves the speech power spectrum, but also to the extent to which the "harmonics" created by clipping are masked by the original speech signal. Specifically, clipping "harmonics" (see eq.(9-81)) which fall into formant regions will tend to be masked by the formants and the deleterious effects of clipping will be less than objective estimations -- ignoring masking effects -- might predict.

In this respect the effects of zero conversion by differentiation before clipping are twofold. First, as noted earlier, the intelligibility of the clipped signal is apparently improved by increasing the amount of information (i.e., zero crossings) which is "perfectly" sampled by the clipping operator. Secondly, the higher order formants are increased in amplitude so as to facilitate more substantial masking of "clipping harmonics" falling into the relevant spectral region.*

However, the detailed consideration of the effects of masking phenomena on the intelligibility of clipped speech are reserved for future studies.

9.7 Clipping: A Zero Based Model

9.7.1 Clipping as a Manipulator of Complex Zeros

We have shown -- experimentally, and to some extent theoretically -- that, for vowel-like signals, clipping followed by re-bandlimiting to the original signal bandwidth does not usually materially affect the RZ signal, $s_{RZ}(t)$. Thus, from *a time domain point of view,* clipping can be considered to be a

*Repeated differentiation, however, will eventually degrade the original speech spectrum to the extent that the benefits afforded by the increase in the number of zero crossings available to the "clipping sampler" are nullified.

member of that class of operations which *significantly* affects only the CZ signal, $s_{CZ}(t)$. That is, *clipping is a complex zero manipulator.*

We have also noted that those pre-clipping operations which yield the most highly intelligible clipped speech signals are those which tend to convert complex zero pairs into real zeros. This action allows more information to be "preserved" by the clipping operator, in the sense that the RZ's (zero crossings) are preserved by the clipping-bandlimiting operation.

It is nevertheless clear that unrestricted manipulation of *one* CZ pair could significantly change the character of the spectrum of the signal. The examples discussed and illustrated (Figs. 8.19 and 8.21) in sec. 8.6.5 vividly illustrate this assertion. What then, we ask, is special about the nature of the clipping-bandlimiting operation as a CZ manipulator?

First, we note that the n_C complex zero pairs possess $2n_C$ degrees of freedom -- the real and imaginary component of one member of each CZ pair. The bandwidth of $s_{CZ}(t)$ is $n_C\Omega/2\pi$Hz so that the $\{Cz_k\}$ are specified by $2n_C+1$ numbers -- n_C complex Fourier coefficients and the real d.c. component. Again, the "extra" parameter is lacking in the CZ signal because the n_C CZ pairs and the $2n_R$ RZ's specify s(t) only to a multiplicative constant.

i) <u>The Real Time CZ Positions</u>

In this chapter we have observed, *experimentally*, that the zeros of vowels are "regular" in real time. We have also tried to justify this *theoretically* by noting that

a) vowel waveforms consist basically of the sum of a few damped sinusoids and exponentials and that this type of signal does not have "huge" excursions which would result from zero gaps and that

b) the vowel formant structure may be sufficient to assure -- via interpretation of the Principle of the Argument -- that the zeros of vowel-like models are regular in real time.

We have likewise noted that, between the RZ's of clipped then bandlimited signals, ripple appears *if* the RZ's are farther apart than, approximately, a period at the ripple frequency. The "ripple", a manifestation of Gibb's phenomenon, is associated with regularly spaced CZ pairs.

Thus, primarily because of the formant structure of vowel waveforms, the zeros of vowel waveforms are distributed "uniformly" in real time. Because of the "constant amplitude" nature of the combined clipping-bandlimiting operator, the CZ's of BL $\{C\, s(t)\}$ are distributed "uniformly" between the (almost) unchanged RZ's of BL $\{C\, s(t)\}$. Therefore, *experimentally* and for the *theoretical* reasons noted (all of which depend upon the formant structure of vowels), the real time positions of CZ's are *effectively* unchanged by the action of clipping and re-bandlimiting. For this reason, clipping apparently has, *effectively,* no degree of freedom to manipulate the CZ's in real time.

However, *there still remain* n_C *degrees of freedom in* $s_{CZ}(t)$ -- *the* n_C *complex zero imaginary time positions.*

ii) <u>The Imaginary Time CZ Positions</u>

We observed in sec. 9.3.3 that, *experimentally,* most of the CZ's of the vowels factorized were located within (roughly) 0.35 milliseconds of the real time axis. This figure is about

1/30 of the average pitch period, 10 msec. Note that the vowels were bandlimited to 3 KHz and therefore contained about 30 zero pairs per period, depending, of course, on the actual voicing frequency. We attempted to justify this observation *theoretically* by applying a modified version of the Lehmur-Schur Algorithm to the three-tone Fourier series polynomial. Using the data of Peterson and Barney for English vowels, we found that -- based upon formant amplitude and frequency information -- the range of maximum distance of CZ's from the real time axis in three-tone models is 0.17-0.36 milliseconds. These figures certainly concur with those experimentally observed.

We similarly noted that, *experimentally*, the CZ's of the clipped, bandlimited signal are "near" the real time axis. In fact, we purposefully matched the vertical scales of the BL {C s(t)} root maps to emphasize that the imaginary CZ positions before and after clipping are certainly within the same range, specifically less than about 0.5 milliseconds. We attempted to justify this observation *theoretically* by pointing out that, in order to produce the ripple characteristic of BL {C s(t)}, the CZ's must be "near" the real time axis. Again, (due to Gibb's phenomenon and the "constant" amplitude nature of the clipped signal) we noted the characteristic "arced" configuration of the CZ's which produce a bandlimited rectangular waveform.

Therefore, both before and after clipping -- for different reasons -- the CZ's are "near" the real time axis. For this reason the clipping operator is somewhat restricted in its ability to manipulate the imaginary parts of the CZ's. We emphasize that this restriction is not nearly as stringent as that apparently imposed upon the real time positions of the CZ's before and after clipping.

In summary, *from a time domain viewpoint,* the clipping-bandlimiting operator has

i) *effectively* little or no ability to modify the $\{\tau_\ell\}$

ii) restricted ability to manipulate the $\{\sigma_\ell\}$.

For these reasons we believe that clipping is not as destructive to the spectrum -- and hence, power spectrum -- as its "complete destruction of all amplitude information except for polarity" might suggest.

9.7.2 Clipping as a Spectral Smearing Operation

The amplitude spectrum of the vowel /u/ is almost -- as far as formant structure is concerned -- unrecognizeable after the clipping-bandlimiting operation. Clipping does, after all, have some ability to manipulate CZ's, *particularly* their imaginary parts, the $\{\sigma_\ell\}$. As previously noted, the less the number of RZ's, the greater the number of CZ's available for manipulation. For this reason, as we pointed out in chapter 5, pre-clipping CZ conversion results in more intelligible clipped speech. Additionally, the post-clipping "robustness" of the speech sound is related to the percentage RZ's (sec. 5.1.3).

From a frequency domain viewpoint, the connection between RZ-CZ balance and post-clipping power spectrum preservation becomes apparent if we re-examine the product convolution relationship,

$$s_{RZ}(t) \cdot s_{CZ}(t) \leftrightarrow \{Rz_k\} * \{Cz_k\} .$$

When the RZ's predominate, the bandwidth of $s_{RZ}(t)$ is "wide" and that of $s_{CZ}(t)$ is "small". By the discrete convolution relationship, eq. (8-19b), $\{c_k\}$ then results from the "smearing" of $\{Rz_k\}$ by $\{Cz_k\}$. Now $s_{RZ}(t)$ is physically somewhat like $s(t)$ in that it has the same zero crossings. Thus, when RZ's predominate we *might* expect that the amplitude spectrum of $s_{RZ}(t)$ is "like" that of $s(t)$ in that it could exhibit peaked structure which, when "smeared" by the convolution operation, would result in the formant structure of $\{|c_k|\}$. This effect can be observed in Fig. 9.16 although in no sense can we say that the RZ's predominate in undifferentiated /e/. However, subsequent experiments involving spectral deconvolution of $\{Cz_k\}$ from $\{c_k\}$ by $\{Rz_k\}$ of the first, second and third derivatives of the vowel pitch periods used in sec. 9.4 have shown that as the proportion of RZ's increases, $\{|Rz_k|\}$ acquires a "peaked" structure.

Conversely, because $s_{CZ}(t)$ is not "like" $s(t)$, a predominatly CZ vowel signal should not be expected to exhibit "peaked" structure in its amplitude spectrum. Again, this behaviour has been noted in limited, qualitative experimental studies.

10 ZEROS II: THE SUFFICIENCY OF REAL ZEROS AS WAVEFORM DESCRIPTORS-- A NEW APPROACH TO THE USE OF ZERO CROSSINGS FOR OBJECTIVE ESTIMATES OF SPECTRAL PARAMETERS

In the introduction to chapter 8, we contended that three basic questions remain unanswered concerning the role of zero crossings in speech recognition and processing.

The first, concerning the effects of clipping on the power spectrum, has been explored in chapter 9.

The second queried the quantitative nature of the information contained in the zero crossings of a speech signal specifically. An answer to this question was also proferred in chapter 9. That is, in bandlimited signals, zeros occur at the Nyquist rate and the percentage of zeros which are real--i.e., zero crossings-- might be regarded as an indication of the amount of signal information actually carried by zero crossings. The fact that a *special* type of "real zero interpolation", bandlimited clipping, yields a signal whose apparent intelligibility far exceeds that which might be predicted on the basis of percentage information carried by zero crossings is, we feel, attributable to

i) the sufficiency of the spectral characteristics of the original speech signal in assuring that the zeros of vowel-like signals are both regular in real time and close to (or on) the real time axis and

ii) the special nature (from a zero-based viewpoint) of the "rectangular" interpolating waveform.

The third question concerned the existence of transformations which ensure that almost all the information contained in a bandlimited signal is available in its zero crossings. In chapter 8 we observed that differentiation and sine wave addition convert complex zeros to real zeros and therefore, after a finite number of differentiations or the addition of a sine wave carrier of correct frequency and "sufficient amplitude", zero crossings will occur at the Nyquist rate.

A signal having such an RZ rate is *completely* determined to a multiplicative constant by its zero crossings and, following clipping, may be reconstructed (to a multiplicative constant) by *Real Zero Interpolation* (sec. 8.4.2).

In this chapter we will consolidate the role of zero crossings as carriers of information in speech signal processing specifically. First, in sec. 10.1, we review a conjecture made by I. J. Good concerning the information lost by clipping a Gaussian signal. Then, in sec. 10.2, we will examine some bounds on the RZ rate of SSB translates of lowpass signals as established by Voelcker. In sec. 10.3 Good's contention regarding signal specification by zero crossing is explored.

Finally, in sec. 10.4, we show that because of the formant structure of vowels, the zero crossings of vowel-like signals may contain more objective information about the *amplitude spectrum* of the vowel than might be inferred directly from the arguments given in sec. 8.1.3 regarding RZ rate and information.

In particular, we show that, under certain conditions, the amplitude spectrum of a vowel-like signal is completely "encoded" within the RZ component of that signal.

10.1 Good's Conjecture

I. J. Good presented [G-9] "an intuitive argument for the measurement of the fraction of information that is lost, if any, when Gaussian noise is clipped."

He observed that white Gaussian noise, bandlimited to $[W_1,W_2]$ Hz-- a bandwidth $W_2-W_1 = W$ Hz--possesses $2W\tau$ degrees of freedom in time τ. Thus, the noise is completely determined by its values at "an enumerable number of instants whose mean density is 2W per second, even if the instants are not uniformly spaced." [G-6, p. 35]

10.1.1 A Gaussian Noise Example

Good noted that the expected number of zeros per second for a white Gaussian noise signal bandlimited to $[W_1,W_2]$ Hz is (sec. 6.2, eq. (6-2))

$$2\cdot\frac{1}{\sqrt{3}}\left[\frac{W_2^{\,3}-W_1^{\,3}}{W_2-W_1}\right]^{\frac{1}{2}} . \tag{10-1}$$

He further noted that, since $2(W_2-W_1)$ observations are required per second to determine the signal completely, "it seems reasonable to say that the zero crossings provide a fraction

$$I = \frac{1}{\sqrt{3}}\left[\frac{W_2^{\,3}-W_1^{\,3}}{(W_2-W_1)^3}\right]^{\frac{1}{2}} \tag{10-2}$$

of the entire information in the noise." For example, when $W_1=0$, the zero crossings provide only $1/\sqrt{3}$ of the information.

