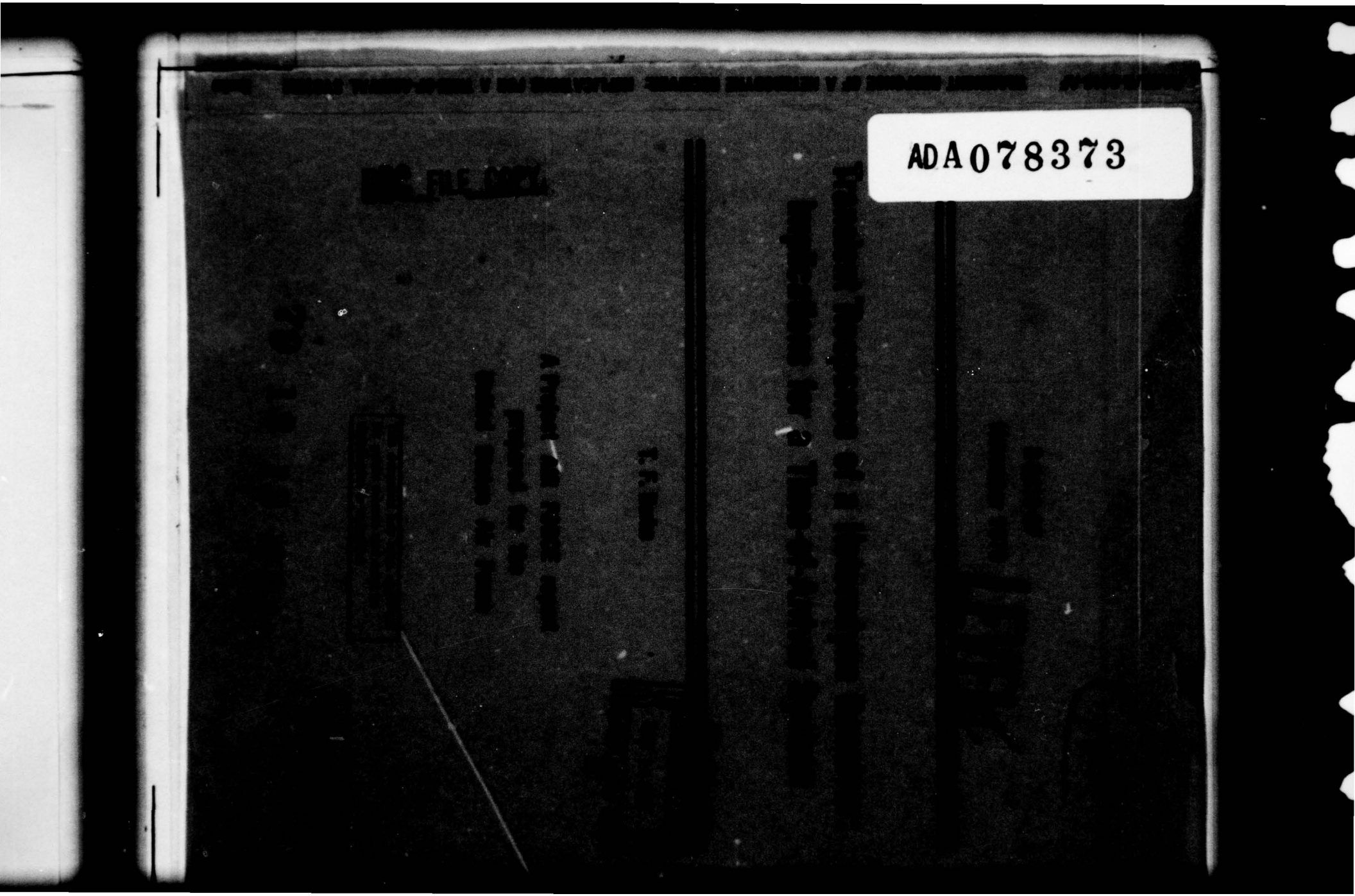
AD-A078 373 RAND CORP SANTA MONICA CA F/G 17/3
TRANSIENT RESPONSE OF A HETERODYNE RECEIVER: IMPLICATIONS FOR A--ETC(U)
NOV 79 T F BURKE F49620-77-C-0023
UNCLASSIFIED RAND/R-2418-AF NL

1 OF 3
AD-
A078373


ADA078373

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: R-2418-AF
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): Transient Response of a Heterodyne Receiver: Implications for a Time-of-Arrival System
5. TYPE OF REPORT & PERIOD COVERED: Interim rept.
6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): T. F. Burke (Thomas Finley Burke)
8. CONTRACT OR GRANT NUMBER(s): F49620-77-C-0023

9. PERFORMING ORGANIZATION NAME AND ADDRESS
The Rand Corporation
1700 Main Street
SAnta Monica, California 90401
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Requirements, Programs & Studies Group (AF/RDQM)
Ofc, DCS/R&D and Acquisition
HQ USAF, Washington, D. C. 20330
12. REPORT DATE: November 1979
13. NUMBER OF PAGES: 213

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): RAND/R-2418-AF
15. SECURITY CLASS (of this report): UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for Public Release; Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
No Restrictions

18. SUPPLEMENTARY NOTES

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

Transient Response	Time-of-Arrival Systems	Antennas
Demodulation	Delay Time	Radar Signals
Wave Propagation	Time Lag	Position(Location)
Superheterodyne Receivers	Direction Finding	

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

296 600

DD FORM 1 JAN 73 1473

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A directional antenna is a filter whose response varies with direction. A modulated signal transmitted from such an antenna produces different far-field waveforms in all directions (except for possible symmetry). The response of a receiver to the transmission will vary with position in the antenna pattern of the emitter. This effect is ordinarily negligible but it could limit the ultimate performance of any system that accepts off-axis signals and relies upon the details of waveform. In a leading-edge TOA system used to determine the location of a pulsed emitter, the several receivers necessarily lie at different angular positions from the emitter. This analysis indicates that in a simple, idealized TOA system that would otherwise be entirely free from error, the waveform effect arising from a directional emitter can lead to errors of several dozen meters in the computed location. (Author)

R-2418-AF
November 1979

Transient Response of a Heterodyne Receiver: Implications for a Time-of-Arrival System

DDC
DEC 17 1979
E

T. F. Burke

A Project AIR FORCE report
prepared for the
United States Air Force

Rand
SANTA MONICA, CA. 90406

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

The research reported here was sponsored by the Directorate of Operational Requirements, Deputy Chief of Staff/Research, Development, and Acquisition, Hq USAF, under Contract F49620-77-C-0023. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

Library of Congress Cataloging in Publication Data

Burce, Thomas Finley, 1918-
Transient response of a heterodyne receiver.

([Report] - Rand Corporation ; R-2418-AF)
"A Project Air Force report."
1. Radio--Receivers and reception--Heterodyne reception. I. Title. II. Series: Rand Corporation. Rand report ; R-2418-AF.
AS36.R3 R-2418 [TK6553] 080s [621.3841'352]
ISBN 0-8330-0139-6 79-17243

The Rand Publications Series: The Report is the principal publication documenting and transmitting Rand's major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

Published by The Rand Corporation

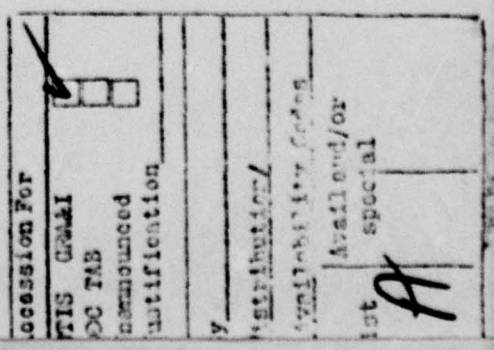
PREFACE

As part of its research under Project AIR FORCE, The Rand Corporation conducts studies of newly emerging technology and considers the possible application of new systems or devices to Air Force missions. In the course of this work, a question arose whether there might not be obscure effects, ordinarily overlooked, which would limit the achievable performance of some systems--for example, certain effects, other than those that are generally understood, that would hamper or limit the accuracy with which the occurrence times of events can be measured in electronic systems. This inquiry led to the publication of Rand Report R-1819-PR, which pointed out that a little-known feature of the behavior of antennas might cause significant error in TOA (time-of-arrival) systems.

The study reported here analyzes that topic in further detail. The achievable accuracy of TOA systems could be limited by this antenna effect. Consequently, we examined how the output response of a typical receiver varies as it is placed in different angular locations from a simple directional pulsed emitter such as a radar. This study treats the exact transient response of a heterodyne receiver to an incoming pulse--so far as we know, the first such analysis to be reported. Numerous features of the analysis will be of interest to the electronics community as a whole, and some features will be surprising to most analysts.

Theoretical difficulties make it impossible to calculate exactly the severity of the antenna effects on TOA measurements. The study results are thought to underestimate the severity, but that opinion is scarcely more than informed conjecture. Careful and well-designed experiments would be needed to clarify the actual effects.

The research reported here results from the project "Technology Applications Research," a component of the Project AIR FORCE Supporting Research program.



SUMMARY

A previous report* pointed out that, as a result of a widely overlooked but inescapable feature of antenna behavior, a directional antenna that is transmitting a modulated signal will deliver different signal waveforms in different directions. This is because such an antenna is an angle-dependent frequency-selective device, as is evidenced from the frequency dependence of the directivity pattern. It follows that a given receiver, when moved to different angular locations with respect to such a source, will exhibit an output response that differs not only in amplitude, but in waveform. It was suggested that such changes in receiver response might cause significant errors in some electronic systems--notably the so-called leading-edge time-of-arrival (TOA) systems. To investigate the matter, it is necessary to examine the extent to which the receiver response can be expected to change as a result of the changes of input waveform. This report presents the results of a detailed study of that question.

It was found necessary to carry out an exact analysis of the transient response of a simplified (but otherwise typical) heterodyne receiver. To our knowledge, this is the first report of such an analysis, and analysts may find it of professional interest; certain aspects of the transient response contradict well-known rules concerning circuit behavior. The analysis is quite lengthy because of the several circuit stages through which the signal passes, and the mathematical expressions are cumbersome. Moreover, the detailed results are remarkably intricate. However, these results can be understood qualitatively to arise from a phenomenon that has not previously been understood to occur at all, but which underlines in the following way the antenna effect on a TOA system.

The analysis shows that, at all angles off the source axis, the incoming waveform can be regarded as the sum of two signals: the one that is ordinarily "expected," and an "extra" signal that arises from

---

*T. F. Burke, *A Little-Known Effect of Antenna Size on Signal Waveform*, The Rand Corporation, R-1819-PR, August 1975.

the antenna effect. These two signals excite, in the receiver circuits, two simultaneous responses. For a rapidly rising incoming pulse, the "extra" signal excites a response that is essentially the impulse response of the receiver, whereas the "expected" signal excites the response to a step-modulated carrier. These two responses are different, and they interfere with each other (ahead of the envelope detector) to produce a resultant that may be different from both. The resultant waveform varies considerably with small changes of the phases and amplitudes at which the two responses are excited. It is this that leads to the intricate behavior found in the analysis.

The interference between the two circuit responses is most pronounced when the two are excited at comparable amplitudes. This happens whenever the receiver is at the edge of a source side lobe, near but not in a steady-state pattern null. At such locations, the shape of the receiver output waveform changes considerably with very small shifts of angular position, and also changes with the phase of the local oscillator in the receiver. When the receiver is near the peak of any side lobe, the "expected" response swamps the "extra" response and the receiver response is very nearly the same as that on the source axis. On the other hand, if the receiver is in a pattern null of the source, the "expected" signal is absent. There the "extra" response alone is excited.

For these reasons, the leading edge of the receiver output changes shape and appears to move back and forth in time. The apparent time at which the signal might be said to arrive depends upon how "arrival" is defined and measured, but for the numerical examples considered in this study, a few shifts as large as 60 nanoseconds--corresponding to a displacement of 20 meters--were encountered. In a typical TOA system that was modeled by a computer simulation, these arrival time shifts led to some errors as large as 48 meters in the reported location of the source; location errors of 10 meters were found in some 10 to 20 percent of the measurements. Theoretical difficulties prevent accurate calculation of these effects, even for a highly idealized system, so these numerical values should not be regarded as firm; there are plausible reasons to expect the effects to be larger in practical devices. Careful experimentation would be needed to explore these questions more fully.

It is noteworthy that the troublesome effects occur in and close to all the pattern nulls of the emitter; they do not become progressively more severe with increasing angle from the source axis. In some respects the most drastic effects arise at the edge of the main lobe. Inasmuch as the number of source pattern nulls tends to increase with increasing source directionality, it can be expected that these effects will be most troublesome when the source is highly directional because of the greater likelihood that a receiver will happen to be in an unfavorable location. Even though each such region spans a small angle near a null, there may be many nulls and a TOA system employs several receivers; the likelihood that one or more receivers is troubled is not negligible.

The two receiver responses that are excited by the incoming "expected" and "extra" signals are almost entirely determined by the receiver design, and are scarcely influenced by the pulse waveform of the source or by its antenna design details. Consequently the amount by which "arrival" tends to shift back and forth is controlled almost entirely by the design of the TOA receiver itself. Thus, the *amount* of error made in locating the emitter tends to be controlled by the TOA system design and operation, but the *likelihood* of experiencing any given amount is controlled primarily by the directionality of the source.

Although the analysis reported here is straightforward, it is lengthy and the mathematical expressions are complicated. Although some readers may want to examine some of the steps, few are likely to read the entire analysis. For that reason, the detailed analysis is relegated to Part III, which occupies most of the bulk of this report. Part II presents a more detailed but nonmathematical overview of the topics covered briefly in this summary. Part II describes and presents the results of two computer simulations of a TOA system, assesses the results, and discusses the question of whether means might be found to mitigate these effects. Briefly, it seems unlikely that much can be done through design of the receiver circuits, even though the design tends to control the magnitude of the system errors. Rather, it appears that these effects are most likely to be mitigated, if necessary, by

various operational procedures that might be adopted. Part I is a brief introduction to the analysis.

In addition to Parts I, II, and III, there are several appendixes that present a brief elementary review of the methods of transient analysis and treat in further detail several topics that relate to the analysis.

ACKNOWLEDGMENTS

The author appreciates numerous helpful discussions with, and suggestions from, Edward Bedrosian throughout the course of this study. M. Lakatos not only wrote all the computer programs used in the study, including those used to generate the figures, but she also detected and corrected a number of algebraic errors in the analysis and participated in the interpretation of the computer output.

CONTENTS

PREFACE iii

SUMMARY v

ACKNOWLEDGMENTS ix

Part

I. INTRODUCTION 1

II. QUALITATIVE OVERVIEW 5
1. The Emitter 6
2. TOA Systems 11
3. Arrival Time 13
4. The Signal as the Sum of Two Signals 17
5. Interference Between Two Receiver Responses 23
6. Two Experiments 26
7. Assessment 37
8. Mitigation 40
9. Conclusions 43

III. DETAILED ANALYSIS 45
1. Introduction to the Analysis 46
2. Symbols, Numerical Values, and the Like 48
3. Source Signal on Axis; F_0 53
4. Source Signal Off Axis; $F_{1,2}$ 59
5. $F_{1,2}$ Regarded as a Sum 65
6. Receiver Front End 72
7. Heterodyne Stage; $F_{3,4}$ 73
8. Phase Effects; the Spectrum of $F_{3,4}$ 79
9. Transient Response of the IF Strip 85
10. Output of the IF Strip; $F_{5,6}$ 90
11. Envelope Detector; F_7 96
12. Examples of Receiver Output, F_7 99
13. Synthesis of the Output Waveform 109
14. Time of Arrival; D_0 and D_1 112
15. Dependence of D_0 and D_1 on Z 119
16. First Computer Experiment 124
17. Tuned Front End; Q_3 126
18. Dependence of Arrival Time on Q_3 132
19. Dependence of Arrival Time on Z when $Q_3 = 10$ 138
20. Transition from Impulse- to Stepped-Carrier Response 144
21. Second Computer Experiment 147

Appendix

A. TRANSIENT ANALYSIS 151
B. SIMPLE BANDPASS FILTERS 157
C. SIMPLE LOWPASS FILTERS 169
D. LOWPASS-BANDPASS EQUIVALENCE 173

E. BUTTERWORTH BANDPASS FILTERS 179
F. ANTENNA THEORY 189
G. IMPULSE RESPONSE OF THE ANTENNA 195
H. THE APERTURE AS A FILTER 203
I. DESIGN OF THE ENVELOPE DETECTOR 205
J. TOA SYSTEM ERROR 210

PART I

INTRODUCTION

I. INTRODUCTION

It is commonly supposed that a directional transmitting antenna, such as a radar, sends the same modulated signal waveform in all directions, and that the signal differs only in amplitude from place to place. A previous report* pointed out that this is an approximation. A directional antenna behaves like a frequency-selective filter whose selectivity differs with angle. Consequently, the far-field waveform differs with off-axis angle, and the changes of waveform are a direct consequence of the directionality itself.

Inasmuch as a receiver will receive different input waveforms at different angular locations in the field of the transmitter (aside from mere changes of signal strength), the receiver output signal will change accordingly. Thus, this obscure antenna effect could, at least in principle, influence the behavior of a variety of different kinds of electronic systems. The extent to which it might affect performance--say, in terms of accuracy--depends, among other factors, on the extent to which the system relies on detailed waveform and on the degree to which the system uses off-axis signals. In most ordinary systems, such as radars and communication links, little use is made of precise waveforms and only signals in the main lobe are employed. Thus, although the waveform varies even within the main lobe, such systems are not likely to be affected appreciably.

Some kinds of systems, however, notably side-looking radars and time-of-arrival (TOA) systems, are intended to use off-axis signals and are therefore more likely to be affected significantly. This antenna effect could limit the achievable resolution of a side-looking radar, and it could limit the achievable accuracy of a TOA system. Consequently, it was decided to examine how the output response of a typical receiver will vary as it is placed in different angular locations from a simple directional pulsed emitter such as a radar. This report presents the results of that study.

---

*T. F. Burke, *A Little-Known Effect of Antenna Size on Signal Waveform*, The Rand Corporation, R-1819-PR, August 1975.

Although the analysis itself is more general, the study treats in detail the case of a so-called leading-edge TOA receiver, i.e., any one of the several system receivers that seek to measure the exact time at which the leading edge of a transmitted pulse arrives. If three or more receivers make such measurements, and if the receiver locations are known, then the location of the emitter can be computed. This technique could be used to observe the location of a friendly vehicle, such as a missile in flight; or it could be used to map the locations of hostile emitters, and perhaps to direct vehicles against them, or to avoid them. The need for accuracy varies widely among the various possible uses of the TOA technique; some applications would require that the emitter location be determined to within, say, a few dozen feet, whereas others might not require accuracy other than several hundred feet. The study was undertaken to estimate where, in such a range, the effect is likely to be. Would the antenna effect lead to trivial error in all cases? Severe error in all cases? Or something in between? As might be expected, the study results indicate system errors big enough to be significant in some applications but not in others.

The TOA scheme requires that the several TOA receivers be fairly widely spaced in angle about the emitter. Thus, at least two of the receivers must perform their measurement on signals far from the axis of the emitter. Moreover, the measurement of arrival time must be accurate because the emitter location is computed from the small differences among arrival times at the several receivers. These two features together tend to make the TOA technique particularly vulnerable to this antenna effect and explain the emphasis here on such systems.

During the study it was found that the usual approximate methods that are used to calculate the transient response of a circuit are inadequate in this case. In fact, such an approximate analysis loses the entire effect. It was necessary to carry out an exact analysis of the transient response of a whole (simplified and idealized) source/receiver combination. To our knowledge, this is the first such analysis to have been carried out; it discloses some curious effects that should be of general interest in the electronic community. The analysis is,

unfortunately, quite lengthy because of the need to follow the signal, step by step, through several successive electronic stages. Further, the equations become extremely cumbersome because the exact treatment is more intricate than the customary approximate treatment. Part III, which constitutes most of the bulk of the report, presents the entire analysis and several figures illustrating exact waveforms at various places in the circuit.

The equations in Part III are so cumbersome that it is impossible to obtain any insight by reading them. Indeed, it is impossible to gain a general overview of the processes at work from an examination of the waveforms shown graphically. Thus, although the analysis in Part III is the foundation, the more useful result of the study is a purely qualitative and descriptive picture that is given in Part II.

Although the subject matter is inherently technical and involves the design and behavior of electronic circuits, Part II is minimally technical. There are no equations, and the text places minimal demands on the nontechnical reader. Part II explains in approximate terms the underlying process that causes a TOA measurement to be influenced--probably more than most people would expect--by the antenna effect. Part II describes, in successive subsections, the highly simplified pulsed transmitter, the way in which the far-field waveform changes with angle, and the way in which the response of any single receiver changes with angular location. Finally, the case of three such receivers, used in combination to form a TOA system, is considered. Two computer "experiments" are described, and the TOA system errors found in these experiments are displayed graphically. Part II ends with the need for experimental confirmation, possible measures to mitigate these effects, and the study's summary conclusions.

The reader should bear in mind that the study treats a hypothesized, idealized system that would be entirely free of error were it not for the antenna effect under study. There is no random noise, no interfering signal, no multipath effects, and no propagation anomalies. All such troubles that might arise in hardware would combine with the effects treated here.

PART II

QUALITATIVE OVERVIEW

II-1. THE EMITTER

DESCRIPTION

The emitting antenna is assumed to be a large uniformly illuminated plane rectangular aperture--a most rudimentary directional source and one that is not often used in practice. This particular choice is made because the uniform rectangle leads to the most tractable mathematics. Inasmuch as the analysis eventuates in expressions that are barely tractable, this simplest starting point was a practical necessity. However, more realistic designs for the source antenna, such as a circular or elliptical shape and illumination shading to reduce side lobes, lead to results that differ only in numerical detail. No important insight is lost by considering a uniform rectangle, and as is seen later, other apertures would not lead to appreciably different phenomena.

The most fundamentally important obstacle to the study is met here at the outset. The term "aperture," widely used in antenna theory, is troublesome. If the antenna under discussion happens to be a very long slowly tapering horn, one can, with no serious ambiguity, say that "the aperture" is the plane surface at the mouth of the horn. For nearly all other kinds of directional sources, such as the commonplace parabolic dish, there is no such unique geometric surface that can be taken to be "the aperture." The concept is, in most cases, an abstraction that is adopted in the course of using an approximate method of analysis.

More important, the tacit assumption that all of the radiated signal can be regarded as emerging through a defined bounded aperture is an approximation that violates the laws of physics. It is impossible to have finite excitation across the entirety of a finite aperture and zero excitation beyond the edge. Nevertheless, this is indeed the assumption that is made here because the problem of calculating the exact behavior of any real directional antenna has not been solved. These matters are discussed further in Appendixes F and G.

The behavior of the source antenna is discussed here, and in Part III, in the approximate terms that are used almost universally for such treatment. The approximate theory works well enough for ordinary

purposes to describe the main lobe and the first few side lobes, i.e., out to angles that may be less than 10 degrees from the axis. The theory fails at larger angles and is meaningless in the rear hemisphere, but these shortcomings are not serious for most purposes. In this study, we are concerned with the detailed behavior at large angles, and the theoretical shortcomings *are* serious, but nothing can be done about them. It is this obstacle that prevents a definitive numerical analysis. The results obtained here must be taken with a grain of salt, and regarded as illustrative estimates rather than being numerically dependable.

We will suppose that the emitter transmits a simple on-off pulse whose duration is long enough so that we can ignore any events arising from the end of the pulse and can concentrate attention on the beginning. If this simple pulse is assumed to be the signal found at large distance on the antenna axis, then, according to the approximate antenna theory, this is also the waveform that excites every portion of the uniform plane radiating aperture. This too is an unphysical approximation, but it allows us to estimate the signal that will be found at other locations away from the axis.

Real devices necessarily have finite bandwidth, and it is not possible to turn a pulse on instantly to full amplitude. The amplitude must rise gradually, but the rise time can be quite brief. If the emitter is a fairly typical ordinary radar, the pulse rise time will be considerably less than the time required for the circuits in a typical receiver to respond. The receiver too must have finite bandwidth, and this bandwidth is not likely to be greater than that of the pulsed source. Indeed, most receivers are designed to reject undesired noise and other interference, and the rise time of the receiver circuits is usually much longer than the rise time of a radar pulse. Although the pulse rise time of the source is carried along explicitly in the analysis reported in Part III, it is convenient here to discuss the pulse as if it rose instantly to full amplitude. In radio parlance such a signal is called a stepped carrier, and that term will be used in the following discussion.

THE ANTENNA EFFECT

We consider next the question of what signal the idealized emitter transmits in other directions off the source axis.

If the source were transmitting an endless unvarying sinusoidal carrier signal, then the waveform at all places in space would also be that same endless unvarying sinusoid. However, at various angles from the axis the amplitude of the sinusoid would differ. The amplitude would be highest on axis and lower everywhere else. At numerous particular angles (depending on the wavelength of the sinusoid and the size of the aperture) the amplitude would be zero; these angles are called pattern nulls or minima. At angles between nulls the amplitude rises, passes through a maximum value that is less than the axial value, and then falls to zero again in the next null. These regions between nulls are called side lobes. For a simple aperture such as that considered here, the maximum amplitude reached in each side lobe is lower than that reached in the lobe next nearer to the axis. Our theory is incapable of describing the lobe structure at angles more than 90 degrees from the axis, and generally fails to portray accurately the lobe structure of real antennas beyond the first few side lobes.

The foregoing, which describes the so-called lobe pattern or directivity pattern of the emitter, is restricted to the steady-state situation wherein the source transmits an endless sinusoid at one particular frequency. The concept of a lobe pattern is in fact out of place in a discussion of modulated signals such as the stepped carrier. If the transmitted signal varies in any way whatever--including being turned on or off--then it contains numerous different frequencies, and each different frequency has associated with it a different lobe pattern.

Despite this fact, nearly all discussions of the signals found in the far field of a modulated emitter suppose that the same waveform is found at all far-field locations, and that they differ from place to place only in having different amplitudes. The amplitudes are supposed to be those described by the directivity pattern at the carrier frequency. This picture of the situation tacitly supposes that there is a single lobe pattern associated with a modulated signal.

Such a description of the far-field signal is an entirely adequate approximation in most ordinary circumstances for several reasons:

1. Most modulated signals contain frequency components that occupy a bandwidth that is not a large fraction of the carrier frequency. Consequently, the lobe patterns of the various frequencies do not differ much.
2. Most ordinary electronic systems, such as radars and microwave links, rely primarily on signals very close to the antenna axis where the lobe patterns differ very little.
3. Most systems are tolerant of minor changes of waveform, and the small waveform effects that are overlooked by this simple picture are negligible.

Nevertheless, this picture, wherein the waveforms are assumed to differ only in amplitude from place to place, overlooks the minor effect that gives rise to this study.

To understand the waveforms that arrive at off-axis locations, we must consider the events that occur as the pulsed signal commences to arrive at such a location. For convenience we can set a clock so that the time $t = 0$ is the instant when arrival would commence if we were on the axis at the same distance from the center of the source aperture. Off axis, one end of the emitting aperture is a little closer than is the middle of the aperture. A signal arriving from that near end must travel a slightly shorter distance and consequently arrival commences earlier than $t = 0$. However, when this arrival commences, it is only the nearer end of the aperture that contributes; there has not yet been time for the more remote parts of the aperture to contribute. Consequently, the signal starts out relatively feeble off axis, whereas on axis it starts right out at full strength.

At a slightly later time, but still earlier than $t = 0$, there has been enough time for the signal to come in from somewhat more remote portions of the near half of the source. Meanwhile, of course, the earlier arriving signal from the nearest portion has continued to come in. These contributions arriving from different parts of the aperture require different travel times. They start out at the aperture all in step but arrive out of step because each portion travels a different distance. Because the various contributions are out of step, they fail to add up to the waveform that is found on axis.

When t = 0, the contribution from the center of the source begins to arrive. At that moment the near half of the aperture is contributing, but the other half is not. Inasmuch as the out-of-step components do not add up to a replica of the on-axis signal, the amplitude at t = 0 need not be half of the ultimate amplitude; indeed, it can be far greater.

It is only at a time later than t = 0 that there has been sufficient time for the most remote part of the source to contribute. Thus, the pulse that arrives off axis builds up over a time interval, whereas that on axis arrives, all in step, at full amplitude. The leading edge of the pulse exhibits an entirely different shape off axis.

If we move a still greater angular distance from the axis and consider the process of arrival there, we note that the near end of the source is even closer than it was. Consequently, arrival begins earlier than it did at the first off-axis location. Moreover, the more remote end of the aperture is even farther away than before and contributes later. Thus, the duration of the interval over which the leading edge builds up increases with the off-axis angle. The "out-of-step" arrangements are different among the various incremental contributions, so the leading edge has a different duration and shape at each angle.

The duration of the buildup is presumably greatest at right angles to the source axis. There the duration is equal to the time required for the signal to travel the length of the aperture.* At off-axis locations the incoming signal begins relatively early by an amount that does not exceed the transit time across half the source aperture, but it is only that first arrival that is displaced that much. Indeed, much of the signal leading edge arrives relatively late. Thus, the effect

---

*We are unable to discuss what happens at angles beyond the right angle. Such a discussion would entail the problem of how the signal finds its way around the end of the aperture and into the region behind, which would require that we understand how the radiation behaves throughout all space. We do not understand this phenomenon because we cannot solve the equations. Indeed, because we do not really understand what happens in the forward region, the discussion above is not realistic. In other words, we do not know enough to describe this buildup process accurately, and thus we cannot calculate the shape of the leading edge correctly. Nevertheless, the description above is approximately correct and it is useful to consider its implications in some detail.

is not at all to displace the waveform forward by a small amount, but rather to produce a different shape at different angles.

II-2. TOA SYSTEMS

In a TOA system three or more receivers make accurate measurements of the signals that they receive from an emitter.* If the source (for example, an ordinary radar) emits a pulsed signal, then in a leading-edge TOA system each receiver measures the time at which the leading edge of a pulse appears to arrive (Time Of Arrival). Three such time measurements, together with knowledge of the locations of the receivers, suffice to calculate the location of the emitter. Similar methods can, of course, be used to devise various kinds of navigation systems. A system of three receivers used to locate an uncooperative emitter would be one member of a generic class of systems that use arrival time data (see Appendix J).

Such systems appear to be particularly sensitive to the antenna effect for a combination of three reasons:

1. The waveform deformation discussed here is associated with modulation, and the deformations that occur are temporally localized in the vicinity of a modulating event. The turn-on of a fast-rising pulse is an especially abrupt kind of modulation. The shape of the leading edge will change appreciably with angular location.
2. The three receivers must subtend a sizable angle at the source to obtain favorable geometry. If they are bunched together

---

*Throughout this report, as in the study, the receivers are assumed to contain nondirectional receiving antennas; all of the antenna effects that are considered arise in the directional emitter. If the receiver and transmitter of a system use directional antennas, then these effects will arise in both ends of the link unless the receiver axis is aimed at the source and the source axis aimed at the receiver. The effects attributable to antenna directivity are bilateral and occur equally during transmission and reception.

so they are at nearly the same angle from the source (and thus obtain nearly the same leading-edge waveform), then very small measurement errors, perhaps caused by noise or hardware faults, can lead to large errors in the computed location of the source. The difficulty is analogous to the errors that arise in an optical rangefinder whose baseline is too short. The need to be widely spaced in angle makes it certain that at least two of the receivers will be far from the axis of the source and will therefore experience the antenna effect.

3. Even with favorable receiver geometry, the time measurements must be quite accurate if the computed location of the emitter is to be usefully accurate. The requirement for system accuracy varies, of course, among possible applications, but in some cases a timing disparity at any one receiver as small as 10 or 20 nanoseconds may lead to significant system error.

There are alternative ways to carry out the TOA task that are generally less likely to be troubled. For example, if the source emits pulses the three receivers could carry out cross-correlations over the entire lengths of the pulses they receive. Such an approach, which can be looked upon as making better use of the entire energy content of a pulse, will be less perturbed by deformation at the front and back ends of the pulses. Nevertheless, the deformations at the ends will degrade the correlation, and this method is not entirely immune to the effect. Furthermore, this method is considerably more complicated and calls for more hardware. Such alternative methods have not been considered in this study, although the analysis reported in Part III could be applied.

II-3. ARRIVAL TIME

DEFINITION

It is important to have a clear understanding of the term "arrival time" of a pulse. Confusion in terminology can lead to substantial misunderstanding.*

The clocks used at the three receivers must be synchronized, but the actual clock readings themselves are of no import; significance attaches only to the *differences* among the times observed at the receivers. Consequently, we are free to set the clocks to any convenient reading.

We consider here a system that would be perfect were it not for the one complicating effect under study. By supposing that the system operators take perfect account of all the other details needed to make the system work, we can discuss timing disparities as displacements from what any one receiver *should have* observed. If all the receivers are free of such a disparity, then the system incurs no error. Similarly, if all were to experience exactly the same disparity, then no system error would occur; such an event would correspond merely to changing the settings of all the clocks by the same amount. System error arises only when the measurements made by the receivers differ from the values that each *should have* obtained by an amount that is not the same for all. The discussion is thus simplified because it is not necessary to consider how far each receiver really is from the source (and therefore the true time at which each really does observe arrival). This arrangement can be viewed as equivalent to setting the

---

*Section II-1 considered the process wherein the arrivals of signals from different portions of the emitter combine to produce the radiated waveform at a point in space. That waveform cannot be observed in any real receiver because such a receiver would require infinite bandwidth. A real receiver must impose some frequency selectivity, and the waveform that emerges at the receiver output terminals must differ from that in the radiation field. A decision concerning the apparent arrival time of the signal must be based on the receiver output waveform. The two waveforms are related to each other by the transient response characteristics of the receiver, and events in one are not simply related to events in the other.

clocks at the receivers to different times to take account of their different distances from the source. All will obtain readings equal to the values they would have received if they were equidistant from the source. (A system cannot really be operated this way because the distances from the source are not known. This artifice merely allows us to discuss errors in the computed location of the source without being concerned with how the computation is actually carried out.)

If all the receivers received the same incoming waveform, then all would deliver the same output response at the terminals where "arrival" is observed. No difficulty would arise as to what is meant by "arrival." In principle, the signal emerging from the receiver could be compared with a "standard"; when the two coincide perfectly over their entire length, that instant could be taken to be "the" time of arrival. Alternatively, the difference between the arrival times at two receivers could be determined by sliding one signal back and forth in time until the two coincide perfectly. The time shift needed to make them coincident is then the time difference, and no ambiguity arises. Discussion of TOA systems is often couched in terms that tacitly suppose that these are the circumstances that would prevail were it not for hardware faults, noise, and other waveform corrupting influences.

We consider here a system that is so idealized that those corruptions do not occur. Nevertheless the different receivers will, because of the antenna effect, deliver different output waveforms. There are no times at which the different outputs will match a "standard" waveform, and there is no time shift that will cause two of them to match. "Time of arrival" of a signal is, under these conditions, not amenable to unique definition. Rather, the instant that is to be regarded as "the" arrival time must be defined by the occurrence of some particular event, and that event must be chosen arbitrarily from any number of alternative candidates. In general, different definitions of "arrival" will yield different results. No particular definition that might be adopted is known to be the best; indeed, it is not evident that such a universal optimum choice exists. All definitions appear to have advantages and drawbacks whose relative import probably changes with the operational circumstances.

During most of this study, the arrival time of a pulse was taken to be the instant when the output signal from the receiver first reaches one-half of the peak amplitude that it next reaches. This choice is entirely arbitrary; for example, the fraction could as well be 3/5 as 1/2. It is, however, one that can be implemented in straightforward hardware and it has the advantage that the receiver output does tend to pass through a peak a short time after the output passes upward through half that peak.

Whatever definition of "arrival" is adopted should be reasonably free of dependence upon the gross signal strength of the incoming signal, because such signal strength is unreliable for several practical reasons (including the dependence on source-receiver distance). It might not be necessary in practice to adopt a definition that is completely independent of signal strength; in Part III a weakly dependent alternative method is mentioned. However, this "1/2 peak" definition is, in principle, independent, and we need not here be concerned with signal strength per se.

In the detailed analysis, it is convenient to adjust the clocks as follows: a receiver that is on the source axis will commence to receive the beginning of the incoming signal at exactly t = 0. Arrival, defined in terms of the 1/2 peak, occurs a little later; how much later depends on the receiver design details (and on the rise time of the source pulse). For the particular receiver design adopted in this study, and for the numerical values that were chosen, arrival on the source axis occurs when the clock reads about 116 nanoseconds. The numerical value is, in itself, inconsequential and merely reflects a convenient setting of the clock. This is the time when a receiver *should* report arrival. Moreover, if the receiver is on the source axis, the analysis supposes that it *does* report this value; there is no waveform deformation on axis, and the system is otherwise free of error.

At other angular locations, the receiver *should* report the same arrival time: 116 nanoseconds. However, as explained above in Section II-1, the incoming signal really commences earlier than t = 0 because one end of the source aperture is a little nearer. That does not necessarily mean that arrival will be measured to be earlier than

116 nanoseconds. The time at which the 1/2 peak logic decides that arrival occurs depends upon how the receiver responds to the incoming waveform. Depending on that waveform, which changes with angle (and with other parameters we have not yet mentioned), the arrival may occur later or earlier than 116 nanoseconds. In a typical system arrangement, if the middle receiver reported arrival at 116 nanoseconds and the two outer receivers reported arrival at 94 nanoseconds, the calculated source location would be in error by about 13 meters (for a symmetric triad of receivers that subtends 90 degrees at the source).

Finally, some people find it more convenient to grasp the significance of time intervals expressed in equivalent distances rather than in nanoseconds (using the speed of light: 0.3 meters per nanosecond). Consequently, we will express the arrival time that a receiver *should* observe as 34.8 meters instead of 116 nanoseconds. In the numerical example mentioned above, the arrival time difference of 22 nanoseconds among the receivers can be expressed as 6.6 meters. (It should be noted that this timing disparity of 6.6 meters is less than, and not simply equal to, the system error of 13 meters. The relationship involves not only the disparities among the receivers, but also the geometry of the receivers with respect to the source. These relationships are given in Appendix J.)

CALCULATED ARRIVAL TIMES

The detailed mathematical analysis was used with a digital computer to calculate the entire receiver response at a wide variety of angular locations and, from those responses, the precise arrival time that a receiver would report at each angle. (This was done for a particular set of numerical values for the many parameters. The same values were used throughout; they reflect only one set of conditions, thought to be reasonably typical.)

The many values of arrival time (hundreds) thus obtained are very confusing, and some aspects of the results remain confusing at this writing because no useful generalization of their behavior was discovered. There were arrival time shifts in a single receiver as great as about 20 meters. Despite the early arrival of the first signal component

off axis, some arrival times are reported to occur later than on axis because of the waveform changes. (Note that shifts as large as 20 meters are much greater than half the width of the source aperture.)

The dependence of arrival time on off-axis angle is quite complex; the times do not shift gradually as the angle increases. Instead, if the receiver is near the peak of any side lobe the arrival time is about the same as it is on axis. Departures from that value occur whenever the receiver approaches the edge of any side lobe. Inasmuch as a highly directional emitter has many side lobes closely spaced in angle, the arrival time tends to fluctuate very rapidly with quite small changes in the angle. With three receivers at widely different angular positions, it is not at all unlikely that one (or more) will find itself near the edge of a side lobe; it is this event that leads to appreciable system error.

In the numerical example cited, a time disparity of 6.6 meters led to an error of 13 meters in the source location. It is evident from this that a disparity of 20 meters can produce a system error of nearly 40 meters. Combining disparities from all three receivers results in an even larger system error; a few system errors nearly equal to 50 meters were encountered in the computer experiments discussed later.

II-4. THE SIGNAL AS THE SUM OF TWO SIGNALS

The detailed numerical values of arrival time that emerge from the analysis result from complicated equations and a variety of factors that influence the results. The expressions are so complicated that one cannot infer any useful generalizations from inspection, nor do visual inspections of waveforms provide much general insight. Seemingly trivial changes in waveform may cause large consequences, and seemingly substantial changes may have small consequences. The study results may appear frustrating because one is unable to gain a simplifying overview as to cause and effect. It seems as though even a minor change of any parameter might change the results in an inestimable way, and one lacks a sense of what is important or why.

If we set aside the actual numerical values, and especially the more intricate aspects of their behavior, it is possible to understand in qualitative terms the principal processes that are at work. They are *qualitatively simple, even though quantitatively intricate*, and this qualitative insight is undoubtedly the most useful result of the study. The starting point for this qualitative understanding is that the incoming signal to the receiver can be regarded as two signals that arrive together. (Formally, we invoke linear superposition.) The two will be called the "expected" signal and the "extra" signal. It is the latter that is the troublemaker.

It will be recalled from Section II-1 that, in common practice, the signal at any point in the far field of the emitter is taken to have exactly the same waveform as that found on the axis, but to appear at lower amplitude. In turn, the receiver response that is excited is expected to be the same at any place in the field except for a corresponding change of amplitude. If, as here, the definition of "arrival" ignores amplitude, then the expected arrival time will be the same everywhere. No timing disparities will exist, and a TOA system will make no error in reporting the location of the source. We will call this the "expected" situation, and the signal waveform found on the emitter axis the "expected" waveform. The "expected" signal at any location is the axial waveform, appearing at whatever amplitude is implied by the steady-state lobe pattern at the carrier frequency.

It will also be recalled that the signal that does arrive at any place off the source axis is not the same as this "expected" signal. Rather, the real signal commences early, when the "expected" signal has not even begun, and the real signal builds up more or less gradually as different portions of the source contribute. During that brief interval the waveform is not at all the same as the "expected" signal because all the various incremental contributions are out of step. It is not until the entire source aperture is contributing that the real signal is the same as the "expected" signal. Thereafter, there is no difference between the two.*

---

*To simplify the picture, we here set aside the finite rise time of the source pulse.

The "extra" signal is completely absent on the antenna axis, but is present at *all* off-axis locations. The amplitude of the "extra" signal is not particularly feeble; the amplitude is generally comparable with that of the "expected" signal. However, the duration of the "extra" signal, although longer at larger angles, is always quite brief. The waveform of the "extra" signal is different from the early portion of the "expected" signal. (See Figs. III-3 through III-6.)

TRANSIENT RESPONSE OF THE RECEIVER

The "expected" signal is taken, for this qualitative discussion, to be a stepped carrier--that is, a steady sine wave that is turned on instantly to full amplitude.* Any particular receiver will respond to this input in a way characteristic of the design. Indeed, a description of the output waveform elicited by this (or any other) particular transient input is as complete a description of the receiver as is the more commonplace (complex) frequency response. (Either can, in principle, be calculated from the other.)

When we attach to the output terminals of the receiver a circuit that employs some kind of logic to report "arrival time," then that circuit will report such a time in response to the stepped carrier. The reported time is dependent on the definition of "arrival," and is also dependent on the receiver design. Various designs respond differently to a stepped carrier, and they will yield different arrival times for the same input.

The receiver design analyzed in this study is considerably simpler than most real receivers (mathematical difficulty makes it impractically tedious to treat realistic designs). For this particular design, and the 1/2 peak arrival logic, a stepped carrier is reported to arrive at 34.8 meters (about 116 nanoseconds after the stepped signal begins to enter the receiver). More realistic receiver designs would probably yield later reported times with the same 1/2 peak logic. It is nearly impossible to calculate this time for real receivers, and no rules of

*To simplify the discussion, we suppose that the signal starts at an upward zero crossing of the carrier.

thumb are available to provide a good estimate, but the value could be measured on the bench with no great difficulty. Values of 50 or 60 meters seem plausible.

Another characteristic transient response of the receiver is the so-called impulse response. Indeed, this is generally looked upon as the fundamental description of the receiver; it is the impulse response that the analyst uses to calculate any other transient response. An impulse can be looked upon here as an extremely brief spike: a signal that rises very quickly to a peak value and falls back to zero equally quickly (formally, infinite height and zero duration).

The impulse response of the receiver rises to a peak value sooner than does the stepped carrier response, and it dies away in a sequence of decreasing peaks. The impulse response is zero when the stepped carrier response reaches its peak.* The two responses cannot be made to be the same, and they necessarily lead to different reported arrival times. How much they differ depends on the design, but there must be a difference for any realistic arrival time logic.

For the receiver design and 1/2 peak logic considered here the arrival time of the impulse response is about 20.7 meters (about 69 nanoseconds after the delivery of the impulse). This arrival time, and the 14.1 meter difference between the two values, is a characteristic of the receiver design. It is not related in any way to the attributes of the pulsed emitter. The *magnitudes* (but not the frequency of occurrence) of the system errors made by a TOA system in locating the emitter are proportional to this difference between the two arrival times. The difference--14.1 meters in this case--is called here the *scale length* of the receiver, designated in the analysis by the symbol S.

There are no rules of thumb to suggest how S varies with changes in the design of the receiver (or with changes in the arrival logic). It seems plausible, but is scarcely more than guesswork, that realistic designs will exhibit larger values of S--because they contain many more

---

*The receiver is a bandpass device. The response to a stepped carrier is not the step response, but the two are related inasmuch as the latter is approximately the envelope of the former. The impulse response is the derivative of the step response, not of the stepped carrier response.

tuned circuits than the four in the IF strip (intermediate-frequency amplifier) that was analyzed. If so, then the study indicates that the system errors found in the computer experiments should be scaled upward in proportion. That is the principal reason that the results reported here are thought to underestimate the severity of the effects.