Expanding (10-2),

$$I = \frac{1}{\sqrt{3}} \left[\frac{W_2^2 + W_1 W_2 + W_1^2}{(W_2 - W_1)^2} \right]^{\frac{1}{2}} . \qquad (10\text{-}3)$$

For I = 1,

$$\frac{W_2}{W_1} = (7 + \sqrt{33})/4 = 3.186 . \qquad (10\text{-}4)$$

Therefore, Good contended, the noise is *overdetermined* by its zero crossings if

$$W_2 < 3.186\ W_1 . \qquad (10\text{-}5)$$

Thus "when the noise is overdetermined by its zeros, then an adequate proportion of the zeros will, in particular, determine the remaining zeros. Hence in narrow-band noise we would expect to find a strong correlation between the lengths of adjacent zero crossing intervals. This is borne out by looking at examples of narrow-band noise."

10.2 A Zero Based Exposition of Good's Conjecture

In sec. 8.3.2 we noted that a periodic signal s(t) band-limited to $n_1\Omega/2\pi < |f| < n\Omega/2\pi$ Hz, can be written as

$$s(t) = \text{Re}\ [m(t)]$$

$$= \text{Re}\ [e^{jn_1\Omega t}|m_{LP}(t)|e^{j\phi_{LP}(t)}] \qquad (10\text{-}6)$$

and exhibits a number of zero crossings per period, $2n_R$, such that

$$2n_1 \leqslant 2n_R \leqslant 2n \quad . \tag{10-7}$$

It can be similarly shown (see [V-10]) that the number of zero crossings of an *SSB signal* is dependent upon the phase characteristics of the *lowpass* signal which we may regard as having been translated to yield the SSB signal.

For example, if $m_{LP}(t)$ is MP,

$$m_{MP,f_o}(t) = e^{j2\pi f_o t} \prod_{i=1}^{n_W} [1-a_i \cdot e^{j\Omega(t-\tau_i)}], \; a_i<1 \tag{10-8a}$$

$$= e^{j2\pi f_o t} \cdot |m_{MP}(t)| \cdot e^{j\phi_{MP}(t)} \tag{10-8b}$$

where

$$M_{MP,f_o}(f) = 0 \text{ for } f_o>f>f_o+W, \tag{10-9}$$

and $n_W = 2\pi W/\Omega$. Then

$$\phi(t) = 2\pi f_o t + \sum_{i=1}^{n_W} \phi_i(t) \tag{10-10a}$$

$$= 2\pi f_o t + \phi_{MP}(t) \quad . \tag{10-10b}$$

Note that for $f_o = 0$ [lowpass signal] the zeros of cos $\phi(t)$--and hence s(t)--occur whenever $\phi_{MP}(t)$ goes through a multiple of $\pm(2p-1)\pi/2$ radians, p an integer. As f_o increases from 0, it is clear that, for some *critical* f_o such that

$$\phi'(t) = 2\pi f_o + \phi'_{MP}(t) > 0 \quad \text{, for all } t, \tag{10-11}$$

the passage of $\phi(t)$ through odd multiples of $\pm\pi/2$ radians will be governed entirely by the carrier. That is, for f_o greater than some critical value, $\phi(t)$ will be a monotone increasing function. Thus, although the number of zero crossings per period of $\mathrm{Re}[m_{MP,f_o}(t)]$, $n_R(f_o)$, is bounded by

$$2f_o \leqslant n_R(f_o) \leqslant 2(f_o+W) \ , \tag{10-12}$$

for f_o "large enough"

$$n_R(f_o) \to 2f_o \quad . \tag{10-13}$$

If $m_{LP}(t)$ is NMP--that is, it contains a *mixture* of UHP and LHP zeros--then

$$m_{NMP,f_o}(t) = e^{j2\pi f_o t} \cdot |m_{NMP}(t)| \cdot e^{j\phi_{LP}(t)} \tag{10-14a}$$

$$= e^{j2\pi f_o t} \cdot |m_{MP}(t)| \cdot |m_{MaxP}(t)| \cdot e^{j[\phi_{MP}(t)+\phi_{MaxP}(t)]} \tag{10-14b}$$

where

$$M_{NMP,f_o}(f) = 0, \quad f_o > f > f_o+W \quad . \tag{10-15}$$

Then $\phi(t) = 2\pi f_o t + \phi_{MP}(t) + \phi_{MaxP}(t) \quad .$ (10-16)

Now, as per sec. 8.2.4,

$$\int_0^T \phi'_{MaxP}(t)\, dt = 2\pi n_{UHP} \ , \tag{10-17}$$

where n_{UHP} is the number of UHP zeros per period of $m_{NMP}(t)$. Hence,

$$2(f_o+n_{UHP}) \leq n_R(f_o) \leq 2(f_o+n_{UHP}+n_{LHP}) \ , \tag{10-18}$$

where $n_{UHP}+n_{LHP}=W$. As in the MP case, there exists a critical frequency, f_o, such that

$$2\pi f_o + \phi'_{MaxP}(t) + \phi'_{MP}(t) > 0. \tag{10-19}$$

$$\text{Therefore } n_R(f_o) \rightarrow 2(f_o+n_{UHP}) = 2(f_o+W[n_{UHP}/(n_{UHP}+n_{LHP})]) \tag{10-20}$$

as f_o becomes "large enough."

It follows that

i) a periodic lowpass signal s(t) of bandwidth ±W Hz will have at *least* $2n_{UHP}$ real zeros per period and at *most* $2W = 2(n_{UHP}+n_{LHP})$ real zeros per period, where n_{UHP} and n_{LHP} are the number of UHP and LHP zeros of the analytic counterpart, m(t), of s(t).

ii) SSB translation of s(t) will eventually—when f_o exceeds some critical value— yield a signal with $2(f_o+n_{UHP})$ zero crossings (RZ's) per period.

The relationship of these results to Good's conjecture can be exposed by rewriting (10-1) with $W_1=f_o$ and $W_2=f_o+W$:

$$n_R(f_o) = 2\left[\frac{(f_o+W)^3-f_o^{\ 3}}{3[(f_o+W)-f_o]}\right]^{\frac{1}{2}} \tag{10-21a}$$

$$= 2\ [f_o^{\ 2} + f_oW + W^2/3]^{\frac{1}{2}} \tag{10-21b}$$

$$= 2f_o[1 + W/f_o + W^2/3f_o{}^2]^{1/2} \quad (10\text{-}21c)$$

$$= 2\ [f_o + W/2 + W^2/6f_o + \ldots\ldots] \ . \quad (10\text{-}21d)$$

Observe that as f_o becomes "large",

$$n_R(f_o) \to 2\ [f_o + W/2] \ . \quad (10\text{-}22)$$

If it is reasonable to suggest that white noise has equal numbers of UHP and LHP zeros, then the W/2 term corresponds to the $W[n_{UHP}/(n_{UHP}+n_{LHP})]$ term in (10-20) and the higher order terms in W correspond to LHP zeros which do not cause zero crossings of $\phi(t)$ when f_o exceeds some critical value [V-10].

10.3 Application of Good's Conjecture to Bandpass Periodic Signals

Good's conjecture implies that, provided the number of zero crossings, per period, of a bandpass periodic signal is greater than twice the actual signal bandwidth, the complete set of signal parameters (i.e., Fourier coefficients) can be extracted, in some manner, from the zero crossing positional information. In sec. 10.2 we noted that SSB modulation of a lowpass signal with carrier frequency f_o results in a signal which possesses a *minimum* of

$$2(f_o+n_{UHP}) \quad (10\text{-}23)$$

zero crossings per period. Here, n_{UHP} is the number of UHP zeros in the analytic version $[s(t) + j\hat{s}(t)]$ of the lowpass signal, s(t). Therefore, SSB modulation of a speech signal of bandwidth ± W Hz such that the number of zero crossings per period is not less than 2W should enable a complete "recovery" of the signal parameters via zero crossing information.

A direct approach is to write the conventional expression for $s_{\omega_o}(t)$,

$$s_{\omega_o}(t) = s(t)\cdot\cos\omega_o t - \hat{s}(t)\cdot\sin\omega_o t$$

$$= \mathrm{Re}\{e^{j\omega_o t}\sum_{k=0}^{n} c_k\cdot e^{jk\Omega t}\} \quad . \tag{10-24}$$

Letting $t = t_m$, a zero crossing of $s_{\omega_o}(t)$, then,

$$0 = \tfrac{1}{2}[c_n^*\cdot e^{-j(\omega_o+n\Omega)t_m} + \; . \; . \; . \; . \; . \; + c_1^*\cdot e^{-j(\omega_o+\Omega)t_m}$$

$$+ c_1\cdot e^{j(\omega_o+\Omega)t_m} + \; . \; . \; . \; . \; . + c_n\cdot e^{j(\omega_o+n\Omega)t_m}$$

$$+ 2c_0\cdot\cos\omega_0 t_m]. \tag{10-25}$$

In matrix form, with $X_m = e^{-jt_m}$

$$c_0\begin{bmatrix} -2\cdot\cos\omega_o t_1 \\ -2\cdot\cos\omega_o t_2 \\ -2\cdot\cos\omega_o t_3 \\ \vdots \\ \vdots \\ -2\cdot\cos\omega_o t_{2n} \end{bmatrix} = \begin{bmatrix} X_1^{-(\omega_o+n\Omega)} & \cdots & X_1^{-(\omega_o+\Omega)}, X_1^{(\omega_o+\Omega)}, & \ldots & X_1^{(\omega_o+n\Omega)} \\ X_2^{-(\omega_o+n\Omega)} & & & & X_2^{(\omega_o+n\Omega)} \\ X_3^{-(\omega_o+n\Omega)} & & & & X_3^{(\omega_o+n\Omega)} \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ X_{2n}^{-(\omega_o+n\Omega)} & \ldots & \ldots & \ldots & X_{2n}^{(\omega_o+n\Omega)} \end{bmatrix} \begin{bmatrix} c_n^* \\ \vdots \\ c_1^* \\ c_1 \\ \vdots \\ c_n \end{bmatrix} \tag{10-26}$$

If the determinant of the $\underline{\underline{X}}$ matrix is non-zero, then we can solve for the Fourier coefficients, {c}. Good has shown [G-7], [G-10] that this is generally so. Thus, theoretically, we can use the zero crossing positions of $s_{\omega_o}(t)$ to resynthesize s_{ω_o} and, by "demodulation", $s(t)$.

It should be noted that certain classes of signals exist which provide a counter-example to Good's conjecture. The obvious example is AM-type signals where the only information conveyed by the zero crossings is the carrier frequency.

10.4 Overspecification in Vowel-like Signals

We have seen that SSB modulation of bandlimited signals theoretically allows, under the conditions outlined in sec. 10.3, complete reconstruction of the signal--to a mutiplicative constant--using zero crossing information only. A necessary condition is that the SSB signal translate possess a number of zero crossings greater than twice its lowpass bandwidth.

However, an important question remains concerning the nature of the information contained in the zero crossings of vowel-like signals. That is the following:

The percentage of zeros per period which are real--i.e., zero crossings--is usually on the order of 25% or less in speech vowels (sec. 9.3.3). In chapter 9 we offered an explanation of why the *power spectrum* of such signals is less altered by clipping than might be expected if clipping is simply considered as a member of that class of transformations which is capable of affecting *only* the complex zero signal. We demonstrated that clipping, effectively, has very little freedom in manipulation of the real parts of the complex zeros. Furthermore, we contended that, for vowel-like signals, the imaginary parts of the complex zeros before and after clipping are somewhat related. Before clipping, the CZ's are "near" the real time axis because of relationships which depend upon the formant structure (sec. 9.5.2). After clipping and bandlimiting the CZ's must remain "near" the real time axis in order to produce the ripple which is characteristic of the clipped, bandlimited signal (sec. 9.4.1). We also noted that, intuitively and observationally, the greater the percentage of real zeros, the higher the expected post-clipping intelligibility.

However, we have not yet proferred an explanation as to why *objective* estimates of speech spectral parameters made using only zero crossing information *are apparently capable of conveying what seems to be an inordinate amount of information.* In particular, zero crossing histograms (ch. 6, sec. 6.6) exhibit features quite analogous to formant structure.

Thus we ask whether it is possible that the real zeros—the zero crossing interval sequence—of a vowel-like signal might contain information concerning the complex zero component of that signal. In the next two sections we show that, for vowel-like signals, *under certain conditions* $s_{CZ}(t)$ *may effectively be almost completely derived from* $s_{RZ}(t)$.

10.4.1 Matrix Formulation

The basic equations relating s(t), $s_{RZ}(t)$ and $s_{CZ}(t)$ to their Fourier series expansions are given in sec. 8.1.3. These relationships may be expressed in matrix form as follows:

$$\underline{\underline{c}} = \underline{\underline{Rz}} \cdot \underline{\underline{Cz}} \tag{10-27}$$

where $\underline{\underline{c}}$ and $\underline{\underline{Cz}}$ are (2n+1) element column vectors consisting of the Fourier series coefficients of s(t) and $s_{CZ}(t)$, respectively,

$$\underline{\underline{c}} = \begin{bmatrix} c_n \\ c_{n-1} \\ \vdots \\ c_1 \\ c_0 \\ c_{-1} \\ \vdots \\ c_{-n+1} \\ c_{-n} \end{bmatrix} \quad (10\text{-}28a) \qquad \underline{\underline{Cz}} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ Cz_{n_C} \\ \vdots \\ Cz_1 \\ Cz_0 \\ Cz_{-1} \\ \vdots \\ Cz_{-n_C} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}}\right\} n_R \\ \left.\vphantom{\begin{matrix} Cz_{n_C} \\ \vdots \\ Cz_1 \\ Cz_0 \\ Cz_{-1} \\ \vdots \\ Cz_{-n_C} \end{matrix}}\right\} 2n_C+1 \\ \left.\vphantom{\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}}\right\} n_R \end{matrix} \qquad (10\text{-}28b)$$

and $[\underline{\underline{Rz}}]$ is the $(2n+1)\times(2n+1)$ square matrix

$$\begin{bmatrix} Rz_0 & Rz_1 & \cdots & Rz_{n_R} & 0 & \cdots & & & 0 \\ Rz_{-1} & Rz_0 & Rz_1 & \cdots & Rz_{n_R} & & & & \vdots \\ \vdots & & Rz_0 & & & & & & 0 \\ \vdots & & & Rz_0 & & & & & Rz_{n_R} \\ \vdots & & & & Rz_0 & & & & \vdots \\ \vdots & & & & & Rz_0 & & & \vdots \\ Rz_{-n_R} & & & & & & Rz_0 & & \vdots \\ 0 & & & & & & & Rz_0 & Rz_2 \\ \vdots & & & & & & & Rz_0 & Rz_1 \\ 0 & 0 & 0 & 0 & Rz_{-n_R} & \cdots & & Rz_{-1} & Rz_0 \end{bmatrix} . \qquad (10\text{-}29)$$

$\longleftarrow n_R+2n_C \longrightarrow \qquad \longleftarrow n_R+1 \longrightarrow$

Note that $\underline{\underline{[Rz]}}$ is Hermitian; that is, $\underline{\underline{[Rz]}}^t = \underline{\underline{[Rz]}}^*$.

$\underline{\underline{[Rz]}}$ represents the information contained in the signal $s_{RZ}(t)$.

Equation (10-27) actually represents a set of (2n+1) equations in (2n+1) unknowns and therefore has a unique solution provided that $|\underline{\underline{Rz}}| \neq 0$. For vowel-like signals many of the c_k can be considered to be zero; that is, the formant structure significantly dominates the spectrum. Specifically, for the three-tone model of vowels, all c_k *except those at the tone frequencies* are zero. In effect, this means that (10-27) represents (2n+1) equations, 6 of which are non-homogeneous with the remainder being homogeneous. Thus the existence and number of *linearly independent* $\underline{\underline{Cz}}$ vectors may be studied via rank constraints.[1]

Effectively, the gaps in the $\underline{\underline{c}}$ vectors may impose dependencies of the $\underline{\underline{Cz}}$ vector on the values of $\underline{\underline{Rz}}$. The problem becomes clearer if (10-27) is rewritten in recursive form.

10.4.2 Deconvolution

The problem of finding $\underline{\underline{Cz}}$ given $\underline{\underline{c}}$ and $\underline{\underline{Rz}}$ can also be formulated as one of deconvolution of $\{c_k\}$ with $\{Rz_k\}$. As per sec. 8.5.1 ii,

---

[1] This approach was suggested by A. Requicha.

$$
Cz_k = \begin{cases} c_{-(n_R+n_C)}/Rz_{-n_R}, & k=-n_C \\[2ex] \dfrac{c_{-n_R+k} - \sum\limits_{j=-n_C}^{k-1} Cz_j \cdot Rz_{k-j-n_R}}{Rz_{-n_R}}, & -n_C<k\leq 0 \\[2ex] Cz_{-k}^{*}, & 0<k\leq n_C . \end{cases} \tag{10-30}
$$

Now, let $s(t) = a_1\cos(2\pi F_1 t+\phi\ )+a_2\cos(2\pi F_2 t+\phi_2)+a_3\cos(2\pi F_3 t+\phi_3)$,

(10-31)

where $F_1=n_1\Omega/2\pi$, $F_2-F_1=n_2\Omega/2\pi$ and $F_3-F_2=n_3\Omega/2\pi$. (10-32)

ϕ_1, ϕ_2, and ϕ_3 are approximated as per sec. 9.5.3. This is a three-tone vowel model.

Examination of (10-30) shows that, for this model, if

$$
n_C+1 = n-n_R+1 < n_3 \tag{10-33}
$$

then *all* components of $\{Cz_k\}$ may be derived using only $\{Rz_k\}$ and $c_{-(n_R+n_C)}$: viz.,

$$
Cz_k = \begin{cases} c_{-(n_R+n_C)}/Rz_{-n_R}, & k=-n_C \\[2ex] \dfrac{- \sum\limits_{j=-n_C}^{k-1} Cz_j \cdot Rz_{k-j-n_R}}{Rz_{-n_R}}, & -n_C<k\leq 0 \\[2ex] Cz_{-k}^{*}, & 0<k\leq n_C . \end{cases} \tag{10-34}
$$

Furthermore, if $c_{-(n_R+n_C)}$ is assumed to have unit amplitude and zero phase angle, then

$$\tilde{C}z_k = \begin{cases} 1/Rz_{-n_R} , & k=-n_C \\ \dfrac{-\displaystyle\sum_{j=-n_C}^{k-1} \tilde{C}z_j \cdot Rz_{k-j-n_R}}{Rz_{-n_R}} , & -n_C<k\leq 0 \\ \tilde{C}z^*_{-k} , & 0<k\leq n_C , \end{cases} \tag{10-34}$$

where $\{\tilde{C}z_k\}$ are estimates of $\{Cz_k\}$. Thus

$$\tilde{C}z_k = Cz_k/\ddot{c}_{-n} \tag{10-35}$$

so that

$$\tilde{c}_k = \sum_{n=\max\{-n_R,k-n_C\}}^{\min\{n_R,k+n_C\}} Rz_n \cdot \tilde{C}z_{k-n} \tag{10-36a}$$

$$= c_k/c_{-n} \quad .$$

Therefore, $$|\tilde{c}_k| = |c_k|/|c_{-n}| \tag{10-37}$$

and *the power spectrum of s(t) is determined to a multiplicative constant entirely by deconvolution of the RZ signal with unity.*

In effect, this has demonstrated that if the percentage of zeros (in a three-tone periodic signal) which are real is sufficiently large, i.e., for a given $F_3 = nF_o$, from (10-33)

$$\text{if} \quad n_R > F_2/F_o + 1 \tag{10-38}$$

then *the RZ signal contains sufficient information to reconstruct the power spectrum of s(t) to a multiplicative constant.*

For this method to be implemented the location of F_3--the effective signal bandwidth--must be known. The number of zero crossings per period is countable. However, the value of $n_3 = (F_3 - F_2)/F_o$ is usually unknown and, except for the fact that

$$|\tilde{C}z_{-k}| = |\tilde{C}z_k| \quad , \quad |k| < n_C \tag{10-39}$$

there is no way of knowing when to stop the deconvolution. In practice, then, although it may be theoretically possible that the power spectrum of a three-tone signal be calculated (to a multiplicative constant) entirely from the RZ signal, it may not be possible to do so because of a lack of information. Nevertheless, under the aforementioned conditions, tracking of the *location* of F2, F_2, and the *location* of F3, F_3, would enable --using zero crossing information, i.e., $s_{RZ}(t)$--"exact" estimation of the *location* of F1, F_1, and the *amplitudes* (to a multiplicative constant) of F1, F2, and F3.

As we have mentioned before (sec. 9.3.5), the three-tone signal is unrealistic in the sense that each sinusoidal component lacks the damping which is present in actual vowel signals because the poles are not on the jω axis, but slightly to the left [F-2, p. 51]. However, the same arguments can be extended to a more realistic model involving, for example, a 3-component reresentation for each formant. The phase angle of the components on either side of the spectral component nearest F_3 can be effectively evaluated using the formant resonator model (sec. 6.3.3). The 3-component F3 complex can then be deconvolved with $\{Rz_k\}$ as before.

11 CONCLUSIONS, MAJOR PROBLEMS, AND RECOMMENDATIONS FOR FURTHER RESEARCH

11.1 Zero Crossings, the Intelligibility of Clipped Speech, and Objective Estimation of Speech Spectral Parameters

11.1.1 Voiced Sounds

In chapter 3 we briefly reviewed the spectral and time-domain characteristics of the sounds of speech. We noted that voiced sounds, including vowels, are quasi-periodic and are most accurately represented over a pitch period by a finite Fourier series. We also observed that voiced speech sounds are characterized by spectral features (formants) which are (experimentally) sufficient to enable a high degree of correct perceptual classification when peripheral cues such as onset, duration and context are absent. In addition, we showed that formant positions possess meaningful physiological correlates and that manipulation of formant positions results in changes in the identity of the perceived vowel. Thus, we argued that preservation of overall spectral structure is, at least, desireable for retention of intelligibility.

After reviewing Licklider's classic experiments on the intelligibility of clipped speech, we noted that Licklider concluded that the intelligibility of clipped speech could be justified by observing that "although many details of the [speech spectral] pattern are changed by infinite clipping, the general . . . structure . . . is by no means rendered unrecognizeable . . . only the details of the intensity-frequency-time pattern are modified." Thus Licklider implicitly accepted the assumptions concerning intelligiblity and power spectrum preservation which we felt necessary to establish, in some detail, by reference to extant experimental results.

Using the concepts associated with zero-based periodic signal models, we observed (in chapter 8) that zero crossings generally permit only a partial description of a bandlimited periodic signal. The total information necessary (and sufficient) for complete signal specification is apparently shared by the zero crossings (RZ's) and the complex zeros (CZ's) which, via the product formulation, specify the signal completely, to a multiplicative constant. We further noted that certain operations (e.g., differentiation and sine wave "carrier" addition) tend to convert CZ pairs to RZ's and thus provide more information *in the form of zero crossings.*

Thus, the fact that pre-clipping differentiation (which affects only the quality but not the intelligibility of the *original* speech signal) yields a *clipped* speech signal which is more intelligible and/or of higher subjective quality than the clipped then bandlimited original speech signal may be attributed, in part, to the fact that a greater percentage of zeros -- in the form of zero crossings -- are available for "perfect" sampling by the clipping-bandlimiting operator. Similarly, Licklider's passing remark concerning the improvement in post-clipping intelligibility resulting from overly large ultra-sonic bias and/or highpass filtering at 250 Hz may be explained in terms of zero conversion processes.

At the conclusion of chapter 5 we contended that clipping preserves other waveform attributes in addition to zero crossing (RZ) positions. In chapter 9 we showed experimentally, and to some extent using the theory of polynomials, that the nature of the clipping-bandlimiting operator is such that the real time positions of the complex zero pairs -- which are waveform attributes -- are *effectively* preserved by "clipping". Furthermore, the ability of the clipping-bandlimiting operation to manipulate the imaginary positions of the CZ's is somewhat restricted because of the constant amplitude nature of clipping and the fact that the formant structure of vowels is sufficient to ensure that their CZ's are "near" the real time axis.

In chapter 6 we reviewed, in some detail, a number of schemes for obtaining objective estimates of speech spectral parameters from zero crossing measurements. We emphasized that methods which effectively count zero crossing rates give poor estimates of formant frequencies unless pre-filtering is used to isolate the formants. Even then, the possible errors are as great as those encountered in spectral "peak picking" formant trackers. In effect, *pre-filtering is a method of increasing the number of zero crossings available as information carriers.* That is, the waveform emerging from each bandpass filter has a number of zero crossings bounded as per equation (10-7). It follows that the total number of zero crossings utilized in the estimates may exceed the number which we know is capable of specifying the signal *completely* (to a multiplicative constant). Thus, the question of how to process speech, *using zero crossing methods*, so as to obtain the best possible estimates of the speech spectral parameters may be answered rather straightforwardly. That is to say, there is a minimum number of zero crossings (depending on the highest frequency present in the speech signal) which are sufficient to completely reconstruct the signal and therefore (via the discrete Fourier transform) *completely* specify the speech spectrum.