THE "NEW" PHENOMENON

Under typical conditions, the rise time of a pulsed source is very brief in comparison with the rise time of the receiver.* When the receiver is on the axis of the source, the incoming signal excites a response that is essentially the stepped carrier response, and the reported arrival time is 34.8 meters.

If, however, the receiver is located in any null of the source lobe pattern, the "expected" signal is absent, and only the "extra" signal arrives. The "extra" signal is so brief (particularly in the first few pattern nulls) that it excites a receiver response that is essentially the impulse response. In those locations the reported arrival time is 20.7 meters.

At all other receiver locations both signals arrive and both responses are excited. The "new" phenomenon that was found in this study is that, at *all* off-axis locations, a pulsed directional emitter excites (very nearly) the impulse response of the receiver as well as the stepped carrier response (very nearly) that is ordinarily expected. Although couched here in terms of a pulsed signal, the phenomenon can be generalized to any kind of modulation: each modulating event will tend, off axis, to excite the impulse response to some extent. The net response of the receiver to this simultaneous excitation of two responses can exhibit surprising departures from the response that is ordinarily thought to occur.

---

*The duration of the source pulse is usually longer than the source rise time. The receiver bandwidth is usually about equal to the reciprocal of the pulse duration, not of the source rise time. Moreover, the receiver contains numerous tuned circuits used to obtain skirt selectivity, and these slow the rise of the receiver.

The response that is excited by the "extra" signal is not exactly the impulse response, but is very nearly so. Thus, the receiver tends to be almost entirely insensitive to the exact waveform of the "extra" signal; brevity is the salient feature except insofar as waveform and duration influence the amount of energy that is delivered. Such matters as the shape of the source aperture and the illumination distribution have scant influence on the receiver output waveform that is excited in pattern nulls, and correspondingly little influence on the reported arrival time. The difference between the 34.8 and 21.7 meter arrivals--i.e., the scale length--is a property of the receiver and is virtually independent of the emitter.

To the extent that the response to the "expected" signal is not exactly the stepped carrier response, and the response to the "extra" signal is not exactly the impulse response, the rise time of the source pulse does influence the two receiver responses. So too does the fact that the duration of the "extra" signal increases with off-axis angle. When either of these responses is excited alone, these variables make *inconsequential differences in the output*. When both responses are excited at once that is not true, and the changing duration of the "extra" signal with angle leads to a steady trend of numerically different behavior near each successive null. It is features such as this that are confusing in detail but whose origins can be understood qualitatively in fairly simple terms.

The "extra" signal arises from the fact that the source aperture has finite width--the same feature that, for steady sinusoidal excitation, leads to the existence of the lobe pattern. Directionality and the "extra" signal are inseparable; the existence of one requires the existence of the other. Our inability to calculate the exact behavior of real directional antennas prevents us from calculating the correct lobe pattern, just as it prevents calculation of the exact transient behavior. Nevertheless, antennas do exhibit directionality that must be accompanied by "extra" signals. The qualitative fact that this "new" phenomenon occurs is on firmer ground than the numerical results of the study.

II-5. INTERFERENCE BETWEEN TWO RECEIVER RESPONSES

Consider now the receiver response in all the angular locations where both signals arrive and both responses are excited. Throughout the receiver circuits, up to but not including the envelope detector, the impulse response and the stepped carrier response are oscillatory signals at, approximately, the center frequency of the IF strip. Both responses exhibit phase modulation, so that their zero crossings are not evenly spaced. Moreover, the amplitudes of both responses vary in time but in different ways. Two such oscillatory signals will reinforce or compete with each other, moment by moment; they interfere. If, as is the case here, two interfering signals are amplitude- and phase-modulated, then the interference is complicated and is acutely sensitive to the strengths at which they were excited, and also to the precise times at which they were excited. The resultant of the interference can exhibit a waveform that is different from either alone, and the resultant can be highly dependent on trivial changes in either of the two interfering waveforms. It is the intricacy of this detailed interference process that causes the numerical results obtained in the study to be so complicated and confusing. It is fortunate that the process can be understood qualitatively in simpler terms.

The interference yields waveforms that, after passage through the envelope detector, exhibit arrival times that differ from the arrival time of either of the two basic receiver responses alone. Most of these times lie between the impulse and the stepped-carrier arrival times, but some of the arrivals are earlier than 20.7 or later than 34.8 meters. Thus, the extreme range over which the reported arrival time can change *exceeds* S; the scale length S does not bound the limits over which the time can vary in any one receiver.

When the receiver is near the top of any side lobe, the amplitudes of the expected and "extra" signals are comparable. The expected signal excites a stronger receiver response than does the "extra" signal because it delivers more energy to the IF strip within its rise time. Under these conditions, the interference is a one-sided contest between

the two responses. The feeble impulse response excitation cannot much alter the stronger stepped-carrier response, and the resultant waveform does not differ very much from the stepped-carrier waveform alone. This is why the measured arrival time is about the same over much of the width of each side lobe as it is on-axis.

As the receiver location approaches the edge of any lobe, the amplitude of the "expected" signal falls. This amplitude reduction offsets the long temporal duration of the "expected" signal, and the two characteristic receiver responses are excited at more nearly comparable strength. The interference between them can be strong, and the resultant can exhibit quite a different waveform. It is here, near but not in the nulls, that the most drastic changes of arrival time occur. Remarkably large shifts can occur with very small changes in the angular location--as much as 10 meters or more in less than one minute of arc.

The most extreme shifts of arrival time occur near the first pattern null between the main lobe and the first side lobe, but they occur only in a tiny angular interval very close to the null. At larger angles the shifts that are found tend to be less drastic, but sizable shifts are found over a larger fraction of the width of each lobe. This change is a reflection of the change in the duration of the "extra" signal with increasing angle, and exemplifies the complexity of the detailed results. On balance one might reasonably estimate that, for a highly directional source, significant time shifts are found over perhaps 10 percent of the angular interval between 0 and 90 degrees. (This is at best a matter of opinion as to what is significant, and such judgment depends on many factors including the system's intended use.) Such an estimate suggests that there is something like a 25 percent probability that at least one receiver out of three will be found to be in a troublesome location. Thus it should not be supposed that these troubles occur too infrequently to matter; there are three receivers, and there may be many nulls.

PHASE EFFECTS

The final complication is the effect of two pertinent phases that work in concert to change the measured arrival time that would otherwise be expected at any particular angle.

One is the emitter carrier phase--the phase that the source carrier oscillator had at the instant the source pulse began. In many pulsed sources no effort is made to hold this source phase constant, even though such control is usually possible, and the phase changes more or less randomly from one pulse to the next--possibly over a large range. It would not tax an uncooperative emitter operator to make this phase vary.

The second pertinent phase is the phase of the receiver's own local oscillator at the moment the signal arrives. This phase depends on the total transit time of the signal from source to receiver. A change of only 1/4 wavelength out of that whole distance would change this phase by 90 degrees; that much change could arise from a change in the average index of refraction as well as from a small change in the receiver location. It seems probable that, in practice, this local oscillator phase will vary randomly from one pulse to the next in any one receiver and also randomly from one receiver to another on any one source pulse.

The two phases influence the outcome in (very nearly but not quite) the same way and they operate in concert. Practically speaking, they can be regarded as interchangeable; if either one varies randomly it is immaterial whether or not the other is held constant.

If either of these phases is changed, the precise waveforms of the stepped-carrier response of the receiver and of the response to the "extra" signal will change. (The waveform of the impulse response does not change with phase.) The previous discussion treated these two responses as if they were fixed quantities, but that is not quite true. For either response alone, in the absence of interference between them, the effect of phase is small: the reported times of arrival of either response change about 0.2 meter (about the same for both responses) as the phases change.

When both responses are excited and appreciable interference beween them occurs, even these small effects of phase on the individual waveforms become important. Near the edges of the side lobes, where interference plays a strong role, phase changes cause large shifts in the reported arrival time.

These phase effects are exacerbated by a feature of transient behavior that is not present in the steady state: the *amplitude* of the receiver output in response to a very brief input signal depends strongly on the phase of the local oscillator. For example, in the receiver design considered here the receiver output caused by the "extra" signal in the first pattern null of the source changes 14 dB with local oscillator phase. At more remote pattern nulls, where the angle is larger and the duration of the "extra" signal is longer, this effect of phase on amplitude is much less. That is why the largest effects are found near the first pattern null.

These changes caused by random phases mean that the arrival time reported by a receiver at any one location may turn out to have any value over a considerable span. The value reported on any one pulse is a matter of chance, and the value on the next pulse may differ. Two receivers symmetrically placed at exactly the same angle on each side of the source axis might report arrival times on ten successive pulses with no two of the twenty values in agreement. Moreover, if one of them were moved in angle by as little as 1/4 of a degree, the two sets of reported numbers might not even overlap in their ranges.

As is evident from the foregoing, the effects are so complicated that it is very nearly impossible to sum up in any simple way the net result on the working of any real system. Some feeling for the severity of the effects is provided by the results of two computer experiments described next.

II-6. TWO EXPERIMENTS

TWO ANALYSES

Before describing the two computer experiments, we discuss why two were needed and how they differ. We also address an important problem that was recognized only after the understanding of the "new" phenomenon was obtained.

Receivers are designed to respond to signals that lie within a chosen bandwidth about the frequency to which the receiver is tuned, and to reject strongly other signals that lie only a little outside the chosen bandwidth. The desired frequency response is a more or less flat response over the chosen bandwidth and very steep falloff--called skirt selectivity--at the edges. Steep skirt selectivity requires the use of numerous tuned circuits, and these are incorporated in the IF strip that follows the heterodyne detector.

The portion of the receiver ahead of the heterodyne detector--in radio parlance, the front end--also exhibits some frequency selectivity. However, the tuned frequency of the front end (along with the local oscillator) must be capable of being changed to different frequencies to cover the band of frequencies that the receiver is designed for. The front end contains few tuned circuits, and the frequency selectivity is not comparable to that of the IF strip. Nearly all of the selectivity of the receiver is in the IF strip, and it is common practice to ignore the selectivity of the front end when considering the response of the receiver.

Because the analytic difficulties of this study would be practically unmanageable if the analyst undertook to deal with more than a few tuned circuits, the IF strip considered contains only four. For the same reason, and because the front-end selectivity is negligible, it was omitted entirely. The first computer experiment used a receiver in which the front-end sections (including the receiving antenna) are credited with infinite flat bandwidth.

The qualitative insights described in Sections II-4 and II-5 may not have been gained if a selective front end had been considered at the outset. Nevertheless, recognition that the "extra" signal excites the impulse response of the receiver cast doubt on the justification for omitting front-end tuning. It became necessary to incorporate front-end selectivity (which adds very considerably to the mathematical burdens and computer costs) and to carry out virtually a second complete study. The receiver used in the second computer experiment was otherwise the same as the first receiver.

The argument that necessitated the second study stems from the fact that the brevity of the "extra" signal is an essential feature if it is to excite the impulse response of the receiver. A tuned circuit of even limited selectivity cannot pass so brief a signal. If a very brief input is delivered to such a circuit, the circuit will continue to ring (the analogy to a bell is not inappropriate) after the input has ended. The output from the tuned circuit will endure for as long as the ringing lasts, and that duration is made longer if the selectivity of the circuit is increased. Such a tuned circuit can be said to stretch an input pulse.

It was necessary to inquire whether a tuned front end, with reasonable selectivity such as might appear in a real receiver, would stretch the "extra" signal enough to remove most of the effects that showed up in the first analysis. There was no choice but to carry out the second analysis.

The results of the second study are discussed more fully below, and in detail in Part III. Briefly, the front-end tuning modifies the results--in some ways considerably. On balance, there is some improvement in TOA system accuracy, but not much. The front-end tuning does not upset the qualitative picture that was presented above; in fact, that picture helps us to understand some of the fresh complications that the tuning produces.

FIRST COMPUTER EXPERIMENT

When a receiver is at any fixed angle with respect to the emitter axis, the reported arrival time of a pulse can have any value within a span of values that depends on the angle. Variation within that span of arrival times depends on the accidental values of the phases that prevailed for that pulse.

If the receiver remains in that same location and observes a great many pulses, the randomness of the phases will result in a distribution of arrival times. The effect of the phases operates sinusoidally, with the result that the distribution of arrival times is the frequency distribution of a sinusoid; measurements at one or the other end of the span are most probable, and values at the middle of the span are least

probable. The average of a great many measurements is the middle value, but that is the value least likely to be observed on a single pulse.

Both the average value and the size of the span change with the angular location of the receiver, but in entirely different ways. These quantities (average and span) change rapidly in angle as the receiver moves across any one side lobe, and the way in which they do so changes gradually as the receiver moves across one lobe after another. For the receiver with a wide-open front end, a formula was found that describes fairly well how the *average* value changes with angle. That formula was used in the first experiment. The formula is certainly not accurate, but the computer experiment is not seriously misleading because of the inaccuracies. Despite considerable effort, no formula was found to describe adequately the way the span of arrival times changes with angle.

The formula for the average over all phases was important because it allowed the computer to come up with an approximate average arrival time at any angle whatever without literally calculating through the equations to find the true value. Each such computation is costly, and it would be too expensive to obtain the average at a great many angles. With the formula, the computer came up with 120,000 average values at small cost. Such large numbers of results provide "good statistics" and lead to a smooth curve.

The first experiments simulated the operation of a simple three-receiver TOA system working, one at a time, against each of four emitters. In each trial the receivers formed a symmetric triad that subtended 90 degrees total angle (two 45 degree angles) at the source. The source was rotated, starting out with the axis aimed at the center receiver, and turning through 45 degrees to end up with the axis aimed at one of the outer receivers.* The source was stopped at each of

---

*The triad is symmetric and random effects of phase are presumed to be removed by perfect averaging. Under these conditions mirror image results would be obtained in turning the other way through 45 degrees to aim at the other receiver. The restriction to 45 degrees is imposed because the antenna theory being used is exceedingly questionable at large angles and inapplicable beyond 90 degrees from the axis. When the axis is aimed at one of the outer receivers, the other outer receiver is 90 degrees from the axis. If one were to decrease the angle subtended by the receivers, the source could be turned through

10,000 evenly spaced angles in that 45 degree span. At each such source rotation angle the receivers can be regarded as having observed the arrival of a very large number of pulses. Each receiver reported the *average* arrival time for its own location at that source angle. The three average values were then used to compute the system error in finding the source location for that source rotation angle. (The computed locations trace a complicated path as the source rotates.)

In most instances, the system made some error in locating the source, in spite of the averaging, because the antenna effect is not removed by averaging; only the influence of random phases is removed by averaging over many pulses at one angle. The experiment involves a good deal of idealization, including especially this perfect averaging over many pulses. A source that is truly rotating would prevent such averaging. Even if the source stands still it might be unrealistic to suppose that the receivers can also stand still and that the horizontal gradient of index of refraction does not change. The change in the *average* arrival time with angle is so rapid that such averaging may be impossible in practice; angular shifts of a small fraction of a degree might spoil the average. Nevertheless, the experiment ignores such difficulties.

This experiment was carried out four times against sources having different directionality (half-power beamwidths of 1.66, 1.25, 1.01, and 0.84 degrees; the corresponding aperture widths are 30.5, 40.5, 50.5, and 60.5 wavelengths). The 10,000 values of system error from each such trial were used to obtain a cumulative distribution of system error. These four cumulative distributions are shown in Fig. II-1. Each curve plots the fraction (of 10,000) of the samples for which the radial miss distance exceeded the number of meters shown.

The four curves have substantially the same shape. There is nothing at all in these first experiments that is random. The peculiar shape of the curves is a consequence of the particular systematic effects at work here, including the precise directivity pattern of the

---

a larger angle, but the smaller subtense would magnify the system errors because of less favorable geometry. The choice of 45 degrees is arbitrary, made to balance these conflicting interests.

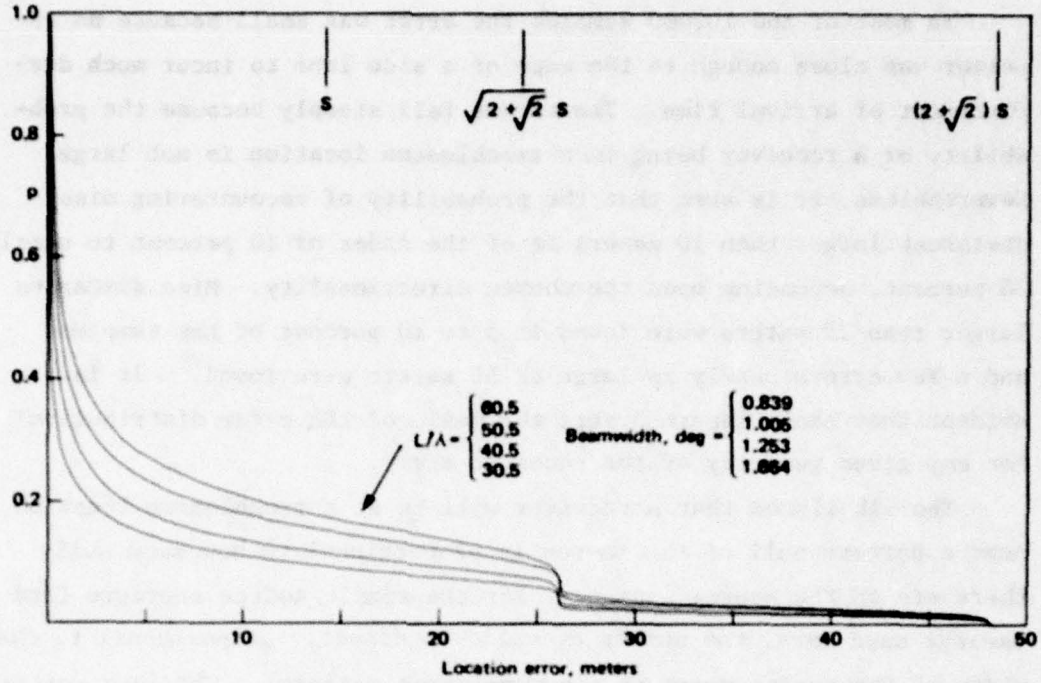


Fig. II-1 — Distribution of system errors versus directionality of the emitter. TOA receivers have no RF selectivity. Q_3 absent. All phase effects removed by averaging.

source, the transient response of the receiver, the definition of arrival, the geometry of the receiver array, and various numerical values.

The three tick marks at the top of Fig. II-1 are helpful in understanding the shape of the distribution curves. The tick on the left is at 14.1 meters; this is the scale length S for the design receiver. No notable feature of the curves corresponds to this particular distance because S does not reflect the geometry of the receiver array.

The middle tick is at 1.848S = 26 meters, and the tick on the right is at 3.414S = 48.1 meters. These two values reflect the receiver geometry (2 × 45 degrees) as well as scale length S, and it is seen that the curves show distinct breaks at those places (see Appendix J). The two major regions reflect whether one or more than one receiver is in a troublesome region, and which receiver is in such a region. (The role of the center receiver differs from that of the outer receivers.)

In most of the 10,000 samples the error was small because no receiver was close enough to the edge of a side lobe to incur much displacement of arrival time. The curves fall steeply because the probability of a receiver being in a troublesome location is not large. Nevertheless, it is seen that the probability of encountering miss distances larger than 10 meters is of the order of 10 percent to nearly 20 percent, depending upon the source directionality. Miss distances larger than 25 meters were found in 5 to 10 percent of the samples, and a few errors nearly as large as 50 meters were found.* It is evident that scale length S sets the scale of the error distribution for any given geometry of the receiver array.

The likelihood that a receiver will be at a troublesome location near a pattern null of the source is proportional to how many nulls there are in the source pattern. For the simple source aperture (and theory) used here, the number of nulls is directly proportional to the width of the source aperture (in wavelength measure). The four curves are arranged vertically in order of source directionality. The distribution curves have the curious property that the major features of the shape are controlled by attributes of the TOA system hardware design and array geometry, but the height of the curve (the likelihood of experiencing any given amount of error) is influenced by the directionality of the source.

SECOND COMPUTER EXPERIMENT

If the first results involving a receiver with a wide-open front end had indicated system errors of only a couple of meters, there would have been no reason to undertake a whole second analysis and second experiment. However, the results shown in Fig. II-1 do not appear to be negligible, especially if the more complicated circuit design of

---

*The presumption of perfect averaging over phases, together with the formula used to approximate this average, restricts the largest possible radial error to $(2 + \sqrt{2})S$. More precise calculation of the detailed arrival time would yield somewhat larger maximum error. In the absence of phase averaging, appreciably larger miss distances could occur, albeit at low probability.

real receivers should turn out to exhibit larger scale length S. The first results made it necessary to investigate whether inclusion of front-end selectivity would mitigate the effects and reduce the system errors to negligible size.

The addition of even one more tuned circuit in the receiver increases the analytic burden considerably, and with it, the cost of running each case in the computer. Further, the addition of still another variable to a list that already includes several makes it more difficult to explore the consequences of changing the numerical values. Nevertheless, the results of the second analysis and experiment are probably indicative of what can be expected from conventional front-end selectivity.

In all the numerical work in this study, the pulsed source was credited with a bandwidth of about 10 percent. This choice establishes the rise time of the source pulse and is a fairly typical value for sources such as radars. Although the second analysis is general, only front-end bandwidths of 5, 10, and 20 percent were examined numerically; the majority of the calculations used 10 percent. This 5 to 20 percent range of bandwidths seemed not only to afford some insight into the effects of the front end, but also to reflect the likelihood that a receiver designer would use a bandwidth roughly comparable to that of the source he expects to work against.

The front-end selectivity changes the numerical results of arrival time and the dependence of arrival time on off-axis angle and phase, even though such selectivity does not upset the qualitative picture of how the effects arise in the receiver. On the whole, the fractional width of each source side lobe over which substantial effects are found is diminished. However, the way in which this fractional width changes from lobe to lobe is different and more complicated; consequently, the improvement is only moderate and is offset by an increase in the span over which the arrival time is changed by phase. Thus, although the likelihood of a receiver being in a troublesome location is reduced, the amount of trouble is increased unless all phase effects can be averaged out. The interplay of these trends does not depend in a simple way on the choice of receiver bandwidth; various seeming contradictions

are found between the 5, 10, and 20 percent cases, depending on what trend is considered and in which side lobe. The effects of receiver selectivity are complex and were not explored fully.

It will be recalled that with a wide-open front end the apparent arrival time in any side lobe is very nearly equal to the on-axis arrival time unless the receiver is near the edge of the lobe. That is why most of the errors shown in Fig. II-1 are small. It is a major effect of adding front-end tuning that this desirable trait is lost. Instead, the apparent arrival times in the various lobes increases gradually with increasing angle. The shift is not large--a couple of meters in this analysis--but it causes system errors of a few meters to be almost inescapable. Even when all the receivers are safely in side lobes, well removed from lobe edges, they obtain arrival times that differ by a meter or two. When these differences combine with the geometry of the receiver array, system errors of a few meters are much more prevalent.

No formula was found in the first analysis to approximate the span over which phase changes shift the arrival time up and down from the average. In the second analysis no attempt was made to find such a formula to describe the more complicated phase effects that arise with front-end tuning. However, considerable effort was made to find a formula to approximate how the *average* arrival time varies with angle. It was such a formula that allowed the first experiment to be run at moderate cost with thousands of samples. In spite of the effort, no such formula was found in the second analysis. If one could be found, it would be fairly complicated and might be highly specialized to the case considered.

Lack of such a formula meant that a second experiment would rely on the complete analytic equations, with the result that the computer would have to calculate the exact arrival time for each receiver and each phase situation at each source orientation. It was out of the question to run off tens of thousands of values. In fact it was not possible to run enough pulses to yield the average over phases for each receiver and each source orientation. Consequently, the second computer experiment simulated a different, probably more realistic, situation but with fewer samples.

In the second experiment only one source directionality was considered ($L/\lambda = 50.5$; this corresponds to the second curve down in Fig. II-1), and the receiver front-end bandwidth was 10 percent (about equal to the source bandwidth). As before, the receiver array was symmetric and subtended 90 degrees total angle at the source. This time, because of phase randomness, the source orientation was rotated through ±45 degrees, starting with the axis aimed at one outer receiver and ending with the axis aimed at the other outer receiver. Only 100 equally spaced source orientation angles were sampled because of the high cost of this work. At each such source orientation only one source pulse was considered, and a random number was drawn to represent the phase of the source carrier at that pulse time. Then random numbers were drawn to represent the phases of the local oscillators in the three receivers for that pulse. The computer calculated the exact arrival time reported by each receiver as influenced by these phases and by the angular location of that receiver in the directivity pattern of the source. These three reported times were used to calculate the error that the system made on that particular source pulse.

There was no averaging over numerous pulses at each source orientation in the second experiment. This is undoubtedly a more realistic portrayal of how results would turn out against a source that rotates while pulsing. However, the accidents of phase are as likely to be helpful as harmful, and there is a considerable tendency for factors to "average out" in a gross sense. Certainly the absence of phase averaging makes larger errors possible, but larger errors did not happen to show up in 100 pulses. The results of the second experiment are shown in Fig. II-2.

The general shape of the curve, as would be expected, is much like those in Fig. II-1. It is also to be expected that the sharp break at 1.848S would be smoothed out in the absence of phase averaging, and the paucity of samples blurs the curve, but even so a suggestion of such a break can be seen. The largest error found in the second experiment was smaller than the largest in the first experiment, but this is a consequence of inadequate sampling.

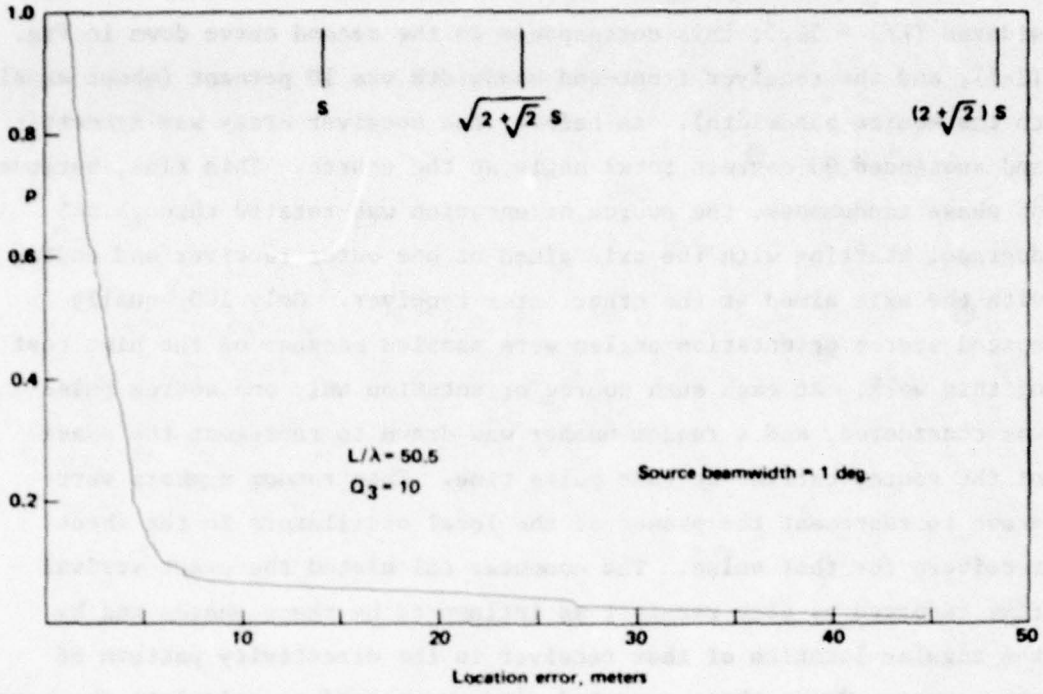


Fig. II-2 — Distribution of system errors. TOA receivers have 10 percent RF bandwidth, random phases.

At the left one finds that errors as large as 4.5 meters are far more likely in the second experiment, a result of the small drift of arrival time from lobe to lobe. It is only for errors larger than about 5 meters that the presence of front-end tuning is beneficial. The probability of very large errors is reduced because the fractional widths of the troublesome regions are reduced. The effect of the front-end selectivity appears to be a partial trade between large errors and small. How much of an improvement this is thought to be is not only a matter of opinion, but also is dependent on the application of the system and the tolerances placed upon it.

The results of the second analysis can be understood qualitatively to result from a moderate amount of stretching the "extra" signal. This is enough so the response that is excited in the IF strip no longer approximates quite accurately the impulse response. However, the response that is excited differs considerably from the stepped-carrier

response, and interference between the two occurs. It is this detailed difference in the response waveforms, and thus in the interference, that causes the effects to differ appreciably in numerical detail.

The second analysis did not examine the effect of very narrow front-end bandwidth (such as 1 percent). Use of a narrow front-end bandwidth might not harm the signal-to-noise ratio provided external noise is dominant. A very narrow bandwidth would stretch the "extra" signal still more, but it is not obvious that this would reduce the undesirable effects of interference. It might, of course, lead to some further tradeoff between large and small errors; whether the effect of such a trade would be beneficial is an open question. A narrow front end would tend to slow down somewhat the rise steepness of the receiver, and that would tend to increase the arrival time errors caused by noise. On balance, there is not a strong argument to suggest that a narrow bandwidth would be helpful, but neither is it clear that it would not be.

II-7. ASSESSMENT

Firm conclusions on the exact consequences of this antenna effect are not possible because of severe theoretical difficulties. The severity undoubtedly depends on the details of hardware design--details that make analysis exceedingly difficult. Moreover, severity is a matter of degree that depends on the performance expected of a system. No flat conclusion that the antenna effect is harmless or serious would be appropriate.

The analysis and the computer experiments indicate numerical results that could be deemed significant for some applications. Although the numerical results are only approximate, the qualitative picture of why the results arise, and of what variables are pertinent, appears to be on firm ground. Thus it seems worthwhile to consider what steps could be taken to establish the reality of the situation implied by the analysis and to reduce the uncertainty in the numerical values.

It is the writer's opinion that, aside from one or two extensions of the analysis (e.g., to examine other choices of the values of various parameters), further analytic effort would not be worthwhile. The fundamental difficulty of calculating the true transient responses of antennas would vitiate the worth of more elaborate analyses. For example, the shapes of the curves seen in Fig. II-1 depend on such minutiae as the shapes of the edges of the wide-angle side lobes of the source, and these are not amenable to analysis. Better understanding of this whole topic can only be had from experiment.

The easiest and least expensive experiment to conduct would be a study of the transient response characteristics of various receivers. Even if no receivers intended for TOA systems are available, any number of representative receivers whose design features are relevant could be examined.

The basic data would be the stepped-carrier and impulse responses of the receiver (with appropriate time resolution and control of the pertinent phases). With those two responses, the scale length S of that receiver could be calculated (or measured directly), and the dependence of S on various definitions of arrival could be examined. No doubt some definitions are less advantageous than others, but little is known about this question and analysis is extremely difficult for real receiver designs.

The investigation could, without much greater difficulty, be extended to examine the interference between these two responses when both are excited in the receiver. Such data would clarify the effects of phases, insertion times, and excitation amplitudes on the net output of the receiver.

Unfortunately, the role of the source antenna, which is the cause of all this, cannot be studied in the laboratory. Appendix H discusses the inadequacy of antenna theory; we do not know enough about antenna behavior to simulate antennas in the laboratory.

Understanding the role of the source antenna will require observation of antennas on a suitable antenna range. We know of no existing data that would suffice because there has been no reason to look for the phenomena that are important in this case. Indeed, to the extent

that we are here concerned with transient events close to or in the nulls in remote side lobes of the source, the needed data are the sort that are ordinarily avoided in most work on a range.

Two approaches to such experimentation come to mind; both have advantages, and a thorough program might pursue both. In one, the experiments could study the behavior of specially constructed antennas--preferably having smooth simple geometry--with the aim of obtaining better understanding of how the antenna structure influences its transient response. This, if successful, would have predictive value inasmuch as we might be able to infer the behavior of different kinds of antennas. Such results might also be helpful by offering a bridge leading to better theoretical descriptions of antennas. In the other, the approach would be to make measurements on antennas of practical interest. It would probably offer less predictive insights but give much more definitive data on problems of immediate concern. The two approaches would be mutually helpful.

In either of these approaches, the data could be taken either as amplitude and phase versus frequency or as transient waveforms (preferably in response to an impulse-like excitation) at a variety of angular positions. A number of experimental problems arise in either case because it would be necessary to achieve good dynamic range, well above noise and extraneous reflections, even though the measurement is made in weak-signal directions. The difficulties of avoiding multipath effects and reflections from the main lobe are obvious.

It is not necessary to investigate these matters in a complete TOA system. A study of how the phenomena vary with change in position of a single receiver would suffice and would avoid the many extraneous and confusing effects that arise among a multiplicity of receivers. If the source antenna under study on the range were one of practical interest, then it would be well to use a counterpart receiver; one could thus study the variation of receiver response, and thence of arrival time, with angle. It would be better, when studying special simplified antennas, to use a very wide band receiver such as a crystal video.

II-8. MITIGATION

If experiment bears out the results of this analysis, and if the effects are sufficient in magnitude, it will be appropriate to consider steps to reduce the severity of the effects. It does not appear likely that the effects can be removed entirely unless the source can be made omnidirectional. (Even then multipath effects, if they arise, could lead to some of the same phenomena.)

This study indicates that the system errors in a TOA system scale in proportion to scale length S. This length cannot be made equal to zero in real designs, but it does vary with design precepts and with numerical choices. S probably is made larger as steeper skirt selectivity is imposed. One ordinarily views steeper selectivity as good, and requirements for steepness are often bounded largely by economic limits or manufacturing tolerance. If, as it now appears, steepness has harmful effects, it might be worthwhile to reconsider the requirements placed on receivers.

Virtually nothing is known concerning how S varies with the design style of the filter circuits (e.g., Butterworth, Tschebychev, etc.), or even with the number of poles in the network.* Even less is known about such items as the use of discrete finite transversal filters. Indeed, there appears to be here a whole realm of circuit theory that has scarcely been entered.

It would be useful to examine these matters experimentally. Such work, if done, should be accompanied by a study of possible definitions of what constitutes arrival. Advantageous definitions might be related to design precepts for the filters. It may be possible to design usable receivers with small scale lengths.

---

*The value of S for various bandpass filter designs can undoubtedly be judged with adequate accuracy from the impulse and step responses of the lowpass equivalent circuits (provided that those equivalents exist). However, the interference process that governs the arrival time everywhere except in pattern nulls and on axis cannot be judged from these lowpass responses.

Any circuit necessarily has an impulse response and a stepped-carrier response and these must differ. It is hard to imagine that there is much more that can be done in the design of the receiver except to see if S can be reduced. However, it is probably possible to decrease the exacerbations that are added by phase if one is willing to use what amounts, virtually, to two or more receivers in parallel. Two parallel circuits that use quadrature samples from the same local oscillator, two matched IF strips, and two envelope detectors can reduce the effect of phase. That approach can be generalized to a three-phase receiver wherein enough data would be available to solve for most of the phase parameters. Whether such hardware proliferation would be worthwhile is not clear. It must be recalled that such measures would only serve to reduce the effects of phase; the basic problems would remain, as is evident from Fig. II-1 where phase effects have been removed by perfect averaging.

The most promising avenues whereby these effects might be mitigated are probably to be found in operational schemes that might be applied to the system as a whole. For example, if the lobe pattern of the source were known, various steps could probably be taken to discard arrival time data obtained in or near pattern nulls. In practice that might not be at all easy to accomplish because of limited time to study the source--especially if the source fails to rotate in a helpful way. It should be noted that the amplitude of the "extra" signal is not especially feeble, even though the signal is brief. The amplitude of the transient response of the receiver certainly does not go to zero in pattern nulls, and the output amplitudes of troublesome responses are not dependably lower than those of innocuous responses. It appears that amplitude is a highly doubtful parameter to depend on for the elimination of bad data.

Much the same can be said concerning the waveform of the receiver output. It is true that the impulse response found in the nulls of the source is unique, and could be discriminated against. However, an examination of the output waveforms in Part III shows that many of the troublesome waveforms simply seem to arrive early and do not differ greatly in shape from the axial response. No doubt a matched detector,

matched to the receiver response on the source axis, is the theoretical optimum detector. However, it is not clear that such a detector would help very much (as compared with the 1/2 peak logic), especially if one recognizes a need to use a limited dynamic range of the receiver output and to be tolerant of some noise and multipath.

Much can be said about averaging, and there is no doubt that appropriate averages can reduce the severity. For example, it could be true that the center of gravity of all calculated source locations over a full 360 degree rotation of the source lies on the true source location. This sounds plausible but might not be true; we do not know enough about the signal in the rear hemisphere of the source. In any case, such an opportunity to observe 360 degrees of source rotation may not arise. The partial improvement associated with averaging over phase in each receiver has been mentioned. So too has the considerable doubt that this could be done in practice by averaging many pulses (a three-phase receiver would avoid these problems and could remove most phase effects on each single pulse).

Partial averages taken over some number of calculated source locations may be helpful, but such schemes are certainly not panaceas. At the very least it is necessary to be cautious as to just what numbers are to be averaged. For example, averaging a set of arrival times in each receiver is not equivalent to averaging a set of calculated source locations. In many such schemes it might be even more worthwhile to devise rules for casting out a subset of the data instead of averaging the whole set. There is limitless opportunity to devise schemes to manipulate data in such ways as these.

Finally, it is generally thought best for the receiver array to subtend as large an angle as possible at the source. In this analysis it appears that larger deviations of arrival time (or higher probability of a given deviation) occur at wider angles off axis. This seems to be particularly so when the receiver front end is tuned, suggesting that large subtense may not be best, and that there may be some optimum subtense that is less than is presently perceived.

II-9. CONCLUSIONS

The study reported here was entirely analytic and considered simplified idealizations of transmitters, receivers, and systems that would be free of error were it not for the particular antenna effect of interest. The conclusions that the study reached require experimental confirmation that is not known to exist in available reports. Although the analysis is as exact and complete as theory permits, the numerical results are only approximate because the true behavior of directional antennas cannot be calculated with the necessary accuracy. Moreover, the behavior to be expected of practical hardware will differ in detail, perhaps considerably, from the behavior of the simpler devices that were analyzed. No rules of thumb are known that might predict the behavior to be expected of such hardware, and it would be impractical to calculate such behavior even in those cases for which it might theoretically be possible. There are reasons to believe that the numerical results given here may underestimate the severity of the effects. If so, the amount of underestimation is unknown.

Subject to these qualifications, the study results support the following conclusions:

1. A pulsed signal from a directional emitter excites, in a receiver that is not on the emitter axis, the impulse response (very nearly) of the receiver as well as the response that is ordinarily expected. This impulsive excitation has not previously been recognized to occur.
2. The output waveform from such a receiver, when in an off-axis location, results from the interference within the receiver between two characteristic transient responses. Such interference can lead to output waveforms that differ appreciably from those that are usually expected.
3. The waveform changes that occur can have adverse influence on the performance of electronic systems, especially if the system is designed to use waveform details. Among electronic

systems of present interest, leading-edge TOA systems appear to be the most vulnerable to disturbance from this antenna effect.

4. The analysis, together with computer simulation of a simple TOA system, indicates that such a system may make radial errors up to 50 meters or so in the calculated location of the emitter; radial errors of the order of 10 meters or more may occur for 20 percent or so of the received pulses.
5. The numerical details such as those cited above depend on a great many details and numerical values, and should be regarded as illustrative but not definitive.
6. The severity of the effects on a TOA system are influenced significantly by the design of the TOA receivers as well as by the operation of the whole system. The likelihood that a given system will incur any given magnitude of error increases with the directionality of the emitter.
7. Experimental confirmation of these effects, together with measurements of their true magnitudes, would require careful work on a good antenna range. The information that is needed cannot be obtained on a laboratory bench or by calculation.
8. It may be possible to mitigate (but not eliminate) the severity of the antenna effect on a TOA system by various methods, notably by suitable operational procedures.

In addition to the foregoing, the detailed analysis treats the exact transient response of a heterodyne receiver to a pulsed signal. This has not, to our knowledge, been reported previously. The analysis should be of interest to circuit designers and analysts. Some aspects of the analysis itself can be regarded as subsidiary conclusions.

PART III

DETAILED ANALYSIS

III-1. INTRODUCTION TO THE ANALYSIS

This analysis was undertaken to learn how a typical heterodyne receiver would respond to the off-axis pulse waveforms discussed in R-1819-PR, and to estimate how the "little-known" antenna effect would influence the accuracy of a leading-edge TOA system. (Such TOA systems are only an illustrative example; similar consequences might appear in future electronic systems if their performance relies on details of signal waveform.)

It was not foreseen at the outset that the analysis would be especially difficult. However, the usual approximate methods were found to lead to mathematical difficulties. To overlook these or to patch in an ad hoc remedy would be equivalent to inserting a new transient artifact at the leading edge of the signal. Such an artifact would cast doubt on all the subsequent results and would vitiate the entire study. On several occasions it became necessary to back up and rework the analysis to eliminate the cause of the trouble. Ultimately it became necessary to go all the way back to the initial source waveform; the quite conventional waveform adopted in R-1819-PR was found to embody an approximation that did no harm there but that cannot be used here. The radiated signal from the source should not be described as an amplitude envelope multiplying a fixed carrier because such representation would imply a finite output at zero frequency. Consequently, this is not an analysis of the waveforms discussed in R-1819-PR.

The usual approach to the transient analysis of bandpass circuits is to treat the response of an equivalent lowpass circuit to the envelope waveform--a method that is approximate but usually adequate because most such analyses are concerned with quasi-steady-state conditions rather than with the details of transients (see Appendix D). Once deprived of the traditional amplitude envelope description, the analysis became committed to an exact treatment of the transient response of the entire bandpass system of source and receiver.

This analysis does not break any new ground in terms of the analytic method used. In every instance the response of a bandpass circuit

is obtained by convolution of the input signal with the bandpass impulse response. So far as we know, the impulse response of a four-pole Butterworth bandpass filter is reported here for the first time; determination of that impulse response was a major task during the study. Moreover, the study examines the progress of a transient waveform through a considerable number of frequency-selective stages, and several very interesting phenomena turn up.

The analysis is lengthy, and it is unlikely that many readers will check all the steps in detail. The work has been subjected to a number of tests for self-consistence (for example, it was shown that convolution of the Butterworth impulse response with an arbitrary steady sine wave yields the correct transfer function), and has been reviewed by others. Moreover, the results appear to "make sense" physically. They are believed to be correct.

Five appendixes review briefly the use of convolution with the impulse response to obtain the transient response, and present the step and impulse responses of various single-stage bandpass and lowpass circuits. The use of the lowpass equivalent to approximate the bandpass response is sketched and the shortcomings are mentioned. One appendix discusses the design of Butterworth bandpass filters and their lowpass equivalents. These appendixes are intentionally inelegant and are not rigorous or complete; their purpose is only to provide a brief review and to set forth various formulas used in the analysis. No mention is made of more powerful and elegant methods such as the Laplace transform.

Now that the analysis is complete and the results are shown not to stem from analytic artifacts, it is evident that comparable future studies could be carried out more easily by judicious use of some approximations. For example, there is no doubt that scale length S can be estimated quite accurately from the step and impulse responses of the lowpass equivalent circuit (*if* that equivalent exists; it often does not). On the other hand, the mutual interference between the bandpass impulse response and the stepped-carrier response cannot be described by the lowpass responses because phase details are essential to that interference. The sum frequency terms at the heterodyne output

can be discarded from the input to the IF strip if the analyst does not consider the very early portion of the transient response. It is unsafe to judge this possibility wholly from the time constant of the sum frequency exponential; characteristic time constants of the other circuits, including the envelope detector, must be considered as well.

The transmitter and receiver that are analyzed are highly simplified and idealized; a great many features present in real devices are omitted entirely or represented by less elaborate circuits. All of the usual analytic assumptions concerning component ideality and linearity are made. All tuned circuits are assumed to be separated from their surroundings by ideal isolation amplifiers; there is no coupling between any of the normal modes of the system.

No consideration is given to noise or to other interfering signals (but the need for receiver selectivity and the impossibility of observing infinitesimal signals are recognized). No fluctuations of the index of refraction are considered, nor is multipath propagation. It is supposed that the clocks are perfect and that time measurements can be made with perfect precision and accuracy. It is assumed that the receivers are stationary or moving very slowly (no Doppler or first-order relativity effects) and that their relative locations are known perfectly in a TOA system. It is intended that the system would be entirely free of error were it not for the effects considered here. In reality, many of these troublesome factors will create errors and difficulties. It would be rash to suppose that they will not interact with the phenomena considered here; synergism is possible and the various consequences need not combine as if they were independent.

III-2. SYMBOLS, NUMERICAL VALUES, AND THE LIKE

Absolute signal level is of no direct concern here, in the absence of signal/noise considerations. The power radiated by the source and the source-receiver distance are unspecified. However, the receiver is unquestionably in the far field of the source (use of the sinZ/Z

description of the source directivity pattern is a tacit assumption that the distance is infinite).

At each step, as the signal progresses through the several sections of the source/receiver system, the signal amplitude is renormalized. In every case, the amplitude of that signal approaches unity as time increases when the receiver is situated on the source axis; elsewhere the steady-state amplitude is less as a consequence of the source directionality. Where dB are used they are defined by

$$dB = 10 \log_{10} \left[(signal\ amplitude)^2\right]$$

Thus, in every case, 0 dB corresponds to the steady-state level on the source axis. Renormalization of the amplitude can be regarded as an assumption that any desired gain is available and that the isolation amplifiers are adjusted to deliver this signal level.