The question of how to force the speech signal to exhibit the requisite number of zero crossings is clear, but does not present a practical solution. Multiple differentiation is associated with noise problems and when the required zero crossing count is obtained, it is the multiply differentiated signal which is completely specified by its zero crossings.* Sine wave addition entails similar problems. When the amplitude is sufficient to convert all CZ's to RZ's, the resultant signal is effectively a sine wave whose zero crossing positions are "phase modulated" by the original signal. In both cases, the total zero count necessary for complete signal specification is *apparently* identical to the number of Nyquist samples required for the same purpose.

We say *apparently* because Good's conjecture states that the number of zero crossings actually needed is numerically equal to twice the signal *bandwidth* rather than twice the *highest frequency* present in the signal, as the product formulation implies. As noted in sec. 10.2, SSB modulation provides a signal with the requisite number of zero crossings -- according to Good -- if the carrier frequency becomes "large enough". Such a signal contains CZ's as well as RZ's. *Good's conjecture implies that these CZ's are entirely specified by the RZ's.* At present, methods of recovering these CZ's are under investigation. Voelcker has shown experimentally that CZ recovery is, in some instances, entirely possible.

The notion that CZ's may be completely determined by RZ's has its analogy in the lowpass speech signal case. Again,

*Thus, real zero encoding of salient perceptual features of the original signal is not necessarily equated with preservation of original signal attributes (e.g., the original amplitude spectrum).

we noted (in chapter 6) that various researchers have experimentally demonstrated that the information contained in the zero crossing intervals of speech signals, especially vowels, can be displayed (as a histogram) so as to exhibit information similar to that seen in a short-term speech spectrogram. Furthermore, Focht and Scarr have presented convincing evidence that all zero crossing intervals of vowels may not be *equally* informative (the SEF concept, for example). (These ideas are in consonance with the findings of Stover; he reported that deletion of all but the *first* 3 msec. of *each* pitch period of a voiced sound leaves a highly intelligible signal.)

Finally, in chapter 10 we showed that for three-tone vowel models, at least, the zero crossings (via $s_{RZ}(t)$) may contain "encoded" information which specifys the power spectrum of the signal to a multiplicative constant. Thus, it is apparent that *for highly structured signals such as voiced speech,* the statement that the amount of information carried by the zero crossings is proportional to the percentage RZ's is not strictly correct. The RZ's *and* CZ's are determined by the signal spectral structure (via the Fourier series polynomial) and it appears that the RZ's (as per the arguments of 10.4.2) do contain CZ information.

In 1959 A.J. Fourcin demonstrated that, from an information theoretic point of view, the (experimentally determined) long-term probability of finding a zero crossing interval length $n\tau$ in a time quantized, clipped differentiated speech sample (quantizing interval $\tau = 10^{-4}$ sec) was such that the zero crossing interval information is transmitted at about 80% of the maximum rate possible for a *two-level, time quantized signal* [F-13].

It appears that the zero crossings of speech waveforms are distributed so as to produce efficient transmittion of information *via a zero crossing mode.* Again, this implies that in speech signals the RZ's are highly CZ dependent.

11.1.2 Consonants

As noted in sec. 3.4, the consonants are characterized by changing, rather than relatively stable, vocal system configurations and spectra. In addition, the characteristics of the signal models which *realistically* describe consonants are varied and generally different from those of vowels. Fricatives and stops, for example, are most aptly described as noise-like and the long-term experimental amplitude distribution for the consonants is Gaussian (sec. 3.5.1).

We believe that the methods employed by Dukes and Fawe (sec. 5.3) in an attempt to explain the intelligibility of clipped speech (e.g., the arcsine law) provide a realistic basis for the belief that the power spectrum of speech sounds which may *realistically* be represented by bands of Gaussian "white" noise may not be significantly altered by clipping. However, the details are far from complete and the treatment of sounds which involve voicing *and* noise production (the voiced stops) deserves attention.

11.2 Zero Crossing-Related Speech Processing Schemes

Various schemes have been implemented in order to increase the naturalness of the clipped speech waveform for speech transmission purposes. We have already described the attempt of Sobolev ([S-17], sec. 8.4.2) to use a modified rectangular waveform for zero crossing interpolation. In addition, schemes for *augmenting* zero crossing information have been investigated.

Mathews showed that transmission of the amplitudes of a speech signal at its extrema (i.e., at the times of the zero crossings of the signal's derivative) as well as the times of the extrema produces a subjectively better signal. The penalty paid is a threefold increase in channel capacity over that required to transmit the zero crossings of s'(t) alone [M-7]. Spogen [S-20] attempted, with little success, to use envelope information to weight the clipped speech signal. Further attempts involved amplitude sampling the original speech signal at times at extrema and holding that amplitude until the following extremum. In this manner, a signal was obtained, which when filtered, produced a waveform significantly more intelligible than clipped speech.

In a sense, these schemes simply supplement the RZ information (of s'(t)) with information culled from the CZ component (of s(t)) in a somewhat arbitrary manner.

11.3 Problems and Recommendations for Future Research

11.3.1 Phase Distortion

No report (known to us) on speech recognition and processing using zero crossings has seriously mentioned the most critical obstacle in this field, *phase distortion*. Experimentally, speech passed through an all-pass network (e.g., a Hilbert transformer) and then clipped is as intelligible as speech which is clipped without deliberate phase distortion. In addition, limited experiments (again using a Hilbert transformer) have shown that the long-term (3-5 minutes) average rate of zero crossings is approximately unaffected by the network.

However, our own experience and that of others,[1] has shown that those schemes based on zero crossing interval distributions are extremely sensitive to *any* change in the phase characteristics of the speech transmission components. The problems of phase distortion appear then to present an insurmountable barrier to the use of zero crossing histograms in automatic speech recognition machines. After all, the power spectrum is *unaffected* by phase distortion. However, the success of the schemes of Teacher *et al.* and others in using zero crossing information for automatic recognition suggests that phase distortion is simply an unsolved problem which should constitute an area of future research.

11.3.2 Zero Crossings and Spectral Estimation

Zero crossings have been used in two ways for automatic speech recognition: they have yielded estimates of formant position and, via interval histograms, patterns representative of the original signal.

However, it has been demonstrated (experimentally) that spectral features alone are not sufficiently invariant to give high rates of automatic recognition if, for example, more than one speaker must be recognized. In any case, the usefulness of zero crossings in this respect apparently depends upon their ability to yield formant frequency estimates in a *simpler* manner than more conventional methods. Peterson and Hanne, for instance, have shown that even under optimum circumstances (i.e., an isolated formant) simple zero crossing spectral estimates are subject to the

---

[1] Personal Communication, R. W. Scarr.

same large error as "peak picking" techniques. Thus, the question arises as to whether zero crossing techniques are relatively complicated methods of estimating feature parameters which may be evaluated more directly by conventional DFT-FFT operations. The utlity of zero crossings as"sufficient statistics" in speech recognition schemes is, we feel, intimately related to Good's conjecture regarding the information contained in zero crossing interval sequences of structured signals.

Appendix A: Bounds on the Imaginary Parts of Complex Zeros -- the Lehmur-Schur Algorithm

We derive a rather close approximation for the minimum radius ($r < 1$) at which the zeros of the Fourier series polynomial which represents the three-tone vowel model are found.

First we state the algorithm upon which the proof is based. Then we work through an example which demonstrates the use of the algorithm. This example suggested the manipulation which allows the above mentioned bound to be derived.

A.1 The Lehmur-Schur Algorithm

The Lehmur-Schur algorithm is used to determine whether or not a zero of a polynomial lies within the unit circle on the w plane [L-9, R-2, pp. 355-359] .

Given

$$f(w) = a_n w^n + a_{n-1} w^{n-1} + \ . \ . \ . \ . + a_0 \qquad \text{(A-1)}$$

then define

$$f^*(w) = a_0^* w^n + a_1^* w^{n-1} + \ . \ . \ . \ . \ . + a_n^* \ . \qquad \text{(A-2)}$$

Further define an operator

$$T[f(w)] = a_0^* \cdot f(w) - a_n f^*(w) \ , \qquad \text{(A-3)}$$

so that

$$T[f(0)] = a_0^* \cdot a_0 - a_n \cdot a_n^* \qquad \text{(A-4a)}$$

$$= |a_0|^2 - |a_n|^2 \ . \qquad \text{(A-4b)}$$

Note that $T[f(w)]$ has no term in w^n,

$T\{T[f(w)]\}$ has no term in w^{n-1},

so that $T^j[f(w)] = T\{T^{j-1}[f(w)]\}$ has no term in w^{n+1-j} or higher.

Let k be the smallest integer for which $T^k[f(0)] = 0$. The basic theorem is as follows [R-2, p. 355]:

Suppose $f(0) \neq 0$. If, for some h such that $0<h<k$, $T^h[f(0)]<0$, then $f(w)$ has *at least* one zero inside the unit circle. If instead $T^i[f(0)] > 0$ for $1<i\leq k$ *and* $T^{k-1}[f(w)]$ is a constant, then no zero of $f(w)$ lies inside the unit circle.

We are concerned exclusively with self-inversive polynomials so that

$$T[f(0)] = |a_0|^2 - |a_n|^2 = 0 \quad . \tag{A-5}$$

Thus, for useful results we must apply the Lehmur-Schur algorithm to the function $f(rw)$, $r<1$, and establish whether $f(w)$ has a root within the circle $|w| = r$. This is the key to our method.

A.2 Demonstration: the Two Component Square Wave

Factorization of the polynomial representing the two component square wave.

$$f_1(w) = \frac{1}{3}(jw^6 + j3w^4 - j3w^2 - j1) = 0 \tag{A-6}$$

reveals zeros at $w = \pm 1$, $\pm j\,1.9319$, $\pm j\,0.5176$. Application of the L-S algorithm to $f(0.8w)$ should therefore yield a positive result:

$$f_1(0.8w) = j0.26w^6 + j1.23w^4 - j1.92w^2 - j1 \tag{A-7a}$$

$$f_1^*(0.8w) = jw^6 + j1.92w^4 - j1.23w^2 - j0.26 \tag{A-7b}$$

$$(A\text{-}7a)x(a_0^*) = -0.26w^6 - 1.23w^4 + 1.92w^2 + 1.0 \quad (A\text{-}8a)$$

$$(A\text{-}7b)x(-a_n) = 0.26w^6 + 0.50w^4 - 0.32w^2 - 0.068 \quad . \quad (A\text{-}8b)$$

Add (A-8a) and (A-8b), giving

$$f_2(w) = -0.73w^4 + 1.62w^2 + 0.93 \quad (A\text{-}9a)$$

and $$f_2^*(w) = 0.93w^4 + 1.62w^2 - 0.73 \quad . \quad (A\text{-}9b)$$

$$(A\text{-}9a)x(0.93) = -0.68w^4 + 1.48w^2 + 0.87 \quad (A\text{-}10a)$$

$$(A\text{-}9b)x(0.73) = 0.68w^4 + 1.18w^2 - 0.55 \quad (A\text{-}10b)$$

Add (A-10a) and (A-10b), giving

$$f_3(w) = 2.66w^2 + 0.31 \quad (A\text{-}11a)$$

and $$f_3^*(w) = 0.31w^2 + 2.66 \quad . \quad (A\text{-}11b)$$

$$(A\text{-}11a)x(0.31) = 0.83w^2 + 0.31^2 \quad (A\text{-}12a)$$

$$(A\text{-}11b)x(-2.66) = -0.83w^2 - 2.66^2 \quad (A\text{-}12b)$$

Add (A-12a) and (A-12b), giving $0.31^2 - 2.66^2 < 0$.

Therefore, there is a root of $f(0.8w)$ inside the unit circle or a root of $f(w)$ inside the circle $|w| = 0.8$, as expected. It is, of course at $w = \pm j\ 0.5176$.

Evaluation of the set of derived functions corresponding to $r = 0.55$ yields a constant term of $0.94^2 - 1.12^2 < 0$ after the same number of operations as above. A similar evaluation for $r = 0.48$ yields a constant term of $0.97^2 - 0.80^2 > 0$. Thus, as the "test" radius approaches the radius at which the *actual root of* smallest magnitude is located, the sign of the constant term remaining after p-1 "T" operations (where p is the number of non-zero terms in the original self-inversive polynomial) changes from a negative quantity (indicating at least one root inside of the test radius) to a positive quantity.

Algebraically, our problem is to find the radius at which the remainder term is identically equal to zero after the prescribed number of operations. We shall demonstrate the algebraic derivation of our criterion using a three-tone model having "formants" at 400, 1000 and 2500 Hz (assuming a fundamental voicing frequency of 100 Hz). The tone complexes have been SSB modulated upwards 100.N Hz, where N is a positive integer (or zero for the lowpass signal).

i.e., $$f(w) = a_3 r^{50+2N} w^{50+2N} + a_2 r^{35+2N} w^{35+2N} + a_1 r^{29+2N} w^{29+2N} + a_1^* r^{21} w^{21} + a_2^* r^{15} w^{15} + a_3^* \quad , \qquad \text{(A-13a)}$$

where $a_1 > a_2 > a_3$.

Then

$$f^*(w) = a_3 w^{50+2N} + a_2 r^{15} w^{35+2N} + a_1 r^{21} w^{29+2N} + a_1^* r^{29+2N} w^{21} + a_2^* r^{35+2N} w^{15} + a_3^* r^{50+2N} \quad . \qquad \text{(A-13b)}$$

$$\text{(A-13a)}\times(a_3) = a_3^2 r^{50+2N} w^{50+2N} + a_2 a_3 r^{35+2N} w^{35+2N} + a_1 a_3 r^{29+2N} w^{29+2N} + a_3 a_1^* r^{21} w^{21} + a_3 a_2^* r^{15} w^{15} + |a_3|^2 \qquad \text{(A-14a)}$$

$$\text{(A-13b)}\times(a_3 r^{50+2N})$$

$$= a_3^2 r^{50+2N} w^{50+2N} + a_2 a_3 r^{65+2N} w^{35+2N} + a_3 a_1 r^{71+2N} w^{29+2N} + a_3 a_1^* r^{79+4N} w^{21} + a_3 a_2^* r^{85+4N} w^{15} + |a_3|^2 |r^{100+4N} \quad . \qquad \text{(A-14b)}$$

Subtract (A-14b) from (A-14a):

$$r^{2N} a_3 a_2 (r^{35} - r^{65}) w^{35+2N} + r^{2N} a_1 a_3 (r^{29} - r^{71}) w^{29+2N}$$

$$+a_3a_1^*(r^{21}-r^{79+4N})w^{21} + a_3a_2^*(r^{15}-r^{85+4N})w^{15}$$

$$+ a_3^2(1-r^{100+4N}) \quad . \tag{A-15}$$

Therefore,

$$f_2(w) \simeq a_3a_2r^{35+2N}w^{35+2N}+a_1a_3r^{29+2N}w^{29+2N}$$

$$+a_3a_1^*r^{21}w^{21}+a_3a_2^*r^{15}w^{15} + |a_3|^2 \tag{A-16}$$

if $r^{35}>>r^{65}$, $r^{29}>>r^{71}$, $r^{21}>>r^{79+4N}$, $r^{15}>>r^{85+4N}$, $1>>r^{100+4N}$

or, equivalently, r^{30}, r^{42}, r^{58+4N}, r^{70+4N} and r^{100+4N} are all much less than unity. Because we are concerned with self-inversive polynomials, $r < 1$. If $r = 0.91$, $r^{30} \simeq 0.05$, and as r becomes much smaller, r^{30} -- and all higher powers of r -- become negligible compared to unity. The observed minimum value of r for actual vowels was $r \simeq 0.72$, corresponding to a σ of 0.5 msec. with $\Omega = 2\pi.100$ rad/sec. Thus, if the zero structure of the three-tone model is similar to that of the actual vowel -- as far as minimum radius at which a complex zero may be found -- the approximations of (A-16) should be valid.

If reduction of (A-16) by the "T" operations is continued, then using assumptions similar to those in (A-16) (i.e., for $r < 0.9$, high powers of r become very small) we finally find that the radius at which the zero of least magnitude is found is given by

$$r \simeq [|a_2| \cdot |a_3|]^{-1/15} \tag{A-17}$$

where 15.100 Hz is the separation of F_2 and F_3. More generally, if the distance between F2 and F3 is $p(\Omega/2\pi)$ Hz -- $5<p<16$ for vowels, generally - then

$$r \simeq [|a_2| \cdot |a_3|]^{-1/p} . \tag{A-18}$$

The estimate becomes better as the degree of SSB modulation or signal translation increases from 0 Hz and for a given degree of SSB modulation is best for larger p. We emphasize that the estimated r is a function only of p so that as SSB modulation is applied, the estimate remains the same but the actual root of least magnitude approximates the estimate more precisely.

List of References Consulted

[A-1] R. Ahmed, "Vowel recognition in clipped speech," *Nature*, vol. 181, p. 210, January 1958.

[A-2] R. Ahmend and R. Fatechand, "Effects of sample duration on the articulation of sounds in normal and clipped speech," *J. Acoust. Soc. Am.*, vol. 31, pp. 1022-1029, July 1959.

[A-3] W.A. Ainsworth, "Relative intelligibility of different transforms of clipped speech," *J. Acoust. Soc. Am.*, vol. 41, pp. 1272-1276, June 1967.

[A-4] S. Amari, "A theory of adaptive pattern classifiers," *IEEE Trans. Electronic Computers*, vol. EC-16, pp. 299-307, June 1967.

[B-1] G. Békésy, *Experiments in Hearing*. New York: McGraw-Hill, 1966.

[B-2] C.G. Bell, H. Fujisaki, J.M. Heintz, K.N. Stevens and A.S. House, "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Am.*, vol. 33, pp. 1725-1736, December 1961.

[B-3] J.S. Bendat, *Principles and Applications of Random Noise Theory*. New York: John Wiley, 1958.

[B-4] W. Bezdel, "Discriminators of sound classes for speech recognition purposes," 1967 Conf. on Speech Communication and Processing, Bedford, Mass.

[B-5] W. Bezdel and J.S. Bridle, "Speech recognition using zero crossing measurements and sequence information," *Proc. IEE*, vol. 116, pp. 617-623, April 1969.

[B-6] W. Bezdel and H.J. Chandler, "Results of an analysis and recognition of vowels by computer using zero-crossing data," *Proc. IEE*, vol. 112, pp. 2060-2066, November, 1965.

[B-7] W. Bezdel and R.W. Scarr, "Approaches to speech recognition equipment based on zero crossings and other speech features," Colloquium on Some Aspects of Speech Recognition for Man Machine Communications, London, April 1968. *IEE Colloquium Digest No. 1968/3.*

[B-8] R. Biddulph, "Short-term autocorrelation analysis and correlatograms of spoken digits," *J. Acoust. Soc. Am.*, vol. 26, pp. 539-541, July 1954.

[B-9] R.B. Blackman and J.W. Tukey, *The Measurement of Power Spectra.* New York: Dover, 1959.

[B-10] W.A. Blankinship, "Method of time normalization for speech," *J. Acoust. Soc. Am.*, vol. 34, p. 242, February 1962.

[B-11] R.E. Bogner, "Frequency division in speech bandwidth reduction," *IEEE Trans. Communication Technology*, vol. COM-13, pp. 438-451, December 1965.

[B-12] F.E. Bond and C.R. Cahn, "On sampling the zeros of band-limited signals," *IRE Trans. on Information Theory*, vol. IT-4, pp. 110-113, September 1958.

[B-13] F.E. Bond and C.R. Cahn, "A relationship between zero crossings and Fourier coefficients for bandwidth limited functions," *IEEE Trans. on Information Theory*, vol. IT-6, pp. 51-52, March 1960.

[B-14] R.E. Bonner, "Pattern recognition with three added requirements," *IEEE Trans. on Electronic Computers*, vol. EC-15, pp. 770-781, October 1966.

[B-15] F. Bonsall and M. Marden, "Zeros of self-inversive polynomials," *Proc. Amer. Math. Soc.*, vol. 3, pp. 471-475, June 1952.

[B-16] R. Bracewell, *The Fourier Transform and Its Applications.* New York: McGraw-Hill, 1965.

[B-17] E.M. Braverman, "Experiments on machine learning to recognize visual patterns." *Automation and Remote Control*, vol. 25, pp. 315-327, March 1962.

[B-18] D. Braverman, "Theories of Pattern Recognition," in *Advances in Communications Systems*, A.V. Balarishman, Editor. London: Academic Press, 1965.

[B-19] E.M. Braverman and A.G. Arkadev, *Teaching Computers to Recognize Patterns.* London: Academic Press, 1966.

[B-20] H.J. Bremerman, "Pattern recognition, functionals and entropy," *IEEE Trans. on Bio-Medical Engineering*, vol. BME-15, pp. 201-207, July 1968.

[B-21] S. Leroy Brown and Lisle L. Wheeler, "A mechanical method for graphical solution of polynomials," *J. Franklin Institute*, vol. 231, pp. 223-243, March 1941.

[B-22] J.D. Bruce, "Discrete Fourier transforms, linear filters and spectrum weighting," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, pp. 495-499, December 1968.

[B-23] W. Bürck, P. Kotowski, and H. Lichte, "Development of pitch sensations," *Elektrische Nachrichen-Technik*, vol. 12, p. 326, 1935.

[C-1] C.R. Cahn, "A note on signal to noise ratio in bandpass limiters," *IRE Trans. on Information Theory*, vol. IT-7, pp. 39-43, January 1961.

[C-2] E.C. Carterette, "A simple linear model for vowel perception," Symposium on Models for the Perception of Speech and Visual Form, Boston, November 1964.

[C-3] S.H. Chang, "Portrayal of some elementary statistics of speech sounds," *J. Acoust. Soc. Am.*, vol. 22, pp. 768-769, November 1950.

[C-4] S.H. Chang, G.E. Pihl and M.W. Essigman, "Representations of speech sounds and some of their statistical properties," *Proc. IRE*, vol. 39, pp. 147-153, February 1951.

[C-5] S.H. Chang, G.E. Pihl and J. Wiren, "The intervalgram as a visual representation of speech sounds," *J. Acoust. Soc. Am.*, vol. 23, pp. 675-679, November 1951.

[C-6] S.H. Chang, "Two schemes of speech compression system," *J. Acoust. Soc. Am.*, vol. 28, pp. 565-572, July 1956.

[C-7] E.C. Cherry, *On Human Communication*. New York: John Wiley, 1957.

[C-8] E.C. Cherry, M. Halle and R. Jakobson, "Toward a logical description of languages in their phonetic aspects," *Language*, vol. 29, pp. 34-46, January 1953.

[C-9] E.C. Cherry and V.J. Phillips, "Some possible uses of single sideband signals in formant-tracking systems," *J. Acoust. Soc. Am.*, vol. 33, pp. 1067-1077, August 1961.

[C-10] N. Chomsky, *Language and Mind.* New York: Harcourt, Brace and World, 1968.

[C-11] H. Cramer and M.R. Leadbetter, *Stationary and Related Stochastic Processes--Sample Function Properties and Their Applications.* New York: John Wiley, 1967.

[D-1] J.L. Daguet, "Speech compression CODIMEX system," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-11, pp. 63-71, March-April 1963.

[D-2] J.E. Damman, "Application of adaptive threshold elements to the recognition of acoustic phonetic states," *J. Acoust. Soc. Am.*, vol. 38, pp. 213-223, August 1965.

[D-3] W.B. Davenport, "A study of speech probability distributions," *Technical Report #148*, Research Laboratory of Electronics, M.I.T., 1950.

[D-4] W.B. Davenport, "An experimental study of speech probability distributions," *J. Acoust. Soc. Am.*, vol. 24, pp. 390-399, July 1952.

[D-5] W.B. Davenport and W. Root, *An Introduction to the Theory of Random Signals and Noise.* New York: McGraw-Hill, 1958.

[D-6] K.H. Davis, R. Biddulph and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Am.*, vol. 24, pp. 637-642, November 1952.

[D-7] P. Delattre, "The physiological interpretation of sound spectrograms," *Publications of the Modern Language Association of America*, vol. 66, pp. 864-875, September 1951.

[D-8] P.C. Delattre, A.M. Liberman and F.S. Cooper," Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Am.*, vol. 27, pp. 769-773, July 1955.

[D-9] L.N. Delves and J.N. Lyness, "A numerical method for locating the zeros of an analytic function," *Math. of Computation*, vol. 27, pp. 543-560, October 1967.

[D-10] P.B. Denes, "On the statistics of spoken English," *J. Acoust. Soc. Am.*, vol. 35, pp. 892-904, June 1963.

[D-11] P.B. Denes and M.V. Mathews, "Spoken digit recognition using time-frequency pattern matching," *J. Acoust. Soc. Am.*, vol. 32, pp. 1450-1455, November 1960.

[D-12] R.O. Duda and H. Fossum, "Pattern classification by iteratively determined linear and piecewise linear discriminant functions," *IEEE Trans. on Electronic Computers*, vol. EC-15, pp. 220-232, April 1966.

[D-13] H. Dudley, "Speech analysis by waveform," M.I.T. Lincoln Lab preprint, 1965.

[D-14] H. Dudley and S. Balashek, "Automatic recognition of phonetic patterns in speech," *J. Acoust. Soc. Am.*, vol. 30, pp. 721-732, August 1958.

[D-15] J. Dugunjii, "Envelopes and pre-envelopes of real waveforms," *IRE Trans. on Information Theory*, vol. IT-4, pp. 53-57, March 1958.