In some of the appendixes, units of volts or volt-seconds are assigned to some quantities for clarity. The choice is arbitrary and microvolts would do as well; moreover, current units might have been chosen. It is tacitly assumed that such voltages are observed by an infinite impedance voltmeter. Similarly, the reader is free to suppose that all the waveforms under discussion are expressed in volts, or microvolts, as he may choose.

Treatments of transient behavior commonly use the symbol u(t) to represent a unit step at time t. That usage is followed in some of the appendixes, but is generally omitted because it clutters the page. The reader should bear in mind that almost all the time functions encountered here have such a unit step, not necessarily at t = 0. The presence of the step coefficient should be evident from the discussion and, in many instances, from the limits of integration. Need for the symbol $\delta(t)$ to represent a unit impulse delivered at time t occurs less often, principally in the appendixes. This symbol should not be confused with phase angles δ_0, δ_2, and δ_4.

Throughout the symbol ω is used to represent angular frequency in radians per second. When not subscripted, this represents an arbitrary

independent variable; subscripts 0, 1, 2, 4 are used for specific fixed constant values associated with tuned circuits. In the appendixes, subscripts A, B, and R are similarly used for arbitrary fixed values that should not be confused with the numerical subscripts. In the text, the word "frequency" is generally used for brevity when angular frequency is intended. The usage is clarified in the few instances where confusion might arise.

Numerous quantities such as phase angles are defined in terms of the sides and angles of right triangles. The lengths of the sides shown in the sketch do not reflect true proportions, and the lengths of some sides may be negative. Care must be exercised regarding the quadrants of trigonometric functions and the inverse functions.

The double bracket is used for brevity to denote an expression to be evaluated at limits of integration:

$$\left[\!\left[f(x) \right]\!\right]_A^B \equiv f(B) - f(A)$$

Graphs showing waveforms are marked in time intervals of periods of the source carrier frequency, or of periods of the IF center frequency, as appropriate. In addition, many are also marked in equivalent distances in meters. The distance representation may be easier to grasp physically than the times (which are generally only a few nanoseconds). Moreover, the distance representation is, in many instances, invariant under changes of the center frequencies (or nearly so) whereas times expressed as periods are not.

In keeping with the foregoing, arrival times and the like are commonly discussed in meters rather than nanoseconds; the units are interchangeable in every such case through the speed of light, which is here taken to be 3×10^8 meters/second.

The analysis itself is general as regards the numerical values of the various parameters. Thus, the angular frequency of the source carrier is ω_0 and the angular center frequency of the receiver IF strip is ω_1, and no particular specification of their values or of the ratio

is made. Similarly, the bandwidths of various circuits are specified by subscripted Q and the value of the Q is not restricted. (However, Q is usually restricted by $Q > 1/2$ because critically damped and overdamped circuits require a different mathematical form.) In every instance, the quantity Q is defined with respect to an associated frequency, and the analysis is specific to that particular choice of frequency; those choices are explained where they are made. In this respect, the analysis is not general; other choices of tuning of two circuits (those involving Q_0 and Q_3) would alter the mathematical expressions themselves, not merely the numerical values.

To foster the view of interchangeability of time and distance, and also for brevity in writing cumbersome expressions, symbols U and V are used through much of the analysis. They are defined:

$$U \equiv \omega_0 t$$

$$V \equiv \omega_1 t$$

where t represents time in seconds and ω_0 and ω_1 are the RF and IF center frequencies. A change of 2π in the numerical value of U can be understood to represent a time change of one RF period or a distance of one RF wavelength. Such a change in V indicates a time interval of one IF period or a distance of one IF wavelength. (There is, of course, no propagating signal at the IF.)

All numerical examples and all graphs shown here employ the following numerical values:

$$\omega_0 = (2\pi)(3.2) \times 10^9 \text{ radians/second}$$

$$\omega_1 = (2\pi)(80) \times 10^6 \text{ radians/second}$$

In other words, the source carrier frequency is 3.2 GHz and the receiver IF center frequency is 80 MHz. Consequently,

$$U = 40V$$

Further,

$$Q_0 = \sqrt{(3\pi)^2 + \tfrac{1}{4}} = 9.438$$

$$Q_1 = 8$$

Q_0 specifies the effective bandwidth of the source to be about 10.5 percent. Q_1 specifies the bandwidth of the IF strip to be 10 MHz. (These are full bandwidths at half power.)

The quantity Q_3, when used, specifies the bandwidth of the RF portion of the receiver ahead of the heterodyne detector. Most numerical values represent either $Q_3 = 10$ or Q_3 nonexistent.

The quantities Q_2 and Q_4 pertain to the tuned circuits of the IF strip and are defined by the choice of the other parameters together with the choice of the Butterworth design.

The quantity Z is defined

$$Z = \pi \frac{L}{\lambda} \sin\theta = \frac{\omega_0 L}{2c} \sin\theta$$

where L = width of the source antenna

λ = plane wave wavelength at the RF frequency

θ = angle measured away from the boresight axis of the source. It is unfortunate that this variable, which is a surrogate for angle and is prominent throughout the discussion, does not have a recognized name. It is simply called Z in the text and is often used in lieu of "angle θ."

For the simple source considered here, the source directivity pattern is taken to be sinZ/Z. Pattern nulls occur at all integer values of Z/π (except $Z = 0$, which is boresight) and the quantity Z/π is most useful in the discussion. It does not seem worthwhile to define a new symbol to represent Z/π.

Finally, it was advisable to maintain a wide dynamic range in the numerical work. The equations contain numerous terms of comparable

magnitudes but different signs. Roundoff and truncation errors can reduce seriously the nominal accuracy of computer arithmetic, but it is often very difficult to estimate the severity of the errors. All numerical values and graphs shown here were obtained with IBM FORTRAN double-precision arithmetic, and all input constants were expressed to 16 significant decimal digits. Numerous checks indicate that the dynamic range of the results is considerably more than 120 dB; chances are that it exceeds 200 dB, but that is hard to prove. Future work might be done with single-precision arithmetic, but the analyst should watch for symptoms of roundoff and truncation error.

III-3. SOURCE SIGNAL ON AXIS; F_0

In R-1819-PR, it was assumed that the beginning of the emitted pulse, when observed in the far field on the source axis, could be described by

$$F(t) = \left|1 - e^{-K\omega_0 t}\right| \sin(\omega_0 t + \psi);\ t \geq 0$$

Numerical examples in that report used $K = 1/6\pi$ to represent a fairly typical rise time in an ordinary pulsed source. Here ψ is an arbitrary phase angle that need not be, and often is not, coherently locked to the pulse repetition rate. It is not unusual for ψ to vary from pulse to pulse. In some sources the pulsed circuit tends to favor excitation with some particular phase, but even in these the phase is loosely constrained and tends to vary over some range. Although ψ can be controlled, and sometimes is, it is unrealistic to limit this analysis to a particular value of the phase. ψ will appear throughout this report as an undetermined quantity that is regarded as having a constant value during any one pulse but that can vary from pulse to pulse.

Consider now the quantity G(U), the integral of F(t) over time $0 \geq t \geq U/\omega_0$:

$$G(U) = \frac{1}{\omega_0} \int_0^U \left| 1-e^{-Ku} \right| \sin(u+\psi)du$$

$$\text{As} \quad U \to \infty, \; G(U) \to \frac{1}{\omega_0} \left[\frac{K^2\cos\psi - K\sin\psi}{1 + K^2} - \cos(U+\psi) \right]$$

The term in cosU has zero average value, and the average of G(U) or, equivalently, of F(t) approaches

$$\frac{K}{\omega_0(1+K^2)} \left| K\cos\psi - \sin\psi \right|$$

This quantity is equal to zero only for two values of ψ given by

$$\tan\psi = K$$

Thus, waveform F(t) above will have a non-zero average value unless ψ is restricted to one of these two special values, contrary to the remarks above that ψ usually varies.

To deliver a waveform whose long-term average value is non-zero, a network must possess a non-zero response at zero frequency. No radio antenna can possess such a response at zero frequency in the radiation field, so F(t) cannot represent a far-field radiated waveform.* It will be seen that networks such as bandpass filters also cannot deliver waveforms of this type.

The envelope above is precisely the step response of a lowpass RC filter whose RC time constant is $1/(K\omega_0)$ (see Appendix C). That filter is, in turn, the lowpass equivalent of a bandpass filter whose center

---

*It cannot be thought that the source can "make it up" at the other end of the pulse. The antenna and radiation field can hardly be thought to await being informed of when the pulse will end. Rather, it is seen below, the waveform undergoes small local phase modulation to avoid accumulation of an average value.

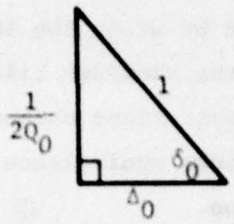
frequency is ω_0 and whose Q is 1/(2K) (see Appendix D). Thus waveform F(t) is the result obtained by using the lowpass equivalent method to approximate the output of the bandpass filter with input $u(0)\sin(\omega_0 t+\psi)$. Many of the envelope representations encountered in practice reflect a similar application of lowpass equivalence to treat bandpass behavior, but this is an approximation.

To preserve as nearly as possible the simple functional form of F(t) while removing the zero-frequency difficulty, it is only necessary to put $u(0)\sin(\omega_0 t+\psi)$ through a bandpass filter. That is easy enough--only a simple convolution is needed--but in the present instance an inconvenient difficulty arises, and we will adopt a stratagem to circumvent it.

The "ringing" frequency of a simple bandpass stage that is tuned to ω_0 is not equal to ω_0. Instead, the tuned circuit "rings" at its own lower natural frequency (see Appendix B). Thus, if the bandpass filter is centered on ω_0, the output response to the stepped carrier will contain two terms running at two different frequencies: the steady-state term driven by the carrier at ω_0, and a transient term running at the filter's own normal mode frequency.

In this instance, it is burdensome to accept at the outset terms running at two different frequencies. Many terms are going to arise even if this is avoided, and the analytic burden would be excessive with an extra frequency. We will adopt the simple stratagem of tuning this bandpass filter very slightly higher than the carrier to make the normal frequency equal to the carrier frequency. The frequency offset is small and inconsequential for reasonable values of Q, and the stratagem does not impose any important limitation on the interpretation of the analysis. It does, however, limit the generality of the equations; in a more general treatment the tuned frequency of the bandpass circuit would be an independent parameter.

Let the Q of this bandpass filter be Q_0, and let Δ_0 and δ_0 be defined (see Appendix B):



We choose to tune the center frequency of the filter to ω_0/Δ_0; the impulse response of the filter is then

$$h(t) = \frac{\omega_0}{\Delta_0^2 Q_0} e^{-\frac{\omega_0 t}{2\Delta_0 Q_0}} \cos(\omega_0 t + \delta_0)$$

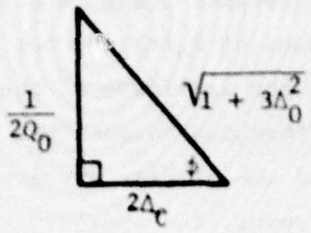
The small frequency offset used to obtain ω_0 in the cosine means that the filter exhibits a small insertion loss at frequency ω_0; an input carrier of unit amplitude will not emerge with unit steady-state amplitude. To compensate and maintain normalization, the input signal is taken to be

$$u(0) \frac{\sqrt{1 + 3\Delta_0^2}}{2\Delta_0} \sin(\omega_0 t + \psi)$$

Convolution of this input with the impulse response of the filter yields the waveform

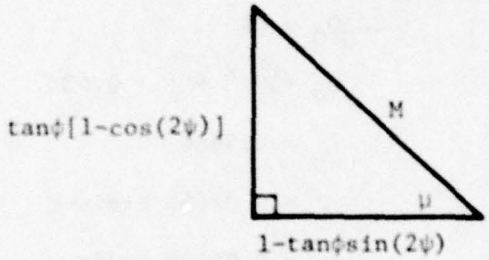
$$F_0(t) = \sin(U+\psi+\phi) - \frac{e^{-\frac{U}{2\Delta_0 Q_0}}}{\Delta_0} \left[\frac{\sqrt{1 + 3\Delta_0^2}}{2\Delta_0} \sin(U+\psi+\delta_0) - \frac{\cos(U-\psi+\delta_0-\phi)}{4\Delta_0 Q_0} \right]$$

where, as defined previously, $U = \omega_0 t$ and where



$$2\tan\phi = \tan\delta_0$$

Inasmuch as ψ is regarded as a constant during any one pulse, even though it may vary from pulse to pulse, it is convenient to define quantities M and μ:



With these, F_0 can be expressed more compactly:

$$F_0(t) = \sin(U+\psi+\phi) - \frac{Me}{\Delta_0}^{-\frac{U}{2\Delta_0 Q_0}} \sin(U+\psi+\delta_0+\mu-\phi)$$

The average value of F_0 is zero for all choices of ψ.

The difference between this waveform and the envelope form, F(t), reflects a characteristic distinction between the lowpass equivalent and the exact bandpass results, and deserves brief comment. Both forms contain two terms describing the steady-state and transient portions of the output. In the lowpass equivalent form both terms have the same phase, ψ, as well as the same frequency, and the amplitudes of the terms are independent of the phase. The zero crossings of F(t) are evenly spaced. In $F_0(t)$ two counterpart terms occur, but they run at the same frequency only because the stratagem arranged for that. Even so, the two have different phases. In view of the varying amplitude

of the transient term, the phase difference leads to a shifting phase of $F_0(t)$ with time; the zero crossings of $F_0(t)$ are not evenly spaced. Moreover, M is a function of ψ, and the amplitude of the transient term depends on ψ. These small differences between the two waveforms are ordinarily of no consequence and the simpler envelope form is usually an adequate approximation. However, the approximation overlooks phase effects that may be significant in transient circumstances.*

Clearly, the quantity $1/(2\Delta_0 Q_0)$ in $F_0(t)$ plays the role of K in the envelope representation F(t). In R-1819-PR, the value $K = 1/6\pi$ was used in numerical examples. Because that is a reasonable value to attribute to real pulsed emitters, that value is adopted here too:

$$2\Delta_0 Q_0 = 6\pi$$

$$Q_0 = \sqrt{9\pi^2 + \tfrac{1}{4}} = 9.438$$

$$\Delta_0 = 0.9986$$

$$\phi = 0.02652 \text{ radians}$$

$$\delta_0 = 0.05300 \text{ radians}$$

The tuned frequency of the bandpass filter is only about 0.1 percent above ω_0--a trivial offset. This choice of Q_0 attributes to the source a half-power bandwidth of 10.6 percent.

The waveform $F_0(t)$ is taken to be the pulse waveform found on the source axis in the far field, and is the starting point for the rest of this analysis.†

---

*If one forms the square root of the sum of two quadrature versions of F_0--that is, for ψ and $\psi+\pi/2$--the resulting function does not depend on the choice of ψ and consequently is the envelope of F_0. In that formal sense, F_0 does possess an envelope. However, F_0 cannot be described by an amplitude function multiplying a fixed carrier; the carrier must be phase-modulated. Thus F_0 does not possess an envelope as that term is commonly meant.

†The question of the impulse response of the entire antenna (as distinct from the impulse response of the aperture) is discussed in Appendix G. It is remarked there that the choice of a one-stage bandpass filter in the source is unsatisfactory. This choice leads to an antenna impulse response that commences with a step on the axis, but

III-4. SOURCE SIGNAL OFF AXIS; $F_{1,2}$

As in R-1819-PR, we adopt the Huygens-Fresnel-Kirchoff model of the source antenna, wherein the source is regarded as a hole in an infinite opaque screen, and we adopt the approximate Kirchoff boundary conditions (see Appendix F). To avoid intractable mathematics we consider a uniformly illuminated rectangular aperture (or line) of length L. Other aperture shapes and distributions of illumination lead to results that differ in detail but not in major features. We consider only field points in the principal plane, parallel to edge L and containing the normal to the center of the aperture.

Insofar as signals in the far field are concerned, it is assumed that each incremental strip of the source aperture of width $d\ell$ is excited by illumination

$$F_0\left(t + \frac{R}{c}\right)\frac{d\ell}{L}$$

where R is the distance from the center of the aperture to the field point. All field points under consideration will be supposed to be at the same distance R from the center of the source (or, if different, that the difference has been allowed for in the clock time and signal amplitude). Thus, the clock setting is such that the initial signal arriving from the center of the source arrives at t = 0.

It is supposed that distance R is so great that all points in the aperture may be regarded as equidistant from the field point on the axis. Then the contributions from all the incremental source aperture regions arrive with the same transit time, and the far-field axial waveform is $F_0(t)$.

---

that commences at zero everywhere else--a physically implausible situation. This formal shortcoming would be removed by assuming a more complicated bandpass circuit in the source--one whose impulse response starts at zero. However, the formal improvement does not seem to be sufficient to justify the added analytic burden of treating a more complicated filter. The shortcoming is thought not to affect adversely the results obtained here.

However, at angle θ from the axis the signal increment from the near edge of the aperture travels a shorter distance and starts earlier at time

$$t \approx -\frac{L\sin\theta}{2c}$$

whereas the signal increment from the more remote edge arrives late and does not commence until

$$t \approx +\frac{L\sin\theta}{2c}$$

At the field point there is a time interval

$$-\frac{Z}{\omega_0} \leq t < \frac{Z}{\omega_0}$$

where $Z = \frac{\pi L}{\lambda}\sin\theta$, during which successive incremental portions of the aperture begin to contribute to the net signal. It is not until time Z/ω_0 that the entire aperture is contributing; at that moment the contribution from the far edge commences, and that contribution must still undergo its rise, so starting transient conditions continue even after $t = Z/\omega_0$.

The incremental contributions must be summed with due regard for the differing transit time delays that each has experienced. During the early interval the resultant waveform is given by

$$F_1(t) = \frac{1}{L}\int_{\frac{-ct}{\sin\theta}}^{L/2} F_0\left[\omega_0\left(t + \frac{\ell\sin\theta}{c}\right)\right] d\ell; \quad -\frac{Z}{\omega_0} \leq t \leq \frac{Z}{\omega_0}$$

and during the subsequent period, when the entire aperture is contributing, the resultant is given by

$$F_2(t) = \frac{1}{L}\int_{-L/2}^{+L/2} F_0\left[\omega_0\left(t + \frac{\ell\sin\theta}{c}\right)\right] d\ell;\ t \geq \frac{Z}{\omega_0}$$

The two integrals differ only in their limits, reflecting the time-dependent span of the aperture that participates during F_1.

Upon substituting

$$z = \omega_0\left(t + \frac{\ell\sin\theta}{c}\right)$$

these integrals become

$$F_1 = \frac{1}{2Z}\int_{0}^{\omega_0 t+Z} F_0(z)\,dz$$

$$F_2 = \frac{1}{2Z}\int_{\omega_0 t-Z}^{\omega_0 t+Z} F_0(z)\,dz$$

They are simple integrals; when evaluated at their limits they yield

$$F_1 = \frac{1}{2Z}\left[-\cos(U+\psi+Z+\phi) + Me^{-\frac{(U+Z)}{2\Delta_0 Q_0}}\cos(U+\psi+Z+\mu-\phi)\right];\ -Z \leq U \leq Z$$

$$F_2 = \frac{\sin Z}{Z}\sin(U+\psi+\phi)$$

$$+ \frac{M}{2Z}\left[e^{-\frac{(U+Z)}{2\Delta_0 Q_0}}\cos(U+\psi+Z+\mu-\phi) - e^{-\frac{(U-Z)}{2\Delta_0 Q_0}}\cos(U+\psi-Z+\mu-\phi)\right];\ U \geq Z$$

The waveform off axis can be regarded as two successive segments, F_1 and F_2. Note, however, that when $U = Z$,

$$F_1 = F_2$$

$$\dot{F}_1 = \dot{F}_2$$

Thus, the two segments are smoothly joined. It is a convenience of mathematical representation to describe them as separate segments, and the distinction is helpful in discussion, but there is only one waveform present. That waveform, taken as a whole, is called $F_{1,2}$ hereafter. The "extra" signal discussed earlier is taken up in the following section; it is not equal to F_1.

Note that as

$$t \to \infty \ , \ F_2 \to \frac{\sin Z}{Z} \sin(U+\phi+\phi)$$

This is the familiar result since sinZ/Z is the amplitude directivity pattern that is customarily attributed to this source. The pattern is the steady-state manifestation of the same geometry that, in transient circumstances, gives rise to the more complicated behavior seen in F_1 and F_2.

It is important to notice that sinZ/Z appears only in the steady-state term. The whole precursor F_1 and the transient terms of F_2 carry the coefficient 1/2Z and show no evidence of the lobe structure. In nulls of the pattern, when sinZ = 0, the signal described by these other terms continues to appear; the null is not a null for transients. Moreover, although the duration of this null signal is generally brief, the peak amplitude is not feeble--it is comparable with the steady-state amplitude at the top of the next side lobe.

The precursor is most prominent in pattern nulls, but F_1 is present at *all* off-axis locations. The duration of F_1 is $2Z/\omega_0$, which increases with off-axis angle. At 90 degrees from the axis, the F_1 segment lasts for L/λ periods of the carrier frequency. (This assumes, of course, the highly doubtful validity of the Kirchoff approximate boundary conditions. It is more probable that the impulse response of the aperture

has a long oscillatory tail, in which case the duration of F_1 will be longer; see Appendixes F and G.)

The appearance of $F_{1,2}$ at various angular locations is shown in Figs. III-1 and III-2. In these figures, the function $Z \cdot F_{1,2}$ is plotted; thus, the figures show the waveform as it would appear on an oscilloscope with the vertical gain adjusted to remove the amplitude shift caused by Z. The first figure shows the axial waveform F_0 and the waveforms in three other locations not far from the axis. The second figure shows the waveforms at wider angles; values of Z/π of order 50 denote angles near 90 degrees from the axis of a highly directional source.

The appearance of waveform $F_{1,2}$ depends, of course, on the choice of phase angle ψ. However, that dependence is not prominent and is not worth showing in several figures. All of the curves shown in these figures are for $\psi = 0$ (for which case $M = 1$ and $\mu = 0$).

When Z is small, the duration of the signal in the pattern nulls is so short that it does not build up to much amplitude. At larger values of Z, the F_1 amplitude is seen to build up to a quasi-steady value. After the transition from F_1 to F_2 the amplitude changes smoothly to the steady-state value associated with the directivity pattern.

All the curves are aligned vertically, and $t = 0$ is marked by a vertical line. The signal is seen to commence at increasingly negative time as Z becomes larger. To avoid confusion, the transition from F_1 to F_2 is not marked; that transition occurs as far to the right of $t = 0$ as the start is to the left of $t = 0$. The transition is, in every case, smooth. Also to avoid confusion, the horizontal axis is not marked off in units of time or distance, but cycles of the carrier frequency can be counted. If the carrier frequency is taken to be 3.2 GHz, then one period corresponds to 9.375 centimeters and the entire length of the horizontal axis shown in these figures is 8.4375 meters. Note, however, that the appearance of these curves does not depend on the numerical value of ω_0.

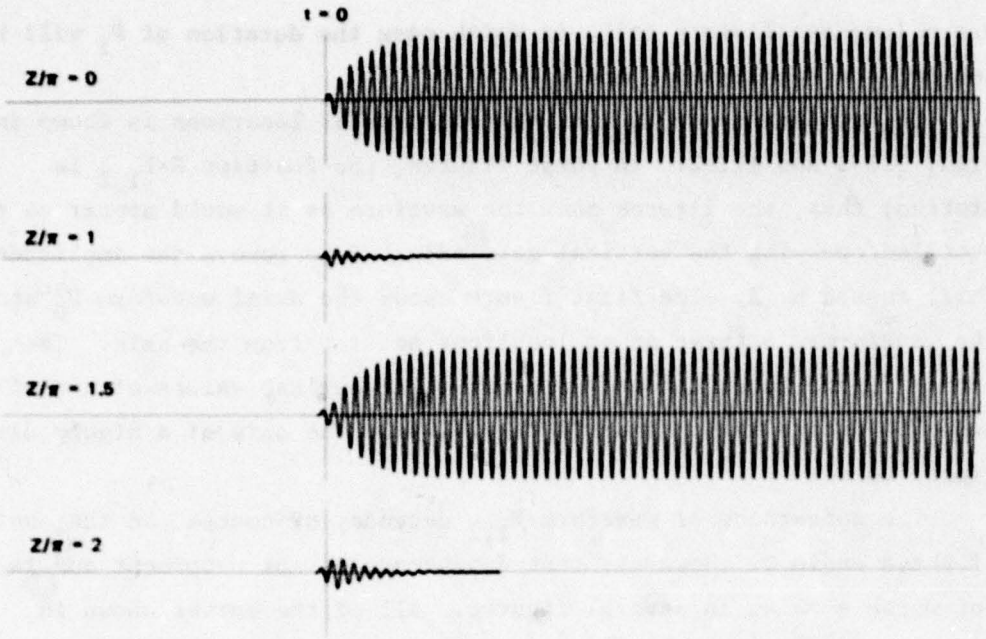


Fig. III-1 — Radiated waveforms on the source axis and in nearby locations. Note the signal in pattern nulls.

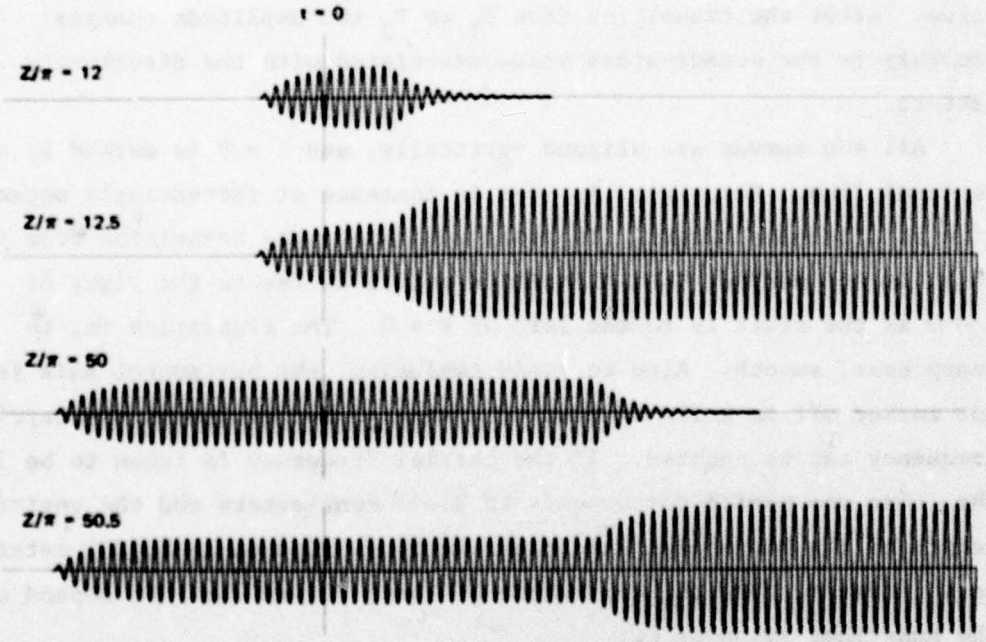


Fig. III-2 — Radiated waveforms at intermediate and large angles from the axis. Note progressively earlier beginning and increasing duration of segment F_1.

III-5. $F_{1,2}$ REGARDED AS A SUM

In general, discussions of the far-field signal suppose that the signal is

$$\frac{\sin Z}{Z} F_0$$

That is, that the off-axis signal has exactly the same form as the axial signal and is merely that signal at an amplitude that reflects the directivity pattern. We may say that this is the signal that is generally "expected."

It is useful to regard the correct signal, $F_{1,2}$, as the sum of this "expected" signal and an unexpected "extra" signal that occurs because of the antenna effect. No analytic use will be made here of this decomposition of $F_{1,2}$, but the decomposition is helpful in obtaining a conceptual grasp of the receiver response. We define the "extra" signal to be equal to

$$F_{1,2} - \frac{\sin Z}{Z} F_0$$

The regions over which F_0, F_1, and F_2 are defined dictate that three separate segments of the "extra" signal be recognized.

When $-\frac{Z}{\omega_0} \leq t \leq 0$, F_0 has not yet begun and the "extra" signal is equal to F_1.

When $0 \leq t \leq \frac{Z}{\omega_0}$, F_0 has begun and F_1 continues, so the "extra" signal is

$$F_1 - \frac{\sin Z}{Z} F_0$$

When $t \geq \frac{Z}{\omega_0}$, F_0 continues, F_1 has ended, and F_2 has begun, so the "extra" signal is

$$F_2 - \frac{\sin Z}{Z} F_0$$

These three segments form a continuous waveform, but the first derivative is not generally continuous across the first junction.

The shape of the "extra" waveform is fairly complicated, and the way it changes with Z is not easy to describe. Because the waveform is thought to have explanatory value, a number of examples are given in Figs. III-3, III-4, III-5, and III-6. All of these curves are drawn for $\psi = 0$; here, as in $F_{1,2}$, the dependence on ψ is not worth showing. The quantity plotted is Z times the "extra" signal, to remove the gross amplitude change associated with Z. As in Figs. III-1 and III-2, the curves are aligned vertically, $t = 0$ is marked by a vertical line, and all have the same horizontal and vertical scales.

It is easiest to begin with large values of Z. Figure III-3 shows the waveform at the tops of the 49th and 50th side lobes and in the adjacent nulls. The waveform consists of two bursts of the same shape, but the two bursts are out of phase by π when Z/π is a half-integer (i.e., near the peaks of the side lobes). This phase shift, together with the changing amplitude, leads to a complex transition region beginning at $t = 0$. Despite the phase shift, the two bursts should not be thought to cancel each other; this signal delivers energy.

In the pattern nulls (Z/π integer), the two bursts are in phase and no transition whatever occurs. Indeed, the "expected" signal is zero in the null, and the "extra" signal is $F_{1,2}$ itself. It is evident by inspection that the "extra" signal in the null delivers about the same amount of energy as it does at the tops of the side lobes.

Even near $Z/\pi = 50$, the entire duration of this extra signal is brief (but not quite negligibly so) compared with the rise time of the receiver. Consequently the receiver is unable to follow, and tends to respond, very nearly, as if the incoming signal were an impulse. Given this tendency, together with the nearly equal energy content of these four waveforms, it can be seen that they will excite nearly the same receiver response.

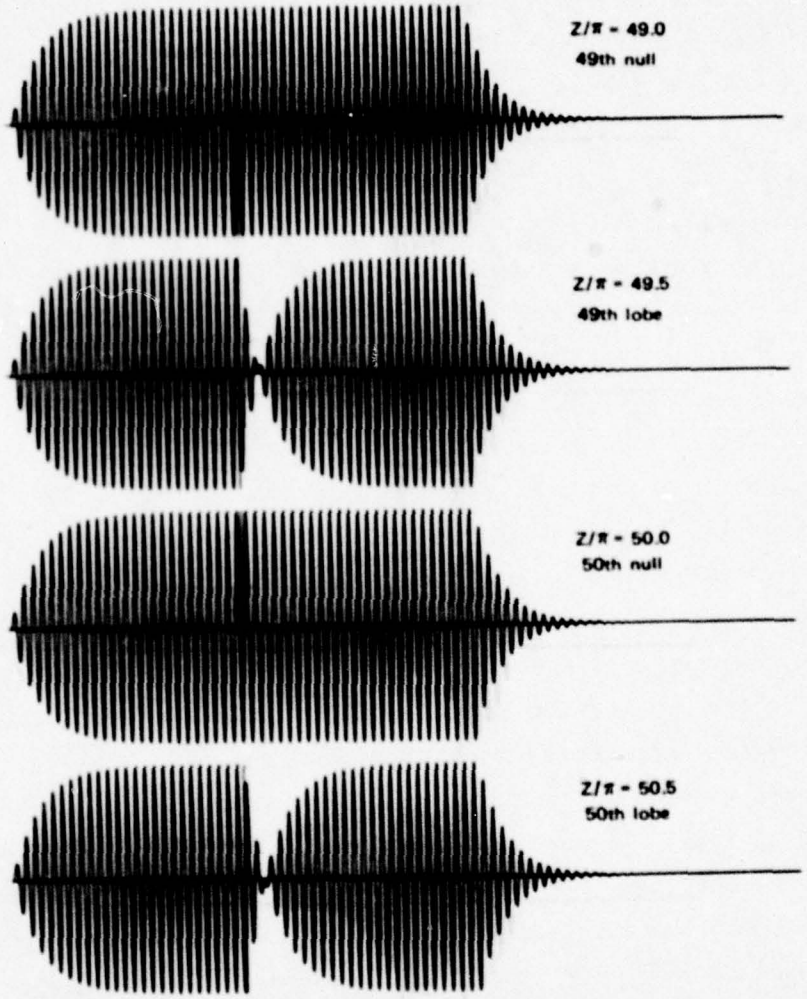


Fig. III-3 — The "extra" signal at large angles from the source axis

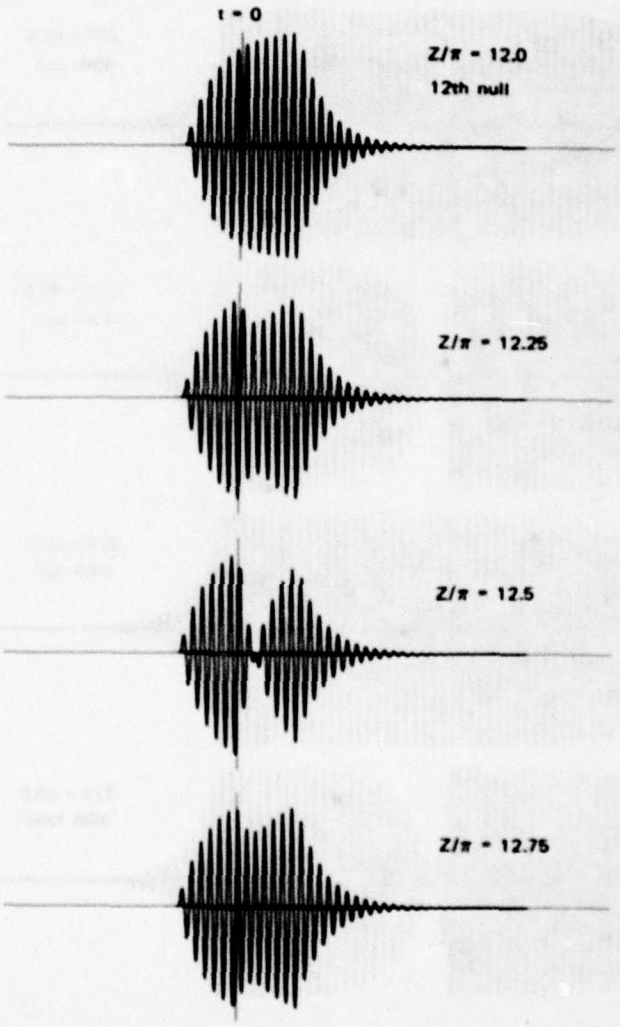


Fig. III-4 — The "extra" signal in the 12th side lobe

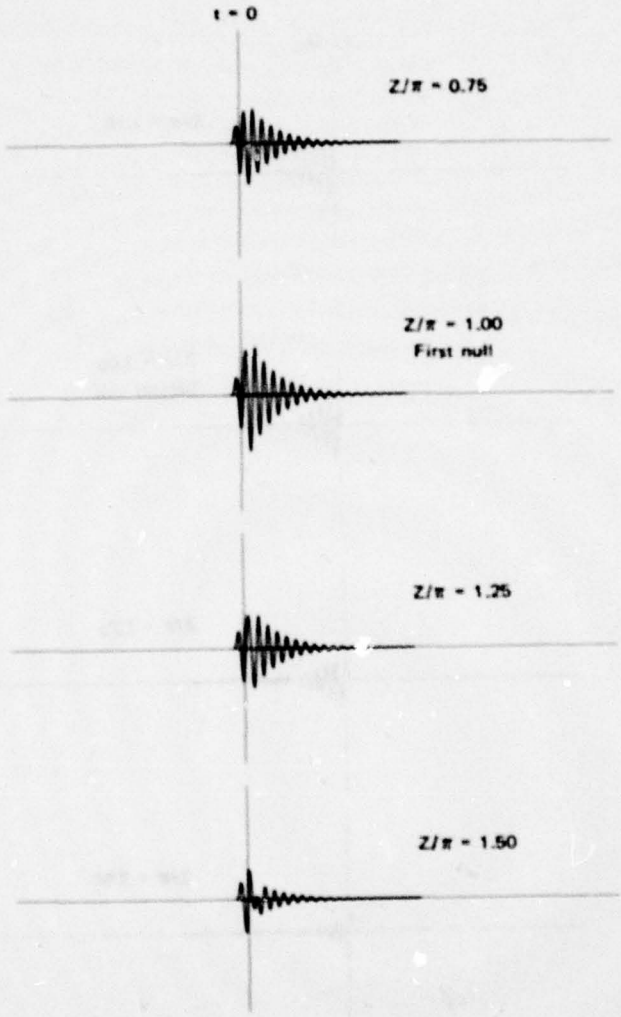


Fig. III-5 — The "extra" signal near the first null of the source lobe pattern

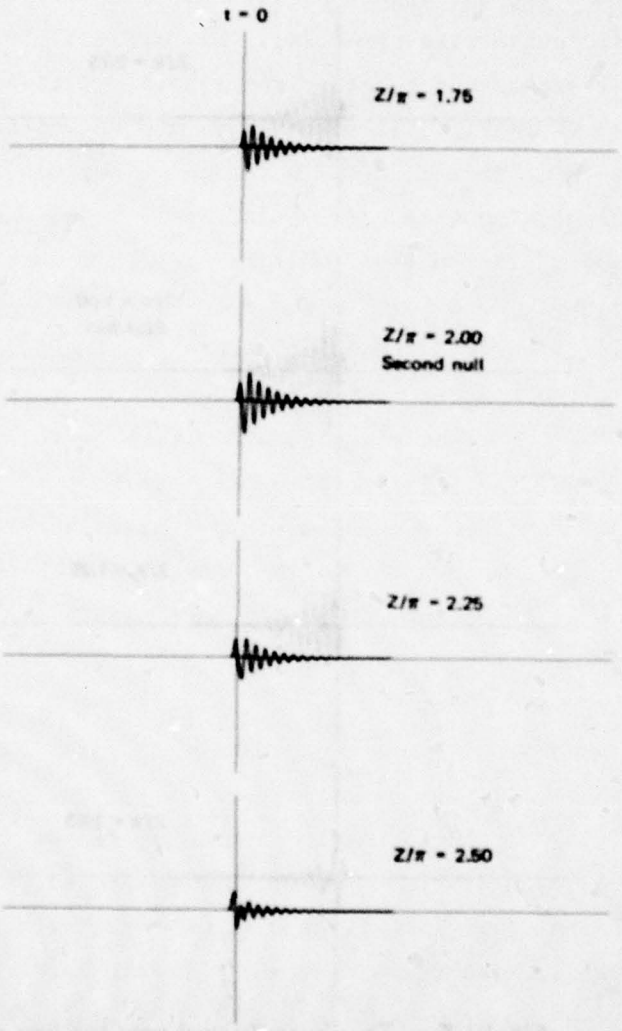


Fig. III-6 — The "extra" signal near the second null of the source lobe pattern

Figure III-4 shows a group of four "extra" waveforms in and near the 12th side lobe; here smaller increments of Z are displayed to show more clearly the change in waveform across the lobe. We see that, in the null at $Z/\pi = 12$, the signal barely has time to rise, and then dies off at the basic pulse rise time rate. As in Fig. III-3, the two bursts of signal are in phase and together comprise $F_{1,2}$ itself. In the other three curves shown in Fig. III-4, there is a phase difference between the two bursts--a difference of π in the third curve. These do not all deliver roughly the same energy, and it is to be expected that the impulse response of the receiver will be excited somewhat more vigorously in the pattern null than in the lobe.

Figures III-5 and III-6 together trace in some detail the evolution of this "extra" signal from a position within the main lobe, across the full width of the first side lobe, and to a position near the top of the second side lobe. The waveform is seen to change rapidly in shape and in peak amplitude because of the phase change between the two bursts and the changing duration as compared with the pulse rise time. All of these waveforms are brief, and excite nearly the pure impulse response of the receiver, but the amplitude of that excitation varies with Z.

All 16 "extra" waveforms shown in these figures can be said to excite, very nearly, the impulse response of the receiver. However, the effective time at which the equivalent impulse might be said to be delivered varies with Z. In contrast, the "expected" signal always arrives at the same time. Thus, when the two arrivals excite their appropriate receiver responses, the interference between the two responses varies with Z in a complex fashion--at least in part because of this shift in the equivalent insertion time of the impulse, but also because of the change with Z of the strength of the excitation. It is because of such matters as this that the detailed behavior analyzed here is complicated, even though the underlying processes are fairly simple.

III-6. RECEIVER FRONT END

It may be questioned whether the source antenna effects that cause $F_{1,2}$ to differ from F_0 might be "undone" by a suitable shape of receiving antenna--for instance, whether a pair of identical plane antennas that are parallel to each other but offset from their axes might "cancel out." They do not; instead, the effects would be doubled. To remove the effect of the source antenna the receiver must contain a dispersive network whose impulse response is the time-reversal of the impulse response of the source--that is, a matched filter. Inasmuch as the impulse response of the source antenna varies with angle, it seems impractical to employ a matched filter in a general-purpose system although it might be done in special circumstances.

No additional insights would be obtained in this study by assuming a directional antenna in the receiver, but the analysis would be made considerably more burdensome. It should be evident that a receiver that observes signals arriving from locations away from the axis of its own directional antenna will experience additional waveform deformations equivalent to the "extra" signal discussed above. To minimize the analytic task it is assumed here that the receiver antenna has perfectly flat phase and amplitude characteristics; the input to the radio-frequency amplifier is assumed to be exactly waveform $F_{1,2}$. That is, it is supposed that the whole receiving system--not merely the receiving antenna--does not disturb in any way the free-field signal produced by the source. From the standpoint of real hardware this is even less realistic than the treatment of the source as a hole in a screen, but it is out of the question to treat the receiver boundary-value problem here.

Section 6 of Part II explained the reason for neglecting the frequency selectivity of the radio-frequency section of the receiver. That was done, initially, as a commonplace approximation that would reduce the analytic task. It was thought at first that the approximation would not yield significantly misleading results, and the study was carried to completion on that basis. Those results will be

considered here first because they offer a simpler picture of the processes that occur in the receiver. To put it briefly, the wide open front end does not stretch the brief duration of the "extra" signal. That duration is, for most values of Z, so short that it excites very nearly the pure impulse response of the receiver. The behavior associated with that response is comparatively easy to comprehend.

Following a discussion of the wide open case, we will return to this point in the analysis and introduce a selective circuit in the front end. The Q of that circuit will be designated Q_3. For now we consider Q_3 to be absent entirely, and assume that waveform $F_{1,2}$ is delivered, unchanged, to the heterodyne detector.

III-7. HETERODYNE STAGE; $F_{3,4}$

The heterodyne detector is a nonlinear device and an exact analysis of its behavior is impractical here. For instance, not all heterodyne detectors are alike. In some the local oscillator (LO) drives the detector back and forth between cutoff and saturation so vigorously that the detector can be regarded as a square-wave device whose zero-crossing times are shifted very slightly by the feeble incoming signal. In that representation the input to the intermediate-frequency amplifier (IF strip) could be treated as a sequence of alternate unit steps up and down with nearly regular spacing. Then the IF strip output could be obtained as a series of step responses. That approach could be implemented in a digital computer, but does not lend itself to analysis.

In other heterodyne detectors, the stage may not be driven so strongly as to resemble a square-wave device. Instead, the stage may be driven back and forth across a curved characteristic, perhaps not all the way to saturation and cutoff. In that case, it is the curvature that offers the necessary nonlinear behavior, and the output of the stage can be obtained as an infinite power series expansion. Such an analysis could be undertaken, but only at great effort due to the difficult integrals. Moreover, the results would be complicated by

the need to provide for a description of the shape of the curved characteristic (that is, by a set of numerical values specifying the linear, quadratic, cubic, quartic . . . coefficients in the expansion.)

In many receivers the situation is made much less tractable because the heterodyne detector is closely coupled to the first stage of the IF strip. The detector then imposes a varying termination impedance on the IF strip and conventional linear filter theory is no longer applicable. Linear superposition fails, and the circuit response to a given input depends on all the antecedent inputs. Analysis of such circuits is a major undertaking even for simple input signals.

The foregoing explains the practical necessity of adopting a conventional approximate description of the heterodyne stage. This is one of only two places wherein the analysis is not formally exact; the other is in the use of numerical integration to obtain the output of the envelope detector. (Other aspects of the analysis are physically unreasonable though analytically exact, as in the Kirchoff boundary-condition assumption, or are impractically idealized as compared to real hardware, but nevertheless treated exactly.)

It is assumed here that the output of the heterodyne stage can be regarded as the square of the sum of the input signal $F_{1,2}$ plus the local oscillator. That is, the curved response characteristic is assumed to be a quadratic function; higher order terms such as the cubic are assumed to be negligible. (The linear term is ignored because it leaves the two signals unaffected and their neglect can be justified on the same bases as are used to neglect the squares of the signals.) The consequences of this approximation cannot be estimated with confidence here; it is hoped they do not lead to appreciable error. This square-law approximation is widely used and works satisfactorily in practice. However, the usual assessment considers quasi-steady-state conditions, and may not be fully applicable to transient conditions. An experimental check of this question might be in order.