[D-16] J.M.C. Dukes, "The effect of severe amplitude limitation on certain types of random signal: a clue to the intelligibility of infinitely clipped speech," *IEE Monograph No. 111R*, pp. 88-97, November 1954.

[D-17] H.K. Dunn, "The calculation of vowel resonances and an electrical vocal tract," *J. Acoust. Soc. Am.*, vol. 22, pp. 740-753, November 1950.

[D-18] H.K. Dunn, "Methods of measuring vowel formant bandwidths," *J. Acoust. Soc. Am.*, vol. 33, pp. 1737-1746, December 1961.

[E-1] P. Erdös, "On the uniform distribution of the roots of certain polynomials," *Annals of Math.*, vol. 43, pp. 59-64, January 1942.

[E-2] P. Erdös and P. Turán, "On a problem in the theory of uniform distribution, I and II," *Akademie van Wetenscheppen, Amsterdam*, vol. 51, pp. 1146-1154 and 1262-1269, no. 9-10, 1948.

[E-3] P. Erdös and P. Turán, "On the distribution of roots of polynomials," *Annals of Math.*, vol. 51, pp. 105-119, January 1950.

[E-4] G.D. Ewing and J.F. Taylor, "Computer recognition of speech using zero crossing information," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-17, pp. 33-36, March 1969.

[F-1] R.M. Fano, "Short-time autocorrelation functions and power spectra," *J. Acoust. Soc. Am.*, vol. 22, pp. 546-550, September 1950.

[F-2] G. Fant, *Acoustic Theory of Speech Production.* 'S-Gravenhage: Mouton and Co., 1960.

[F-3] G. Fant, K. Fintoft, J. Liljencrants, B. Lindblom and J. Martony, "Formant-amplitude measurements," *J. Acoust. Soc. Am.*, vol. 35, pp. 1753-1761, November 1963.

[F-4] A.L. Fawe, "Interpretation of infinitely clipped speech properties," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-14, pp. 178-183, December 1966.

[F-5] W. Feller, *An Introduction to Probability Theory and Its Applications, vol. 1, Third Edition.* New York: John Wiley, 1968.

[F-6] B.A. Fersman, "An experimental investigation of the statistical properties of the music and speech signals which are contained in radio broadcasts," *Soviet Physics-Acoustics*, vol. 3, pp. 292-298, July-Sept. 1957.

[F-7] L.M. Fink, "Relations between the spectrum and instantaneous frequency of a signal," *Problems of Information Transmission*, vol. 2, pp. 26-38, Winter, 1966.

[F-8] J.L. Flanagan, *Speech Analysis, Synthesis and Perception.* Berlin: Springer-Verlag, 1965.

[F-9] H. Fletcher, *Speech and Hearing in Communication.* New York: D. Van Nostrand, 1953.

[F-10] L.R. Focht, "The single equivalent formant," *IEEE Inter-Communications Conference Digest*, p. 106, 1966.

[F-11] J.W. Forgie and C.D. Forgie, "Results obtained from a vowel recognition computer program," *J. Acoust. Soc. Am.*, vol. 31, pp. 1480-1489, November 1959.

[F-12] J.D. Foulkes, "Computer identification of vowel types," *J. Acoust. Soc. Am.*, vol. 33, pp. 7-11, April 1961.

[F-13] A.J. Fourcin, "An investigation into some possibilities for the reduction of channel capacity in speech transmission," *Ph.D. Thesis*, University of London, 1959.

[F-14] A.J. Fourcin, "A note on the spectral analysis of unvoiced sounds," *Proceedings of the Fifth International Congress of Phonetic Sciences*, pp. 287-291, 1964.

[F-15] B. Friedman, "Note on approximating the complex zeros of a polynomial," *Comm. Pure and Applied Math.*, vol. 2, pp. 195-208, June-Sept. 1949.

[F-16] D.B. Fry, "Prospects and problems in mechanical speech recognition," Colloquium on Some Aspects of Speech Recognition for Man Machine Communications, London, April 1968. *IEE Colloquim Digest No. 1968/3.*

[F-17] D.B. Fry and P. Denes, "The solution of some fundamental problems in mechanical speech recognition," *Language and Speech*, vol. 1, pt. 1, pp. 35-58, January-March 1958.

[F-18] O. Fujimura, "Analysis of nasal consonants," *J. Acoust. Soc. Am.*, vol. 34, pp. 1865-1875, December 1962.

[F-19] K. Fukunaga and T. Ito, "A design theory of recognition functions in self-organizing systems," *IEEE Trans. on Electronic Computers*, vol. EC-14, pp. 44-52, February 1965.

[G-1] D. Gabor, "Theory of Communication," *J. IEE*, vol. 93, part III, pp. 429-457, 1946.

[G-2] S.E. Gerber, "The intelligibility of delta modulated speech, *IEEE Trans. on Audio and Electroacoustics*, vol. AU-14, pp. 93-96, June 1966.

[G-3] B. Gold, "Computer Programme for Pitch Extraction," *J. Acoust. Soc. Am.*, vol. 34, pp. 916-921, July 1962.

[G-4] B. Gold and C.M. Rader, *Digital Processing of Signals*. New York: McGraw-Hill, 1969.

[G-5] J.L. Goldstein, "An investigation of monaural phase perception," *Ph.D. Thesis*, University of Rochester, New York, 1965.

[G-6] I.J. Good and K.C. Doog, "A paradox concerning rate of information," *Information and Control*, vol. 1, pp. 113-126, May 1958.

[G-7] I.J. Good and K.C. Doog, "A paradox concerning rate of information; correction and additions," *Information and Control*, vol. 2, pp. 195-197, June 1959.

[G-8] I.J. Good, "Effective sampling rates for signal detection or can the Gaussian model be salvaged?," *Information and Control*, vol. 3, pp. 116-140, June 1960.

[G-9] I.J. Good, "The loss of information due to clipping a waveform," *Information and Control*, vol. 10, pp. 220-222, February 1967.

[G-10] I.J. Good, personal correspondence with H.B. Voelcker, 1968.

[G-11] G.G. Gouriet and G.F. Newell, "A quadrature network for generating vestigial-sideband signals," *Proc. IEE*, vol. 107, pp. 253-260, May 1960.

[G-12] J.S. Griffin Jr., J.H. King, Jr., and C.J. Tunis, "A pattern identification system using linear decision functions," *IBM Systems Journal*, vol. 2, pp. 248-267, September-December 1963.

[G-13] S.C. Gupta, *Transform and State Variable Methods in Linear Systems*. New York: John Wiley, 1966.

[H-1] S. Haavik, "The conversion of zeros of noise," *M.Sc. Thesis*, University of Rochester, 1966.

[H-2] M.P. Haggard, "In defense of the formant," *Phonetica*, vol. 10, pp. 231-233, No. 3-4, 1963.

[H-3] M. Halle and K. Stevens, "Speech recognition; a model and a program for research," *IRE Transactions on Information Theory*, vol. IT-8, pp. 155-159, February 1962.

[H-4] M. Halle, G.W. Hughes and J.-P.A. Radley, "Acoustic properties of stop consonants," *J. Acoust. Soc. Am.*, vol. 29, pp. 107-116, January 1957.

[H-5] H.F. Harmuth, "A generalized concept of frequency and some applications," *IEEE Trans. on Information Theory*, vol. IT-14, pp. 375-382, May 1968.

[H-6] C.M. Harris and M.R. Weiss, "Pitch extraction by computer processing of high resolution Fourier analysis data," *J. Acoust. Soc. Am.*, vol. 35, pp. 339-343, March 1963.

[H-7] K.S. Harris, "Cues for the discrimination of American English fricatives in spoken syllables," *Language and Speech*, vol. 1, part 1, pp. 1-7, January-March 1958.

[H-8] K.S. Harris, H.S. Hoffman, A.M. Liberman, P.C. Delattre and F.S. Cooper, "Effect of third formant transitions on the perception of the voiced stop consonants," *J. Acoust. Soc. Am.*, vol. 30, pp. 122-126, February 1958.

[H-9] J.M. Heintz and K.N. Stevens, "On the properties of voiceless fricative consonants," *J. Acoust. Soc. Am.*, vol. 33, pp. 589-596, May 1961.

[H-10] H. Helmholtz, *On the Sensations of Tone*, reprint of 1877 edition. New York: Dover, 1954.

[H-11] W.H. Highleyman, "Linear decision functions with application to pattern recognition," *Proc. IRE*, vol. 50, pp. 1501-1514, June 1962.

[H-12] D.R. Hill, "Automatic speech recognition: a problem for machine intelligence," chapter 13 of *Machine Intelligence*, edited by N.L. Collins and D. Michie. New York: American Elsevier Publishing Co., 1967.

[H-13] W.A. Hillix, "Uses of two nonacoustic measures in computer recognition of spoken digits," *J. Acoust. Soc. Am.*, vol. 35, pp. 1978-1984, December 1963.

[H-14] W.A. Hillix, M.N. Fry and R.L. Hershman, "Computer recognition of spoken digits based on six nonacoustic measures," *J. Acoust. Soc. Am.*, vol. 38, pp. 790-796, November 1965.

[H-15] K. Hiramatsu, "Zero-corssing information of SSB signal," *J. Acoust. Soc. Japan*, vol. 18, no. 6, pp. 301-309, 1962. (In Japanese)

[H-16] K. Hiramatsu, M. Tatsuji and C.L. Coates, "Analysis of the phase function in single sideband signal and its application to speech," *Technical Report No. 24*, Laboratories for Electronics and Related Science Research, University of Texas, September 1966.

[H-17] H.S. Hoffman, "A study of some cues in the perception of the voiced stop consonants," *J. Acoust. Soc. Am.*, vol. 30, pp. 1035-1041, November 1958.

[H-18] E.M. Hoffstetter, "Construction of time-limited functions with specified autocorrelation functions," *IEEE Trans. on Information Theory*, vol. IT-10, pp. 119-126, April 1964.

[H-19] G.L. Holmgren, "Speaker recognition, speech characteristics, speech evaluation, and modification of speech signals-a selected bibliography," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-14, pp. 32-39, March 1966.

[H-20] A.S. House, "On vowel duration in English," *J. Acoust. Soc. Am.*, vol. 33, pp. 1174-1178, September 1961.

[H-21] C.R. Howard, "Speech analysis-synthesis scheme using continuous parameters," *J. Acoust. Soc. Am.*, vol. 28, pp. 1091-1098, November 1956.

[H-22] J.A. Howard and R.C. Wood, "Hybrid simulation of speech waveforms utilizing a Gaussian wave function representation," *Simulation*, vol. 11, pp. 117-124, September 1968.

[H-23] W.H. Huggins, "System function analysis of speech sounds," *J. Acoust. Soc. Am.*, vol. 22, pp. 765-767, November 1950.

[H-24] W.H. Huggins, "A note on autocorrelation analysis of speech sounds," *J. Acoust. Soc. Am.*, vol. 26, pp. 790-792, September 1954.

[H-25] W.H. Huggins and J.C.R. Licklider, "Place mechanisms of frequency analysis," *J. Acoust. Soc. Am.*, vol. 23, pp. 290-299, May 1951.

[H-26] G.W. Hughes and M. Halle, "Spectral properties of fricative consonants," *J. Acoust. Soc. Am.*, vol. 28, pp. 303-310, March 1956.

[H-27] A.F. Huxley, "Is resonance possible in the cochlea after all?," *Nature*, vol. 221, pp. 935-140, March 8, 1969.

[J-1] A.T. Jones, *Sound.* New York: D. Van Nostrand, 1937.

[J-2] M. Joos, "Acoustic phonetics," supplement to *Language,* vol. 24, 1948.

[K-1] L.R. Kahn, "The use of speech clipping in single side-band communication systems," *Proc. IRE,* vol. 45, pp. 1148-1149, August 1957.

[K-2] A.J. Kempner, "On equations admitting roots of the form $e^{i\theta}$," *Tohoku Math.,* vol. 10, pp. 115-117, 1916.

[K-3] A.J. Kempner, "Concerning the distribution of complex roots of algebraic equations," *Mathematiche Annalen,* vol. 85, pp. 49-59, 1922. (In German)

[K-4] A.J. Kempner, "On the separation and computation of complex roots of algebraic equations," *University of Colorado Studies,* vol. 16, pp. 75-87, no. 2, 1928.

[K-5] A.J. Kempner, "On the complex roots of algebraic equations," *Bull. American Math. Society,* vol. 41, pp. 809-843, 1935.

[K-6] J.H. King, Jr., and C.J. Tunis, "Some experiments in spoken word recognition, *IBM Journal of Research and Development,* vol. 10, pp. 65-79, January 1966.