The local oscillator is taken to be

$$A \cos[(\omega_0+\omega_1)t+\sigma] = A \cos(U+V+\sigma)$$

where A >> 1. That is, the amplitude of the LO input to the detector is very much greater than the amplitude of the incoming signal $F_{1,2}$. Here ω_1 is the center frequency of the IF strip, so this assumes that the receiver is perfectly tuned to the carrier frequency of the incoming signal. Phase angle σ will be regarded as constant. Slight mistuning of the receiver can be thought of as equivalent to making σ proportional to time; in view of the dependence of receiver response on σ that is found below, such a time dependence will be seen to open vistas of additional complications in the performance of a system. In any case, σ must be regarded as random with respect to the clock time used to describe $F_{1,2}$. To control σ with respect to that clock necessitates that the entire transit time from source to receiver be invariant within a fraction of a nanosecond; numerous effects, including fluctuations in the index of refraction, would defeat such an effort. For similar reasons, the phases of the local oscillators of separate receivers must be regarded as mutually random as well as random with respect to the source.

With the square-law assumption, the output from the heterodyne stage is taken to be

$$[F_{1,2} + A\cos(U+V+\sigma)]^2 = F_{1,2}^2 + A^2\cos^2(U+V+\sigma) + 2AF_{1,2}\cos(U+V+\sigma)$$

The square of the local oscillator can be neglected because the response of the IF strip is down a great many dB at zero frequency and at twice the LO frequency. Indeed, even the simple IF strip considered here is down 225.3 dB at twice the LO frequency that is considered in numerical examples. $F_{1,2}$ is a modulated signal, and argument based on the steady-state response of the IF strip at frequency ω_0 does not provide a sufficient basis to neglect $F_{1,2}^2$. That term will be neglected here, and the neglect justified on the basis that the peak amplitude of $F_{1,2}^2$ is about 1 at most, whereas the amplitude of the cross-product term is higher by the very large factor A. Thus, even though $F_{1,2}^2$ is modulated,

it contributes negligibly to the excitation of the IF strip in comparison to the cross-product term that is also modulated. Once it is decided that the heterodyne output can be regarded as only the cross-product term, the coefficient A can be discarded in the interest of normalization. The factor of two is retained.

Just as the input signal is regarded as two successive segments F_1 and F_2, the output from the heterodyne is regarded as two successive counterpart segments F_3 and F_4:

$$F_3 = 2F_1 \cos(U+V+\sigma)$$

$$F_4 = 2F_2 \cos(U+V+\sigma)$$

The entire output will be termed $F_{3,4}$. It is the input to the IF strip that follows the heterodyne stage. As in the case of $F_{1,2}$, the two segments are smoothly joined and $F_{3,4}$ is a smooth waveform.

It is convenient for subsequent analysis and for the discussion to express $F_{3,4}$ *in terms of the sum frequency* $2\omega_0 + \omega_1$ and the difference frequency ω_1:

$$F_3 = \frac{1}{2Z}\,[-\cos(V+\sigma-\psi-Z-\phi) - \cos(2U+V+\sigma+\psi+Z+\phi)]$$

$$+ \frac{Me}{2Z}^{-\frac{(U+Z)}{2\Delta_0 Q_0}}\,[\cos(V+\sigma-\psi-Z-\mu+\phi) + \cos(2U+V+\sigma+\psi+Z+\mu-\phi)];\ -Z \le U \le Z$$

$$F_4 = \frac{\sin Z}{Z}\,[-\sin(V+\sigma-\psi-\phi) + \sin(2U+V+\sigma+\psi+\phi)]$$

$$+ \frac{Me}{2Z}^{-\frac{(U+Z)}{2\Delta_0 Q_0}}\,[\cos(V+\sigma-\psi-Z-\mu+\phi) + \cos(2U+V+\sigma+\psi+Z+\mu-\phi)]$$

$$- \frac{Me}{2Z}^{-\frac{(U-Z)}{2\Delta_0 Q_0}}\,[\cos(V+\sigma-\psi+Z-\mu+\phi) + \cos(2U+V+\sigma+\psi-Z+\mu-\phi)];\ U \ge Z$$

At this point it is customary to argue that the sum frequency terms can be neglected because the response of the IF strip is down a great many dB at the sum frequency. Under ordinary quasi-steady conditions, the argument is reasonable and those terms are usually discarded. The sum terms will be retained here because the *transient* response of the IF strip to those terms is not down nearly so many dB.

The role of the transient response to the sum frequency terms is made clear by a consideration of the situation when $U = -Z$. This is the arrival time of the first infinitesimal energy, and $F_1 = 0$, when $U = -Z$. Consequently, F_3 must also equal zero at this time. However, if the sum frequency terms are discarded, F_3 is not equal to zero at this time; instead it undergoes a discrete jump to a finite value. Discarding the sum frequency terms is equivalent to introducing a stepped artifact at the leading edge of the pulse. Such an analytic artifact might be thought to cause the transient phenomena that are subsequently found, and would vitiate the study. It will be found later that those terms in the IF output that arise from the sum frequency terms contain a rapidly decaying exponential factor. They contribute significantly only to the very early portion of the receiver response. Further, their amplitude is comparatively low. Thus, now that the analysis reported here has shown that this is their contribution, they might be omitted from future analyses provided that the very early response is ignored. Inasmuch as the envelope detector may have a rather long "memory," the analyst should be watchful for signs of an artifact if he discards these terms.

It is noteworthy that phases σ and ϕ occur in the combination $(\sigma - \phi)$ in all the difference frequency terms, and in the combination $(\sigma + \phi)$ in the sum frequency terms. Because the difference frequency terms are dominant, it can be foreseen that phase effects will be determined primarily by the difference $(\sigma - \phi)$, but that weaker effects determined by $(\sigma + \phi)$ will be found. To the extent that $(\sigma - \phi)$ governs the behavior, the two phases are interchangeable. There is, therefore, scant reason to debate whether source phase ϕ might be controlled to a particular value or might be random. As explained above, σ must be regarded as random from one pulse to another and random from

one receiver to another. The quantity $(\sigma - \psi)$ will be random. There is, of course, a weak dependence on ψ that is not influenced by σ; this occurs in the dependence of M and μ on ψ. For this reason the receiver response cannot be governed exclusively by $(\sigma - \psi)$, and the weaker dependence on $(\sigma + \psi)$ is in keeping with that fact.

Phases σ and ψ play an important role here, and their effect will be discussed at length in the next section. First, however, we display some representative samples of $F_{3,4}$ at various angular locations. The values $\sigma = \psi = 0$ are used. Figures III-7 and III-8 show $Z \cdot F_{3,4}$ as the signal would appear on an oscilloscope with the vertical gain adjusted to remove the gross effect of Z on amplitude. The curves are all drawn to the same horizontal and vertical scales and are aligned vertically. A vertical line indicates t=0 on each curve. Figure III-7 shows the waveform on the source axis and in three nearby locations. Figure III-8 shows the waveforms at larger angles. Waveforms in pattern nulls and near the tops of the adjacent lobes are shown in both

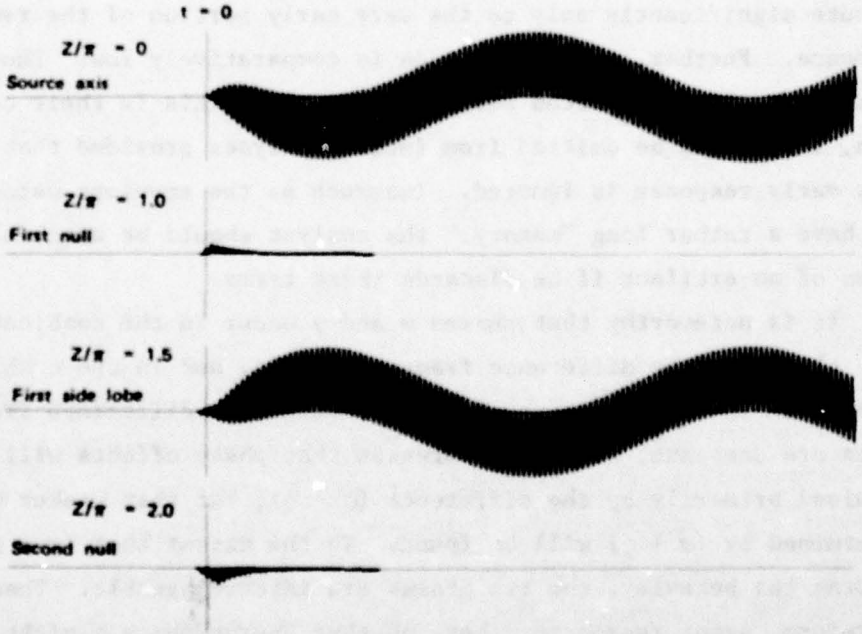


Fig. III-7 — Output waveforms from the heterodyne detector on the source axis and at nearby locations. Phase conditions: $\sigma = \psi = 0$.

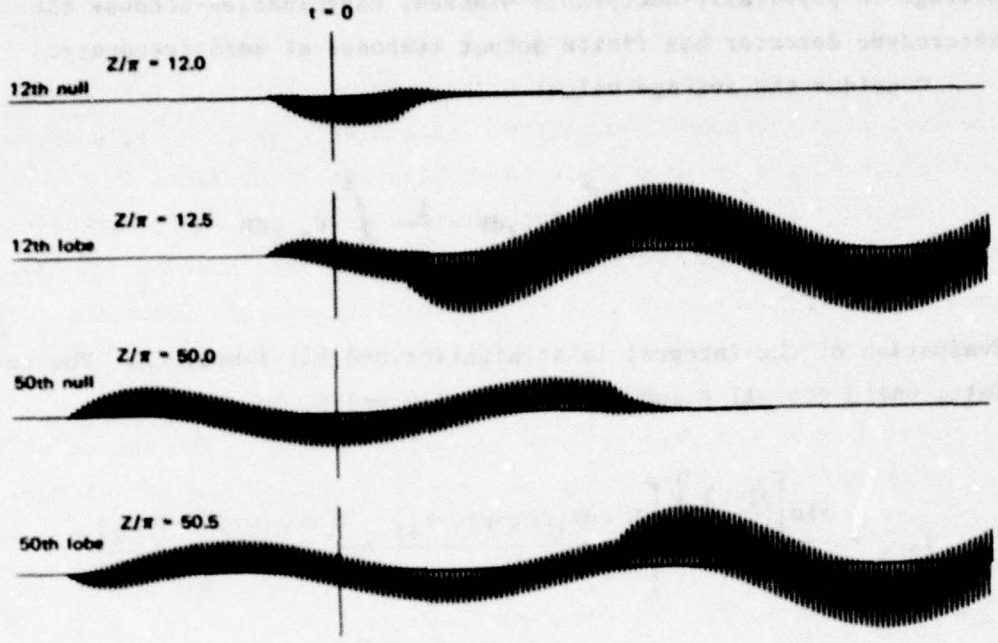


Fig. III-8 — Output waveforms from the heterodyne detector at intermediate and large angles from the source axis. *Phase conditions:* $\sigma = \psi = 0$.

figures. The slow undulations of these waveforms reflect the familiar difference-frequency signal. The rapid oscillations result from the sum-frequency signal that has been retained. It will be seen that the two signals are added together; they do not modulate each other.

III-8. PHASE EFFECTS; THE SPECTRUM OF $F_{3,4}$

Earlier, in choosing a suitable form for F_0, care was taken to avoid a waveform that has a non-zero average value. Here in $F_{3,4}$ we find waveforms that certainly have non-zero average values. Consider, for example, the signal in the first pattern null ($Z/\pi=1$) in Fig. III-7. The signal is almost entirely on one side of the axis and obviously has a substantial average value. It is not so evident to the eye, but is true, that every waveform shown in Figs. III-7 and III-8 has a non-zero average. Here, unlike the situation involving F_0, the non-zero

average is physically acceptable--indeed, unavoidable--because the heterodyne detector has finite output response at zero frequency.

Consider the average value:

$$I_{3,4} = \int_{-Z/\omega_0}^{\infty} F_{3,4}\,dt = \frac{1}{\omega_0}\int_{-Z}^{\infty} F_{3,4}\,dU$$

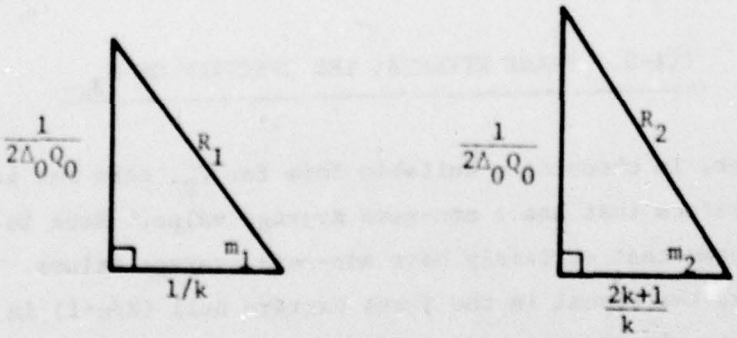
Evaluation of the integral is straightforward but laborious. The result, valid for all Z and all values of σ and ψ, is

$$I_{3,4} = \frac{\sin\left[\left(\frac{1+k}{k}\right)Z\right]}{\omega_0 Z}\left[\frac{M\cos\left(\sigma-\psi-\mu+\phi-m_1\right)}{R_1} + \frac{M\cos\left(\sigma+\psi+\mu-\phi-m_2\right)}{R_2}\right.$$

$$\left. -k\cos(\sigma-\psi-\phi) - \frac{k}{2k+1}\cos(\sigma+\psi+\phi)\right]$$

where $k = \frac{\omega_0}{\omega_1}$

and



For the numerical values used in this report, including k = 40, and for the case ϕ = 0 shown in Figs. III-7 and III-8, and assuming the signal normalization is one volt:

$$I_{3,4} = \frac{\sin\left(\frac{41}{40} Z\right)}{\left(\frac{41}{40} Z\right)} \left[1.799369 \times 10^{-9} \cos(\sigma - 2.7274099)\right] \text{ volt-seconds}$$

Several observations emerge from this result. First, this is dimensionally an impulse; it can be foreseen that this signal will inject a net charge into the IF strip as an impulse would. Second, the dependence of $I_{3,4}$ upon Z is "out of step" with the directivity pattern sinZ/Z; $I_{3,4}$ has a number of nulls, just as the pattern does, but the nulls of $I_{3,4}$ occur at varying positions with respect to the pattern nulls. This will lead to a complicated and asymmetric behavior about the position of each pattern null. Third, and very important, at any location defined by Z the amplitude of $I_{3,4}$ varies sinusoidally with phase σ (and with ψ). Thus, the strength of this input may be positive, negative, or zero depending on σ. Fourth, and rather surprising, the constant, 2.7274 radians, is independent of Z; the dependence upon σ is the same at all locations.

When $Z/\pi = 1$, the value of $I_{3,4}$ is

$$-4.384 \times 10^{-11} \cos(\sigma - 2.7274) \text{ volt-seconds}$$

The peak value when $\sigma = 2.7274$, is 207.16 dB below a unit impulse. This level will be compared to the output of the IF strip in Section III-10, and will be found to agree within a fraction of a dB.

The dependence of $I_{3,4}$ on σ means that the waveform changes with phase, and the appearance of the waveform is of some interest. Figure III-9 shows $F_{3,4}$ in the first pattern null, for $\psi = 0$, and for the two values of σ that maximize and minimize the magnitude of $I_{3,4}$. Comparable changes of waveform with phase occur at other angles.

Even when $I_{3,4}$ is zero an oscillatory signal remains. The entire amplitude of $F_{3,4}$ does not vary sinusoidally with the phase. Rather, $I_{3,4}$ represents the zero-frequency end of the energy spectrum of $F_{3,4}$, and the shape of that spectrum changes with phase. Figures III-10 and III-11 show the energy spectrum over a wide frequency range for four angular positions near boresight. These values of Z were chosen to illustrate the presence or absence of the steady-state

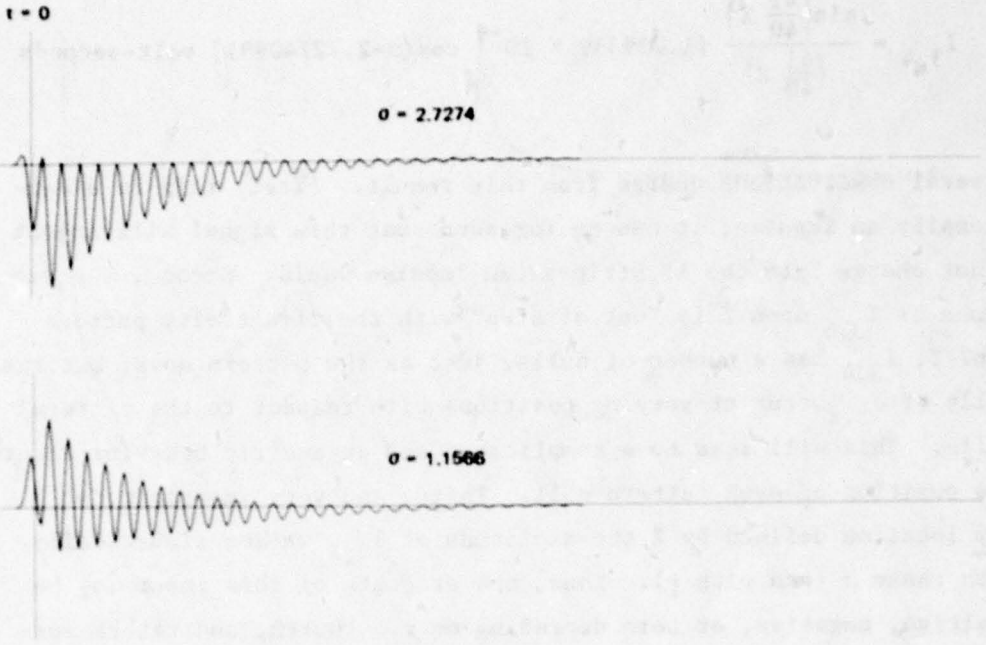


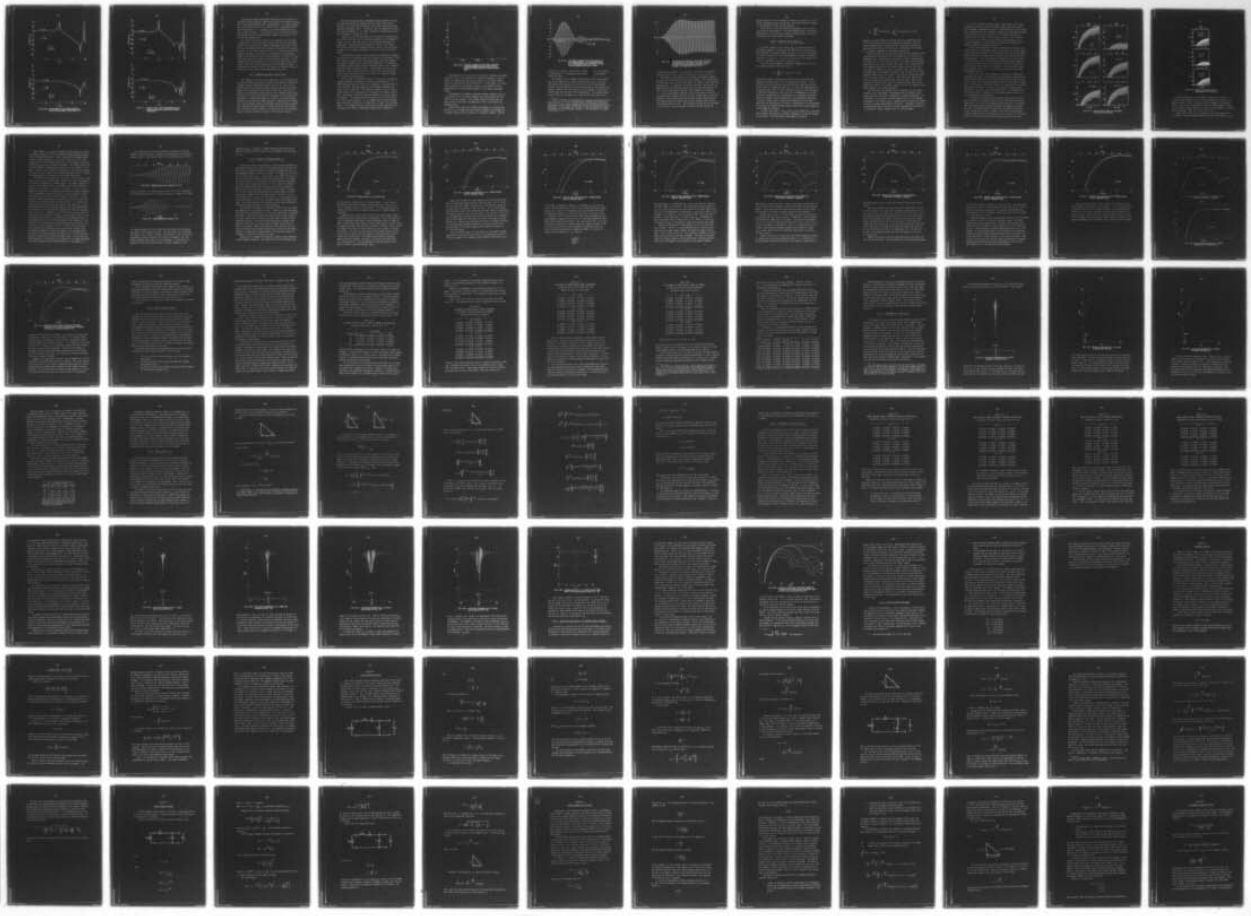
Fig. III-9 — Output waveforms from the heterodyne detector when the receiver is in the first null of the lobe pattern of the source, showing change of waveform with phase, σ, of the local oscillator. Source phase $\psi = 0$.

signal and to indicate that the phase dependence is present at all locations. The scale of dB is the same in the four cases, but the choice of zero dB is arbitrary. In each figure, the two curves indicate the spectra when the phases are chosen to maximize $I_{3,4}$ and to make $I_{3,4}$ zero.

These are energy spectra, not power spectra (the latter would be averaged over infinite time, whereas the energy spectrum is not). When Z/π is an integer, no steady-state signal is present because of the directivity pattern; the signal contains a finite amount of energy and the spectrum is finite at all frequencies. When Z/π is a non-integer, the never-ending steady-state term delivers infinite energy (in infinite time) and the energy spectrum rises to infinity at the sum frequency as well as at the difference frequency. Nevertheless, the shape of the spectrum is otherwise not much different, and the phase effect remains evident.

AD-A078 373 RAND CORP SANTA MONICA CA F/G 17/3
TRANSIENT RESPONSE OF A HETERODYNE RECEIVER: IMPLICATIONS FOR A--ETC(U)
NOV 79 T F BURKE F49620-77-C-0023
UNCLASSIFIED RAND/R-2418-AF NL

2 OF 3
AD
A078373



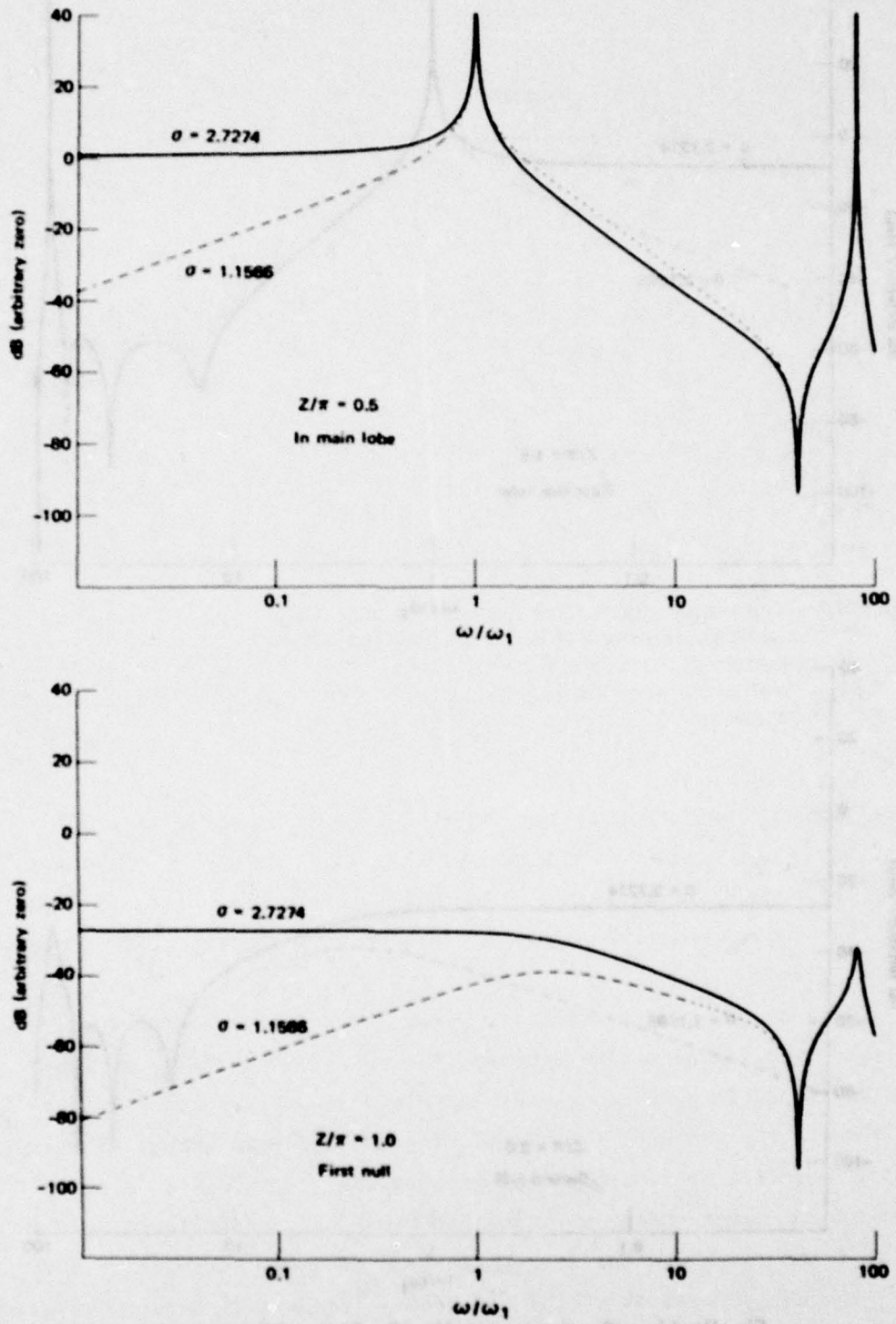


Fig. III-10 — Energy spectra of $F_{3,\phi}$ showing dependence upon Z and upon phase. Source phase $\psi = 0$.

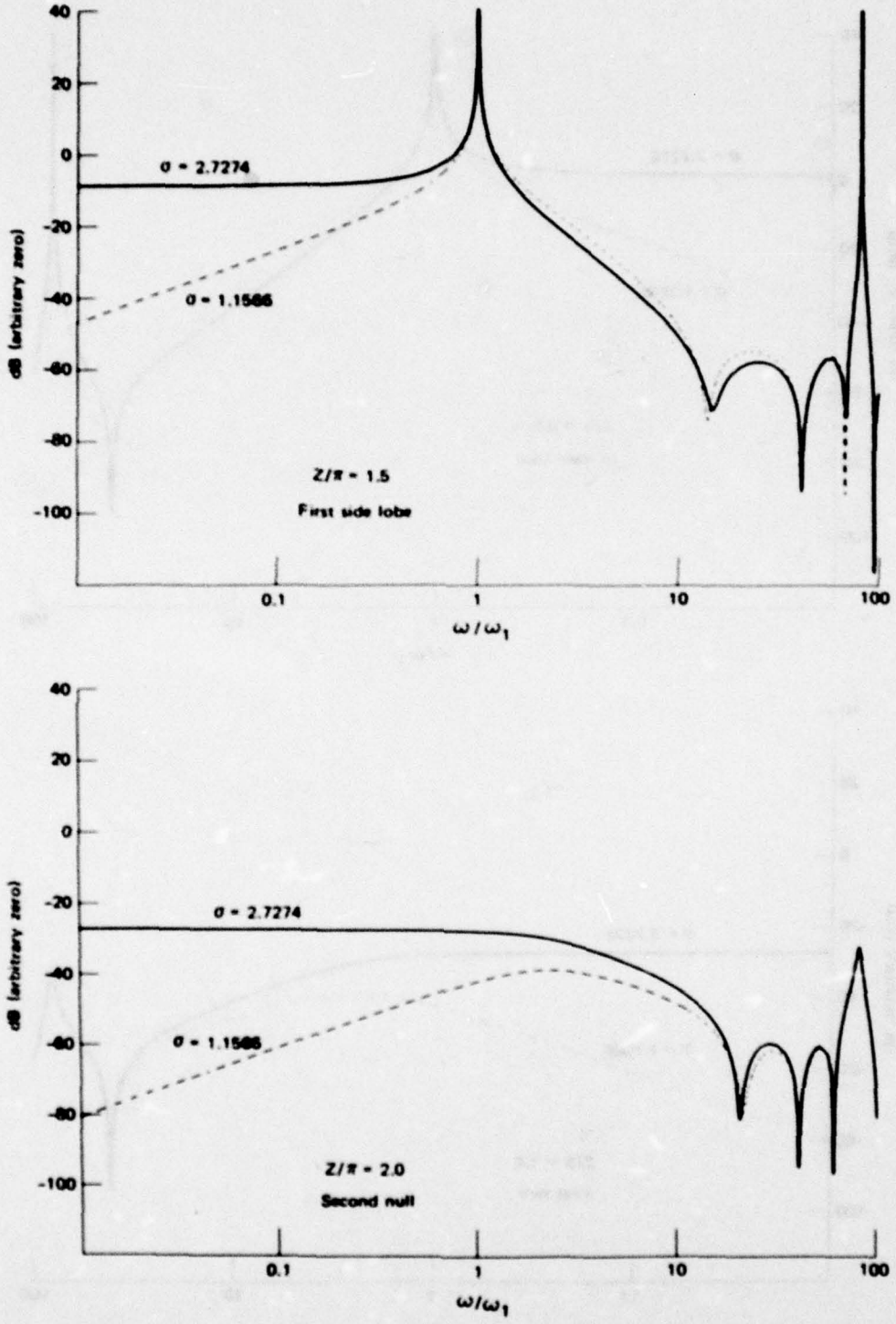


Fig. III-11 — Similar to Fig. III-10. Note additional minima caused by source antenna behaving as spatially varying filter.

Such spectra become increasingly complicated as Z increases because the spectrum reflects, among other things, the frequency selectivity of the source antenna itself (see Appendix H). The high frequency end of the spectrum at $Z/\pi = 2$ is seen to have a few minima that are not present at $Z/\pi = 1$. At larger Z, still more minima occur and move to lower frequency. These minima are not true zeros of the spectrum because a zero for the difference frequency term is not a zero for the sum frequency term.

The total energy in the signal is the integral over the energy spectrum, and it is clear from the figures that the total energy delivered in the pattern nulls changes considerably with phase. (Elsewhere in the pattern the infinite energy in the steady state becomes dominant, but the finite energy delivered by the transient terms still changes with phase.) This change of signal energy, and consequent change in the amplitude of the receiver response, is peculiar to transient conditions and is absent from steady-state conditions. The dependence of output amplitude on local oscillator phase and on source carrier phase will surprise analysts who are accustomed to quasi-steady-state conditions. Much of the complexity encountered later in this report arises from the phase dependence shown here.

III-9. TRANSIENT RESPONSE OF THE IF STRIP

The designs of IF strips vary greatly, but it is not unusual for the strip to contain more than six poles (which amount to individual resonant circuits), and some contain a dozen or so. These are arranged to provide the desired in-band frequency response (not necessarily flat or symmetric) and to provide very steep cutoff at the edges of the band (skirt selectivity) to reject adjacent interference. Typically one or more poles provide a stop-band to reject particular frequencies. Usually some stages are coupled, so the normal modes interact, while others are nearly isolated. Calculation of the bandpass impulse response would be exceedingly difficult and has probably never been done for such a filter. In most cases not even the low-pass equivalent impulse response (if it exists at all) is considered; design is usually specified in terms of the frequency response.

This study required that the bandpass impulse response be known, and that need limited the complexity of the filter that could be treated. The design that was adopted is the fairly simple four-pole Butterworth bandpass filter. The design rules for Butterworth filters are given in Appendix E. The center frequency is 80 MHz and the half-power bandwidth is 10 MHz ($Q_1 = 8$).

Appendix E discusses the method used to obtain the bandpass step and impulse responses, and the complete expression for the impulse response is given. That expression is quite long and will not be repeated here. Appendix E also discusses the design of the lowpass equivalents of Butterworth bandpass filters, and the step and impulse responses of the lowpass equivalent of this four-pole IF strip are given.

The Butterworth design is not especially advantageous in practice, but neither is it deficient except in the absence of stop-bands. There is no reason to believe that this design leads to peculiar behavior that would circumscribe the conclusions reached here. If anything, the design is rather simple and the transient behavior may be smoother than that of more intricate filters. In any event, the rise of the transient output is slowed as more poles are used, and the effects found in more complicated filters are likely to be greater in magnitude than those discussed here.

In this study, the complete absence of noise and of interfering signals mitigates considerably the practical shortcomings of this simple design because the filter is not obliged to cope with those inputs. Moreover, although moderate by conventional standards, the filter is quite selective. The frequency response of the IF strip is shown in Fig. III-12, along with the response of the lowpass equivalent. The latter is shown for general interest, but was not used in the analysis. It is seen that the response of the IF strip is down 225 dB at the sum frequency. It is a common rule of thumb that each pole contributes 6 dB/octave of selectivity, but that is the limiting slope. It is evident from Fig. III-12 that the skirt selectivity at the edges of the band is considerably greater than 24 dB/octave.

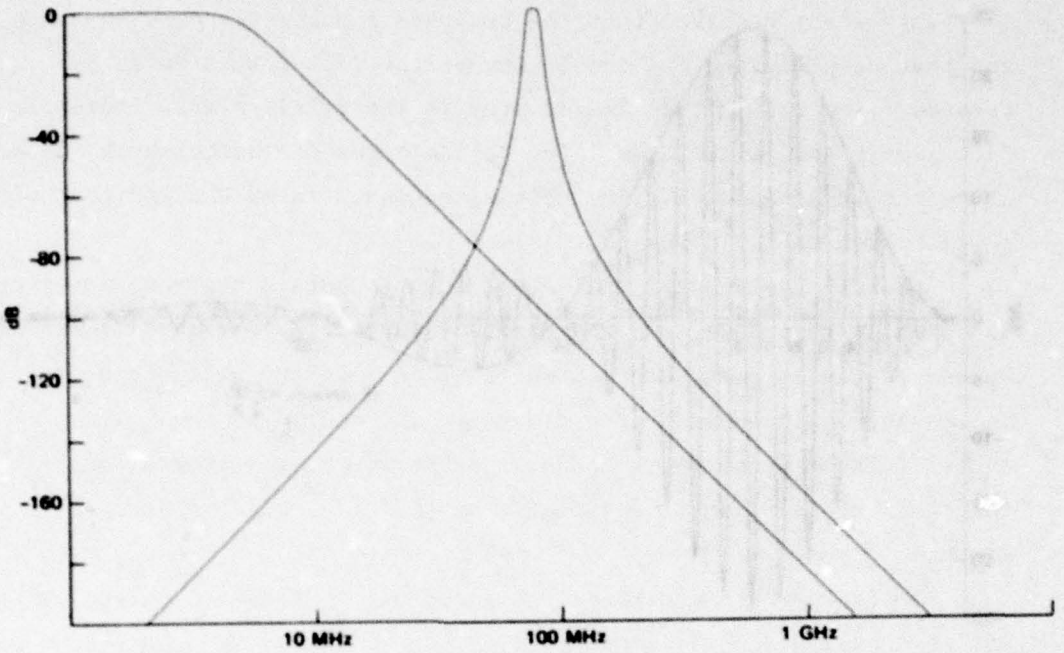


Fig. III-12 — Frequency response of IF strip used in this study (4-pole Butterworth bandpass; center 80 MHz; bandwidth 10 MHz) and of the lowpass equivalent filter

Interpretation of the results of this analysis depends on an understanding of two of the characteristic transient responses of the IF strip: the impulse response and the response to a stepped carrier. These are shown in Figs. III-13 and III-14. In each figure, the oscillatory waveform is the transient response of the bandpass filter; superposed is the corresponding response of the lowpass equivalent filter.

The maximum of the lowpass equivalent impulse response occurs at 7.37 periods of the IF center frequency (92.1 nanoseconds after the insertion of the impulse). The bandpass impulse response does not have a peak at this time; the peaks near 7 and 7.5 IF periods are slightly lower.

Figure III-14 shows the IF output in response to a stepped sine wave (i.e., $\psi = 0$) input at the IF center frequency; the response to a stepped cosine differs in phase but is otherwise similar. This is

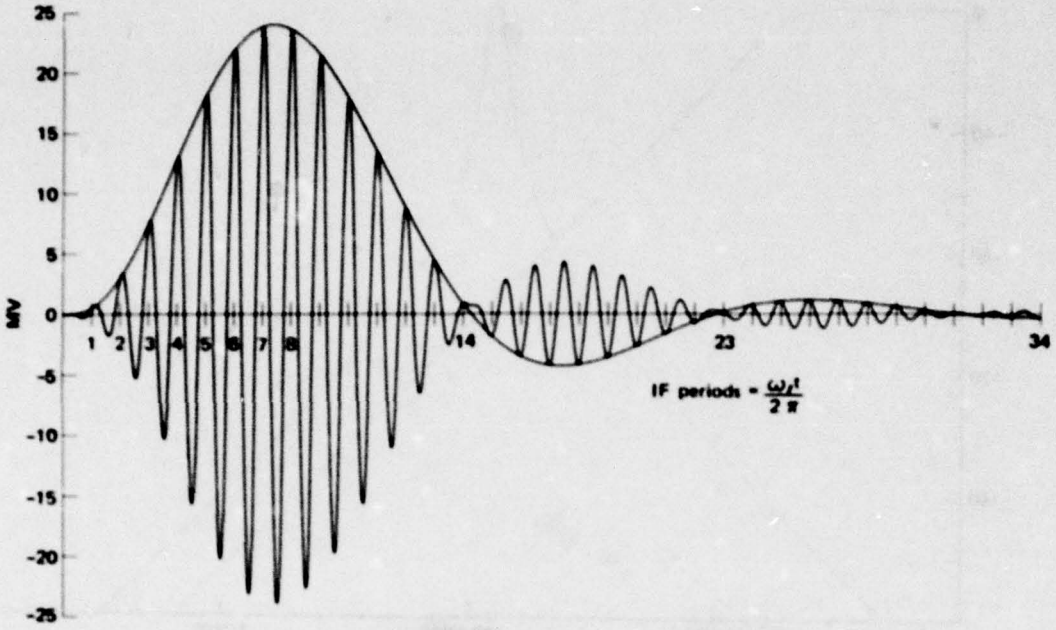


Fig. III-13 — The impulse response of the IF strip and that of the lowpass equivalent. The peak voltage caused by a 1 volt-second impulse is 2.39745×10^7 volts. The zero crossings are not evenly spaced.

not the step response of the filter; we have no use for that response in this analysis.* Superposed is the step response of the lowpass equivalent filter.

The lowpass impulse response shown in Fig. III-13 is the derivative of the lowpass step response shown in Fig. III-14. The first peak of the step response occurs at the first zero of the impulse response, and consequently must occur later than the first peak of the impulse response. The first peak of the step response occurs 14.25 periods later than the step (178.2 nanoseconds). It is characteristic of most filters to overshoot, as is seen at this first peak of the

*This is also not the stepped-carrier response $H_S(t)$ discussed in Section III-13. $H_S(t)$ is the response to $u(0)\sin(\omega_0 t)$ delivered to the receiver input. After passage through the heterodyne stage that input delivers sum frequency components as well as difference frequency components. The response shown in Fig. III-14 is elicited by a stepped sinusoid at the difference frequency. The distinction is trivial except near $t = 0$.

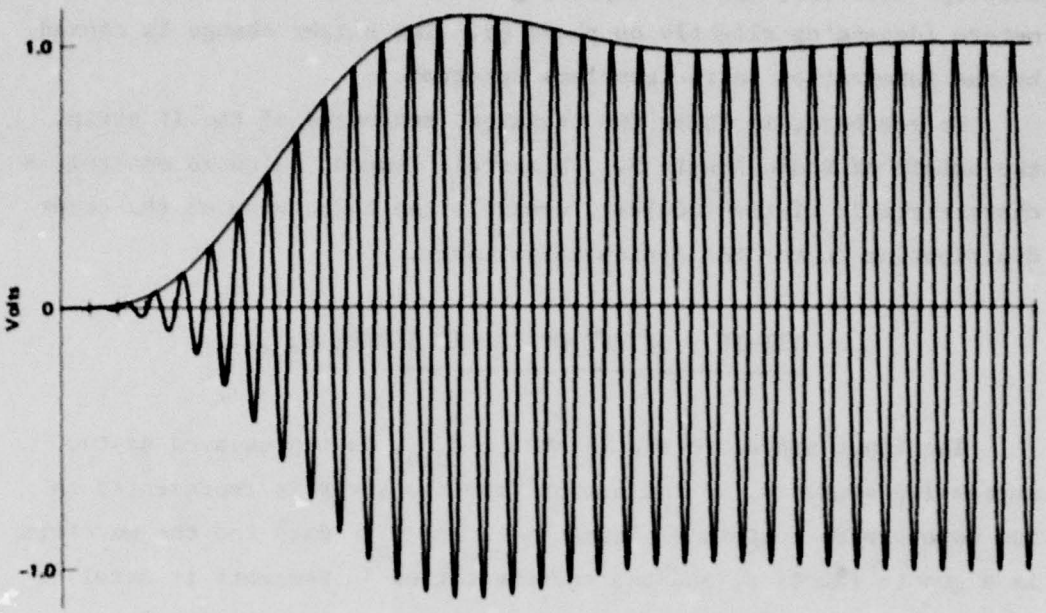


Fig. III-14 The stepped carrier response of the IF strip ($\psi = 0$) and the step response of the lowpass equivalent. The zero crossings are not evenly spaced. Note overshoot of first peak. Time scale same as Fig. III-13.

stepped-carrier response. The overshoot is 0.89 dB for this simple four-pole filter, but is usually somewhat more in real IF strips.

The lowpass curves shown in these figures approximate, roughly, the outputs of the envelope detector that follows. The envelope detector has a rise time of its own and delays the signal, so the peaks come somewhat later. In addition, the envelope detector output at any instant reflects the antecedent inputs and this alters the waveform itself. For a well-designed detector, these changes are not great.

If these two lowpass responses could be subjected to the time-of-arrival logic adopted herein (they cannot because they do not exist anywhere in the receiver), then the impulse response would "arrive" 14.4 meters behind the impulse and the step response would "arrive" 28.3 meters behind the step. The difference between the two "arrivals" is 13.9 meters. When the bandpass responses are put through the

envelope detector, the corresponding difference is about 14.1 or 14.2 meters (depending slightly on phase ψ). The slight change is caused by the integration in the envelope detector.

We see here, in these two transient responses of the IF strip, the origin of scale length S. This scale length, which is entirely a characteristic of the receiver, controls the basic size of the error distribution in the TOA system performance.

III-10. OUTPUT OF THE IF STRIP; $F_{5,6}$

The input signal to the IF strip, $F_{3,4}$, is represented as two successive segments, so the output from the strip is represented as two counterpart successive segments F_5 and F_6. Here too the waveform is a smooth function, and the representation in segments is merely a convenience of notation. As in the previous cases, the waveform as a whole is designated by the double subscript $F_{5,6}$.

The first output segment is obtained by convolution of F_3 with the impulse response of the strip, h(t):

$$F_5 = \int_{-Z/\omega_0}^{t} h(t-\tau)F_3(\tau)d\tau;\ t \leq Z/\omega_0$$

The impulse response contains four terms of similar form but with slightly different values for all the numerical constants. F_3 also contains four terms that can be said to have that same form (in two of the terms the exponential factor is 1). Thus, the integrand above contains 16 terms, each of which contains the product of two sinusoids at different frequencies. When each term is decomposed into sum and difference frequencies, F_5 is represented as 32 elementary integrals. Each such integral yields a single term if an appropriate phase angle is used to combine pairs of terms whose frequencies are the same. When these are evaluated at the limits of integration F_5 is represented by 64 terms, all having the same simple form but with various numerical values among the constants.

At later time, when $t > Z/\omega_0$, input F_3 has stopped and F_4 has begun. Even though F_3 has stopped, the filter continues to "ring" in response to that imput. Thus the second output segment is given by

$$F_6 = \int_{-Z/\omega_0}^{+Z/\omega_0} h(t-\tau)F_3(\tau)d\tau + \int_{+Z/\omega_0}^{t} h(t-\tau)F_4(\tau)d\tau;\ t \geq Z/\omega_0$$

Segment F_4 is represented as six terms, all in the same basic form but with different numerical values among the constants. When the second integral above is broken down into elementary integrals and evaluated at the limits, 96 terms occur. In addition, the 64 terms of F_5 reappear, evaluated at a different upper limit.

All these terms were written out in the hope that combination and cancellation of terms would reduce the expressions for F_5 and F_6 to manageable size. A good deal of collection of terms is possible, but it turned out to be less than hoped for. No way was found to combine terms down to, say, only a dozen or two terms. These expressions require numerical evaluation in a digital computer in any case, and it was decided to abandon the effort to combine terms. Instead all 80 resultant terms were expressed in generic form in terms of the diverse numerical values of the four terms of h(t) and of the numerical values of the ten input terms of F_3 and F_4. Thus, the output from the IF strip was expressed as a double sum of generic terms. That sum was evaluated at the two limits and the difference between those limit values was listed by the computer (and was used by the computer as the input signal to the envelope detector).

The resulting value obtained this way reflects the difference--possibly a very small number--among many terms with opposite sign. Double precision arithmetic is mandatory if reasonable accuracy is to be obtained. A great many tests were applied to check for internal consistency of the results, and it appears that the results are dependable over a dynamic range larger than 120 dB. The range is likely to be considerably better than that--possibly approaching 200 dB--but it is difficult to devise suitable tests at such low levels.

It is seen from the foregoing that a single point on an output waveform $F_{5,6}$ requires that the computer evaluate 160 terms, each of which contains an exponential factor, a sinusoidal factor, and an amplitude coefficient that is a fraction. Generation of the graphic display of the waveform for just one choice of Z, σ, and ψ involves much arithmetic and is costly even in a modern computer. It is impractical to carry out an extensive exploration of the dependence of $F_{5,6}$ on those parameters, to say nothing of exploring other choices of center frequencies, bandwidths, and so on.

All the effects described in Part II and implied in earlier sections of Part III are present in the $F_{5,6}$ waveform, but are not readily visible in graphic displays. The oscillatory character of this signal makes it difficult to judge what the output of the envelope detector will look like. It is not worthwhile to present an extensive display of this waveform; Figs. III-15 and III-16 show abbreviated samples of $F_{5,6}$ over an angular span from boresight to the first side lobe and over a span at large angles from boresight. All these samples show the particular case $\sigma = \psi = 0$; in most cases other choices of phase appear about the same.

Here for the first time waveforms are displayed on logarithmic amplitude scale (dB). At the expense of discarding the algebraic sign (which is lost anyway in the envelope detector) and doubling the number of "loops," this presentation makes possible a wide range of amplitude. All of these waveforms are shown on the same scales of amplitude and time (or distance). It will be seen that the boresight signal approaches 0 dB level, whereas the level at $Z/\pi = 1.5$ approaches -13.5 dB, as they should. Similarly, the wide-angle signals shown in Fig. III-16 approach much lower levels.

On the whole, the shapes of these signals look pretty much the same, although, as discussed below, there are significant differences among them. It is only in the pattern nulls, where Z/π is an integer, that the waveform differs greatly from what is ordinarily expected. Not only is a signal shown to be present in the null, but the level of the signal is not trivial. Indeed, in the 50th null the signal reaches a peak level that is comparable to the level in the first null, and not much below the peak levels in the 49th and 50th side lobes.

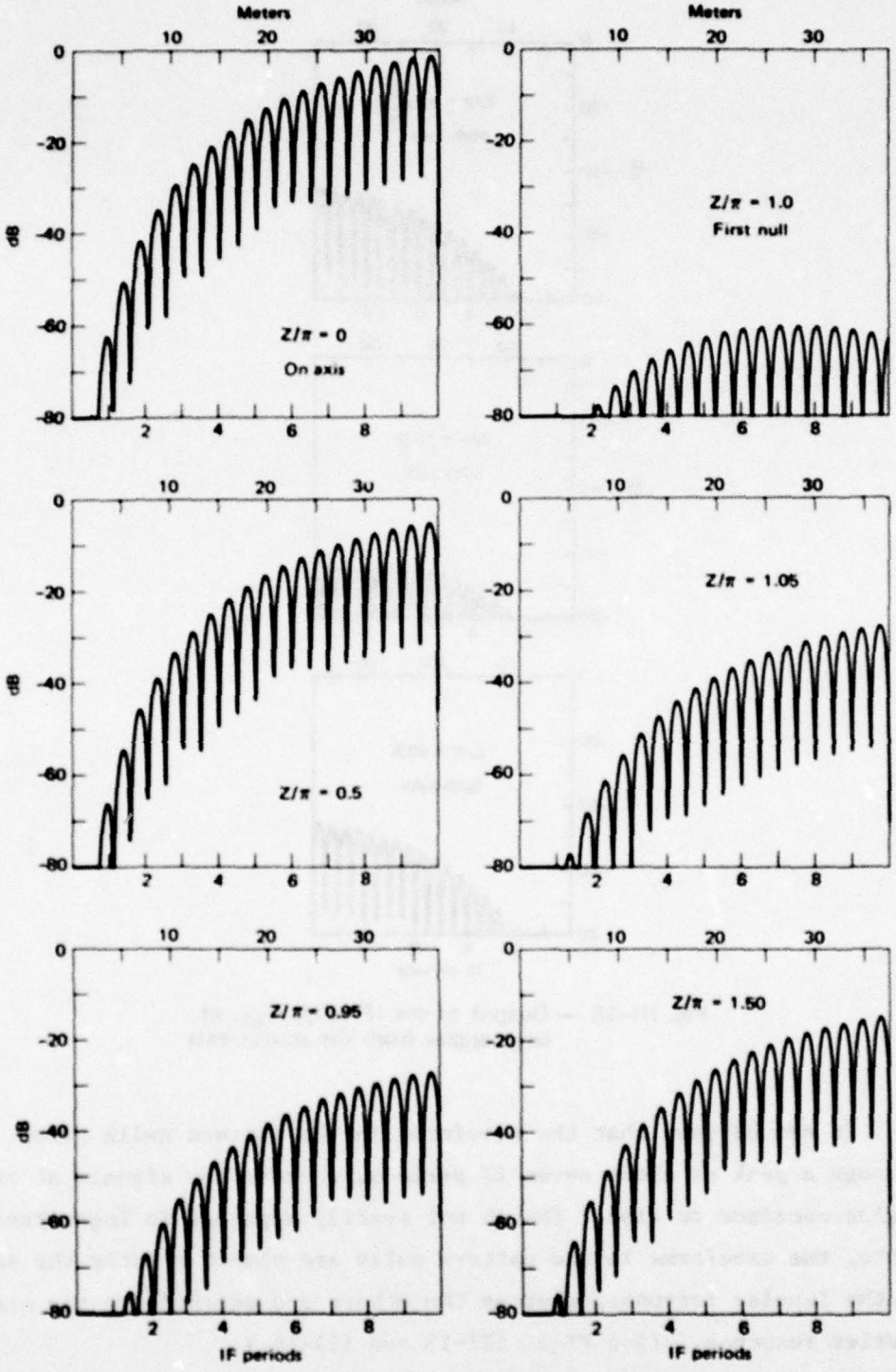


Fig. III-15 — Output of the IF strip, $F_{5,6}$, at locations on and near the source axis

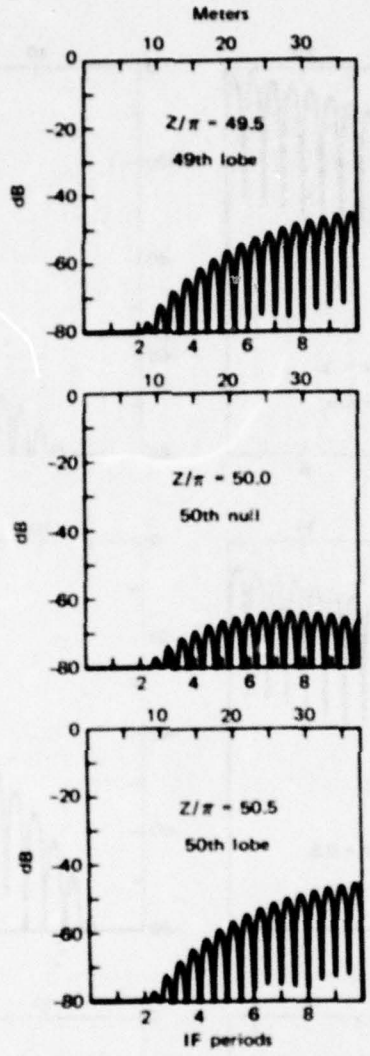


Fig. III-16 — Output of the IF strip, $F_{5,6}$, at large angles from the source axis

It can be seen that the waveforms in the pattern nulls pass through a peak at about seven IF periods, whereas the signals at other angles continue to rise. Though not readily apparent in logarithmic plots, the waveforms in the pattern nulls are almost exactly the same as the impulse response, whereas the others are essentially the stepped-carrier response. (See Figs. III-13 and III-14.)

Figure III-17 shows the $F_{5,6}$ waveform in the first pattern null for phase $\psi = 0$ and for the two choices of phase σ that maximize $I_{3,4}$

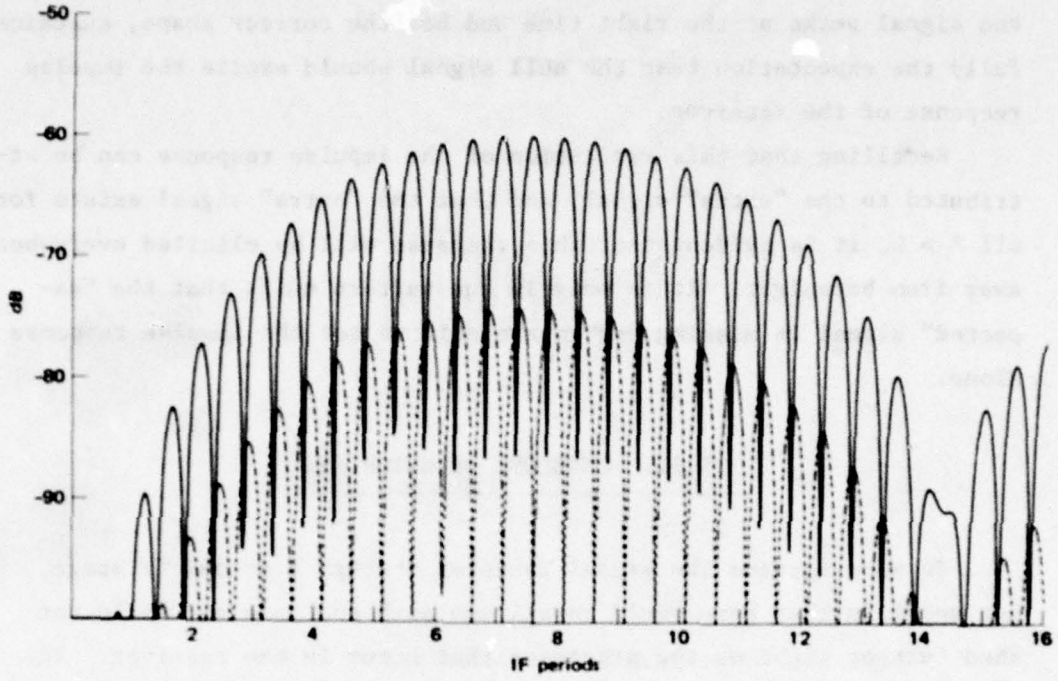


Fig. III-17 — Output of the IF strip in the first pattern null of the source ($Z/\pi = 1.0$) showing change of amplitude with LO phase σ

or make $I_{3,4}$ zero (see Section III-8). It is seen that a phase change of 90 degrees changes the output amplitude of the IF strip by 14.2 dB but the waveform itself is virtually unchanged. This substantial amplitude change with phase alters greatly the resultant of the mutual interference between the impulse response and the stepped-carrier response when both are present at similar amplitude.

It was shown in III-8 that for $Z/\pi = 1$, $\phi = 0$, and $\sigma = 2.7274$, $I_{3,4}$ is equivalent, nearly, to an impulse whose strength is 207.16 dB below a unit impulse. In Fig. III-9, the impulse response of the IF strip was illustrated, and the peak voltage produced by a unit impulse was seen to be 2.397×10^{7} volts--147.60 dB above 1 volt. From these two values it is to be expected that the peak voltage reached by the upper curve in Fig. III-17 should be -59.56 dB. The highest peak listed by the computer is -59.98 dB--a disparity of only 0.41 dB. This excellent agreement in signal level, together with the fact that

the signal peaks at the right time and has the correct shape, sustains fully the expectation that the null signal should excite the impulse response of the receiver.

Recalling that this excitation of the impulse response can be attributed to the "extra" signal, and that the "extra" signal exists for all $Z > 0$, it is evident that this response will be elicited everywhere away from boresight. It is only in the pattern nulls that the "expected" signal is missing and we are able to see the impulse response alone.

III-11. ENVELOPE DETECTOR; F_7

In some systems the signal is taken through a second IF stage, but modeling that here would entail too much analysis and would not shed further light on the processes that occur in the receiver. Accordingly, it is assumed that the IF output, $F_{5,6}$, is passed through an envelope detector and the arrival time of the incoming signal is measured at the output from the detector. The envelope detector consists of a rectifier followed by a lowpass filter.

For most applications the envelope detector design need not receive much attention; usually design is devoted to reduction of the ripple to an acceptable level, and not much attention is given to the rise time of the detector. It is assumed here that in TOA system design the conflicting demands of rise time and ripple reduction should be balanced with care. The design of the envelope detector is discussed at some length in Appendix I.

The design adopted here consists of an ideal full-wave rectifier followed by four cascaded identical fully isolated stages of lowpass RC filtering. The RC time of each stage is 33.3 nanoseconds; that is, the cutoff frequency of each stage is 3/8 of the IF center frequency. In view of the ten MHz IF bandwidth, this choice of cutoff frequency is somewhat higher than usual and is made to shorten the rise time of the filter. The peak-to-peak ripple from a unit amplitude sine wave is about 1 millivolt--too little to be seen in a graphic display, but evident in computer listings.

When signal $F_{5,6}$ is passed through an ideal full-wave rectifier, the output is $|F_{5,6}|$. This is a nonlinear operation and the output is not an analytic function. Linear superposition no longer holds, and the absolute magnitude of $F_{5,6}$ is not equal to the sum of the magnitudes of the individual terms of $F_{5,6}$. In principle, the convolution of $|F_{5,6}|$ with the impulse response of the lowpass filter could be done piecewise, integrating from one zero of the function to the next. In fact, that is impractical because it would require locating the zeros of $F_{5,6}$. There is no feasible alternative but to use numerical integration in the computer to perform the convolution.

The computer has no difficulty, of course, in discarding the algebraic sign of $F_{5,6}$ to obtain $|F_{5,6}|$, and the algorithm to accomplish the numerical integration is simple enough. The impulse response of the four-stage lowpass filter is a simple function (see Appendix C). However, the calculation is inexact in two ways. First, as in all numerical integration, the algorithm moves along in finite increments of the independent variable (time) and a simple assumption is made concerning the shape of the function between those points. In this instance ordinary trapezoidal integration was used and the function was taken to be linear between the sample points. The second, more subtle, error arises from the fact that the true zeros of $F_{5,6}$ were not located, and the zeros usually landed between the sample points. That means that a short segment was assigned the incorrect algebraic sign. The effect is somewhat as if the rectifier were not ideal and varied more or less randomly from one zero crossing to another. Each calculation of a value of $F_{5,6}$ requires so much computation that it was impractical to use tiny time increments in the integration. The increment used to produce the results discussed here was 1/20 of the period of the center frequency of the IF strip (0.625 nanosecond). It is difficult to estimate the size of the errors caused by this approximate treatment, but they are thought to be inconsequential. To judge from internal evidence, the output of the envelope detector appears to have at least 120 dB of dynamic range. At this point the representation of the waveform as two segments is no longer meaningful. The output waveform from the envelope detector is designated F_7.

It is appropriate, before considering F_7 waveforms for various values of Z and the phases, to examine the performance of the envelope detector itself. Figure III-18 shows the oscillatory output of the IF

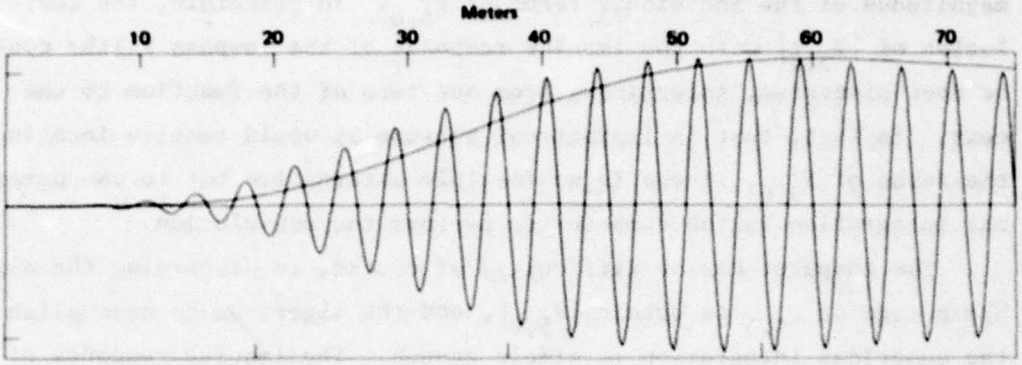


Fig. III-18 — Stepped carrier input to receiver ($\sigma = \psi = 0$)

strip in response to an input stepped carrier ($\psi = 0$) and, superposed, the corresponding output of the envelope detector. Figure III-19 shows

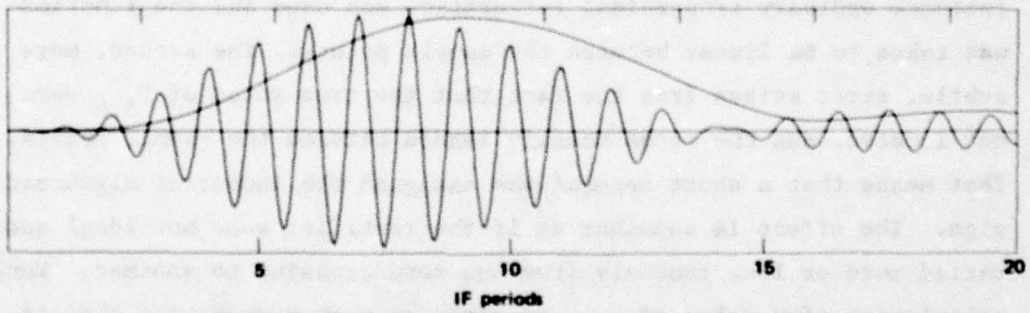


Fig. III-19 — Impulse delivered to receiver ($\sigma = 0$)

the impulse response of the IF strip and, superposed, the corresponding output of the envelope detector. All four curves are on the same time scale. The F_7 output is seen to lag by about 1.5 to 2 IF periods (depending on which portions of the curves are compared), but this gross throughput time shift is of no consequence. On the whole, F_7 appears to come pretty close to what is usually regarded as "the

envelope" of $F_{5,6}$. However, a careful comparison will show that no shift of F_7 to the left will cause it to be tangent to all the peaks of $F_{5,6}$.

III-12; EXAMPLES OF RECEIVER OUTPUT; F_7

This section shows an assortment of receiver output waveforms at various angular locations chosen to illustrate the effects that arise. Each figure is marked in units of time along the bottom edge and in corresponding distances along the top edge. For simplicity, all curves are for source phase $\psi = 0$. In each figure, two curves show the output for two values of phase σ that are 90 degrees apart. The numerical values of σ are of no particular interest and are not mentioned; the values tend to change rapidly with Z. The reason for choosing phases 90 degrees apart and the method used to pick the numerical values will be discussed later in connection with quantities D_0 and D_1. The two curves show approximately the extreme range over which the waveform changes for all possible choices of ψ and σ. The two curves shown in each case do not differ merely by a horizontal or vertical displacement; in every case the shape of the curve changes with phase. For that reason the true boundaries of the region occupied by all possible phases would not be actual waveforms.

The boresight performance is the natural choice to use as a baseline against which to compare the others. Except for amplitude change with directivity pattern, this is the receiver output that is generally "expected." Figure III-20 shows the output of the envelope detector, F_7, on the source axis. It can be seen that there is only a little dependence on phase. For the 1/2 peak definition of "arrival time" adopted in this study, the shift caused by phase is only about 0.2 meter on boresight. These curves scarcely differ from the stepped-carrier response of the receiver shown in Fig. III-18, and the output waveform is essentially a characteristic of the receiver alone with virtually no trait of the source to be seen.

Figures III-21 through III-30 show the receiver output waveforms at other angular locations. The dotted curve is the left-hand curve

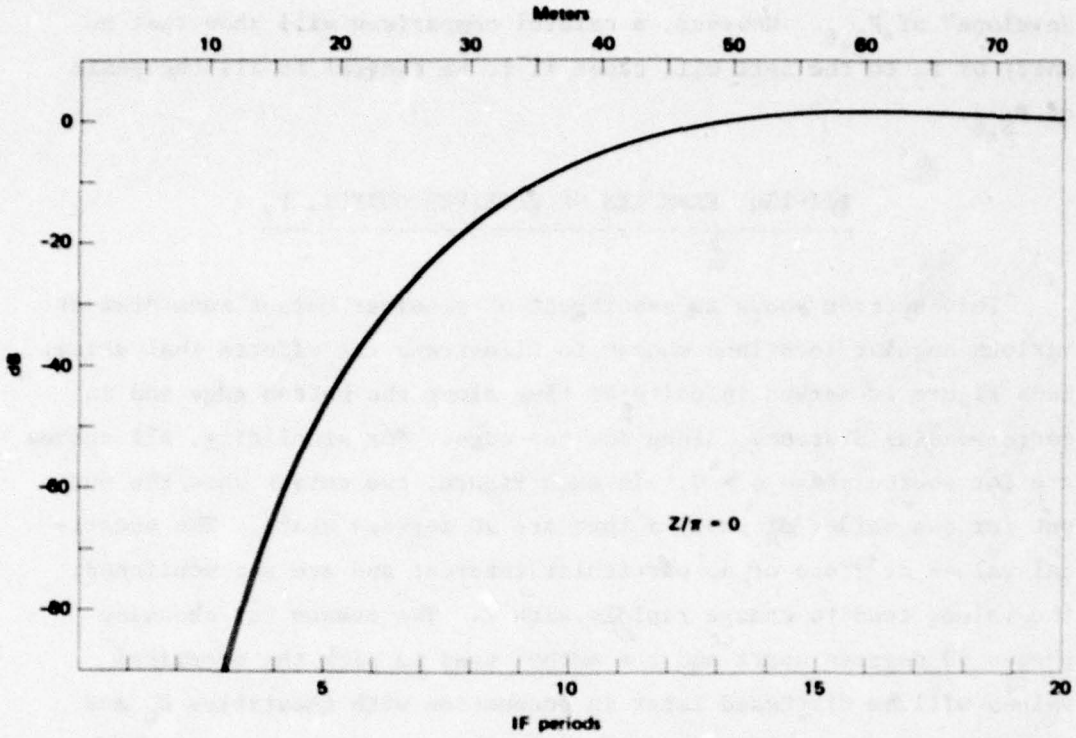


Fig. III-20 — Receiver output, F_γ, on the source axis

of Fig. III-20, repeated here for visual reference. In each case, the dotted curve has been adjusted in amplitude by sin Z/Z, so this is the "expected" output at each location. Of course, when Z/π is an integer the dotted curve is absent because sin Z = 0 and no signal is "expected."

Figures III-21, III-22, and III-23 show the rapid onset of phase effects as the first pattern null is approached. If the source aperture is 50.5 wavelengths wide (about 1 degree beamwidth) then the change of angle θ across these three figures is only 0.0045 degree! In Fig. III-21, the change of arrival time with phase is about 4.3 meters; in Fig. III-22 the change with phase is about 12.7 meters; in Fig. III-23 it is about 15.6 meters. It is true in every case, but especially obvious in Fig. III-23, that the shift with phase depends on the definition of what constitutes "arrival time." It should be noted that at locations near boresight the dotted curve tends to lie inside the swath swept out by phase effects. Thus the measured arrival time may be either later or earlier than the boresight value.

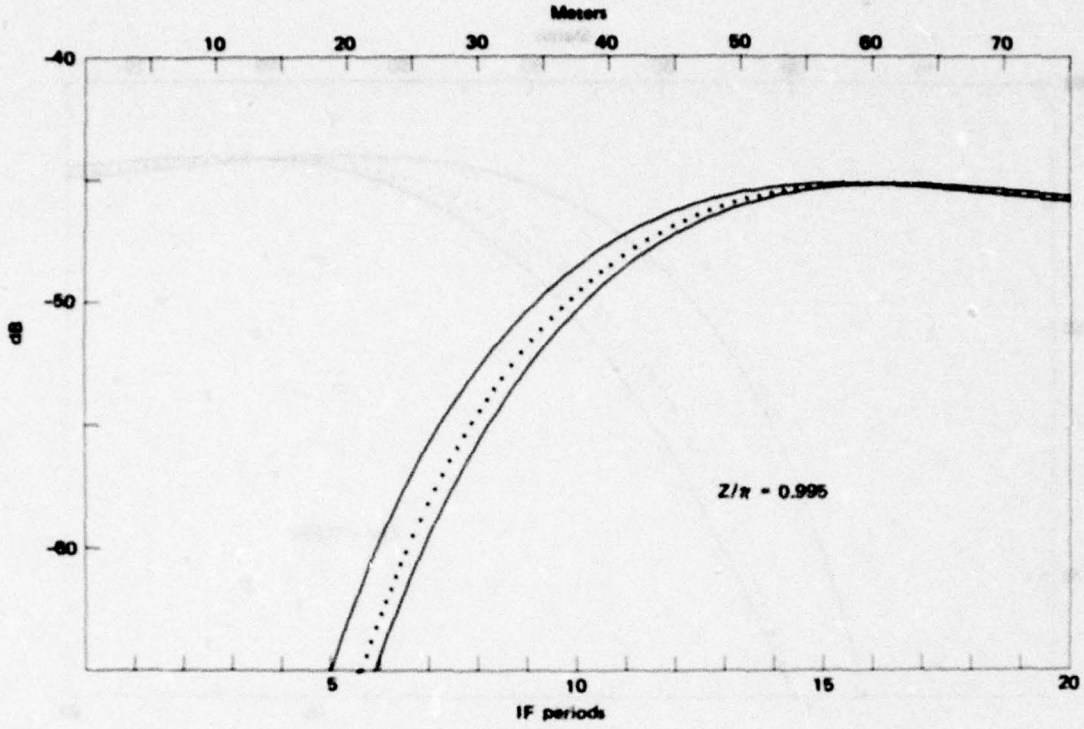


Fig. III-21 — Receiver output waveforms at $Z/R = 0.995$ compared with the "expected" output

Figure III-24 shows the output exactly in the first pattern null. (This is only 0.0011 degree from Fig. III-23 if $L/\lambda = 50.5$.) In the null the "expected" signal is absent and no interference between the stepped-carrier and impulse responses occurs. The two curves shown here have almost exactly the same shape, and that shape is essentially the impulse response of the receiver shown in Fig. III-19. If, as herein, the definition of arrival time relies on waveform and is independent of gross amplitude, then these two curves yield nearly the same arrival time. However, the waveform in the null is entirely different from the boresight waveform, and yields a different arrival time. That difference is the scale length S and is about 14.1 meters for this receiver design.

The two curves in Fig. III-24 are about 14.2 dB apart--the same separation as that shown in Fig. III-17. The difference in level reflects the change in the strength of the "impulse" delivered by $F_{5.6}$

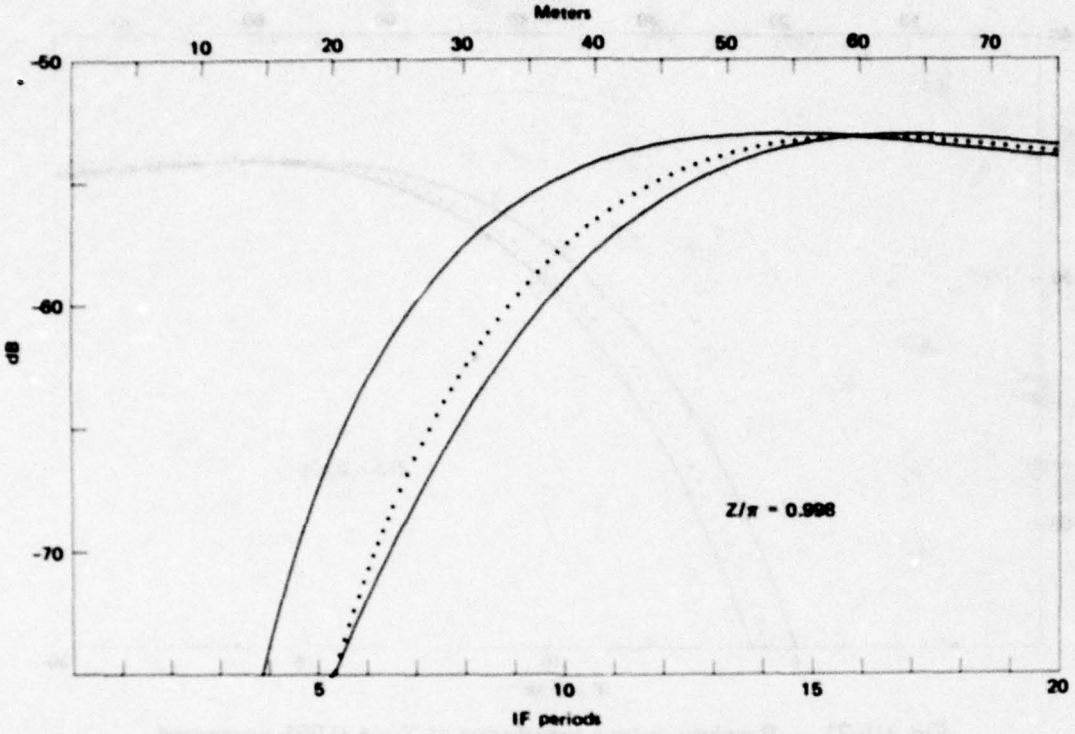


Fig. III-22 — Receiver output waveforms at $Z/\pi = 0.998$ compared with the "expected" output

as the phase is adjusted to maximize or eliminate $I_{3,4}$ (see Section III-8). The two outputs seen in Fig. III-24 arise from the two inputs shown in Fig. III-9; it is remarkable that two waveforms that differ so little can eventuate in two outputs that differ as much as 14 dB.

Figures III-25, III-26, and III-27 show output waveforms at locations roughly 45 degrees from the axis of a one-degree-beamwidth source. Figure III-25 shows the waveforms in the 35th pattern null. As in the first null, the waveform is almost exactly the impulse response of the receiver, and the two curves are displaced in amplitude but not in time; the arrival time scarcely changes with phase. It will be recalled that the amplitude of $I_{3,4}$ varies as

$$\frac{\sin\left(\frac{41}{40}\,Z\right)}{\left(\frac{41}{40}\,Z\right)}$$

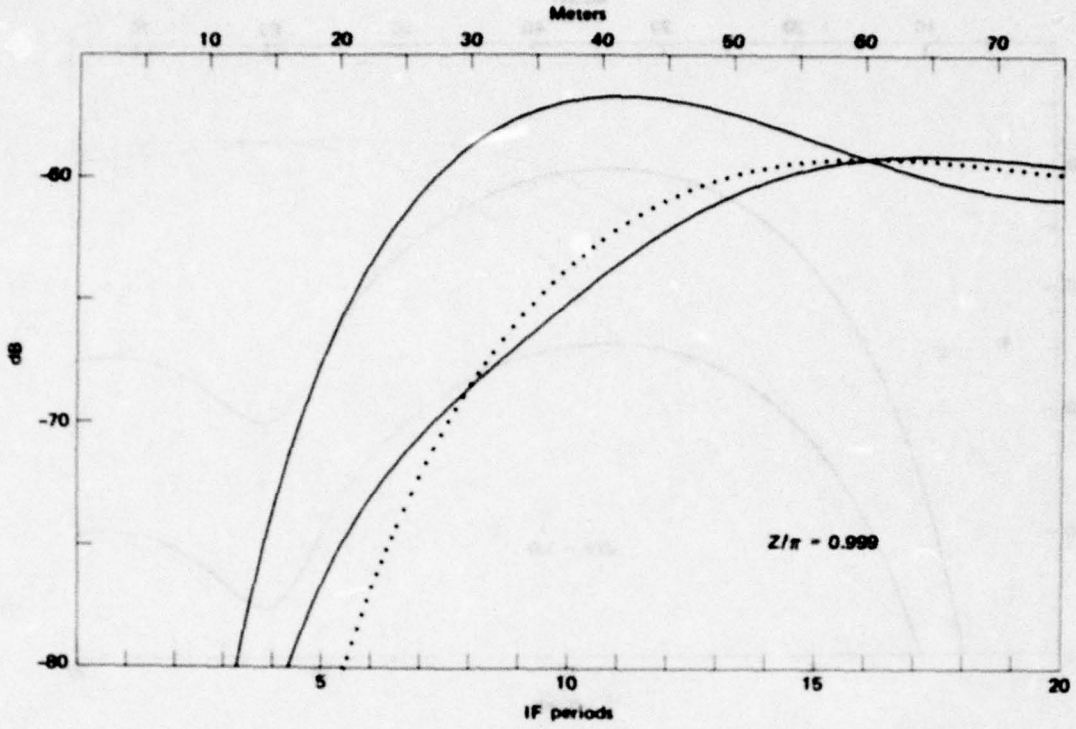


Fig. III-23 — Receiver output waveforms at $Z/\pi = 0.999$ compared with the "expected" output

Here that factor is -0.003395 whereas in the first null the factor is -0.024365--about 17.1 dB higher. The low-frequency energy content of $F_{3,4}$ is a lesser fraction of the total energy in this null than in the first null, and maximizing or eliminating $I_{3,4}$ has little effect on the strength of the "impulse." The two curves in Fig. III-25 are only 1.5 dB apart. On the other hand, the lower curve ($I_{3,4}$ eliminated) is about 9 dB higher than the lower curve in Fig. III-24. In this null the duration of $F_{3,4}$ is considerably longer and delivers more energy.

Figures III-26 and III-27 show waveforms in the 35th side lobe near the null. For the one-degree-beamwidth source, these two positions are 0.047 degree apart--roughly ten times the angular span in Figs. III-21, III-22, and III-23. The dependence on angle θ is still steep, but much less so than near boresight. The width of the swath swept out by phase change is fairly narrow here, in keeping with the fact that the strength of the excitation of the impulse response

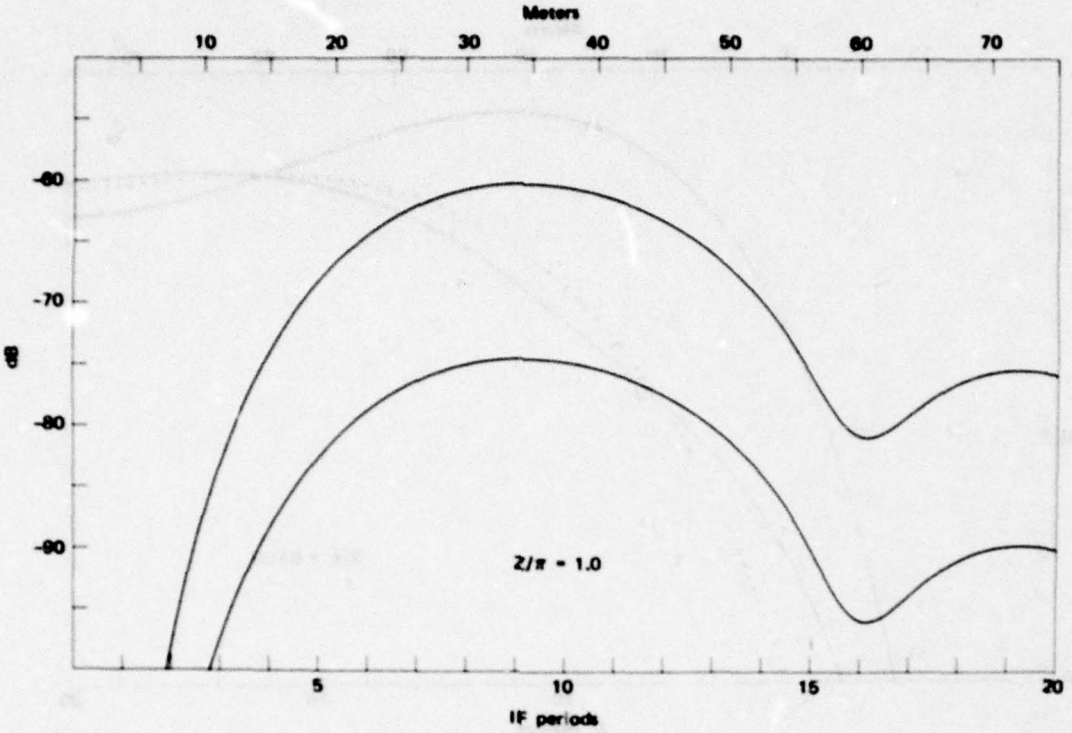


Fig. III-24 — Receiver output waveforms in first pattern null of the source. No output is "expected."

varies only a little with phase. However, the entire swath shifts to the left as the null is neared, leading to early arrival times for all phases. These swaths are less than two meters wide, but the arrival time changes by 9.5 meters (average) between the two locations.

The waveforms at $Z/\pi = 35.1$ (not shown) are fairly near the dotted line and the phase swath is narrow. Presumably a system designer would prefer to regard them as acceptable. The peak level reached at $Z/\pi = 35.1$ is only five dB above the peak shwon in Fig. III-27, and roughly 13 dB above the peak in Fig. III-26. It would seem to be difficult to discriminate dependably against these effects by signal level because level shifts of a half dozen dB and more occur often for many reasons.

Figures III-28, III-29, and III-30 show output waveforms at about 90 degrees from a one-degree-beamwidth source. It is doubtful that the approximate Kirchoff description of the antenna is meaningful at

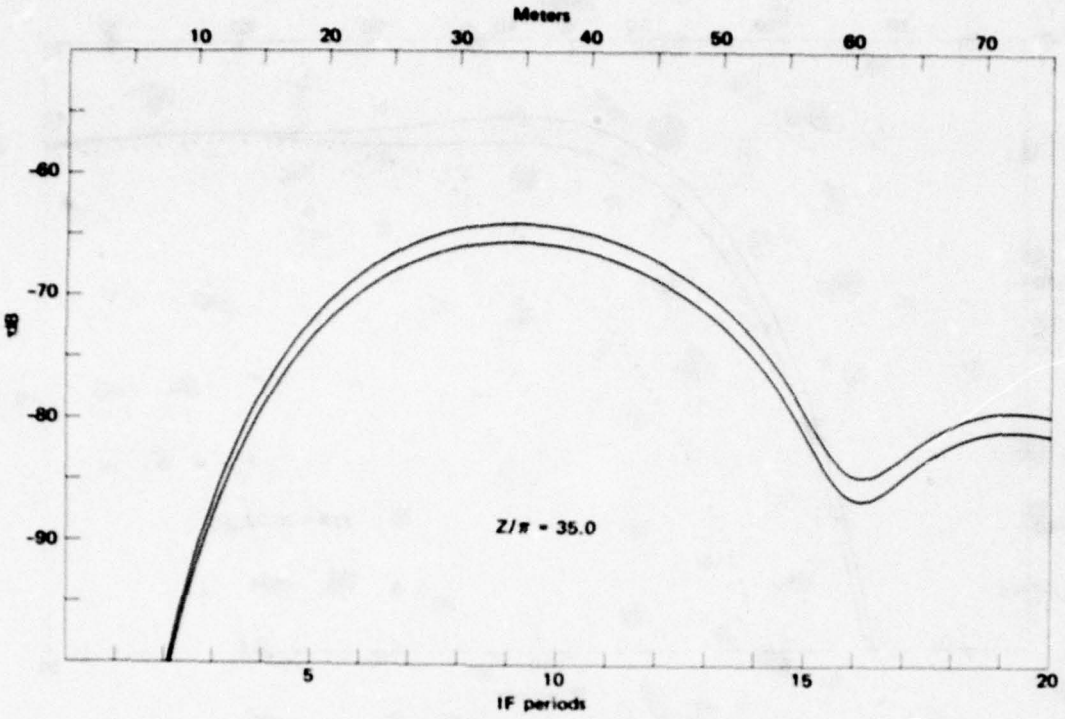


Fig. III-25 — Receiver output waveforms in 35th pattern null of the source. No output is "expected."

such wide angles; they are shown only because no better guess can be offered.

Figure III-28 shows the output in the 50th pattern null. As in the other nulls, the two curves are displaced in amplitude but not in time, and arrival time changes only about 0.2 meter with phase. These two curves are almost indistinguishable from the curves in the 35th null; at large angles the output is much the same in all the pattern nulls. The peak output level in the 50th null is 20 dB below the steady-state level at the top of the 50th lobe, whereas the peak level in the first null is at least 46 dB below the steady-state level at the top of the first side lobe. At large values of Z, it would not be easy to discriminate against even the null signals on the basis of signal level.

Figures III-29 and III-30 show output waveforms in the 50th side lobe near the 50th null. They are similar to the corresponding outputs

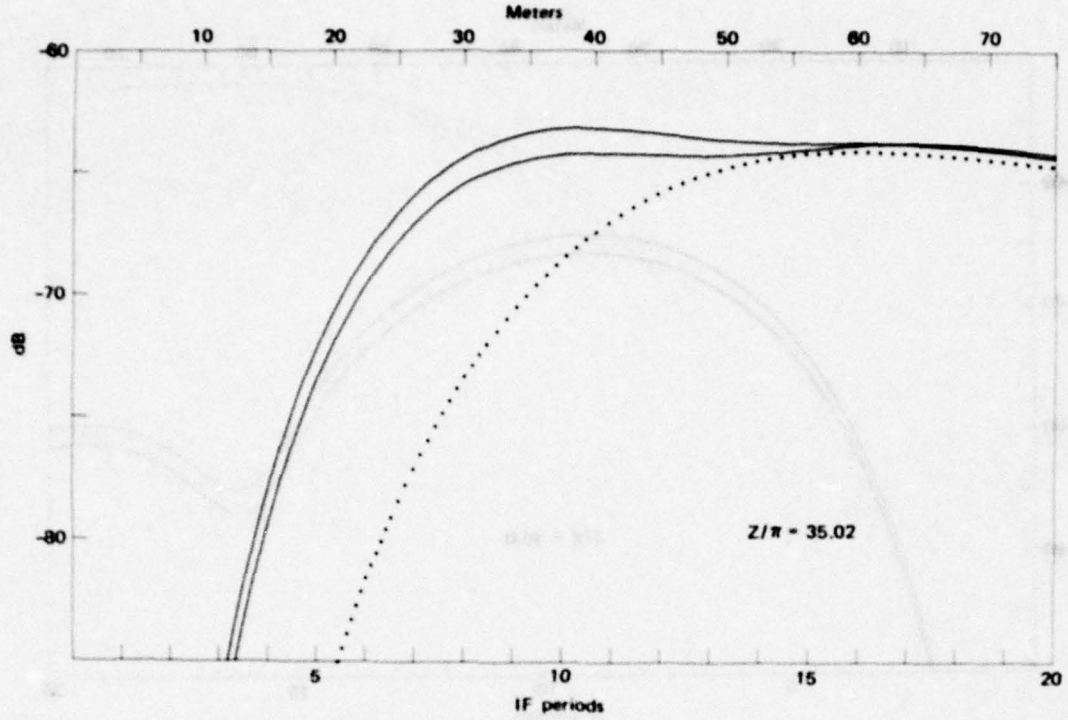


Fig. III-26 — Receiver output waveforms at $Z/\pi = 35.02$ compared with the "expected" output

near the 35th null shown in Figs. III-26 and III-27. The swath swept out by phase effects is fairly narrow, but the swath shifts to early arrival as the null is approached. The angular span between Figs. III-26 and III-27 is 0.25 degree, roughly five times the span between $Z/\pi = 35.02$ and $Z/\pi = 35.05$.

These ten examples are not an exhaustive display of all the phenomena, but they suffice to illustrate the major trends. The output waveform seen in each pattern null is distinct from the other waveforms--not so much, perhaps, in the initial rise as in the fact that the curve passes through a pronounced peak and then dips into a deep minimum. It seems plausible that a filter matched to the boresight waveform could discriminate dependably against the null waveform (or, after identifying it, make allowance for the earlier arrival). That filter in itself would not help this system much because the angular interval over which this particular shape appears is very narrow. Only slightly out of the null the waveform changes.