[K-7] L.E. Kinsler and A.R. Frey, *Fundamentals of Acoustics. 2nd Ed.* New York: John Wiley, 1962.

[K-8] L. Kleinrock, "Detection of the peak of an arbitrary spectrum," *IEEE Trans. on Information Theory,* vol. IT-10, pp. 215-221, July 1964.

[K-9] A. Koestler, *The Act of Creation.* London: Hutchison's, 1964.

[K-10] A. Kohlenberg, "Exact interpolation of band limited functions," *J. Applied Physics,* vol. 24, pp. 1432-1436, December 1953.

[K-11] K.D. Kryter, "Speech bandwidth compression through spectrum selection," *J. Acoust. Soc. Am.,* vol. 32, pp. 547-559, May 1960.

[L-1] P. Ladefoged and D.E. Broadbent, "Perception of sequence in auditory events," *Quarterly J. of Experimental Psychology*, vol. 12, pp. 162-170, August 1960.

[L-2] P. Ladefoged and D.E. Broadbent, "The information conveyed by vowels," *J. Acoust. Soc. Am.*, vol. 29, pp. 98-104, January 1957.

[L-3] H.L. Lane, "Psychophysical parameters of vowel perception," *Psychological Monographs*, vol. 76, no. 44, 1962.

[L-4] S.H. Lavington, "Computer simulation of a speech recognition system," *Proc. IEE*, vol. 116, pp. 1053-1059, June 1969.

[L-5] M. Lecours, "Adaptive spectral analysis for speech-sound recognition," *Ph.D. Thesis*, University of London, 1967.

[L-6] Y.W. Lee, *Statistical Theory of Communication*. New York: John Wiley, 1960.

[L-7] I. Lehiste and G.E. Peterson, "The identification of filtered vowels," *Phonetica*, vol. 4, pp. 161-177, no. 4, 1959.

[L-8] I. Lehiste and G.E. Peterson, "Transitions, glides and dipthongs," *J. Acoust. Soc. Am.*, vol. 33, pp. 268-277, March 1961.

[L-9] D.H. Lehmer, "A machine method for solving polynomial equations, *J. ACM*, vol. 8, pp. 151-162, April 1961.

[L-10] B. Ja. Leven, *Distribution of Zeros of Entire Functions*. volume five of Mathematical Monographs.

[L-11] M.D. Levine, "Feature extraction: a survey," *Proc. IEEE*, vol. 57, pp. 1391-1407, August 1969.

[L-12] K.P. Li, J.C. Damman and W.D. Chapman, "Experimental studies in speaker verification using an adaptive system," *J. Acoust. Soc. Am.*, vol. 40, pp. 966-978, November 1966.

[L-13] J.C.R. Licklider, D. Bindra and I. Pollack, "The intelligibility of rectangular speech waves," *American J. of Psychology*, vol. 51, pp. 1-20, January 1948.

[L-14] J.C.R. Licklider and I. Pollack, "Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech," *J. Acoust. Soc. Am.*, vol. 20, pp. 42-51, January 1948.

[L-15] J.C.R. Licklider, "The intelligibility of amplitude dichotomized, time quantized speech waves," *J. Acoust. Soc. Am.*, vol. 22, pp. 820-823, November 1950.

[L-16] B. Lindblom, "A spectrographic study of vowel reduction," *J. Acoust. Soc. Am.*, vol. 35, pp. 1773-1781, November 1963.

[L-17] D.A. Linden, "A discussion of sampling theorems," *Proc. IRE*, vol. 47, pp. 1219-1226, July 1959.

[L-18]
[L-19] N. Lindgren, "Machine recognition of human language: part I - Automatic speech recognition," *IEEE Spectrum*, vol. 2, pp. 114-136, March 1965. "part II - Theoretical models of speech perception and language," *IEEE Spectrum*, vol. 2, pp. 45-59, April 1965.

[L-20] J.E. Littlewood, "Every polynomial has a root," *London Math. Society Journal*, vol. 16, part 2, pp. 95-98, April 1942.

[L-21] J.E. Littlewood, "On the real roots of real trigonometric polynomials," from *Studies in Mathematical Analysis and Related Topics*. California: Stanford University Press, 1962.

[L-22] J.E. Littlewood, "On the real roots of real trigonometric polynomials II," *J. London Math. Soc.*, vol. 39, pp. 511-532, July 1964.

[L-23] J.E. Littlewood, "The real zeros and value distributions of real trigonometric polynomials," *J. London Math. Soc.*, vol. 41, pp. 336-342, April 1966.

[L-24] B.M. Lobanov, "Automatic separation of hiss sounds from voiced sounds using clipped speech signals," *Telecommunications*, vol. 22, part I, pp. 42-46, November 1968.

[L-25] R.D. Luce, "Amplitude distorted signals," *MIT Quarterly Progress Reports*, pp. 37-41, April 15, 1953.

[M-1] D.M. MacKay, "The 'Active/Passive' controversy," shortened preprint for *Moscow Speech Conference*, 1967.

[M-2] D.M. MacKay, J.B. Millar and M.J. Underwood, "Discriminative value of the digram structure of speech waveforms," shortened preprint for *Moscow Speech Conference,* 1967.

[M-3] H.J. Manley, "Analysis-snythesis of connected speech in terms of orthogonalized exponentially damped sinusoids," *J. Acoust. Soc. Am.,* vol. 35, pp. 464-474, April 1963.

[M-4] H.J. Manley, "Fourier coefficients of speech power spectra as measured by autocorrelation analysis," *J. Acoust. Soc. Am.,* vol. 34, pp. 1143-1145, August 1962.

[M-5] P. Marcou and J. Daguet, "New methods of speech transmission," in *Proc. London Symposium on Information Theory, 1955.* Editor E.C. Cherry. London: Butterworth's Scientific Publications, 1955.

[M-6] M. Marden, *Geometry of Polynomials.* Providence, Rhode Island: American Mathematical Society, 1966.

[M-7] M.V. Mathews, "Extremal coding for speech transmission," *IEEE Trans. On Information Theory,* vol. IT-5, pp. 129-136, September 1959.

[M-8] M.V. Mathews, J.E. Miller and E.E. David, Jr., "Pitch synchronous analysis of voiced sounds," *J. Acoust. Soc. Am.,* vol. 33, pp. 179-186, February 1961.

[M-9] W. Meyer-Eppler, "Realization of prosodic features in whispered speech," *J. Acoust. Soc. Am.,* vol. 29, pp. 104-106, January 1957.

[M-10] G.A. Miller, *Language and Communication.* New York: McGraw-Hill, 1951.

[M-11] R.L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Am.,* vol. 31, pp. 667-677, June 1959.

[M-12] D. Monro, "Sampling programme for 7094 Direct Data Channel," Imperial College, 1967.

[M-13] D. Monro, "Calcomp contour plotter," Imperial College, 1968.

[M-14] L.R. Morris, "A brief guide to the fast Fourier transform," Imperial College, May 1968.

[M-15] P.M. Morse and K.U. Ingard, *Theoretical Acoustics.* New York: McGraw-Hill, 1968.

[N-1] G. Nagy, "Classification algorithms in pattern recognition," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, pp. 203-212, June 1968.

[N-2] K. Nakata, "Synthesis and perception of nasal consonants," *J. Acoust. Soc. Am.*, vol. 31, pp. 661-666, June 1959.

[N-3] N.J. Nilsson, *Learning Machines*. New York: McGraw-Hill, 1965.

[N-4] A. Noll, "Short time spectrum and cepstrum techniques for vocal pitch detection," *J. Acoust. Soc. Am.*, vol. 36, pp. 296-303, February 1964.

[N-5] H. Nyquist, "Certain factors affecting telegraph speed," *Bell System Technical Journal*, vol. 3, pp. 324-346, April 1924.

[O-1] J.D. O'Connor, L.J. Gerstman, A.M. Liberman, P.C. Delattre, and F.S. Cooper, "Acoustic cues for the perception of initial /w,j,r,l/ in English," *WORD*, vol. 13, pp. 24-43, January 1957.

[O-2] H.F. Olson, H. Belar, and E.S. Rogers, "Speech processing techniques and applications," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-15, pp. 120-126, September 1967.

[P-1] J.H. Painter, S.C. Gupta and J.W. Bayless, "Analytic function models of noise and modulation," *NASA Technical Report*, TR-R-314, June 1969.

[P-2] A. Papoulis, *The Fourier Integral and Its Applications*. New York: McGraw-Hill, 1962.

[P-3] A. Papoulis, *Probability, Random Variables, and Stochastic Variables*. New York: McGraw-Hill, 1965.

[P-4] A. Papoulis, "Error analysis in sampling theory," *Proc. IEEE*, vol. 54, pp. 947-955, July 1966.

[P-5] A. Papoulis, "Limits on bandlimited signals," *Proc. IEEE*, vol. 55, pp. 1677-1686, October 1967.

[P-6] J.K. Parks, "Optical correlation detector for the audio frequency range," *J. Acoust. Soc. Am.*, vol. 37, pp. 268-277, February 1965.

[P-7] J.D. Patterson and B.F. Womack, "An adaptive pattern classification system," *IEEE Trans. on Systems Science and Cybernetics*, vol. SSC-2, pp. 62-67, August 1966.

[P-8] A.P. Paul, A.S. House and K.N. Stevens, "Automatic reduction of vowel spectra: an analysis by snythesis method and its evaluation," *J. Acoust. Soc. Am.*, vol. 36, pp. 303-308, February 1964.

[P-9] E. Peterson, "Frequency detection and speech formants," *J. Acoust. Soc. Am.*, vol. 23, pp. 668-674, November 1951.

[P-10] G.E. Peterson, "Studies in speech analysis and synthesis," *University of Michigan Report No. NR 049-122*, August 1966.

[P-11] G.E. Peterson and H.L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, pp. 175-184, March 1952.

[P-12] G.E. Peterson and J.R. Hanne, "Examination of two different formant estimation techniques," *J. Acoust. Soc. Am.*, vol. 38, pp. 224-228, August 1965.

[P-13] E.N. Pinson, "Pitch synchronous time domain estimation of formant frequencies and bandwidths," *J. Acoust. Soc. Am.*, vol. 35, pp. 1264-1269, August 1963.

[P-14] R. Plomp, "The ear as a frequency analyzer," *J. Acoust. Soc. Am.*, vol. 36, pp. 1628-1636, September 1964.

[P-15] R.K. Potter, G.A. Kopp and H.C. Green, *Visible Speech*. New York: D. van Nostrand, 1947.

[P-16] R.K. Potter and G.E. Peterson, "The representation of vowels and their movements," *J. Acoust. Soc. Am.*, vol. 20, pp. 528-535, July 1948.

[P-17] A.J. Prestigiacomo, "Amplitude contour display of sound spectrograms," *J. Acoust. Soc. Am.*, vol. 34, pp. 1684-1688, November 1962.

[P-18] S. Pruzansky and M.V. Mathews, "Talker recognition procedure based on analysis of variance," *J. Acoust. Soc. Am.*, pp. 2041-2047, November 1964.

[P-19] R.F. Purton, "Speech recognition using autocorrelation analysis," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, pp. 235-239, June 1968.

[R-1] L.R. Rabiner, R.W. Schafer and C.M. Rader, "The chirp Z-transform and its applications," *Bell Systems Technical J.*, vol. 48, pp. 1249-1292, May-June 1969.

[R-2] A. Ralston, *A First Course in Numerical Analysis*. New York: McGraw-Hill, 1968.

[R-3] A. Ralston and H. Wilf, *Mathematical Methods for Digital Computers, vol. II, second edition*. New York: John Wiley, 1967.

[R-4] D.R. Reddy, "Segmentation of speech sounds," *J. Acoust. Soc. Am.*, vol. 40, pp. 307-312, August 1966.

[R-5] D.R. Reddy, "Phoneme grouping for speech recognition," *J. Acoust. Soc. Am.*, vol. 41, pp. 1295-1300, May 1967.

[R-6] D.R. Reddy, "Computer recognition of connected speech," *J. Acoust. Soc. Am.*, vol. 42, pp. 329-347, August 1967.

[R-7] A. Requicha, "Contributions to a zero-based theory of bandlimited signals," *Ph.D. Thesis*, University of Rochester, 1970.

[R-8] H.L. Resnikoff, "Sound statistics find the lost consonants," *MIT Technology Review*, vol. 71, pp. 72-73, April 1969.

[R-9] H.L. Resnikoff and G.A. Sitton, "A new type of hearing aid," *Rice University Review*, Fall/Winter 1968, pp. 31-35.

[R-10] S.O. Rice, "Mathematical analysis of random noise," in *Selected Papers on Noise and Stochastic Processes*. Editor, N. Wax. New York: Dover Publications Inc., 1954.

[R-11] A.W. Rihaczek, "Signal energy distribution in time and frequency," *IEEE Trans. on Information Theory*, vol. IT-14, pp. 369-374, May 1968.

[R-12] A.V. Rimskii-Korsakov, "Calculation of the audibility of nonlinear distortions arising in an electroacoustic channel," *Soviet Physics-Acoustics*, vol. 2, pp. 48-59, No. 1, 1956.

[R-13] A.V. Rimskii-Korsakov, "Statistical properties of a radio broadcast signal," *Soviet Physics-Acoustics*, vol. 6, pp. 360-368, January-March 1961.

[R-14] H.E. Rose, "Performance evaluation of clipped speech," 1967 Conf. on Speech Communication and Processing, Bedford, Mass.

[R-15] C.A. Rosen and D.J. Hall, "A pattern recognition experiment with near-optimum results," *IEEE Trans. on Electronic Computers*, vol. EC-15, pp. 666-667, August 1966.

[R-16] P.W. Ross and E.S. Rogers, "The application of pattern recognition techniques to the evaluation of speech recognition systems," 1967 Conf. on Speech Communication and Processing, Bedford, Mass.

[R-17] G.O. Russell, "The mechanism of speech," *J. Acoust. Soc. Am.*, vol. 1, pp. 83-109, October 1929.

[S-1] T. Sakai and S. Inoue, "New instruments and methods for speech analysis," *J. Acoust. Soc. Am.*, vol. 32, pp. 441-450, April 1960.

[S-2] T. Sakai and S. Doshita, "The automatic speech recognition system for conversational sound," *IEEE Trans. on Electronic Computers*, vol. EC-12, pp. 835-845, December 1963.

[S-3] D.J. Sakrison, *Communication Theory: Transmission of Waveforms and Digital Information*. New York: John Wiley, 1968.

[S-4] R.W.A. Scarr, "Zero crossings as a means of obtaining spectral information in speech analysis," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, pp. 247-255, June 1968.

[S-5] I.J. Schoenberg, "Extensions of the theorems of Descartes and Laguerre to the complex domain," *Duke Math. Journal*, vol. 2, pp. 84-94, 1936.

[S-6] M.R. Schroeder, "Vocoders: analysis and synthesis of speech," *Proc. IEEE*, vol. 54, pp. 720-734, May 1966.

[S-7] M.R. Schroeder and B.S. Atal, "Generalized short time power spectra and autocorrelation functions," *J. Acoust. Soc. Am.*, vol. 34, pp. 1679-1683, November 1962.

[S-8] M.R. Schroeder, J.L. Flanagan, and E.A. Lundry, "Bandwidth compression of speech by analytic signal rooting," *Proc. IEEE*, vol. 55, pp. 396-401, March 1967.

[S-9] G.S. Sebestyen, *Decision-making Processes in Pattern Recognition*. New York: the Macmillan Company, 1962.

[S-10] J.L. Sevy, "The effect of multiple CW and FM signals passed through a hard limiter or TWT," *IEEE Trans. on Communication Technology*, vol. COM-14, pp. 568-578, October 1966.

[S-11] H.L. Shaffer, "Information rate necessary to transmit pitch period durations for connected speech," *J. Acoust. Soc. Am.*, vol. 36, pp. 1895-1900, October 1964.

[S-12] S. Sherman, "Generalized Routh-Hurwitz discriminant--an extension of the theorems of Sturm, Routh and Hurwitz on the roots of polynomial equations," *Philosophical Magazine*, vol. 37, pp. 537-551, August 1946.

[S-13] P.N. Sholtz and R. Bakis, "Spoken digit recognition using vowel consonant segmentation," *J. Acoust. Soc. Am.*, vol. 34, pp. 1-5, January 1962.

[S-14] J.E. Shoup, "Phoneme selection for studies in automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 34, pp. 397-403, April 1962.

[S-15] E.G. Shower and R. Biddulph, "Differential pitch sensitivity of the ear," *J. Acoust. Soc. Am.*, vol. 3, pp. 275-287, October 1931.

[S-16] B.M. Siegenthaler, "A study of the intelligibility of sustained vowels," *Quarterly Journal of Speech*, vol. 36, pp. 202-208, April 1950.

[S-17] V.N. Sobolev and V.N. Telepnev, "Simple methods of clipped speech regeneration," *Telecommunications*, vol. 23, pp. 37-44, March 1969.

[S-18] W. Sollfrey, "Hard limiting of three and four sinusoidal signals," *IEEE Trans. on Information Theory*, vol. IT-15, pp. 2-7, June 1969.

[S-19] D.F. Specht, "Generation of polynomial discriminant functions for pattern recognition," *IEEE Trans. on Electronic Computers*, vol. EC-16, pp. 308-319, June 1967.

[S-20] L.R. Spogen, H.N. Shaver, D.E. Baker and B.V. Blom, "Speech processing by the selective amplitude scaling system," *J. Acoust. Soc. Am.*, vol. 32, pp. 1621-1625, December 1960.

[S-21] E.V. Stansfield, "Acoustic transmission lines," Imperial College, December 1968.

[S-22] H.A. Steinberg, "The effect of clipping on the performance of a phased array antenna in an anisotropic background," *IEEE Trans. on Aerospace and Electronic Systems*, vol. AES-5, pp. 103-110, January 1969.

[S-23] K.N. Stevens, "Autocorrelation analysis of speech sounds," *J. Acoust. Soc. Am.*, vol. 22, pp. 769-771, November 1950.

[S-24] K.N. Stevens, "Toward a model for speech recognition," *J. Acoust. Soc. Am.*, vol. 32, pp. 47-55, January 1960.

[S-25] K.N. Stevens, A.S. House and A.P. Paul, "Acoustical description of syllable nuclei: an interpretation in terms of a dynamic model of articulation," *J. Acoust. Soc. Am.*, vol. 40, pp. 123-132, July 1966.

[S-26] W. Stover, "Time domain bandwidth compression," *J. Acoust. Soc. Am.*, pp. 348-359, August 1967.

[S-27] P.D. Strevens, "Spectra of fricative noises in human speech," *Language and Speech*, vol. 3, pp. 32-49, January-March 1960.

[S-28] J. Suzuki, Y. Kadokawa and K. Nakata, "Formant frequency extraction by the method of moment calculations," *J. Acoust. Soc. Am.*, vol. 35, pp. 1345-1353, September 1963.

[T-1] W.P. Tanner, "Theory of recognition," *J. Acoust. Soc. Am.*, vol. 28, pp. 882-888, September 1956.

[T-2] C.F. Teacher, H.G. Kellett and L.R. Focht, "Experimental, limited vocabulary, speech recognizer," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-15, pp. 127-130, September 1967.

[T-3] V.N. Teterev, "The connection between the distribution functions for the zeros in a harmonic signal and its frequency characteristics in the case of extreme limiting," *Telecommunications and Radio Engineering, part 1*, vol. 19, pp. 62-64, July 1965.

[T-4] I.B. Thomas, "The influence of first and second formants on the intelligibility of clipped speech," *J. Audio Engineering Society*, vol. 16, pp. 182-185, April 1968.

[T-5] I.B. Thomas and R.J. Niederjohn, "Enhancement of speech intelligibility at high noise levels by filtering and clipping," *J. Audio Engineering Society*, vol. 16, pp. 412-415, October 1968.

[T-6] J. Thomas, *An Introduction to Statistical Communication Theory*. New York: John Wiley, 1969.

[T-7] R. Thomas, "A real time audio spectral analyzer," *Ph.D. Thesis*, University of London, 1964.

[T-8] W.R. Tiffany, "Vowel recognition as a function of duration, frequency modulation and phonetic context," *J. of Speech and Hearing Disorders*, vol. 18, pp. 289-301, September 1953.

[T-9] G.I. Tsemel, "Determination of invariant criteria for stop consonants on the basis of clipped speech signals," *Akademia Nauk SSSR, Izvestia, Otdelenie Teknicheskikh Nauk. Energetika i Automatica.* vol. 4, pp. 214-215, No. 4, 1959. (in Russian)

[T-10] G.I. Tsemel, "Identification of unvoiced fricative sounds from a clipped speech signal," *Problems of Information Transmission*, vol. 1, pp. 33-40, October-December 1965.

[T-11] D.G. Tucker, "Linear rectifiers and limiters," *Wireless Engineer*, vol. 29, pp. 128-137, May 1952.

[V-1] H.L. Van Trees, *Detection, Estimation and Modulation Theory*. New York: John Wiley, 1968.

[V-2] J.H. Van Vleck and D. Middleton, "The spectrum of clipped noise," *Proc. IEEE*, vol. 54, pp. 2-19, January 1966.

[V-3] A.I. Velichkin, "Statistical investigation of speech transmission," *Telecommunications*, No. 8, pp. 3-10, August 1961.

[V-4] A.I. Velichkin, "Amplitude clipping of speech," *Soviet Physics-Acoustics*, vol. 8, pp. 360-368, October-December 1962.

[V-5] F.J. Vilbig, "The analysis of clipped speech," *U.S. Air Force Report AFCRC-TR-56-120*, 1956.

[V-6] H.B. Voelcker, "Toward a unified theory of modulation-part I: phase-envelope relationships," *Proc. IEEE*, vol. 54, pp. 340-353, March 1966 and "part II: zero manipulation," *Proc. IEEE*, vol. 54, pp. 735-755, May 1966.

[V-7] H.B. Voelcker, notes on Signal Theory, Imperial College, 1967.

[V-8] H.B. Voelcker, "Generation of digital signaling waveforms," *IEEE Trans. on Communication Technology*, vol. COM-16, pp. 81-93, February 1968.

[V-9] H.B. Voelcker, notes on Real Zero Signals, 1967.

[V-10] H.B. Voelcker, notes on Zero Conversion by Frequency Translation, 1968.

[V-11] H.B, Voelcker, notes on Clipping, 1967.

[W-1] F.M. Weiner and D.A. Ross, "The pressure distribution in the auditory canal in a progressive sound field," *J. Acoust. Soc. Am.*, vol. 18, pp. 401-408, October 1946.

[W-2] S. Weinrib, "A digital spectral analysis technique and its application to radio astronomy," *MIT Research Laboratory of Electronics, Technical Report No. 412,* August 1963.

[W-3] M.R. Weiss and C.M. Harris, "Computer technique of high speed extraction os speech parameters," *J. Acoust. Soc. Am.*, vol. 35, pp. 207-214, February 1963.

[W-4] M.R. Weiss, R.P. Vogel and C.M. Harris, "Implementation of a pitch extractor of the double-spectrum analysis type," *J. Acoust. Soc. Am.*, vol. 40, pp. 657-662, September 1966.

[W-5] P.D. Welch and R.S. Wimpress, "Two multivariate statistical computer programs and their application to the vowel recognition programme," *J. Acoust. Soc. Am.*, vol. 33, pp. 426-434, April 1961.

[W-6] E.G. Wever, *Theory of Hearing.* New York: John Wiley, 1949.

[W-7] B. Widrow, "Statistical analyses of amplitude quantized sampled data systems," *Trans. AIEE, part II, Applications and Industry*, vol. 79, pp. 555-568, January 1961.

[W-8] J. Wiren and H.L. Stubbs, "Electronic binary selection system for phoneme classification," *J. Acoust. Soc. Am.*, vol. 38, pp. 1082-1091, November 1956.

[W-9] P.M. Woodward, *Probability and Information Theory with Applications to Radar.* London: Pergamon Press, 1953.

[X-1] R.E. Bogner and J.L. Flanagan, "Frequency multiplication of speech signals," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-17, pp. 202-208, September 1969.

[X-2] J.R. Pierce, "Whither speech recognition?," *J. Acoust. Soc. Am.*, vol. 46, pp. 1049-1051, October 1969.

[X-3] L.R. Wilson, "Asymptotes and bandwidths for spectra at the output of a hard limiter," *Proc. IEEE*, vol. 57, pp. 1676-1677, September 1969.

[X-4] H.R. Ward, "The effect of bandpass limiting on noise with a Gaussian spectrum," *Proc. IEEE*, vol. 57, pp. 2089-2090, November 1969.

[X-5] E.C. Titchmarsh, *The Theory of Functions*. Second Edition. London: Oxford University Press, 1939.

[X-6] D.D. McCracken, *Fortran with Engineering Applications*. New York: John Wiley, 1967.

[Y-1] J.L. Yen, "On nonuniform sampling of bandwidth limited signals," *IRE Trans. on Circuit Theory*, vol. CT-3, pp. 251-257, December 1956.

[Z-1] E. Zwicker, E. Flottorp and S.S. Stevens, "Critical bandwidth in loudness summation," *J. Acoust. Soc. Am.*, vol. 29, pp. 548-557, May 1957.