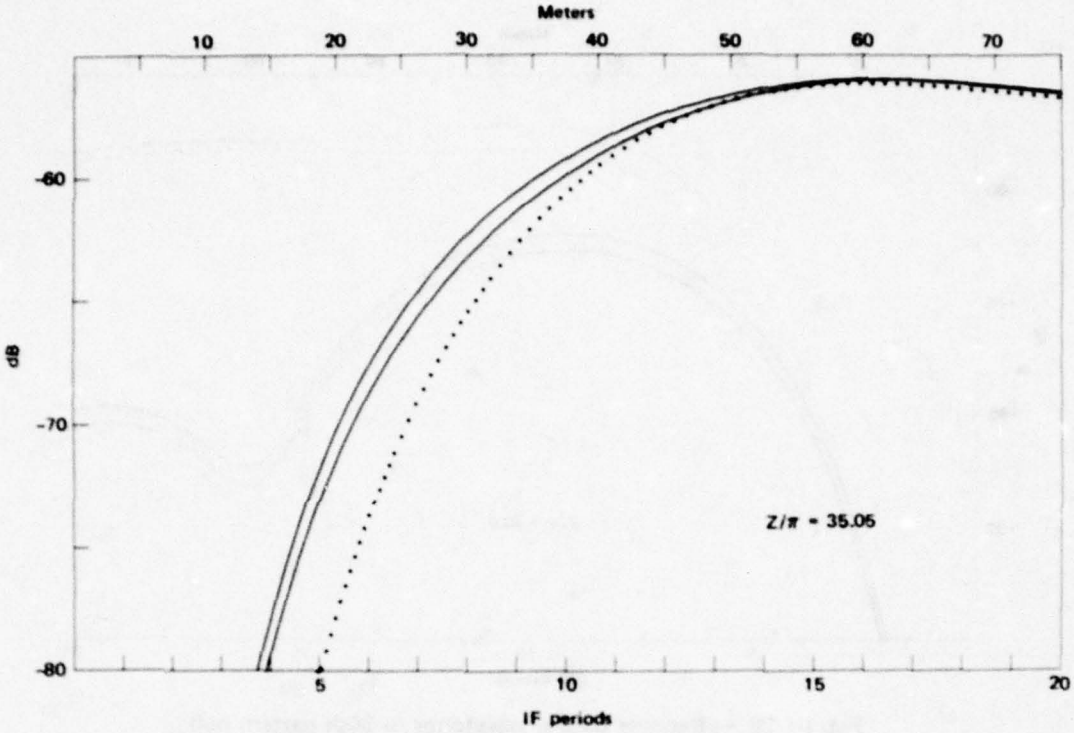


Fig. III-27 — Receiver output waveforms at $Z/\pi \approx 35.05$ compared with the "expected" output

All the waveforms shown in these figures are different, but in most instances the differences are not great. It is not evident that a matched filter could, as a practical matter, discriminate between those that are seriously influenced and those that are scarcely influenced by these antenna effects. As was remarked in Part II, remedies to mitigate these effects might better be found in system operational arrangements than in hardware devices built into the receivers.

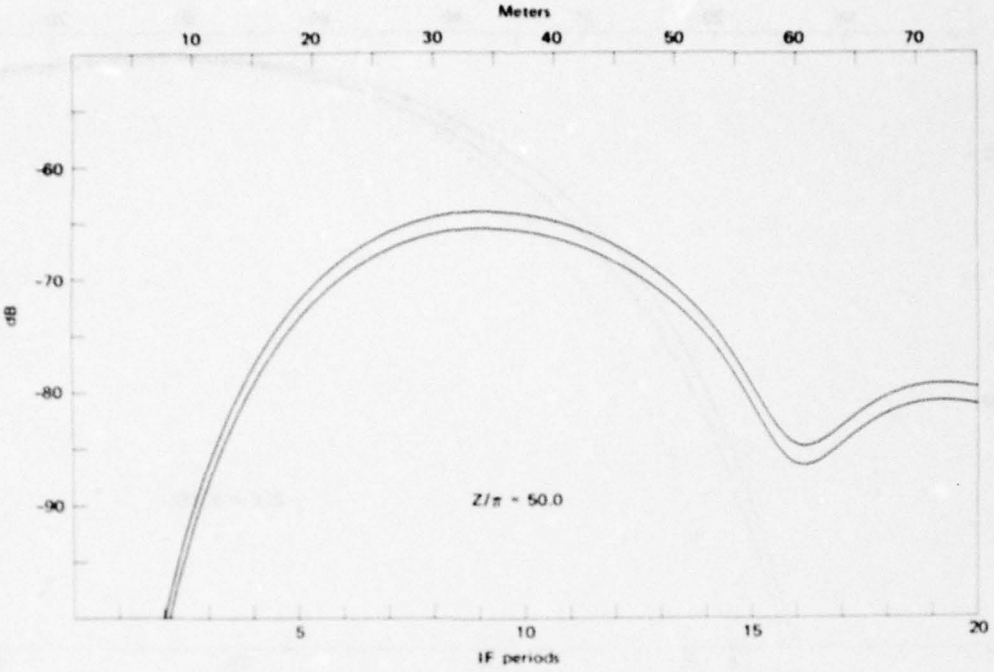


Fig. III-28 — Receiver output waveforms in 50th pattern null of the source. No output is "expected".

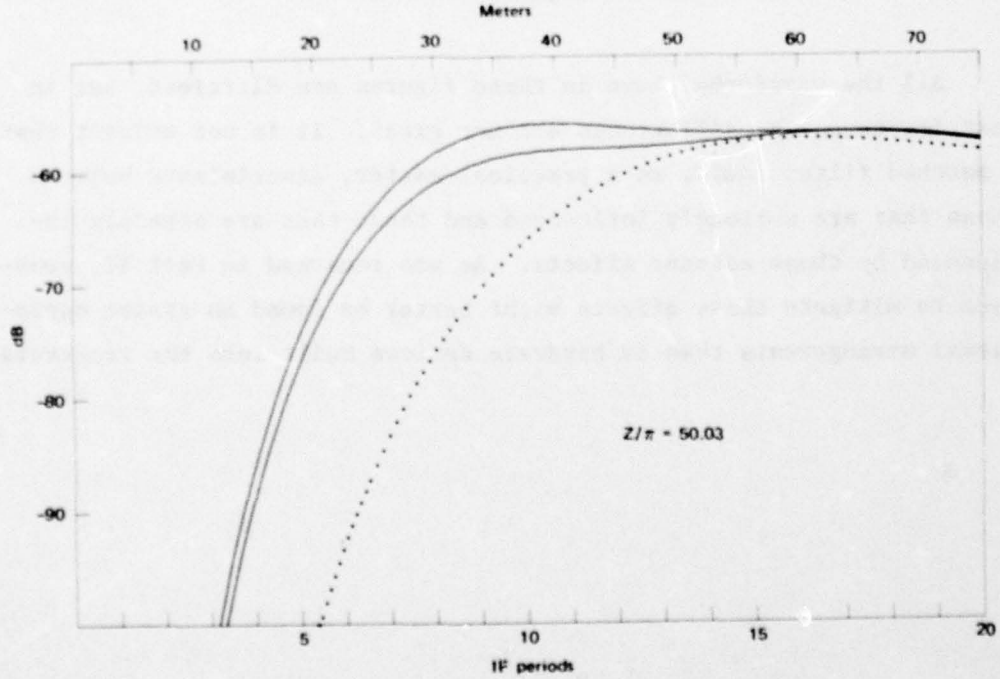


Fig. III-29 — Receiver output waveforms at $Z/\pi = 50.03$ compared with the "expected" output

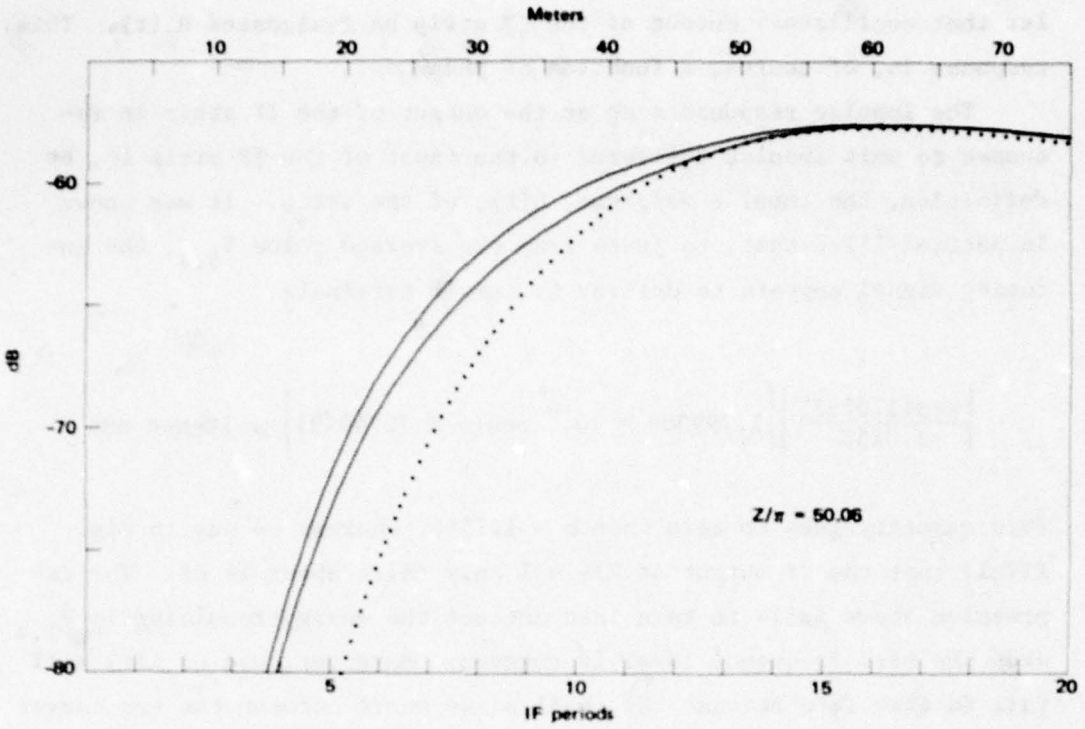


Fig. III-30 — Receiver output waveforms at $Z/\pi = 50.06$ compared with the "expected" output

III-13. SYNTHESIS OF THE OUTPUT WAVEFORM

The analysis described in the previous sections, together with the waveform shapes shown in the examples, provide a persuasive case that the output waveforms can be understood to result from the mutual interference of the two characteristic transient responses of the receiver: the impulse response and the stepped-carrier response. However, that model can be tested directly by attempting to synthesize one of the output waveforms, including the phase effects, from the two transient responses themselves. Such a test omits entirely the effects of the finite rise time of the source pulse.

To conduct that test, the output of the IF strip in response to an input stepped carrier, $u(0) \sin(\omega_0 t)$, at the receiver front end was calculated. Because the examples considered previously used $\psi = 0$, that phase is used here. For convenience in this discussion,

let that oscillatory output of the IF strip be designated $H_S(t)$. This response is, of course, a function of phase σ.

The impulse response seen at the output of the IF strip in response to unit impulse delivered to the input of the IF strip is, by definition, the impulse response, $h(t)$, of the strip. It was shown in Section III-8 that, to judge from the average value $I_{3,4}$, the incoming signal appears to deliver to the IF terminals

$$\left|\frac{\sin(1.025Z)}{1.025Z}\right|\left|1.799369 \times 10^{-9} \cos(\sigma-2.7274099)\right| \text{ volt-seconds}$$

This quantity goes to zero when $\sigma = 1.1566$, whereas we saw in Fig. III-17 that the IF output at $Z/\pi = 1$ only falls about 14 dB. The expression above fails to take into account the energy remaining in $F_{3,4}$ when the zero frequency level is removed. Moreover, use of $h(t)$ will fail to take into account the small phase shift between the two curves shown in Fig. III-17. Thus, a test based upon the pure responses $H_S(t)$ and $h(t)$ cannot be expected to reproduce the correct receiver response precisely.

To take account of the expected excitation levels of the two responses, the signal

$$\left|\frac{\sin Z}{Z}\right|H_S(t) + \left|\frac{\sin(1.025Z)}{1.025Z}\right|\left|1.799369 \times 10^{-9} \cos(\sigma-2.7274099)\right|h(t)$$

was formed. That is, of course, an oscillatory signal present at the IF output. The signal was put through the numerical integration routine used to obtain output F_7 from $F_{5,6}$ to produce the synthesized output from the envelope detector.

The most drastic deformation of waveform shown in the examples seems to be that at $Z/\pi = 0.999$ shown in Fig. III-23, so this value of Z was used in the synthesis. The same two phases, $\sigma = -50$ and $+40$ degrees, were used. The resulting synthesized waveforms are shown in Fig. III-31, which is drawn on the same scales as Fig. III-23. The

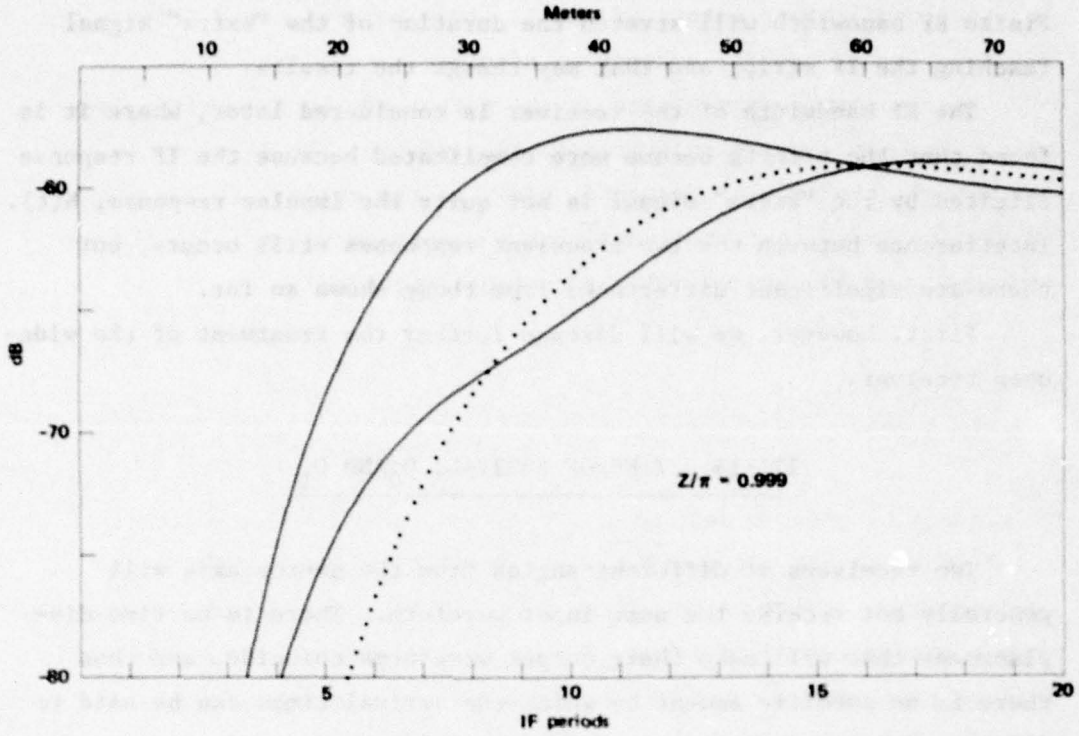


Fig. III-31 — Synthesized output waveform caused by interference between the exact impulse and stepped-carrier responses of the receiver. Compare with Fig. III-23.

synthesized waveforms are very nearly, but not quite, the same as the F_7 waveforms. Thus, except for minor details, the receiver output waveforms can be understood to result from the interference between these two theoretical receiver responses. It is primarily in the relative numerical strength of the equivalent impulse that source characteristics affect the output waveform. That relative strength is dependent on the Kirchoff approximate assumptions used to describe the source behavior, and that is the reason why the results obtained here are not entirely reliable.

Finally, it is important to observe that the analysis so far has considered a wide open front end in the receiver. That front end preserves the brevity of the "extra" signal and ensures that the "extra" signal excites a response that is nearly the true impulse response. Now that it is clear that this is central to the results, it is obvious that the choice of a wide open front end must be reconsidered.

Finite RF bandwidth will stretch the duration of the "extra" signal reaching the IF strip, and that may change the results.

The RF bandwidth of the receiver is considered later, where it is found that the effects become more complicated because the IF response elicited by the "extra" signal is not quite the impulse response, h(t). Interference between the two transient responses still occurs, but there are significant differences from those shown so far.

First, however, we will discuss further the treatment of the wide-open receiver.

III-14. TIME OF ARRIVAL; D_0 AND D_1

Two receivers at different angles from the source axis will generally not receive the same input waveform. There is no time displacement that will make their output waveforms coincide, and thus there is no specific amount by which the arrival times can be said to differ. Various definitions of "arrival" can be devised, and they will generally give diverse values for the time difference between the two receivers. No particular choice of definition is the correct one, and it is not evident that a criterion for selecting the best definition can be found. In a noise-free environment and in the absence of difficulties caused by multipath, perhaps a detector circuit matched to the stepped-carrier response of the receiver would be best. However, it was remarked earlier that, for practical reasons, the matched detector may not be optimum in this case.

During the study detailed consideration was given to three definitions of the time at which "arrival" is said to occur:

- The time at which the detector output, F_7, first equals a fixed value
- The time at which F_7 first equals one-half the ultimate steady-state value
- The time at which F_7 first equals one-half the value reached at the highest subsequent peak.

For brevity these will be termed "fixed level," "sinZ/2Z," and "peak/2."

The fixed-level method fails, of course, to take account of the changing signal level with changing position in the directivity pattern of the source. (Presumably the several receivers could make the small adjustment needed to correct fairly well for diverse source-receiver distances.) However, it can be seen in Figs. III-21 through III-30 that in all cases the early portion of F_7 rises rapidly, and at sufficiently low level the time differences at different angles are not great. Indeed, for infinitesimal signal level they all start at time $-Z/\omega_0$, and that differs with angle by only half the width of the source aperture. Obviously, the fixed-level method would provide quite small timing errors if the level could be set low enough.

Numerous calculated arrival times were obtained for a level fixed at -80 dB. Even if the receivers are as far as 100 km or so from a typical radar, they might be able to set the level this low and still be well above noise. For higher power sources, shorter ranges, and so on, even lower levels--possibly -90 dB or -100 dB might be possible. The situation improves rapidly below -80 dB. TOA system errors resulting from the -80 dB choice of fixed level were generally greater than errors resulting from the peak/2 method. In special circumstances wherein lower fixed levels could be used, the fixed-level method might be advantageous.

The fixed-level method can be used in real time, but the sinZ/2Z and peak/2 methods necessitate some form of time delay because they require a retrospective determination of the time at which the signal attained a fraction of a later value. For signals on the source axis or near the tops of the side lobes these two methods yield nearly the same arrival time because the peak (overshoot) output is only a decibel or so higher than the steady-state (sinZ/Z) output. However as pattern nulls of the source are approached, the sinZ/Z level falls and the sinZ/2Z definition leads to progressively earlier arrival. In the pattern nulls the sinZ/2Z arrival time is undefined but approaches the value $-Z/\omega_0$. On the whole, the sinZ/2Z method appears to be somewhat inferior to the peak/2 method, and it requires a longer time delay.

Thus, although numerous sinZ/2Z values were obtained, along with -80 dB values and peak/2 values, most attention was devoted to the peak/2 arrival times. The following discussion uses the peak/2 definition of arrival.

The dependence of waveform, and hence of arrival time, on phases σ and ψ, as well as on Z, is complex in detail. A thorough exploration of all possible phase combinations at all values of Z would be forbiddingly expensive. We thus adopted an approximate description of the phase dependence that would necessitate less computation. To explain the approximation it is helpful first to consider a more extensive display of arrival times at several angular locations.

Table III-1 shows a 4x4 array of computed peak/2 arrival times, expressed in meters, for signals received on the source axis. The

Table III-1

4X4 ARRAY OF PEAK/2 ARRIVAL TIMES (IN METERS) ON SOURCE AXIS

(Wide-open front end; Q_3 absent; $Z/\pi = 0$)

	ψ, degrees			
σ, degrees	0	45	90	135
0	34.638486	34.993929	34.871192	34.679946
45	34.672875	34.634592	35.002787	34.879024
90	34.863162	34.657495	34.636044	35.002304
135	34.993445	34.855301	34.664570	34.639935

array shows all combinations of the two phases taken in 45 degree increments. A change of 180 degrees in either phase angle merely changes the algebraic sign of $F_{5,6}$, and that sign is discarded in the envelope detector. Consequently the 4x4 array shows a uniformly spaced sampling across the entire phase space.

The six-decimal point precision shown in the table is of no practical significance, but such precision is helpful when trying to fit the values with a mathematical function. However, the finite step size used in the numerical integration to obtain F_7, together with the small ripple present in F_7, introduces small "errors" in the listed

values. It is questionable just how many decimal places are really useful for curve-fitting purposes, but the number is thought to be three or four.

Evidently the dependence of arrival time on phase at Z = 0 is small. The average of the 16 values is 34.792818 meters and the standard deviation is 0.139825 meter. The largest difference in the array is 0.368195 meter.

Tables III-2, III-3, and III-4 show such 4x4 arrays for other values of Z. The format in these arrays is identical to that in Table

Table III-2

4X4 ARRAYS OF PEAK/2 ARRIVAL TIMES (IN METERS)

AT THE EDGE OF THE MAIN LOBE OF THE SOURCE

(Wide-open front end; Q_3 absent)

	Z/π = 0.998		
31.202968	27.440028	31.015130	37.577752
37.483564	31.000005	27.356856	30.968294
31.096299	37.636040	31.129598	27.434619
27.518726	31.142577	37.732765	31.340502
	Z/π = 0.999		
22.425309	23.133843	27.719184	39.114951
38.775132	22.257131	23.082991	27.615210
27.769438	39.092119	22.257235	23.090210
23.145329	27.878482	39.410108	22.425122
	Z/π = 1.000		
20.525235	20.573838	20.789771	20.386740
20.407823	20.530336	20.579614	20.820606
20.854153	20.380926	20.514724	20.601619
20.596175	20.825098	20.367973	20.509482

III-1 but the listing of angles σ and ϕ is omitted to facilitate reading. The reader may wish to refer to the waveforms shown in Figs. III-21 through III-30. Those figures pertain to the first column of each array (ϕ = 0) but, in most cases, show other values of σ.

Table III-3

4X4 ARRAYS OF PEAK/2 ARRIVAL TIMES (IN METERS)

AT THE EDGE OF THE 35*th* SOURCE SIDE LOBE

(Wide-open front end; Q_3 absent)

	Z/π = 35.00		
20.748390	20.465284	20.439796	20.712423
20.705364	20.740457	20.472398	20.444742
20.439174	20.700989	20.761959	20.485039
20.477882	20.434164	20.708065	20.769906
	Z/π = 35.02		
22.522539	21.872016	21.682884	21.633378
21.629059	22.543465	21.892858	21.694037
21.683565	21.625055	22.566589	21.895982
21.875109	21.672406	21.629392	22.545577
	Z/π = 35.05		
32.296843	31.219979	30.548155	31.713535
31.724874	32.302779	31.240303	30.573314
30.593050	31.741951	32.334912	31.278845
31.258517	30.568260	31.730269	32.329004

With 16 numerical values available, 16 undetermined constants can be used to fit the array with appropriate functions of σ and ψ, and there is a limitless assortment of such fitting functions that will fit these 16 values perfectly when 16 constants are adjusted. To describe the phase dependence more fully would require more samples--say, 64 values in an 8x8 array--but even that description would be merely a better approximation. Such an effort--especially if carried out for each value of Z--would be costly and not worthwhile. The dependence of arrival time on these phases is complicated in detail and not worth exploring.

It was decided to settle for a simple approximate description of the phase dependence that would require computation of only three values of arrival time instead of 16 values. The approximation stems from the tendency for dependence primarily on $\sigma - \psi$ and on the near-sinusoidal dependence in each row and column:

Table III-4

4X4 ARRAYS OF PEAK/2 ARRIVAL TIMES (IN METERS)

AT THE EDGE OF THE 50*th* SOURCE SIDE LOBE

(Wide-open front end; Q_3 absent)

	Z/π = 50.00		
20.450801	20.546321	20.668325	20.608712
20.598602	20.444205	20.552596	20.674430
20.660747	20.589200	20.454426	20.561301
20.555052	20.654595	20.599334	20.461043
	Z/π = 50.03		
21.835905	21.812685	23.080019	22.398526
22.358199	21.824243	21.795170	23.054583
23.036576	22.340376	21.835753	21.789064
21.806546	23.061664	22.380604	21.847425
	Z/π = 50.06		
28.702666	30.023009	31.131634	29.854434
29.823451	28.667399	29.995799	31.113755
31.074670	29.784181	28.643617	29.975343
30.002280	31.092290	29.814066	28.678417

$$\text{arrival time} \approx D_0 + D_1 \cos[2(\sigma - \psi + \gamma)].$$

Here D_0, D_1, and γ are constants that can be determined from three values of arrival time. The angle γ is of no great interest,* so the phase dependence may be described approximately by D_0 and D_1.

In this description, D_0 is the value of arrival time that would be obtained by averaging over all possible combinations of σ and ψ, and the extreme values that would occur would be $D_0 - D_1$ and $D_0 + D_1$. Neither of these statements concerning the average or the extreme is exact,

*The values of σ chosen for Figs. III-20 through III-30 were $-\gamma$ and $-\gamma + \pi/2$. Thus, to the extent that this simple model describes the arrival time, the waveforms shown in those figures yield the greatest and least peak/2 arrival times at those values of Z. The extrema of the -80 dB and sinZ/2Z arrival times would occur at other values of σ.

but in most cases the errors are not great. (For $Z/\pi = 0.999$, $D_0 - D_1$ is less than the true minimum by 11.1 meters, but this is a particularly extreme case.)

It should be noted that this approximate description predicts that the distribution of arrival times with random phases will exhibit the frequency distribution of a sinusoid. In other words, times near the extrema (near $D_0 + D_1$ or $D_0 - D_1$) will be equally likely and most prevalent, and times near the average will be considerably less frequent. The distribution is bimodal with cusps at the extrema and a minimum at the average.

Once this simplified description is chosen, it is immaterial which three combinations of σ and ψ are chosen, provided the choice yields a fair sample of the frequency space. All the values of D_0 and D_1 discussed herein (including those taken up later with a tuned front end) were calculated from $\psi = 0$, $\sigma = 0$, 45 and 90 degrees. Other choices would have led to slightly different values of D_0 and D_1, neither better nor worse than these.

We tabulate, for illustrative purposes, the true averages and extrema of the 4x4 arrays that were given earlier, together with the values of D_0 and D_1 obtained from the first three entries in column one of each array:

Z/π	average	minimum	maximum	D_0	$D_0 - D_1$	$D_0 + D_1$
0	34.792818	34.634592	35.002787	34.750824	34.614091	34.887557
0.998	31.817233	27.356856	37.732765	31.149634	24.815479	37.483789
0.999	28.074487	22.257131	39.410108	25.097374	11.161055	39.033693
1.000	20.579007	20.367973	20.854153	20.689694	20.363354	21.016034
35.00	20.594127	20.434164	20.769906	20.593782	20.403114	20.784450
35.02	21.935244	21.625055	22.566589	22.103052	21.470092	22.736012
35.05	31.465912	30.548155	32.334912	31.444947	30.548238	32.341656
50.00	20.567481	20.444205	20.674430	20.555774	20.442400	20.669148
50.03	22.266084	21.789064	23.080019	22.436241	21.830854	23.041628
50.06	29.898563	28.643617	31.131634	29.888668	28.700874	31.076462

The shortcomings of the $D_0 \pm D_1$ description are not, on the whole, serious. It will suffice to indicate reasonably well what consequences ensue for a whole TOA system and whence they arise. Note that, among the maximum and minimum values listed above, we find a set that would produce an error of 51 meters in locating the source with a fairly reasonable TOA system layout (39.410 and 31.132 meters in the two outer receivers about 81 degrees apart, and 20.434 meters in the middle receiver approximately halfway between them).* The arrival time disparities above occur in a single receiver and do not indicate the system error arising from the interplay of these disparities among several receivers.

III-15. DEPENDENCE OF D_0 AND D_1 ON Z

It is tempting to argue that the angular intervals near the source pattern nulls over which the D_0 and D_1 values differ significantly from the boresight values are so narrow as to be negligible. Such an argument is weakened, however, by the fact that there may be a great many such nulls, and also by the fact that the TOA system includes several receivers. For example: if $L/\lambda = 50.5$ (one-degree-source beamwidth) then the angular span from $Z/\pi = 34.98$ to 35.02 is only 0.06296 degrees--a very small span. However, there are 50 pattern nulls between $Z = 0$ and $Z/\pi = 50.5$, and 50 times this span amounts to 3.5 percent of 90 degrees. This suggests a probability of 0.035 that any one receiver might be in such a span, and perhaps a probability of 0.10 that one out of three receivers would be. Such an event might lead to an error of some 20 meters in the calculated source location, depending on the system configuration, and 10 percent probability of that much error may not be insignificant. Consequently, the dependence of D_0 and D_1 on Z deserves more careful consideration.

---

*The few values of Z in the text table do not present a case that is symmetric with 90 degrees subtense. If the source width is taken to be $L/\lambda = 50.5$, then in this example the subtense of 81.298 degrees is made up of 42.740 and 38.558 degrees. This geometry could arise if the source happended to be to one side of the axis of symmetry, and a bit too far away for 90 degrees subtense.

The complicated dependence of D_0 and D_1 on Z is best shown graphically, as is done in Fig. III-32, III-33, and III-34. These figures

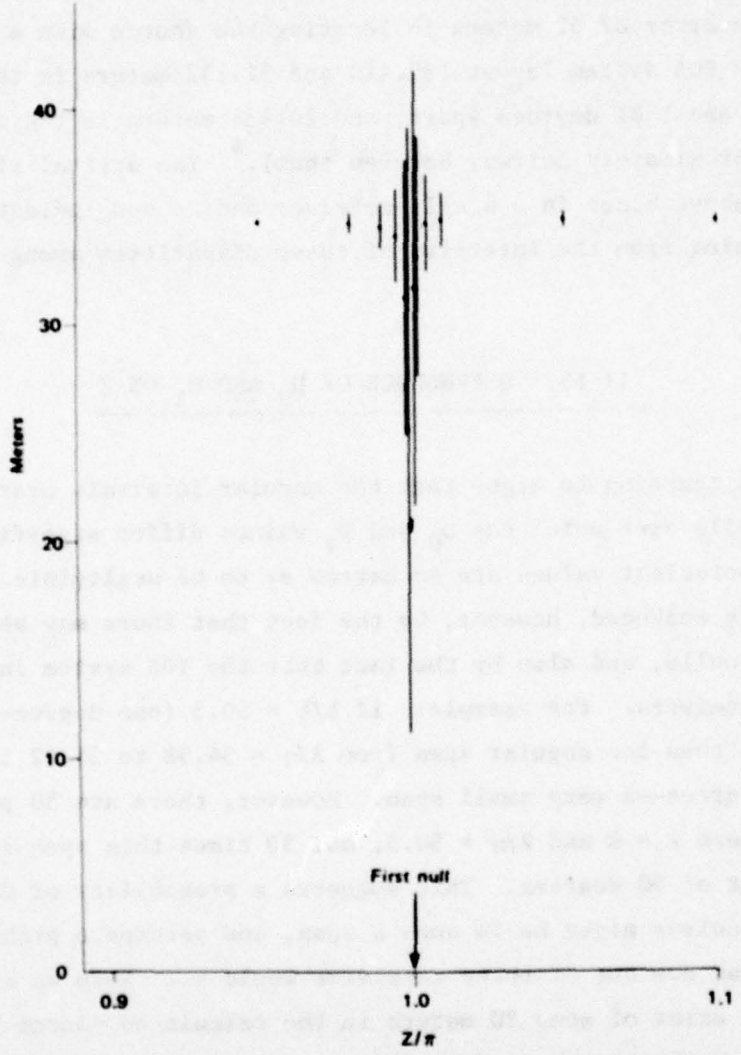


Fig. III-32 — Arrival time expressed as $D_0 \pm D_1$ meters (Q_3 absent); vicinity of first null

illustrate the behavior near the source axis, near $Z/\pi = 35$, and near $Z/\pi = 50$. In each figure D_0 and D_1, expressed in meters, are shown for discrete values of Z/π. At each value of Z/π, a vertical line extends from $D_0 - D_1$ upward to $D_0 + D_1$; thus such a line shows the span

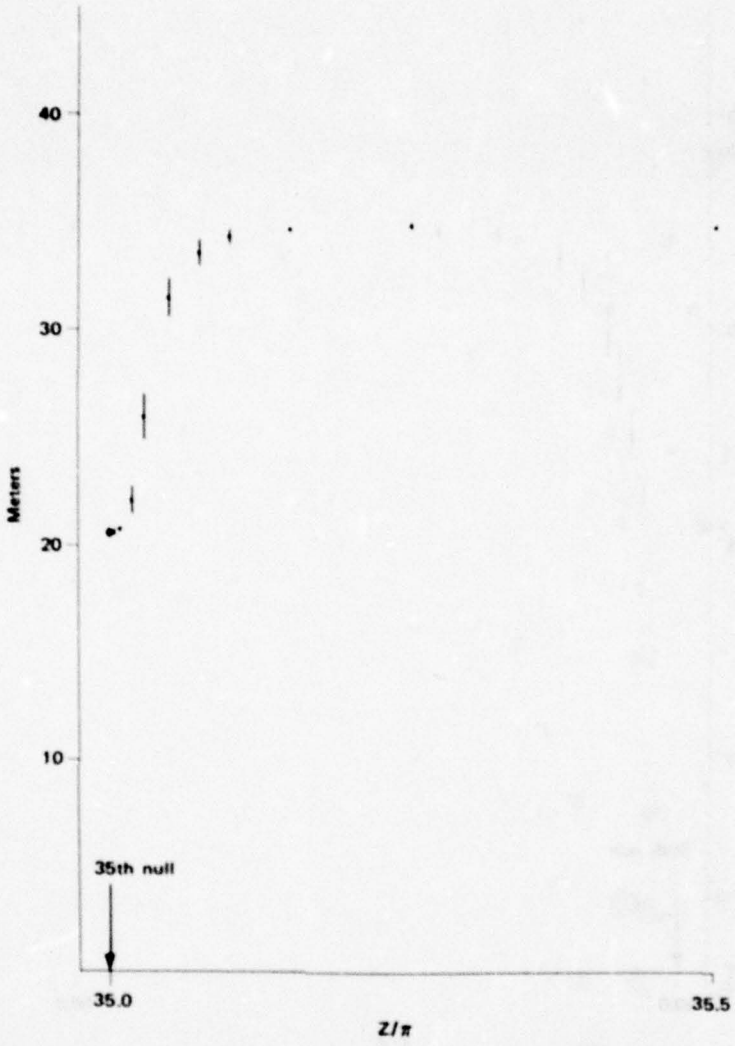


Fig. III-33 — Arrival time expressed as $D_0 \pm D_1$ meters (Q_3 absent); 35th side lobe

of arrival times swept out at that receiver position by all possible phase combinations (insofar as D_0 and D_1 manage to portray that span of times). A heavy dot at the midpoint of each such line shows the average value, D_0. All three figures are drawn on the same scale of time expressed in meters, but the scale of the Z/π axis is not the same. The latter scale change is not important because the corresponding scale of angle θ changes greatly among these figures.

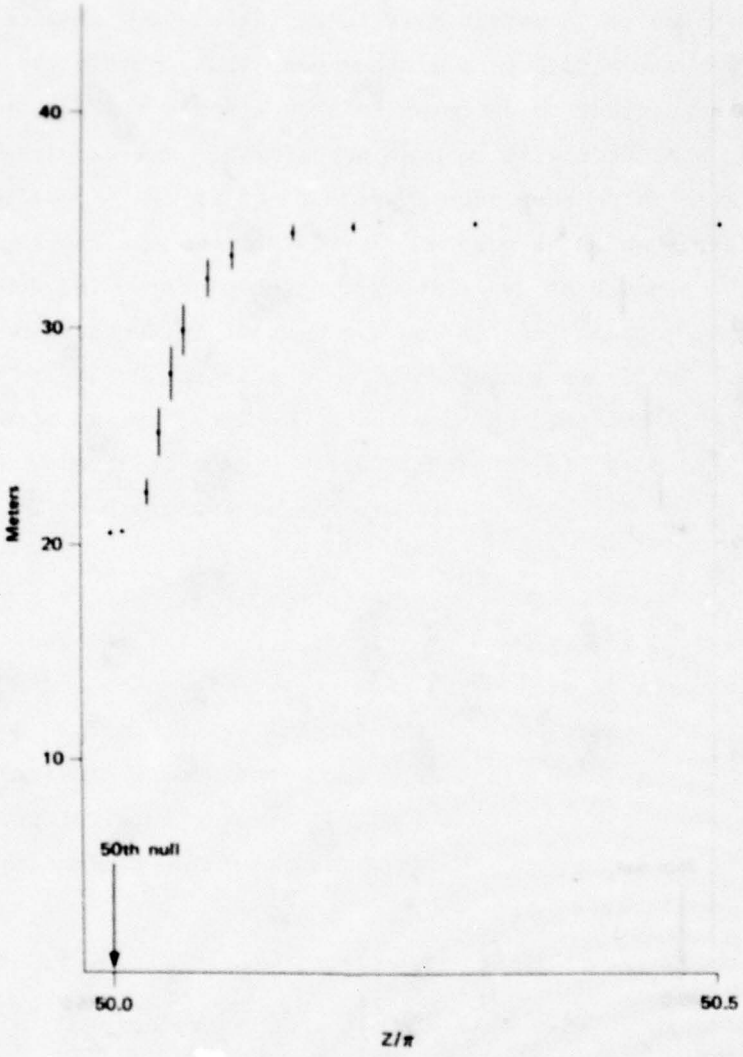


Fig. III-34 — Arrival time expressed as $D_0 \pm D_1$ meters (Q_3 absent); 50th side lobe

The points are not connected in these figures, but it is easy to judge by eye the shape of the dependence of D_0 on Z and the envelope of the shifts caused by phases σ and ψ. That envelope has a complicated shape; it can be seen that D_1 is largest at positions near the pattern null and falls quickly to a small value at the null. Some consideration was given to devising a functional description of the dependence of D_1 on Z, but that was soon abandoned as being too difficult and not worth the effort.

D_0 is very nearly constant over most of the width of each lobe and falls rapidly but smoothly to a minimum near (but not exactly at) the null. It is convenient to describe this as a notch in the curve of D_0 versus Z, and that term will be used henceforth. One can imagine a graph of D_0 stretching continuously from $Z = 0$ to $Z/\pi = 50.5$ (or beyond). The curve would be very nearly flat across the whole graph, but would have a notch at every integer value of Z/π. The depths of all the notches would be nearly the same--about 14.2 meters down from the flat top. The facts that the top of the graph is flat and horizontal, and that all the notches have the same depth, are important, and will be recalled later in connection with a tuned front end. With a tuned front end, the tops slope upward across the graph of D_0, which leads to problems not found in this simpler case.

The D_0 notches are not symmetric (nor is D_1), but the asymmetry is not excessive. To estimate the likelihood of the receivers "falling into" these notches it will do no harm to fit the notch shape with a function that is symmetric about the integer value of Z/π. Moreover, it appears that the notches all have about the same shape and the width of the notch is proportional to Z. These convenient traits make it feasible to devise a fairly simple functional approximation:

$$D_0 \approx 20.5916 + 14.2012 \left[1-e^{-(g^3)}\right] \text{ meters}$$

where $g = \frac{835\xi}{|Z/\pi|}$

and $\xi = \left|Z/\pi - \text{nearest integer}\right|$

This approximation differs appreciably from calculated values of D_0 only in places where the curve is steep. At such places the difference can be regarded as an inconsequential error in angle θ.

III-16. FIRST COMPUTER EXPERIMENT

Section II-6 described in some detail the computer simulation that was used to investigate the performance of a simple TOA system in which the receivers have no front-end tuning. The description is not repeated here.

The approximate formula for the average arrival time that is mentioned in Section II-6 is the expression for D_0 given above. Use of that approximation avoids the need to compute the $F_{5,6}$ waveform, to integrate that waveform numerically to obtain F_7, and to determine the time at which F_7 crosses the peak/2 level, for each value of Z for each separate receiver and for particular values of the phases. The consequent cost saving made possible the sampling of thousands of angles. Moreover, the approximation serves to estimate the system performance when the receivers are assumed to average out the phase effects. To obtain the same kind of result by direct calculation it would be necessary to perform each calculation three times for three phase combinations. The moderate improvement in the accuracy of the simulation would certainly not justify the cost of so much computer work.

The equations for the downrange and crossrange errors of the system (as viewed along the line of sight from the middle receiver to the source) are given in Appendix J. The expressions there assume that the errors are small compared with all three source-receiver distances, but are otherwise exact. It should be noted that, if the errors are small in this sense, then the errors caused by disparities of arrival time do not depend on these source-receiver distances.

Even if the three receivers are disposed at the corners of an isosceles triangle, the two internal angles A and B may not be equal if the source is not located on the axis of symmetry. However, the computer simulation assumes the favorable case that the two angles are equal.

It was explained in Part II that the shortcomings of basic antenna theory restrict the size of the total subtense, and a subtense of 90 degrees was chosen to permit the source orientation to rotate through a total of 90 degrees. The sum must not exceed 180 degrees. (Even then the validity of the results is open to question, as has been noted earlier.)

When the angles A = B = 45 degrees are chosen, it is clear that the downrange errors tend to be larger than the crossrange errors because of the denominators of the two expressions as well as the difference in the role of the middle receiver. Consequently, a scatter diagram of the numerous calculated locations of the source is more or less elliptical, with the long axis in the downrange direction. In the first experiment, all randomness was presumed to be removed by perfect averaging, so the scatter diagram is necessarily mirror-symmetric about the downrange axis (only half of the scatter diagram was computed; the other half was inferred). This mirror symmetry ensures that, although crossrange errors occur at individual source orientations, the average crossrange error is zero if a symmetric span of source orientations is observed.

It might be true that the average downrange error is also zero when the source is observed over a complete rotation, but this need not be so. Certainly the radiation into the rear hemisphere of a directional emitter is not the same as that into the forward hemisphere, so there is no prima facie basis for such symmetry.

The first computer experiment was carried out four times against sources with aperture widths, L/λ, equal to 30.5, 40.5, 50.5, and 60.5. (Half-integer values were used to avoid the possible bias that might be said to result from placing a pattern null at 90 degrees from the axis. There is no other significance to this choice.) The error distributions that were found in these four trials are shown in Fig. II-1. The results can be characterized further by some of the usual statistical parameters:

	Aperture width L/λ			
Parameter	30.5	40.5	50.5	60.5
$\langle \delta X \rangle$	0.79731	1.10126	1.40606	1.71535
$\langle \delta Y \rangle$	0	0	0	0
$\langle \delta R \rangle$	3.45844	4.29961	5.52201	6.33583
σ_X	8.97800	9.88266	11.23458	11.92581
σ_Y	2.38359	2.70731	3.08613	3.37194
σ_R	8.65799	9.36604	10.35492	10.78864

NOTE: $\langle \delta X \rangle$ = mean downrange error, $\langle \delta Y \rangle$ = mean crossrange error, $\langle \delta R \rangle$ = mean radial error (in meters), and the σ parameters are the standard deviations of these parameters.

The average downrange errors lie almost on a straight line, but the intercept is -0.136 meters. (The significance of the negative intercept is not clear, but undoubtedly bespeaks a more complicated relationship as $L/\lambda \to 0$.) Even though the curves in Fig. II-1 fall very steeply on the left side, it can be seen that the average radial error is fairly sizeable, and is considerably larger than the average downrange error. This suggests the improvement to be obtained from suitable averaging of the data. However, it is emphasized that the values shown here reflect not only perfect averaging of phase, but also averaging over a large symmetric span of source orientations. The source lobe pattern exhibits rapid fluctuations over small spans of angle, and averages taken over small spans may fail to yield the improvement shown here. In any case, it is evident that one must be cautious concerning exactly what data are to be averaged.

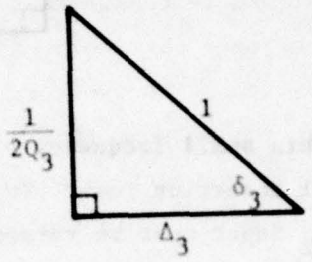
III-17. TUNED FRONT END: Q_3

At this point the original objective of the analysis has been achieved. More computer experiments could be run, but they would hardly be worthwhile. One can estimate with reasonable confidence the change in the error distribution curves with other values of triad angle A, or other values of L/λ, or other scale lengths S. More important, the entire analysis has been brought into question with the recognition that the brevity of the "extra" signal is important in that brevity determines that it is the impulse response that is excited. Reconsideration of the wide open front end was deferred to complete the original work through the first experiment. Now we return to the input signal $F_{1,2}$ and consider the situation when the front end of the receiver has frequency selectivity.

The number of terms in the expression for the IF output $F_{5,6}$ when the signal entering the heterodyne is merely $F_{1,2}$ should suffice to explain why it is necessary to introduce as little new complexity as possible in the front end lest the analysis become wholly infeasible. It was decided to consider only a single bandpass stage between $F_{1,2}$ and the heterodyne detector and, moreover, to tune that stage very

slightly above the carrier frequency. This is the same stratagem that was used in the source to avoid introducing a new frequency.

Let the Q of this tuned RF section be designated Q_3, and define as usual



The resonant frequency is adjusted to $\frac{\omega_0}{\Delta_3}$, and the impulse response of this stage is

$$h(t) = \frac{\omega_0}{\Delta_3^2 Q_3} e^{-\frac{\omega_0 t}{2\Delta_3 Q_3}} \cos(\omega_0 t + \delta_3)$$

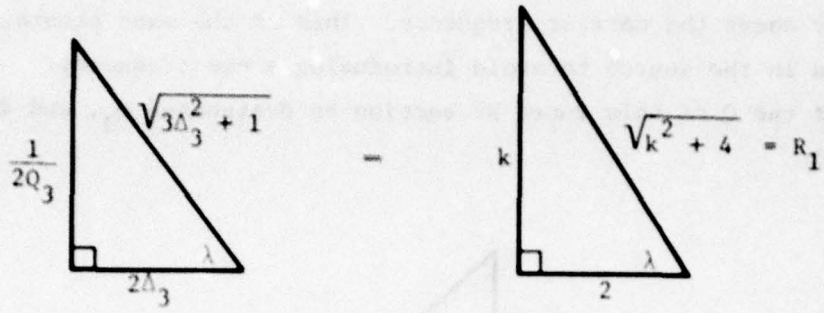
For brevity we define

$$K = \frac{1}{2\Delta_0 Q_0} = \frac{1}{6\pi}$$

$$k = \frac{1}{2\Delta_3 Q_3}$$

and, as usual, $U = \omega_0 t$. Define further*

*This usage of λ should not be confused with λ meaning wavelength. This phase angle λ is the counterpart of ϕ defined for the filter in the source. Also, this usage of symbol k should not be confused with that in Section III-8.



As in the source, this small frequency offset of the resonant frequency introduces a small insertion loss. To restore normalization the amplitude of the $F_{1,2}$ input must be raised by a factor

$$\frac{\sqrt{3\Delta_3^2 + 1}}{2\Delta_3} = \frac{1}{\cos\lambda}$$

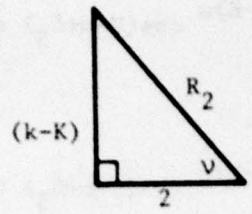
Here, as in previous cases, the signal emerging from the front end and going to the heterodyne stage will be described as two successive segments. (As usual, they are smoothly connected and comprise a single waveform.) To preserve the numbering system wherein the output of the heterodyne was designated $F_{3,4}$, we designate these intermediate segments F_a and F_b, and the entire signal $F_{a,b}$. Then

$$F_a = \frac{ke^{-kU}}{Z\Delta_3\cos\lambda}\left[-\int_{-Z}^{U} e^{ku}\cos(U-u+\delta_3)\cos(u+\psi+Z+\phi)du\right.$$

$$\left. + Me^{-KZ}\int_{-Z}^{U} e^{(k-K)u}\cos(U-u+\delta_3)\cos(u+\psi+Z+\mu-\phi)du\right]$$

$$-Z \le U \le Z$$

Now define



We will avoid consideration of k = K because that requires a special form of the solution.

$$F_a = \frac{ke^{-kU}}{2Z\Delta_3\cos\lambda}\left\{ - \cos(U+\delta_3+\psi+Z+\phi)\left[\left[\frac{e^{ku}}{k}\right]\right]_{-Z}^{U}\right.$$

$$+ Me^{-KZ}\cos(U+\delta_3+\psi+Z+\mu-\phi)\left[\left[\frac{e^{(k-K)u}}{k-K}\right]\right]_{-Z}^{U}$$

$$+\left[\left[e^{ku}\,\frac{\sin(U-2u+\delta_3-\psi-Z-\phi-\lambda)}{R_1}\right]\right]_{-Z}^{U}$$

$$\left. - Me^{-KZ}\left[\left[e^{(k-K)u}\,\frac{\sin(U-2u+\delta_3-\psi-Z-\mu+\phi-\nu)}{R_2}\right]\right]_{-Z}^{U}\right\}$$

When U > Z, signal F_1 has ceased but the filter continues to ring in response to the earlier input. That ringing is described by the expression for F_a above, evaluated when the upper limit is Z. For brevity we designate that function here as $F_a(Z)$. Then the second segment is

$$F_b = F_a(Z) + \frac{2ke^{-ku}}{\Delta_3\cos\lambda}\left[\frac{\sin Z}{Z}\int_{+Z}^{U} e^{ku}\cos(U-u+\delta_3)\,\sin(u+\psi+\phi)\,du\right.$$

$$+ \frac{Me^{-KZ}}{2Z} \int_{+Z}^{U} e^{(k-K)u} \cos(U-u+\delta_3) \cos(u+\psi+Z+\mu-\phi)du$$

$$- \frac{Me^{+KZ}}{2Z} \int_{+Z}^{U} e^{(k-K)u} \cos(U-u+\delta_3) \cos(u+\psi-Z+\mu-\phi)du \Bigg] ; \; U \geq Z$$

$$F_b = F_a(Z) + \frac{ke^{-ku}}{\Delta_3 \cos\lambda} \left\{ - \frac{\sin Z}{Z} \left[\!\!\left[e^{ku} \frac{\cos(2u-U-\delta_3+\psi+\phi+\lambda)}{R_1} \right]\!\!\right]_Z^U \right.$$

$$+ \frac{\sin Z}{Z} \sin(U+\delta_3+\psi+\phi) \left[\!\!\left[\frac{e^{ku}}{k} \right]\!\!\right]_Z^U$$

$$+ \frac{Me^{-KZ}}{2Z} \cos(U+\delta_3+\psi+Z+\mu-\phi) \left[\!\!\left[\frac{e^{(k-K)u}}{k-K} \right]\!\!\right]_Z^U$$

$$- \frac{Me^{-KZ}}{2Z} \left[\!\!\left[e^{(k-K)u} \frac{\sin(U-2u+\delta_3-\psi-Z-\mu+\phi-\nu)}{R_2} \right]\!\!\right]_Z^U$$

$$- \frac{Me^{+KZ}}{2Z} \cos(U+\delta_3+\psi-Z+\mu-\phi) \left[\!\!\left[\frac{e^{(k-K)u}}{k-K} \right]\!\!\right]_Z^U$$

$$\left. + \frac{Me^{+KZ}}{2Z} \left[\!\!\left[e^{(k-K)u} \frac{\sin(U-2u+\delta_3-\psi+Z-\mu+\phi-\nu)}{R_2} \right]\!\!\right]_Z^U \right\}$$

As $U \longrightarrow \infty$, $F_a(Z) \longrightarrow 0$ and

$$F_b \longrightarrow \frac{\sin Z}{Z} \sin(U+\psi+\phi+\lambda)$$

The small phase shift λ appears because this front-end filter was detuned, just as phase ϕ entered from the detuning of the filter in the source.

This is the signal entering the heterodyne detector. If we continue to use designation $F_{3,4}$ for the output of that detector, then the new output is

$$F_3 = 2F_a \cos(U+V+\sigma)$$

$$F_4 = 2F_b \cos(U+V+\sigma)$$

When the limits are inserted in the expressions for F_a and F_b, and F_3 and F_4 are decomposed into sum- and difference-frequency terms, a great many terms result. One can define appropriate constants so every one of these terms can be written in the generic form

$$C_n e^{-g_n V} \cos(j_n V+m_n)$$

and F_3 and F_4 are then represented as the sum of such terms.

The impulse response of the IF strip can be written as a sum of four terms in the same generic form, and the convolution of $F_{3,4}$ yields $F_{5,6}$ as still another sum over terms in that form. That expression is then entered into the same numerical integration routine to produce the new output F_7.

The foregoing process is tedious but straightforward. It requires the definition of a great many quantities such as the angles, m_n, some of which are quite intricate. This plethora of algebra is too extensive and routine to warrant presentation here. However, it will be

evident that the computer time needed to carry out all this arithmetic is considerably more than the not-trivial computation of F_7 when Q_3 is absent.

III-18. DEPENDENCE OF ARRIVAL TIME ON Q_3

It would be informative and possibly useful to explore fully the dependence of arrival time upon Q_3, Z, σ, and ϕ. However, the introduction of Q_3 makes even more costly the computation of each output waveform, and an examination of that four-dimensional space would be too costly. Budget and time limitations compelled a less ambitious study, and it was decided to look first for a dominant effect of Q_3. It seemed possible that the shifts of arrival time with Z and phase might be reduced appreciably, more or less at all Z, and perhaps in proportion to the numerical value of Q_3.

The same peak/2 definition of arrival time was adopted, and 4x4 arrays of arrival times (σ = 0(45)135; ϕ = 0(45)135) were calculated for Q_3 = 5, 10, and 20 at a few values of Z. The results are shown in Tables III-5 through III-9. Tables III-5, III-6, III-8, III-9 show arrival time in meters at Z/π = 0, 1, 0.998, and 35.02. Table III-7 shows the peak level reached by F_7, expressed in dB, in three pattern nulls widely spaced in angle. All five tables employ the same array format as in Tables III-1 through III-4.

The Q of the source is Q_0 = 9.438. Thus the three values of Q_3 considered correspond to receiver RF bandwidths approximately twice, equal to, and half the source bandwidth. For the source carrier frequency used here (3200 MHz), Q_3 = 20 corresponds to an RF bandwidth of 160 MHz whereas the IF bandwidth is 80 MHz. It seems unlikely that a circuit designer would use an RF bandwidth appreciably narrower than the IF bandwidth and also much narrower than the source bandwidth. Q_3 = 20 seems to be near the largest value likely to be used in conventional design practice. Possibly no loss of signal-to-noise ratio would result from use of an RF bandwidth narrower than the source (provided that external noise is dominant), and there might be merit in extending this study to higher values of Q_3. However, higher Q_3

Table III-5

PEAK/2 ARRIVAL TIMES (IN METERS); RECEIVER ON SOURCE AXIS

(Dependence on phases σ and ϕ and on Q_3; $Z/\pi = 0$)

	$Q_3 = 5$		
34.764423	35.090508	35.092820	34.935284
34.925937	34.756831	35.096136	35.098092
35.082914	34.911795	34.758391	35.096919
35.091291	35.077641	34.921142	34.765983
	$Q_3 = 10$		
34.914841	35.156566	35.217852	35.141628
35.131172	34.905650	35.161011	35.221313
35.207939	35.119816	34.909559	35.163956
35.159511	35.204479	35.130272	34.918749
	$Q_3 = 20$		
35.268849	35.413768	35.453376	35.467590
35.447358	35.260141	35.418763	35.456927
35.445292	35.438405	35.266529	35.424244
35.419248	35.441741	35.448638	35.275236

will tend to slow the receiver rise time and flatten the slope at the peak/2 point. Reduced slope would increase the errors caused by noise, and that increase might offset any possible advantage of higher Q_3 found here.

Tables III-5 and III-6 show the results at what might be regarded as principal values of Z: on the source axis where the "extra" signal is absent, and in the first pattern null where the briefest "extra" signal is found. Two general observations can be made concerning these results:

- $Q_3 < 20$ is not a cure. The shift of arrival time remains substantially equal to S = 14.15 meters. (These waveforms are not shown, but they differ very little from the stepped-carrier and impulse responses of the wide open receiver.)

Table III-6

PEAK/2 ARRIVAL TIMES (IN METERS); RECEIVER IN FIRST NULL

(Dependence on phases σ and ψ and on Q_3; $Z/\pi = 1$)

	$Q_3 = 5$		
20.675049	20.658415	20.868300	20.658692
20.652353	20.679840	20.657904	20.870058
20.877173	20.652799	20.690356	20.658805
20.659161	20.875626	20.638353	20.685670
	$Q_3 = 10$		
20.761386	20.840579	21.016261	21.107483
21.105181	20.766684	20.849293	21.022852
21.018847	21.046942	20.776511	20.852081
20.843410	21.012021	21.049951	20.773085
	$Q_3 = 20$		
20.950138	21.091210	21.151143	21.466794
21.461283	20.940812	21.093779	21.157655
21.154210	21.469437	20.948982	21.097844
21.095286	21.147620	21.475016	20.953115

- The dependence on phases σ and ψ is much the same as when Q_3 is absent. In these angular locations, where no mutual interference between two receiver responses occurs, the phase effects are small.

It will be recalled that when Q_3 is absent, the signal level in the first pattern null changes 14.2 dB with phase, but in nulls at larger angles the amplitude change with phase is much less. For example, with Q_3 absent the peak level reached by F_7 at $Z/\pi = 50$ is about -64.85 ± 0.55 dB. That aspect of the receiver response is different when Q_3 is present. Table III-7 shows the phase dependence of peak amplitude in three widely spaced pattern nulls, all for $Q_3 = 10$. In all three the level changes roughly 10 dB with phase. Moreover, the signal level in the nulls falls off considerably at large Z, whereas

Table III-7

PEAK SIGNAL LEVEL (IN dB) IN THREE PATTERN NULLS

(Dependence on phases σ and ψ; $Q_3 = 10$)

	$Z/\pi = 1$		
-63.5635	-61.3238	-64.2511	-71.1017
-71.0847	-63.5152	-61.2373	-64.1371
-64.2225	-71.2687	-63.4475	-61.2242
-61.3104	-64.3376	-71.2853	-63.4954
	$Z/\pi = 35$		
-75.9400	-78.1641	-86.6984	-79.1500
-79.6198	-76.1657	-78.1440	-86.4910
-86.6248	-79.6855	-75.9601	-77.8133
-77.8072	-86.9124	-79.2123	-75.7397
	$Z/\pi = 50$		
-87.6120	-80.2362	-79.4247	-84.3368
-84.2087	-87.5374	-79.9235	-79.1443
-79.2519	-84.6381	-86.8223	-79.7549
-80.0614	-79.5359	-84.7727	-86.8917

when Q_3 is absent that level does not change very much with Z. This substantial change of behavior suggests that the introduction of Q_3 has greater effect than might be inferred from Tables III-5 and III-6. However, even at $Z/\pi = 50$ the output waveforms remain nearly identical to the receiver impulse response. The level changes with phase shown in Table III-7 indicate sizeable changes of the strength at which the receiver response is excited, but not in the response waveform itself. $Q_3 \lesssim 20$ does not "stretch" the "extra" waveform enough to remove the impulsive character.

It can be foreseen from the above discussion that the interference between the stepped-carrier and impulse responses when both are present will be changed by Q_3, even though the individual waveforms are virtually unchanged in shape. In other words, the effects of Q_3 will vary with Z. This is illustrated in Tables III-8 and III-9 which show 4x4

Table III-8

PEAK/2 ARRIVAL TIMES (IN METERS) NEAR THE FIRST NULL

(Dependence on phases σ and ψ and on Q_3; $Z/\pi = 0.998$)

	$Q_3 = 5$		
34.298068	28.176338	30.126203	36.306781
36.175852	34.359345	28.087546	29.974299
29.975102	36.353332	34.368930	27.983108
28.070321	30.127041	36.485152	34.309200
	$Q_3 = 10$		
35.405572	29.439908	29.781337	34.630494
34.598794	35.478672	29.412117	29.729526
29.757635	34.728096	35.575330	29.398732
29.426514	29.809513	34.759523	35.504431
	$Q_3 = 20$		
34.987966	31.427715	30.488902	33.510562
33.518124	35.057702	31.452099	30.477327
30.493435	33.582801	35.125619	31.460811
31.436403	30.514550	33.575340	35.055245

arrays of arrival time at two locations near but not in widely spaced nulls. Comparable arrays when Q_3 is absent were shown in Tables III-2 and III-3. It will be seen that at $Z/\pi = 0.998$, $Q_3 = 10$ has a small beneficial effect, but at $Z/\pi = 35.02$ $Q_3 = 10$ leads to drastically different behavior. In the latter case the average value, D_0, is raised to a value that is near the stepped-carrier arrival time and is a big improvement. On the other hand, D_1, which reflects phase dependence, is increased substantially and that is not helpful.

To consider whether a systematic improvement, of whatever amount, is indicated by the results at these four values of Z, it is helpful to compare the dependence of average arrival times on Z and Q_3. We tabulate the average of all 16 values in each 4x4 array rather than the three-phase parameter D_0:

	Q_3			
Z/π	Absent	5	10	20
0	34.792818	34.966632	35.104020	35.396007
0.998	31.817233	32.198539	32.339762	32.635288
1	20.579007	20.716160	20.927660	21.165895
35.02	21.935244	36.218793	35.951884	35.293878

Table III-9

PEAK/2 ARRIVAL TIMES (IN METERS) NEAR THE 35th NULL

(Dependence on phases σ and ϕ and on Q_3; $Z/\pi = 35.02$)

	$Q_3 = 5$		
35.824976	38.807028	36.127031	33.773650
33.956140	35.822531	38.775401	36.331683
36.513133	33.926182	35.761782	39.011699
39.044210	36.314583	33.747062	35.763590
	$Q_3 = 10$		
34.578630	38.723913	36.731740	33.490391
33.691804	34.591881	38.666608	36.919938
37.107739	33.656387	34.476455	38.849881
38.908741	36.916880	33.455945	34.463203
	$Q_3 = 20$		
33.010710	37.582678	37.099221	33.259249
33.439789	33.028254	37.489746	37.276392
37.473159	33.402527	32.898393	37.628895
37.720932	37.289929	33.222983	32.879185

The small increase with Q_3 across the first three lines is merely a reflection of the increase of the receiver rise time, and is of no consequence in itself. It is the differences seen reading down each column that matter, and no strong trend of improvement with Q_3 is evident in the first three lines. The fourth line indicates a substantial improvement, seemingly better for $Q_3 = 20$ than for $Q_3 = 5$ or 10, but the strength of the dependence on the size of Q_3 is not great. The

considerable increase in the size of the phase effects noted above makes it difficult to judge just how much improvement is found here.

These results indicate that the effect of introducing Q_3 is not uniformly beneficial, and depends upon Z.

III-19. DEPENDENCE OF ARRIVAL TIME ON Z WHEN $Q_3 = 10$

The value $Q_3 = 10$ was adopted for further analysis primarily because this choice matches the bandwidth of the RF section of the receiver to the bandwidth of the source. For economy, the approximate description in terms of D_0 and D_1 (see III-14) was used at nearly all the values of Z that were examined. (The few instances where 16 or more phase combinations were considered show that here, as when Q_3 is absent, the values of D_1 may be misleading when D_1 is large.)

It soon was obvious that the dependence of D_1 on Z is even more complicated when $Q_3 = 10$ than when Q_3 is absent. Hope of devising a mathematical description of $D_1(Z)$ was abandoned early. A substantial effort was made, however, to find a mathematical "fit" for D_0, for two reasons:

- A functional form for $D_0(Z)$ with $Q_3 = 10$ would permit a computer experiment comparable with the first experiment. Such an experiment would afford a measure, in terms of system performance, of the improvement resulting from front-end tuning.
- Such a fit to D_0 when $Q_3 = 10$ could probably be extended to other values of Q_3. If so, the dependence on the value of Q_3 might be determined from a limited sampling of the dependence on Z.

Values of D_0 and D_1 (calculated for $\psi = 0$, $\sigma = 0(45)90$) were obtained at numerous values of Z--more than were obtained with Q_3 absent, and more than will be displayed here. The many values of Z were run to test a very wide variety of "fits" to the data. Many functional forms

were found that could fit $D_0(Z)$ near any particular integer value of Z. No form was found that exhibited a systematic pattern over a wide span of Z. Some orderly trend with Z is needed to allow interpolation to avoid the need to make a separate fit at each pattern null. The shape of the D_0 notch is not only asymmetric about each pattern null, but the shape varies considerably with Z. The full width of the notch at numerous depths was obtained at four values of Z. Even though this width ignores the asymmetry of the notch, no fit could be found. Indeed, even the notch width at one depth--halfway down--appears to vary with Z in a complicated way, and the dependence at other fractional depths is different.

The variation of D_0 and D_1 with Z for $Q_3 = 10$ is illustrated in Figs. III-35 through III-38, counterparts of Figs. III-32, III-33, and III-34. At each value of Z, the vertical line extends from $D_0 - D_1$ to $D_0 + D_1$ and a dot at the middle indicates D_0. The scales of Z/π differ among these figures, but all seven use the same scale of arrival time in meters.

The introduction of Q_3 narrows the widths of the D_0 notches, tending to reduce the probability that the system will make an error greater than any given amount. If the shapes of the notches or the dependence on Z did not change with Q_3, then narrowing the notches would lower the error distribution curves without changing their shape. However, both the shape of the D_0 notch and the dependence of notch width on Z are changed by Q_3. The narrowing of the notches is offset to a degree by an increase in the typical values of D_1. Such increase, if not removed by averaging over phase effects, would probably tend toward somewhat larger system errors and would alter the shape of the distribution curves.

It appears from the results of the second experiment discussed below that the balance between narrowing the notches and increasing D_1 turns out favorably. That is, the effect of Q_3 is to reduce the probability of large errors, and this improvement is not cancelled by D_1. However, a new unexpected phenomenon arises that complicates the situation immensely; it is illustrated in Fig. III-39.

Figure III-39, like the preceding figures, shows arrival time in meters versus Z/π. In this figure, however, only D_0 is plotted, and

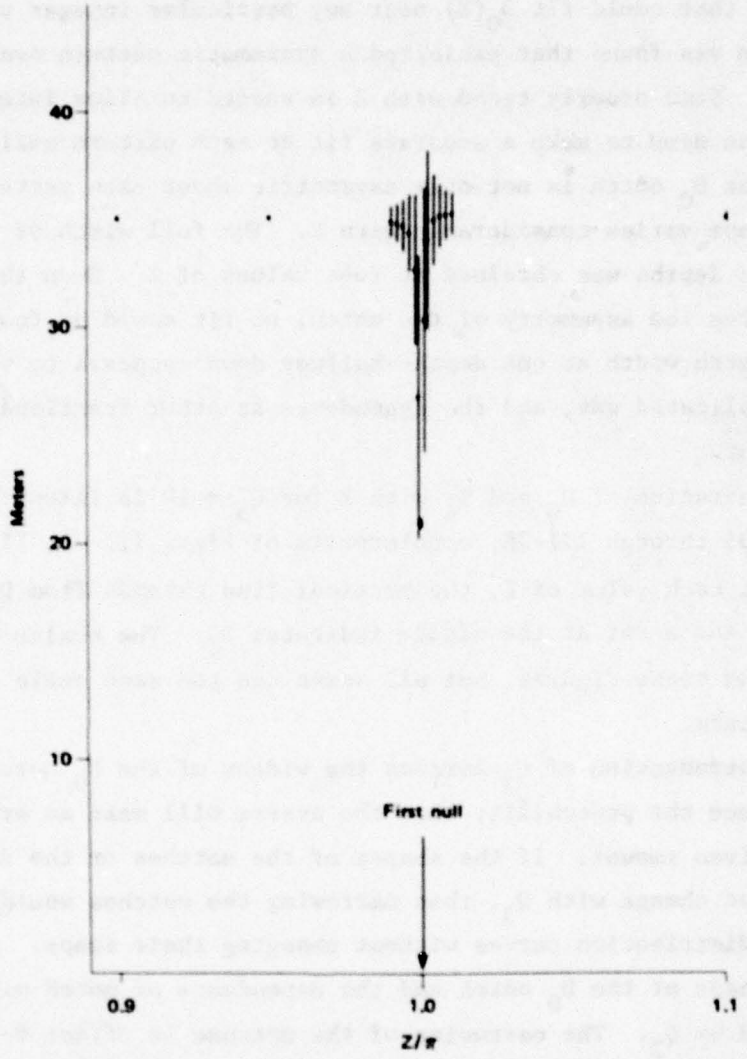


Fig. III-35 — Arrival time, expressed as $D_0 \pm D_1$ meters, near the first null ($Q_3 = 10$)

only at integer and half-integer values of Z/π (that is, in the nulls and at the tops of the lobes). The two rows of dots indicate the tops and the bottoms of the notches. The notch depth is very nearly constant across the figure (the extreme values in the figure are 13.900 and 14.171 meters). But the two rows curve upward. The value of D_0 at $Z/\pi = 50.5$ is 2.0794 meters larger than at $Z/\pi = 0$.

The upper row of dots shows the D_0 arrival time that will be obtained almost everywhere; the lower values are found only in the

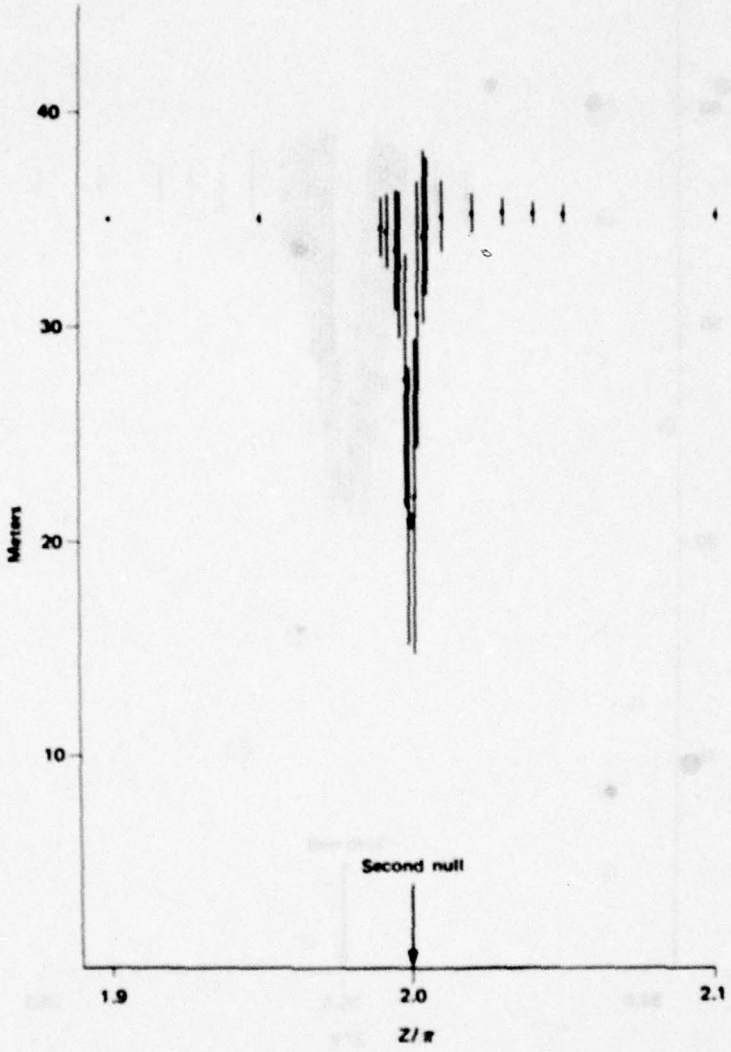


Fig. III-36 — Arrival time, expressed as $D_0 \pm D_1$ meters, near the second null ($Q_3 = 10$)

narrow notches. Thus, the upper row shows the value of D_0 that is most likely to be measured. The curvature indicates an "error" that is virtually inherent: under the best of conditions, and with perfect averaging of phase effects, the different receivers will usually report arrival times that differ by a meter or two because of this curvature. Timing disparities of a meter or two among the receivers will lead to system errors up to perhaps four meters or so (if angle A =

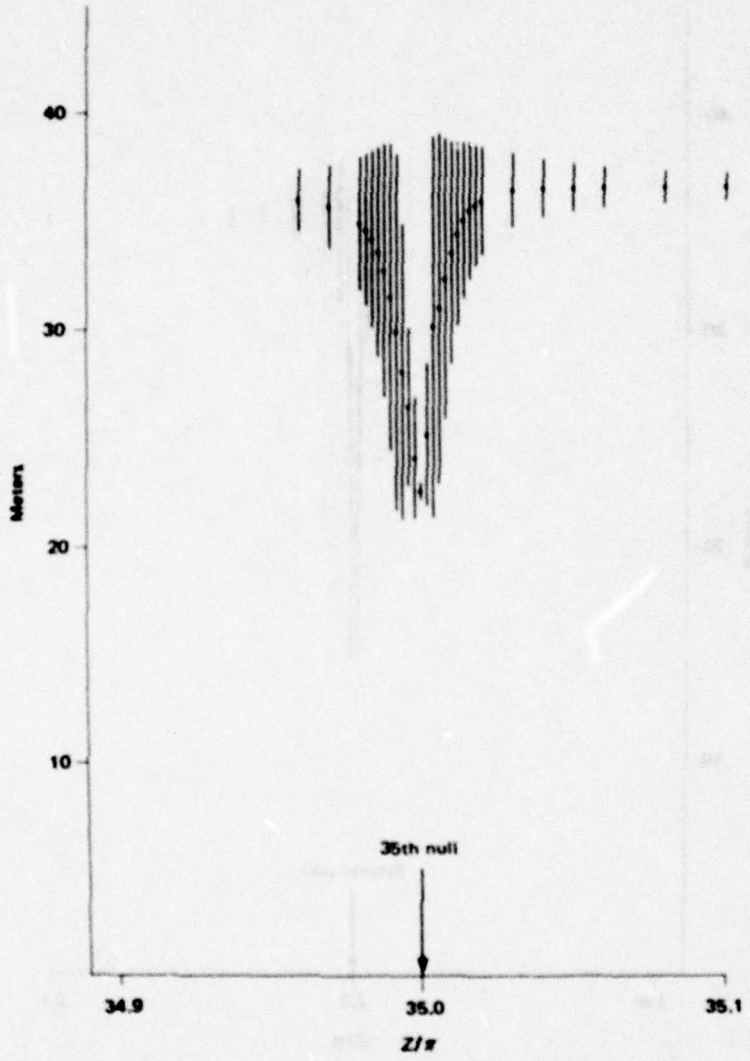


Fig. III-37 — Arrival time, expressed as $D_0 \pm D_1$ meters, near the 35th null ($Q_3 = 10$)

angle B = 45 degrees; see Fig. J-1). Thus the curvature seen in Fig. III-39 leads to a near-certainty of errors of a few meters, a distinct change from the results with wide open receivers; in that case the distribution curves fell very steeply close to the zero-error axis. The plethora of small errors when $Q_3 = 10$ will shift that steep fall to the right a few meters.

The outcome is that $Q_3 = 10$ tends to reduce the probability of large errors, but at the expense of much higher probability of small

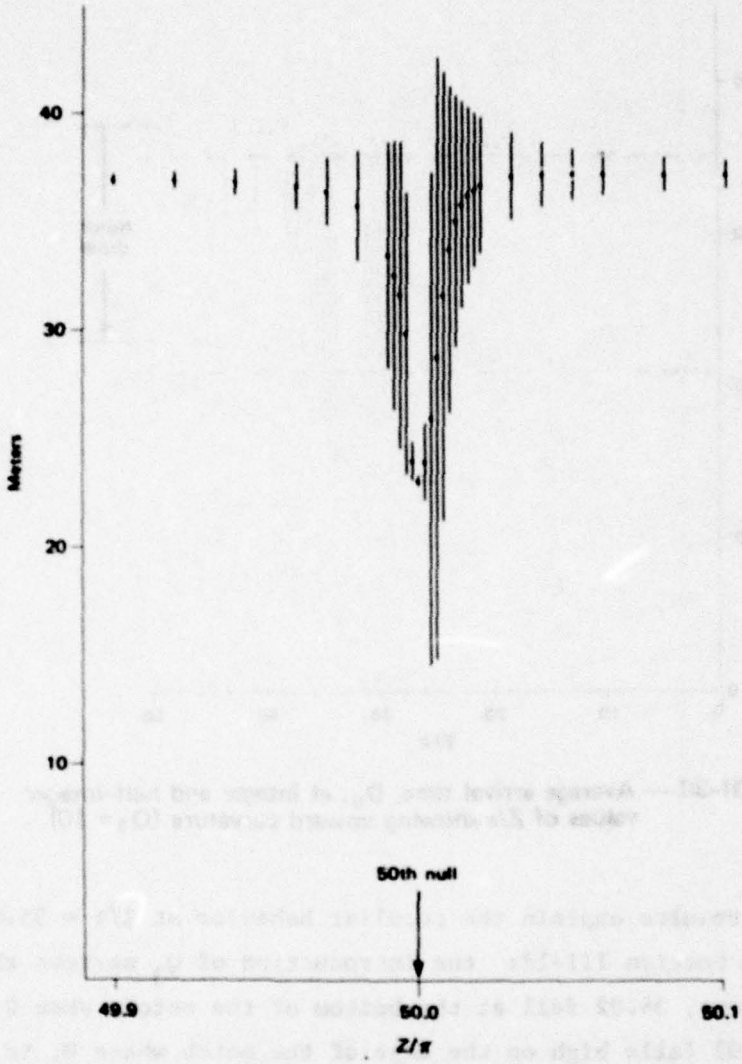


Fig. III-38 — Arrival time, expressed as $D_0 \pm D_1$ meters, near the 50th null ($Q_3 = 10$)

errors. Whether such a change is an improvement will depend on the intended purpose of any particular system and on various operational arrangements. For example, if one can, by some arrangement, discern and discard large errors, then a moderate reduction of the probability of large errors would not be beneficial. For such a system, Q_3 might be deemed harmful. This sort of uncertainty arises in judging whether, or by how much, Q_3 helps.

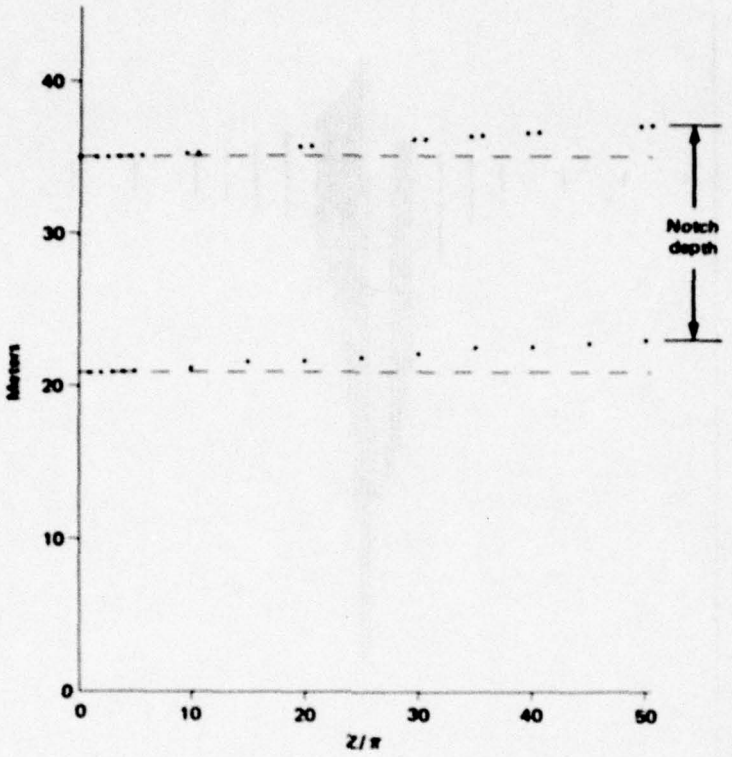


Fig. III-39 — Average arrival time, D_0, at integer and half-integer values of Z/π showing upward curvature ($Q_3 = 10$)

These results explain the peculiar behavior at $Z/\pi = 35.02$ that was seen in Section III-18: the introduction of Q_3 narrows the notch. With Q_3 absent, 35.02 fell at the bottom of the notch; when Q_3 is inserted, 35.02 falls high on the edge of the notch where D_0 is much larger (but so too is D_1). Except for this, no new inferences are available to judge whether or not there is an orderly trend with the value of Q_3. Moreover, a new question has been raised but not answered: how does the curvature of D_0 depend on Q_3?

III-20. TRANSITION FROM IMPULSE- TO STEPPED-CARRIER RESPONSE

Inclusion of a tuned circuit in the front end changes the receiver (ahead of the envelope detector) from a four-pole to a five-pole filter. Nevertheless, some insight can be gained into the results shown above by adopting a qualitative viewpoint wherein Q_3 is regarded merely as

a circuit that changes the duration of brief input signals going to the IF strip. The F_7 waveforms with Q_3 present show that $Q_3 < 20$ does not "stretch" the signals in the pattern nulls enough to change the waveform notably. Even at $Z/\pi = 50$, where the natural length of the "extra" signal exceeds one IF period, and Q_3 acts to extend the duration even longer, the output response scarcely differs from the impulse response of the IF strip. It is of some interest, then, to consider what input signal duration to the IF strip is needed to change the response from the impulse form to the stepped-carrier form.

The transition from one to the other can be illustrated by examining the output response elicited by an RF input that starts at time $t = 0$ and consists of exactly N complete sine-wave cycles of the carrier frequency. Inasmuch as the stepped-carrier response of the receiver with Q_3 absent was already in hand, the dependence of the response on the number N could be investigated readily. All that was needed was to deliver into the F_7 numerical integration routine the $F_{5,6}$ stepped-carrier response together with a delayed response with opposite sign. Phases $\sigma = \phi = 0$ were used.

The output waveforms elicited by six values of N are shown in Fig. III-40. The lowest curve is the impulse response of the wide open receiver. As the value of N increases, the response undergoes a gradual transition leading, when $N = \infty$, to the stepped-carrier response. The successive minima shift to later time as N increases, and we can see the undulations of the stepped-carrier response arise from those minima shifting up and to the right. The peak/2 arrival time shifts smoothly from the value for the impulse response over to the value for the stepped-carrier response. It is this shift that is called scale length S in this study.

Consider now the numerical value of N for these curves. The $N = 40$ curve is essentially the impulse response, even though the input signal lasts for an entire IF period. For $N = 80$, the response differs little from the impulse response (signal level in dB is of no consequence here; it is the shape of the curve that is significant). Even the $N = 160$ curve resembles the impulse response much more than it does the stepped-carrier response. Some 400 RF cycles, equivalent

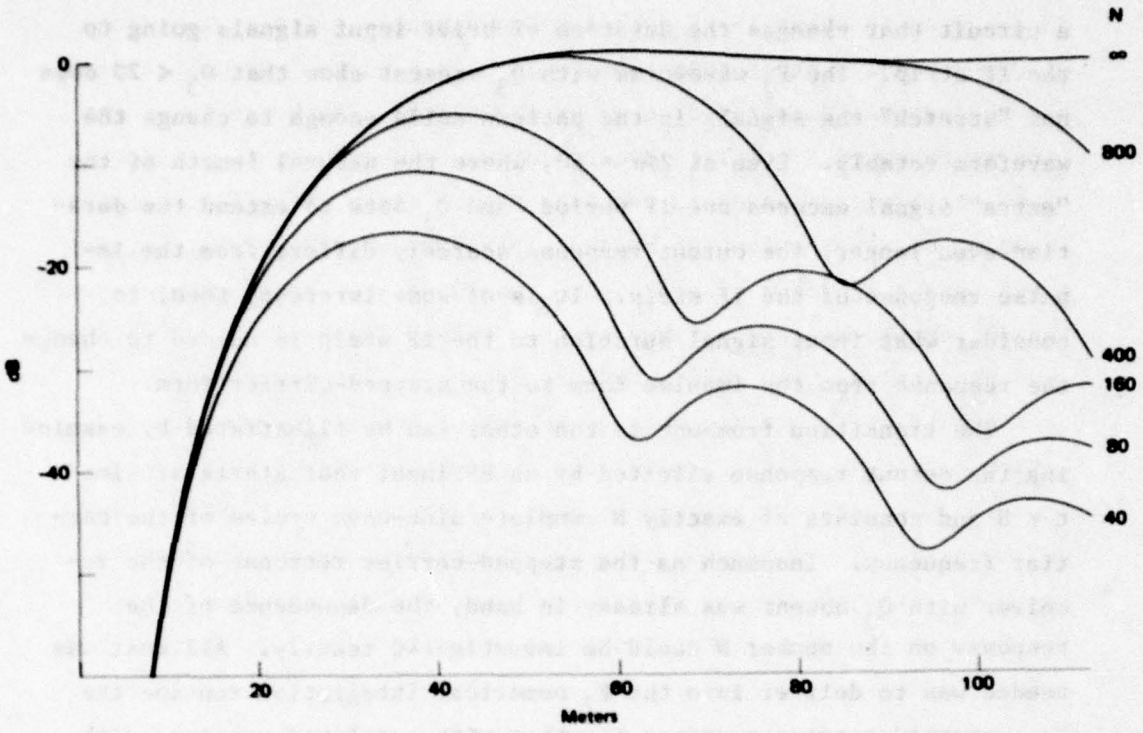


Fig. III-40 — Evolution of the receiver output with duration of a simple input pulse that contains N complete cycles of the RF carrier (wide-open front end)

to ten IF cycles, are needed to shift the peak/2 arrival time to the stepped-carrier value. It is of general interest, aside from TOA considerations, that an input duration of several IF periods elicits a response so nearly the impulse response.

If one overlooks the rise time of the source pulse, determined by Q_0, then at angular location Z the "extra" signal duration is Z/π RF periods. The curves shown here suggest that even at $Z/\pi = 80$ we can expect to see the impulse response.

Consider now the extent to which Q_3 tends to "stretch" this inherent duration. It is a little known but handy rule of thumb that the ringing of a single-pole filter such as Q_3 decays at about 27/Q dB per cycle.* Even for $Q_3 = 20$, the ringing of the RF filter will

---

* $20 \log_{10}\left[e^{-\frac{2\pi}{2Q}}\right] = \frac{-27.288}{Q}$. See Appendix B.

decay 20 dB in about 15 RF cycles. Consequently quite large values of Q_3--probably larger than 100--would be needed to "stretch" appreciably the natural duration of the "extra" signal. Although this examination in terms of N complete cycles certainty overlooks the influence of the true signal waveform reaching the IF strip, it affords an understanding of why the impulse response persists when $Q_3 = 20$. It also gives us a rough idea of how high Q_3 would have to be to change greatly the IF output waveform excited by the "extra" signal.

The arrival time shifts with Z, σ, and ψ that occur when Q_3 is present differ from those when Q_3 is absent because of changes in excitation amplitude of the impulsive response, because of differences in the time dependence of the amplitude, and because of detailed phase differences. The shifts do not occur because of a gross change in the response waveform arising merely from "stretching" the "extra" signal. Consequently there is little basis on which to speculate that high Q_3 is better than low. The interaction of Q_3 with Z, σ, and ψ is likely to be intricate and to depend on the particular design of the IF strip. Possibly the optimum RF bandwidth for some particular system might be found at an intermediate value of Q_3, and might be different for a different application. The dependence on Q_3 was barely touched upon in this study.

III-21. SECOND COMPUTER EXPERIMENT

To assess the improvement, if any, attributable to front-end tuning, it would be best to repeat the first computer experiment, making no change in any of the experimental parameters other than Q_3. Unfortunately the failure to find a usable functional description of the dependence of D_0 on Z with $Q_3 = 10$ prevented this. Replication of that experiment for only one source width would require about 90,000 calculations of exact arrival time. It was necessary to run a limited experiment that differed in several important ways from the first experiment (see Appendix J):

- Only one source width, $L/\lambda = 50.5$, was used.

- o Only 100 source rotation angles Λ between -45 and +45 degrees were run, and only one pulse was considered at each such angle.
- o No approximations were made; at each value of Λ the exact arrival time was calculated for each of the three receivers. Completely random phases ψ and σ were used, with no phase averaging. (In this respect the second trial is probably more realistic than the first, particularly if the source is assumed to rotate while pulsing.)

At each value of Λ, a random number was drawn to determine phase $0 \leq \psi \leq \pi$. This phase is the same for all three receivers since it is supposed that they observe the same pulses. Then three random numbers were drawn to determine the three local oscillator phases $0 \leq \sigma \leq \pi$. The routine to compute $F_{5,6}$ and F_7 and the peak/2 arrival time was used for each pulse in each receiver. Those arrival times were used to compute, for each of the 100 source pulses, the system error in determining the source location. The cumulative distribution of those errors was shown in Fig. II-2 and was discussed in Part II.

The curvature of D_0 shown in Fig. III-39 leads to a high probability of system errors of a few meters. It is not known how this curvature would differ for other more realistic designs of the IF strip. We have no information on how these small system errors depend on Q_3.

The introduction of $Q_3 = 10$ reduces significantly the probability of large system errors. Nevertheless, the general shape of the curve is about the same in the two experiments, and the upper limit is about the same. The following measures should be compared with those listed in Section III-16 (under the column heading 50.5):

$$\langle \delta X \rangle = -1.675 \text{ meters}$$
$$\langle \delta Y \rangle = -0.017 \text{ meters}$$
$$\langle \delta R \rangle = 4.576 \text{ meters}$$
$$\sigma_X = 7.527 \text{ meters}$$
$$\sigma_Y = 2.290 \text{ meters}$$
$$\sigma_R = 6.615 \text{ meters}$$

The tradeoff between fewer large errors and many more small errors produces an average radial error of 4.58 meters, compared with 5.52 meters in the first experiment--hardly a big improvement. How much improvement this is thought to be depends, in part, on the relative importance of large and small errors. That balance probably depends on the application and on a number of operational features of the system.

The uncertainty of these numerical values is reemphasized. In addition to the theoretical shortcomings of antenna theory and the many simplifications of the receiver design, the question of the curvature of D_0 and its dependence on design details is unresolved.

Appendix A

TRANSIENT ANALYSIS

Analysis of the response of a network to nonperiodic excitation has never been popular. Few circuit designers even make much use of approximate transient analysis, and those few almost never undertake exact transient analysis. Consequently, a brief review of the analytic methods used in this study is given in Appendixes A through E.

These appendixes are limited in several respects. Formal mathematical aspects are minimized, and such matters as existence, uniqueness, and convergence are not addressed. Further, only the types of networks considered in this study are covered; for example, distributed networks that possess an infinite number of normal modes are not discussed. Finally, complex notation is used in frequency-domain functions such as transfer functions, but time-domain functions are described entirely as real variables. This is because complex notation was not found to be particularly helpful during the study.

The problem can be summarized thusly: a four-terminal network is excited at one terminal pair by a time-varying signal f(t); what response, g(t), does the excitation produce at the other terminal pair? It will be assumed that the network is passive, linear, bilateral, composed of a finite number of lumped-constant ideal circuit elements, and that the finite speed of propagation of the signal along the circuit branches can be neglected. These are idealizations that are never met exactly in real networks, and significant departures--notably in nonlinearity and in distributed effects--arise in practice.

If f(t) were a single-frequency sinusoid (and, thus, of infinite duration) the response g(t) would be given by

$$g(t) = f(t)\,\Gamma(\omega)$$

where Γ is the transfer function of the network (evaluated at the input frequency) and is routinely calculable from the well-known complex impedances of the circuit elements.

If f(t) is a more complicated function of time, it can be decomposed into its Fourier spectrum.* Then the assumptions of linearity, etc., permit the summation of the individual responses to all the Fourier components, and g(t) can be obtained as a sum or integral over the frequencies. In other words, one can use a Fourier transform to take the time function f(t) over to the frequency domain, obtain the frequency-domain description of the desired response, and then use an inverse Fourier transform to obtain the time-domain response g(t). This can always be done in principle for the signals and networks considered here, but it may be difficult in fact.†

$\Gamma(\omega)$ is generally a complex function whose phase and amplitude vary with frequency. There is at least one frequency at which $|\Gamma(\omega)|$ attains a maximum. That maximum is often less than 1, in which case the network is said to exhibit insertion loss. However, in many treatments of circuit theory $\Gamma(\omega)$ is normalized (perhaps tacitly and without warning to the reader) so the maximum of $|\Gamma(\omega)| = 1$. That is, the insertion loss is discarded. If that form of $\Gamma(\omega)$ is used when going into (and back out of) the frequency domain, the resulting solution for g(t) will differ by a multiplicative constant from that obtained directly in the time domain. This is of no concern when one is interested only in waveforms, and amplifier gain is assumed to be freely available, but it is a source of possible confusion that is rarely mentioned. This situation arises in the following appendixes.

The topic of transient analysis addresses this same problem directly in the time domain. The analysis may or may not be easier to perform, but is generally much less widely familiar than the excursion through the frequency domain. The starting point, if one chooses an

---

*Consideration is restricted to signals for which this decomposition is possible, albeit with some need for care in the treatment of singularities such as steps and impulses.

†The foregoing sketch of analysis in the frequency domain is intended merely to connect up with the techniques with which the reader is likely to be most familiar. Discussion of the use of the Laplace transform would be out of place here since no use is made of Laplace transforms in the study.

elementary approach, is the step response of the network, H(t). Let the symbol u(t) represent a function:

$$u(t) = \begin{cases} 0 \text{ for time} < t \\ 1 \text{ for time} > t \end{cases}$$

This is called the unit step function; it is not analytic because it is not defined at time t. As such it is not a physically realizable signal, but it can be approximated in practice.

H(t) is the output of the network when the input is u(0). Because of causality, $H(t) = 0$ when $t < 0$. H(t) is often equal to zero at $t = 0$, but need not be; if the network is purely resistive, H(t) is a step function (possibly less than a unit step if the network is a voltage divider). H(t) can, in principle, be obtained by direct solution of the differential equations of the network with application of the appropriate boundary conditions, but may be difficult if the network is complicated and is characterized by numerous simultaneous differential equations. In some such cases, it might be easier to obtain H(t) by transformation from the frequency domain. Because u(0) can usually be approximated adequately by a real signal, it is generally possible to measure H(t).

In view of the assumption of linearity, it is evident that the response of the network to a step of amplitude v applied at time T will be vH(t-T). As noted above, the output will be zero for $t < T$ because of causality.

Consider now the response of the network to an arbitrary input f(t). We commence by approximating the input by a sequence of contiguous narrow rectangular pulses--similar to a histogram--and we ask: What is the output from the circuit at the present time T caused by all the input pulses that have heretofore been delivered?

Consider the pulse whose center was at earlier time $\tau < T$. The height of the pulse was $f(\tau)$, and it lasted from time $\tau - (\Delta\tau/2)$ to time $\tau + (\Delta\tau/2)$, where $\Delta\tau$ is the arbitrarily chosen small pulse width. The output at the present time T caused by that pulse is

$$f(\tau)\left[H\left(T-\tau+\frac{\Delta\tau}{2}\right) - H\left(T-\tau-\frac{\Delta\tau}{2}\right)\right]$$

The entire present output of the circuit is the sum over all the past pulses, each of which contributes an output in this form.

Consider the quantity

$$\left[\frac{H\left(T-\tau+\frac{\Delta\tau}{2}\right) - H\left(T-\tau-\frac{\Delta\tau}{2}\right)}{\Delta\tau}\right]\Delta\tau$$

As $\Delta\tau$ is made smaller and approaches the increment $d\tau$, the quantity in brackets approaches the time derivative of H evaluated at time $T - \tau$. (We assume the existence of this limit.) Let us define the function

$$h(t) = \frac{d}{dt} H(t)$$

We note that some care is needed when $t = 0$ if $H(t)$ has a step at $t = 0$; some instances of this will be treated in Appendix B.

Thus, as $\Delta\tau \to d\tau$ the present circuit output attributable to the input increment centered at time τ approaches

$$h(T-\tau)f(\tau)d\tau$$

Inasmuch as the input might have been arriving for an infinitely long time, and the present output must be obtained by summing over the entire past, the present output is given by the integral

$$g(T) = \int_{-\infty}^{T} h(T-\tau)f(\tau)d\tau$$

If the input signal was zero prior to some earlier time, then that earlier time can be used as the lower limit.

This is called a convolution integral, and is the method whereby one obtains the output response directly in the time domain. The

convolution may or may not be easier to carry out than the counterpart Fourier integration. The function h(t), which is defined here as the time derivative of the step response, is called the impulse response. It is needed as the starting point, just as the transfer function $\Gamma(\omega)$ is needed in the frequency-domain approach. Both of these functions, h(t) and $\Gamma(\omega)$, are characteristics of the network; they are a Fourier transform pair, and either can be obtained from the other by a Fourier transform.

The origin of the name of the impulse response, and an alternative view of its character, are made evident as follows. Consider an input to the network that is a rectangular pulse whose duration is Δt, centered at time zero, with height v:

$$f(t) = \begin{cases} 0 \text{ for } t < -\Delta t/2 \\ v \text{ for } -\Delta t/2 < t < +\Delta t/2 \\ 0 \text{ for } t > +\Delta t/2 \end{cases}$$

Let us define

$$I = \int_{-\infty}^{+\infty} f(t)dt = v\Delta t$$

This pulse, which can be regarded as a pair of steps, causes output response

$$v\left[H\left(t+\frac{\Delta t}{2}\right) - H\left(t-\frac{\Delta t}{2}\right)\right] = I\left[\frac{H\left(t+\frac{\Delta t}{2}\right) - H\left(t-\frac{\Delta t}{2}\right)}{\Delta t}\right]$$

Now let Δt approach zero while keeping constant the value I of the integral. The pulse height v increases without limit as the pulse width approaches zero, and the limit that is approached by f(t) is called an impulse. Like the unit step, it is not an analytic function; it too can be approximated by real signals.

As $\Delta t \to 0$, the quantity in brackets on the right approaches the derivative of H, which we have called h. The output approaches

Ih(t). If the area I under f(t) has been chosen to have unit value, then the limiting form of the input pulse is called a unit impulse (or a Dirac delta function) and is commonly designated $\delta(0)$; if delivered at time T, then $\delta(T)$. The quantity called I here, the time integral of f(t), is called the strength of the impulse.

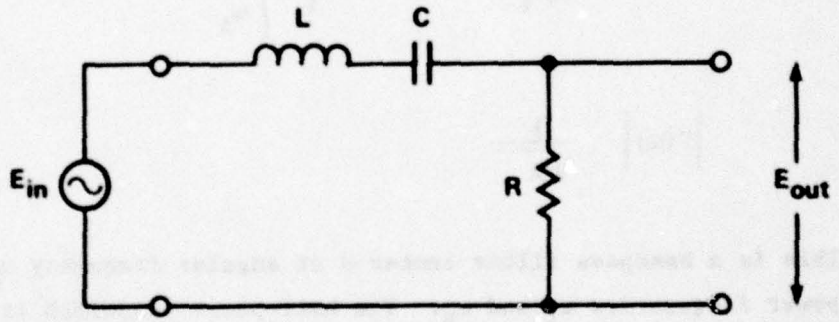
The dimensions of these quantities deserve some attention. For convenience, let us suppose that the network is an electric circuit, the signals are expressed in units of volts, and times are measured in seconds. H(t) was defined as the output response to a unit step input u(0). It would seem, then, that H(t) has dimensions of volts; the impulse response, the time derivative of H, then would have units of volts/second. However, if the input signal f(t) is expressed in volts, the convolution integral would yield a result in volts^2, and that is impossible. Either one must take the view that $f(\tau)$ in the integrand is a dimensionless proportionality factor, or one must decide that H(t) is really dimensionless. The latter choice is more satisfactory; H(t) should be regarded as the *dimensionless ratio* of the output response amplitude to the height of the input step, rather than the response to a 1-volt step. Thus, h(t) has dimensions time^{-1}, which usually shows up in the form of a frequency factor in a coefficient. By taking the Fourier integral over time of h(t), one obtains the dimensionless transfer function $\Gamma(\omega)$. The strength of an impulse, given by the time integral of f(t), has dimensions of volt-seconds.

Appendix B

SIMPLE BANDPASS FILTERS

All of the frequency-selective networks considered in this study are composed of simple LCR or RC series circuits that are completely isolated from the adjacent circuits by isolation amplifiers. It is assumed that the input signal is delivered, in series, from a zero-impedence source, and that the output signal is observed by an infinite impedence voltmeter. Thus, infinite amplifier gain and infinite impedence mismatch are presumed. The resulting overall stage gain is considered to be unity. Inasmuch as the purpose of these appendixes is tutorial, some circuits will be discussed that are not involved in this analysis.

Consider first the singly resonant bandpass filter:



Let

$$\omega_R^2 = \frac{1}{LC}$$

$$Q = \frac{\omega_R L}{R} > \frac{1}{2}$$

The transfer function is

$$\frac{E_{out}}{E_{in}} = \Gamma(\omega) = \frac{1}{1 + jQ\left(\frac{\omega}{\omega_R} - \frac{\omega_R}{\omega}\right)}$$

When $\omega = \omega_R$, $\Gamma(\omega) = 1$ = maximum. When

$$\omega = \frac{\omega_R}{2Q}\left[\sqrt{1 + 4Q^2} \pm 1\right] = \begin{cases} \omega_B \\ \omega_A \end{cases}$$

$$\left|\Gamma(\omega)\right| = \frac{1}{\sqrt{2}}$$

This is a bandpass filter centered at angular frequency ω_R with half-power frequencies ω_A and ω_B. The half-power bandwidth is related to Q by

$$Q = \frac{\omega_R}{\omega_B - \omega_A} = \frac{f_R}{f_B - f_A}$$

Thus, although Q was defined as a figure of merit of the choke at the resonant frequency, Q can alternatively be defined (in this simple circuit) as the reciprocal of the fractional bandwidth.

It is important to note that

$$\omega_A \omega_B = \omega_R^2$$

and $$\omega_A + \omega_B \neq 2\omega_R$$

That is, ω_R is the geometric mean, not the arithmetic mean, of the half-power frequencies. Circuit responses are symmetric on logarithmic frequency scale.

The differential equation for this circuit is usually written

$$L\ddot{q} + R\dot{q} + \frac{q}{C} = E_{in}$$

where $\dot{q}$ is the instantaneous current at any point in the circuit. The complete solution is the sum of two functions. One of these is the complementary function

$$c_1 e^{m_1 t} + c_2 e^{m_2 t}$$

where $m_{1,2}$ are the roots of the algebraic equation

$$m^2 + \frac{\omega_R}{Q} m + \omega_R^2 = 0$$

This function describes the free-running characteristics of the circuit, and contains the two adjustable constants $c_{1,2}$ needed to specify the initial phase and amplitude of the excitation of the (single) normal mode.

The second portion of the complete solution is the particular integral that describes the circuit response to the forcing function. The particular integral is

$$\frac{e}{L}^{m_1 t} \int e^{(m_2 - m_1)t} \int E_{in}\, e^{-m_2 t}\, dt \cdot dt$$

It is convenient to define

$$\Delta = \sqrt{1 - \frac{1}{4Q^2}}$$

If $Q > \frac{1}{2}$, Δ is real. The special case $Q = \frac{1}{2}$ is called the dead-beat or critically damped condition; in that case $\Delta = 0$ and the complementary function has a different form.

In these terms,

$$m_1 = \omega_R \left[\frac{-1}{2Q} + j\Delta \right]$$

$$m_2 = \omega_R \left[\frac{-1}{2Q} - j\Delta \right]$$

To obtain the step response of the filter, we take E_{in} to be a unit step at time zero. Integration of the particular integral is simple and yields

$$\frac{1}{m_1 m_2 L} = C$$

The boundary conditions that, at time zero, $q = \dot{q} = 0$ suffice to determine the constants $c_{1,2}$, and we obtain

$$q(t) = C \left[1 + \frac{m_2 e^{m_1 t} - m_1 e^{m_2 t}}{m_1 - m_2} \right]$$

The current in the circuit is

$$\dot{q}(t) = \frac{Cm_1m_2}{m_1 - m_2}\left[e^{m_1t} - e^{m_2t}\right]$$

$$= \frac{e^{-\frac{\omega_R t}{2Q}}}{L\Delta\omega_R}\sin(\Delta\omega_R t)$$

and the step response of the filter is

$$H(t) = R\dot{q} = \frac{e^{-\frac{\omega_R t}{2Q}}}{\Delta Q}\sin(\Delta\omega_R t)$$

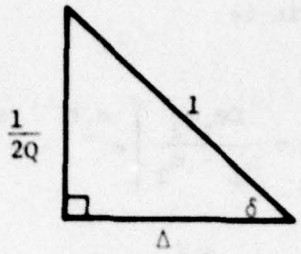
Before going further, it is noted that the sinusoidal factor does *not* ring at the tuned frequency ω_R, but at a somewhat lower frequency $\Delta\omega_R$. This widely overlooked feature of the ringing behavior of a simple tank circuit is a source of much of the analytic complication of this analysis. For instance, this consideration underlies the choice of tuned frequency ω_o/Δ_o for the single-stage bandpass filter attributed to the internal portion of the source.

To finish, we obtain the impulse response of the simple bandpass filter by differentiation:

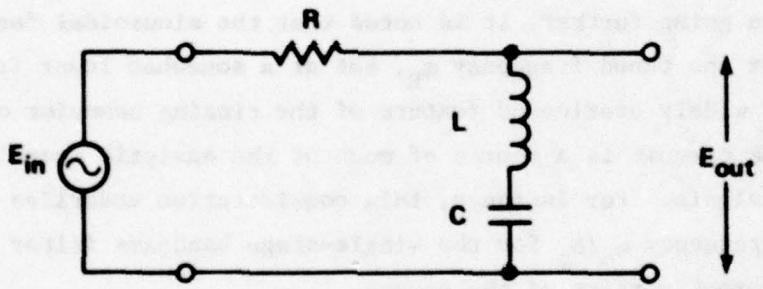
$$h(t) = \dot{H}(t)$$

$$= \frac{\omega_R}{\Delta Q}e^{-\frac{\omega_R t}{2Q}}\cos(\Delta\omega_R t+\delta)$$

where



The analysis described in this report does not consider band-reject circuits (even though such circuits are sometimes used in receivers). However, in the tutorial spirit of these appendixes, it is worthwhile to consider briefly a pitfall that can arise. Consider the band-reject circuit:



This is exactly the same circuit as the one considered above but with the output taken across the reactances instead of the resistance.

One can write out the step and impulse responses of this circuit directly from the results obtained above, because the sum of the voltages around the circuit must be zero at all times. Thus, the step response of this circuit is u(0) minus the step response of the bandpass circuit:

$$H(t) = u(0) - \frac{e}{\Delta Q} e^{-\frac{\omega_R t}{2Q}} \sin(\Delta\omega_R t)$$

$$h(t) = \delta(0) - \frac{\omega_R}{\Delta Q} e^{-\frac{\omega_R t}{2Q}} \cos(\Delta\omega_R t + \delta)$$

These expressions are correct, and the implication that

$$\frac{d}{dt} u(0) = \delta(o)$$

is correct. However there is a pitfall.

Suppose that this circuit had been considered initially, rather than the bandpass circuit. Exactly the same differential equation and initial conditions apply, and the same solution would have been obtained for q(t). In this circuit the step response is

$$H(t) = E_{out} = L\ddot{q} + \frac{q}{C}$$

Performing the indicated differentiation of the solution for q(t) written above yields

$$E_{out} = 1 + \frac{(m_1 + m_2)}{(m_1 - m_2)} \left[e^{m_1 t} - e^{-m_2 t} \right]$$

$$= 1 - \frac{e}{\Delta Q} e^{-\frac{\omega_R t}{2Q}} \sin(\Delta\omega_R t)$$

Here, an incautious routine solution of the differential equation obtains a 1 rather than a step u(0) in the leading term--no serious fault in itself, because the analyst knows full well that E_{out} is zero to the left of the origin--but differentiation of this expression fails to yield the impulse $\delta(0)$ in h(t).

The underlying difficulty is found in the particular integral. Two differentiations strip away the double integration, leaving the integrand $E_{in} = u(0)$.

When the particular integral was evaluated, this integrand was taken to be 1--which led to the correct bandpass step response. However, the correct function, u(0), must be retained here if one more differentiation is to yield the correct impulse response for the band-reject filter. This need for care will arise whenever the desired output function requires more successive differentiations of the solution of the differential equation than the multiplicity of integrations in the particular integral.

Finally, consideration of a bandpass filter composed of two fully isolated single-stage bandpass filters will provide a useful introduction to Butterworth bandpass filters, taken up in Appendix E.

We may define a parameter α such that one of these stages is tuned to $\alpha\omega_1$ and the other stage is tuned to ω_1/α, where ω_1 is the geometric mean of the two tuned frequencies and can be regarded as the center frequency of the filter. The two stages could, in general, have different Q's, but we will consider the case in which the two have the same Q.* One way to obtain the step response of the entire filter is to apply a step to the first stage and then to treat the step response of that stage as the input signal to the second stage. The response of the second stage is obtained by convolution, and is the step response of the entire filter. Evidently this rather primitive approach can, in principle, be used to determine the step response of any filter if the transient responses of the various stages can be determined. That determination is easy if each stage is a fully isolated LCR circuit, but we will find below that the task is not easy even in this case.

Let the stage tuned to $\alpha\omega_1$ be regarded as the first stage. (The choice is immaterial.) From above, the output of this stage is

---

*The Q of each stage is defined in terms of its own tuned frequency, not in terms of the center frequency.

$$\frac{e}{\Delta Q}^{-\frac{\alpha\omega_1 t}{2Q}} \sin(\alpha\Delta\omega_1 t)$$

The response of the filter is obtained by convolving this signal with the impulse response of the second stage:

$$h(t) = \frac{\omega_1}{\alpha\Delta Q} e^{-\frac{\omega_1 t}{2\alpha Q}} \cos\left(\frac{\Delta}{\alpha}\omega_1 t+\delta\right)$$

Thus, the step response of the two-stage filter is

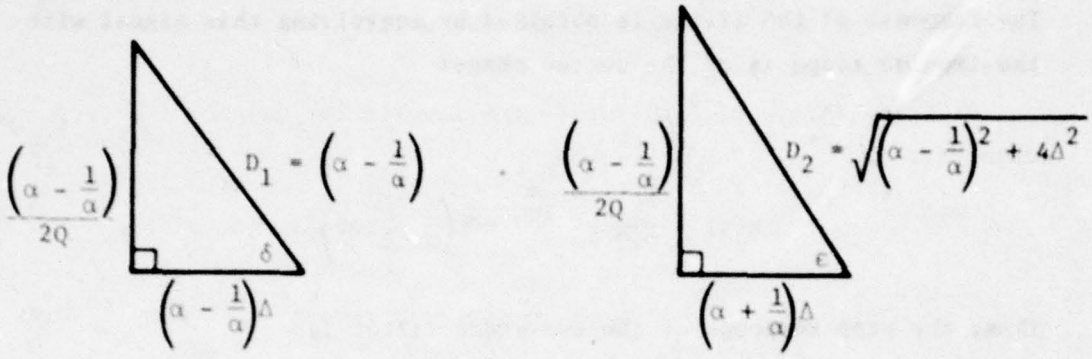
$$H(t) = \frac{\omega_1 e^{-\frac{\omega_1 t}{2\alpha Q}}}{\alpha(\Delta Q)^2} \int_0^t e^{-\left(\alpha-\frac{1}{\alpha}\right)\frac{\omega_1\tau}{2Q}} \cos\left[\frac{\Delta}{\alpha}\omega_1(t-\tau) + \delta\right] \sin(\alpha\Delta\omega_1\tau)d\tau$$

The product of two sinusoids can be decomposed into sum and difference terms, which leads to two integrals of the form

$$\int e^{px} \begin{matrix}\sin\\ \cos\end{matrix}(ax+\phi)dx = e^{px}\left[\frac{p \begin{matrix}\sin\\ \cos\end{matrix}(ax+\phi) \mp a \begin{matrix}\cos\\ \sin\end{matrix}(ax+\phi)}{p^2 + a^2}\right]$$

This form occurs over and over throughout the analysis, and presents no difficulty. We never encounter an integral whose value cannot be written out immediately in this form. The quantity a in this generic form is different, however, in the two simple integrals of this problem; in one it involves the sum $\alpha + 1/\alpha$, and in the other the difference $\alpha - 1/\alpha$. Consequently, the expression for H(t) ends up with terms in two different denominators. This is the crux of all the laborious tedium encountered in this study: one faces the difficulty of finding some way to collect terms with different denominators. Complex notation is no help, for it leads to exactly the same denominators.

It is convenient to define some new quantities that are closely related to Δ and δ, defined earlier:



Quantities counterpart to δ ϵ, D_1, and D_2 will be used in Appendix E.

With these the step response can be expressed:

$$H(t) = \frac{e^{-\frac{\omega_1 t}{2\alpha Q}}}{2\alpha(\Delta Q)^2}\left[\left[e^{-\left(\alpha-\frac{1}{\alpha}\right)\frac{\omega_1\tau}{2Q}}\left\{ -\frac{\cos\left[\Delta\left(\alpha-\frac{1}{\alpha}\right)\omega_1\tau + \frac{\Delta}{\alpha}\omega_1 t\right]}{D_1} - \frac{\cos\left[\Delta\left(\alpha+\frac{1}{\alpha}\right)\omega_1\tau - \frac{\Delta}{\alpha}\omega_1 t - \delta - \epsilon\right]}{D_2}\right\}\right]\right]_0^t$$

When the limits are inserted, the terms can be collected to yield

$$H(t) = \frac{1}{\Delta Q^2 D_1 D_2}\left\{ e^{-\frac{\omega_1 t}{2\alpha Q}}\cos\left(\frac{\Delta}{\alpha}\omega_1 t + \epsilon\right) - e^{-\frac{\alpha\omega_1 t}{2Q}}\cos(\alpha\Delta\omega_1 t - \epsilon)\right\}$$

Note that the step response of the two-stage filter contains two terms, each characteristic of the normal mode of one of the stages, but with the modal phases and amplitudes influenced by their effect on each

other. This will generally be the case, and the task of determining the transient response of a multistage (isolated) filter amounts to finding the amplitude coefficients and the phases. The general form that the transient response will take, including the exponential decrements and the ringing frequencies, can be written by inspection from the several normal modes.

Differentiation of H(t) yields the impulse response of the two-stage filter:

$$h(t) = \frac{\omega_1}{\Delta Q^2 D_1 D_2}\left\{\alpha e^{-\frac{\alpha\omega_1 t}{2Q}}\sin(\alpha\Delta\omega_1 t-\varepsilon+\delta) - \frac{e}{\alpha}e^{-\frac{\omega_1 t}{2\alpha Q}}\sin\left(\frac{\Delta}{\alpha}\omega_1 t+\varepsilon+\delta\right)\right\}$$

As a matter of interest, consider the response of this filter to a steady-state sine wave, $\sin(\omega_1 t)$, at the center frequency. The response is obtained by convolution:

$$\int_{-\infty}^{t} h(t-\tau)\sin(\omega_1\tau)d\tau = \frac{\sin(\omega_1 t)}{1+\left[Q\left(\alpha-\frac{1}{\alpha}\right)\right]^2}$$

As might be expected, the output is not phase-shifted, but there is an insertion loss; neither stage passes frequency ω_1 unattenuated. Note further that this is also the result that would have been obtained from the product of the two transfer functions when $\omega = \omega_1$:

$$\frac{1}{1+jQ\left(\alpha-\frac{1}{\alpha}\right)}\cdot\frac{1}{1-jQ\left(\alpha-\frac{1}{\alpha}\right)} = \frac{1}{1+\left[Q\left(\alpha-\frac{1}{\alpha}\right)\right]^2}$$

No factor has been discarded to normalize these individual transfer functions (they are naturally normalized) and the impulse response obtained above is indeed the Fourier transform of the transfer function of the entire filter when that transfer function is written in this form. However, as noted in Appendix A, it is commonplace to find such a transfer function to be normalized to remove the insertion loss. If the transfer function of the two-stage filter were written in normalized form it would appear as

$$\Gamma(\omega) = \frac{1 + Q^2 \left(\alpha - \frac{1}{\alpha}\right)^2}{1 - Q^2\left[\frac{\omega^2}{\omega_1^2} + \frac{\omega_1^2}{\omega^2} - \alpha^2 - \frac{1}{\alpha^2}\right] + jQ\left(\alpha + \frac{1}{\alpha}\right)\left(\frac{\omega}{\omega_1} - \frac{\omega_1}{\omega}\right)}$$

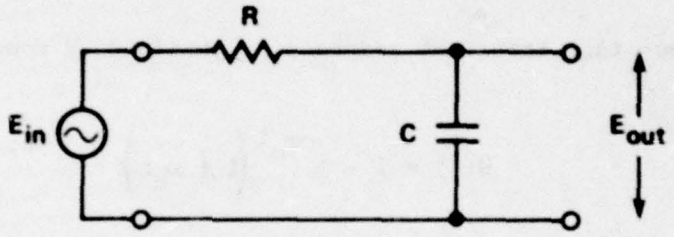
and this differs from the transform of h(t) by the factor in the numerator.

Appendix C

SIMPLE LOWPASS FILTERS

The only lowpass filter used in the study is a multistage RC filter in the envelope detector. Lowpass LRC and RC filters form part of the discussion in Appendix D.

Consider first the fully isolated single RC lowpass stage:



Let

$$\omega_c \triangleq \frac{1}{RC}$$

Then

$$\Gamma(\omega) = \frac{1}{1 + j\frac{\omega}{\omega_c}}$$

$$H(t) = 1 - e^{-\omega_c t}$$

$$h(t) = \omega_c e^{-\omega_c t}$$

When $\omega = 0$, $\Gamma(\omega) = 1 =$ maximum.

When $\omega = \omega_c$, $|\Gamma(\omega)| = \frac{1}{\sqrt{2}}$; the half-power bandwidth is ω_c.

Consider next two cascaded identical isolated RC stages:

$$\Gamma(\omega) = \left(\frac{1}{1 + j\frac{\omega}{\omega_c}}\right)^2 = \frac{1}{1 - \left(\frac{\omega}{\omega_c}\right)^2 + 2j\frac{\omega}{\omega_c}}$$

When $\omega/\omega_c = \sqrt{\sqrt{2} - 1}$, $|\Gamma(\omega)| = \frac{1}{\sqrt{2}}$; the half-power bandwidth is $0.64359\ \omega_c$.

The two-stage transient response is obtained by convolution:

$$H(t) = 1 - e^{-\omega_c t}\left(1 + \omega_c t\right)$$

$$h(t) = \omega_c e^{-\omega_c t}\left(\omega_c t\right)$$

For N cascaded identical isolated RC stages:

$$\Gamma(\omega) = \left(\frac{1}{1 + j\frac{\omega}{\omega_c}}\right)^N$$

When $\omega/\omega_c = \sqrt{2^{1/N} - 1}$, $|\Gamma(\omega)| = \frac{1}{\sqrt{2}}$; the half-power bandwidth moves to lower frequences as more stages are added.

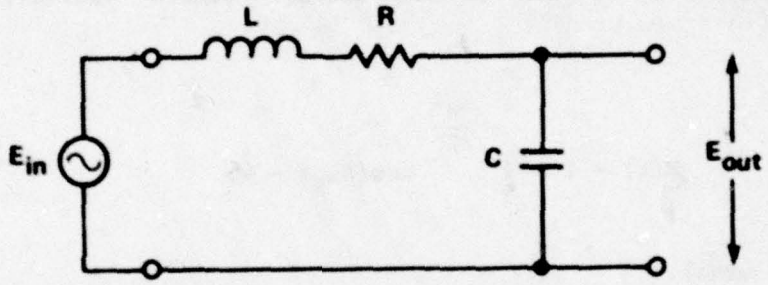
For the N stages:

$$H(t) = 1 - e^{-\omega_c t}\left[1 + \omega_c t + \frac{(\omega_c t)^2}{2!} + \ldots + \frac{(\omega_c t)^{N-1}}{(N-1)!}\right]$$

$$h(t) = \omega_c e^{-\omega_c t}\left[\frac{(\omega_c t)^{N-1}}{(N-1)!}\right]$$

In the study reported here, this impulse response was used to obtain the output of the envelope detector (see Appendix I). The convolution integral was evaluated numerically, with N = 4 and with ω_c = 3/8 of the IF frequency.

Consider now the lowpass LCR filter:



As usual, let

$$\omega_R^2 = \frac{1}{LC}$$

$$Q = \frac{\omega_R L}{R} > \frac{1}{2}$$

This will be recognized to be the same series circuit as in the simple bandpass filter treated in detail in Appendix B, except that here the output is taken across the capacitor instead of the resistor. The transfer function is

$$\Gamma(\omega) = \frac{1}{1 - \left(\frac{\omega}{\omega_R}\right)^2 + j\frac{\omega}{Q\omega_R}}$$

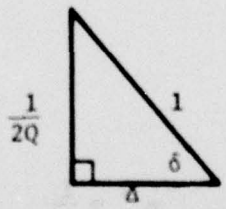
Note that $\Gamma(\omega) = 1 =$ maximum when $\omega = 0$. The half-power frequency is given by the cumbersome expression:

$$\omega_{\frac{1}{2}} = \omega_R\sqrt{\sqrt{1 + \left(1 - \frac{1}{2Q^2}\right)^2} + 1 - \frac{1}{2Q^2}}$$

The solution for q/C obtained in Appendix B can be used directly to obtain the step response of this lowpass filter. Collecting terms, we have:

$$H(t) = 1 - \frac{e}{\Delta}\, e^{-\frac{\omega_R t}{2Q}} \cos(\Delta\omega_R t - \delta)$$

where, as usual,



Similarly, from Appendix B, we obtain the impulse response

$$\frac{\dot{q}}{C} = h(t) = \frac{\omega_R e}{\Delta}\, e^{-\frac{\omega_R t}{2Q}} \sin(\Delta\omega_R t)$$

(Note that this single differentiation of the particular integral does not involve the derivative of the step function.)

Appendix D

LOWPASS-BANDPASS EQUIVALENCE

Calculation of the response of a bandpass filter to a modulated input signal is usually laborious. Not only does the impulse response of the filter contain numerous terms, but the convolution integration generates still more terms that are difficult or impossible to collect. It is virtually universal practice to use an approximate method, and the exact calculation is done only rarely. (In fact, even the approximate transient analysis is carried out infrequently; most circuit design practice considers only the frequency responses of the circuits.) This appendix treats the basis for the approximate method and its shortcomings.

The approximate method involves the concept of a lowpass filter whose behavior can be regarded, at least approximately, as "equivalent" to the behavior of the bandpass filter. Equivalence is usually described in terms of the frequency responses or transfer functions of the two filters: the circuit constants of the lowpass filter are so chosen that, near zero frequency, the functional form of the transfer function is the same as the functional form of the transfer function of the bandpass filter for frequencies near the center frequency.

To illustrate, consider a single-stage bandpass filter whose center frequency is ω_R, and its equivalent lowpass RC filter. The transfer function of the bandpass filter is

$$\Gamma(\omega) = \frac{1}{1 + jQ\left(\frac{\omega}{\omega_R} - \frac{\omega_R}{\omega}\right)}$$

and that of the lowpass filter is

$$\Gamma(\omega) = \frac{1}{1 + j\,\frac{\omega}{\omega_c}}$$

Let $\Delta\omega = \omega - \omega_R$. For frequencies close to the center frequency of the bandpass filter

$$\left| \frac{\Delta\omega}{\omega_R} \right| << 1$$

and the bandpass transfer function is *approximately* equal to

$$\frac{1}{1 + jQ\left(\frac{2\Delta\omega}{\omega_R}\right)}$$

If the cutoff frequency of the lowpass filter is adjusted to

$$\omega_c = \frac{\omega_R}{2Q}$$

then the lowpass transfer function is exactly

$$\frac{1}{1 + jQ\left(\frac{2\omega}{\omega_R}\right)}$$

If the frequency ω in the lowpass filter is regarded as "equivalent" to frequency deviation $\Delta\omega$ in the bandpass filter, then the two transfer functions have the same form. An alternative description is to say that the frequency response of the lowpass filter near zero frequency has the same shape as the frequency response of the bandpass filter near its center frequency.

To assess the approximation above, consider the case $Q = 2$, $\omega = 2\omega_R$. At this frequency, the transfer function of the bandpass filter is

$$\frac{1}{1 + 3j}$$

Here $\Delta\omega = \omega_R$, and at that distance from zero frequency the transfer function of the lowpass filter is

$$\frac{1}{1 + 4j}$$

The equivalence is inexact by 2.30 dB in amplitude and 4.40 degrees in phase. Although the equivalence is only approximate, these errors are not great. Moreover, $Q = 2$ is an unusually low value, and a frequency deviation of an octave from ω_R is unusually large. Thus, the results obtained this way are, for most ordinary purposes, sufficiently accurate and entirely satisfactory--principally because it is usually the quasi-steady state that is considered, and severely transient conditions are of no great interest.

The argument for equivalence assumes that the frequency response of the bandpass filter is (geometrically) symmetric about its center frequency. (The frequency response of the lowpass filter is necessarily symmetric about zero frequency.) The single-stage bandpass filter is symmetric, but many complicated bandpass filters are not. It is commonplace to design bandpass filters so the frequency response slopes one way or the other across the pass band, and to include one or more band-stop circuits to reject particular frequencies that are not symmetrically disposed about the center frequency. An asymmetric bandpass filter does not possess a lowpass equivalent. All lowpass filters have bandpass equivalents, but the converse is not true. All the bandpass circuits considered in this study are symmetric and do possess lowpass equivalents; however, no use is made of those equivalents in the analysis.

The lowpass equivalent method for the approximate analysis of transient behavior is:

- Assume, if necessary, that the input signal to the bandpass filter can be described as a time-varying envelope multiplied by a fixed sinusoidal carrier at the center frequency of the filter.

- o Calculate the output that would result if the envelope were input to the lowpass equivalent filter.
- o Multiply that lowpass output response by the original carrier, and regard the product as the approximate output from the bandpass filter.

The method amounts to assuming that the bandpass filter treats the sidebands in about the same way that the lowpass filter would treat the upper (or lower) sideband if the sideband were shifted near zero frequency.

To illustrate, we consider the response of a single-stage bandpass filter to a step-modulated carrier at the center frequency:

$$u(0)\ \sin(\omega_R t + \psi)$$

The convolution of this input with the impulse response of the single-stage bandpass filter yields the exact bandpass output:

$$\int_0^t h(t - \tau)\ \sin(\omega_R \tau + \psi)d\tau$$

$$= \frac{\omega_R}{\Delta Q} e^{-\frac{\omega_R t}{2Q}} \int_0^t e^{+\frac{\omega_R \tau}{2Q}} \cos[\Delta\omega_R(t - \tau) + \delta]\ \sin(\omega_R \tau + \psi)d\tau$$

$$= \frac{\omega_R}{2\Delta Q} e^{-\frac{\omega_R t}{2Q}} \left\{ \int_0^t e^{+\frac{\omega_R \tau}{2Q}} \sin[\omega_R \tau(1 + \Delta) + \psi - \delta - \Delta\omega_R t]d\tau \right.$$

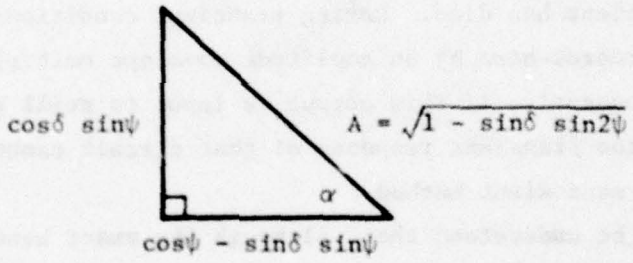
$$\left. + \int_0^t e^{+\frac{\omega_R \tau}{2Q}} \sin[\omega_R \tau(1 - \Delta) + \psi + \delta + \Delta\omega_R t]d\tau \right\}$$

It should be noted that the appearance of the factors $(1 + \Delta)$ and $(1 - \Delta)$ is characteristic of the exact analysis and leads to terms with different denominators. Most of the burden of the exact analysis arises here, and not in carrying out the necessary integrations. Some readers may find it worthwhile to go through the labor of obtaining the final result to acquaint themselves with the manipulations needed to collect terms.

The exact bandpass output is

$$\sin(\omega_R t + \psi) - \frac{e^{-\frac{\omega_R t}{2Q}}}{\Delta} A \sin(\Delta\omega_R t + \alpha)$$

where



The ease with which the lowpass equivalent is analyzed illustrates why the exact analysis is done so rarely. The envelope, in this case, is the step function u(0). When the envelope is input to the lowpass filter, the output is the step response of that filter:

$$1 - e^{-\frac{\omega_R t}{2Q}}$$

Multiplying that output by the original carrier yields the approximate bandpass output:

$$\sin(\omega_R t + \psi) - e^{-\frac{\omega_R t}{2Q}} \sin(\omega_R t + \psi)$$

Comparison of the two results shows them to be similar. Both approach $\sin(\omega_R t + \psi)$, as they must, when $t \to \infty$. However, the exact solution yields a transient term that differs from the approximation in two respects:

- o The amplitude of the transient term depends on the carrier phase ψ.
- o The frequency of the sinusoid in the transient term is the modal frequency of the tuned circuit, $\Delta\omega_R$, whereas the approximate solution yields two terms at the same frequency.

The exact output signal does not have evenly spaced zero crossing until the transient has died. During transient conditions, the output cannot be represented by an amplitude envelope multiplying a fixed carrier. Consequently, if this output is input to still another bandpass circuit, the transient response of that circuit cannot be treated by the lowpass equivalent method.

It should be understood that, although the exact bandpass output cannot be represented by an envelope multiplying a fixed carrier, the output signal does possess an envelope in the formal sense. If one forms the sum of the squares of two outputs with phase ψ and with phase $\psi + \pi/2$, the resulting sum of the squares does not contain ψ. That is, that result is the square of the formal envelope. However, the formal envelope is a fairly complicated function that resembles the lowpass envelope only in general shape.

Finally, it is noted that

$$\text{as } Q \to \infty, \quad \alpha \to 0$$
$$\Delta \to 1$$
$$\alpha \to 4$$
$$A \to 1$$

and the exact output approaches the result obtained by approximation.

Appendix E

BUTTERWORTH BANDPASS FILTERS

An N-pole Butterworth bandpass filter contains N mutually isolated simple LCR bandpass stages and no band-reject stages. The fundamental characteristic of the Butterworth design is the form of the frequency response:

$$\Gamma\Gamma^* = \frac{1}{1 + \left[Q_1 \left(\frac{\omega}{\omega_1} - \frac{\omega_1}{\omega} \right) \right]^{2N}}$$

where Γ is the normalized transfer function, ω_1 is the (geometric) center frequency of the filter and

$$Q_1 = \frac{\omega_1}{\text{full (angular) half-power bandwidth}}$$

The nth derivative of $\Gamma\Gamma^*$ contains in the numerator a factor

$$\left[Q_1 \left(\frac{\omega}{\omega_1} - \frac{\omega_1}{\omega} \right) \right]^{2N-n}$$

Consequently, the first 2N derivatives are zero when $\omega = \omega_1$. For this reason, the Butterworth design is also called maximally flat. This feature offers no special advantage or drawback in practice, but this particular design is not widely used--in part because other designs offer preferred amplitude responses across the passband. Sometimes a group of Butterworth stages are combined with other band-reject stages in IF strips.

If the N stages were to be tuned to arbitrarily chosen frequencies and had arbitrary Q's, the denominator of $\Gamma\Gamma^*$ would be a polynominal.

AD-A078 373 RAND CORP SANTA MONICA CA F/G 17/3
TRANSIENT RESPONSE OF A HETERODYNE RECEIVER: IMPLICATIONS FOR A--ETC(U)
NOV 79 T F BURKE F49620-77-C-0023
UNCLASSIFIED RAND/R-2418-AF NL

3 OF 3
AD-
A078373

END
DATE
FILMED
1 80
DDC

The prescription that all except the first and last terms have zero coefficients furnishes a set of equations that specifies the design completely. If N is an even number, then there are N/2 symmetric pairs of stages. The two members of the pair have the same Q (each Q is defined at the tuned frequency of that particular stage), and the product of their tuned frequencies is ω_1^2; they are tuned to geometrically symmetric frequences about ω_1. If N is an odd number, then there are (N-1)/2 such pairs, and one stage is tuned to ω_1 with Q equal to Q_1.

To discuss the tuning and Q's of the pairs it is convenient to use index j to characterize the two members of a pair. Let

$$j = 2, 4, 6, \ldots \begin{matrix} N \\ \text{or} \\ N-1 \end{matrix}$$

The jth pair has Q equal to Q_j. One member is tuned to frequency $\alpha_j\omega_1$ and the other to frequency ω_1/α_j.

The complex roots of the algebraic equation

$$x^{2N} + 1 = 0$$

lie at angles $\pm\beta_j$ from the negative real axis:

$$\beta_j = \frac{(j-1)}{N} \cdot \frac{\pi}{2}$$

These β_j are highly characteristic of the Butterworth design. They occur as phase angles, and their trigonometric functions appear in amplitude coefficients. For the N = 4 filter used in this study,

$$\beta_2 = \pi/8$$

$$\beta_4 = 3\pi/8$$

Let $c_j = 2\cos(2\beta_j)$. The design of the filter is completely specified by two inconvenient equations:

$$\frac{1}{Q_j^2} = \frac{1}{2Q_1^2} - 2\left[\sqrt{1 + c_j\left(\frac{1}{2Q_1}\right)^2 + \left(\frac{1}{2Q_1}\right)^4} - 1\right]$$

$$\left(\alpha_j - \frac{1}{\alpha_j}\right)^2 = \frac{1}{Q_1^2} - \frac{1}{Q_j^2}$$

The numerical values of the Q_j can be calculated from the first equation, and the values of α_j then obtained from the second.

The equation in α_j is biquadratic. The two negative roots are discarded. The two positive roots admit an immaterial choice between $\alpha_j \geq 1$ and $\alpha_j \leq 1$, corresponding to interchange of the two members of the pair. The choice $\alpha_j \geq 1$ is adopted here:

$$\alpha_j = \frac{Q_j}{Q_1}\sin\beta_j + \sqrt{\left(\frac{Q_j}{Q_1}\sin\beta_j\right)^2 - 1}$$

$$\frac{1}{\alpha_j} = \frac{Q_j}{Q_1}\sin\beta_j - \sqrt{\left(\frac{Q_j}{Q_1}\sin\beta_j\right)^2 - 1}$$

From these roots it follows that

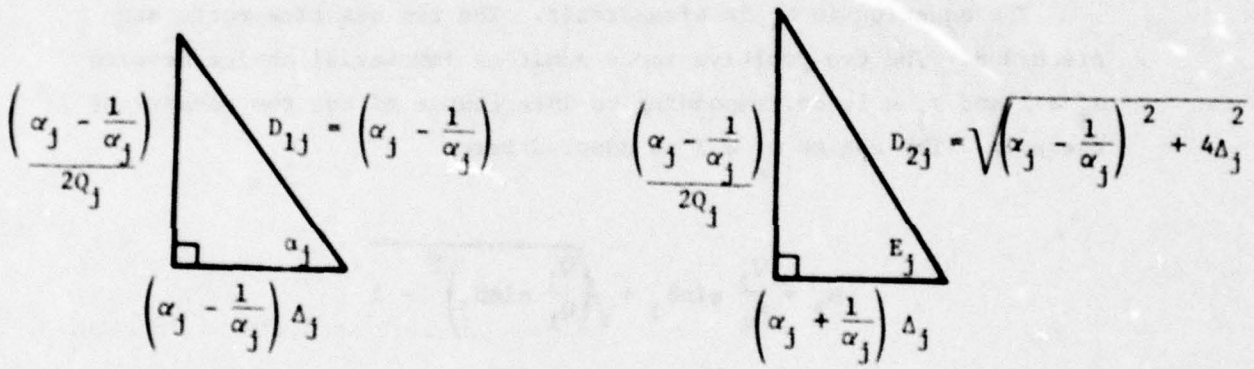
$$\left(\alpha_j + \frac{1}{\alpha_j}\right) = \frac{2Q_j}{Q_1}\sin\beta_j = \frac{Q_j}{Q_1}\sqrt{2 - c_j}$$

$$\left(\alpha_j - \frac{1}{\alpha_j}\right) = \frac{\cos\beta_j}{\Delta_j Q_1}$$

where, as usual,

$$\Delta_j = \sqrt{1 - \frac{1}{4Q_j^2}}$$

It is convenient to generalize, with index j, some quantities that were defined in Appendix B:



It is not difficult to show that

$$\sin(\delta_j + \epsilon_j) = \frac{\alpha_j \Delta_j}{Q_j D_{2j}}$$

$$\sin(\delta_j - \epsilon_j) = \frac{\sin(\delta_j + \epsilon_j)}{\alpha_j^2}$$

The various relationships given above are especially useful in algebraic manipulations aimed at collecting terms during analysis.

For the choices N = 4, Q_1 = 8 adopted here

$$\alpha_2 = 1.0594 \qquad D_{12} = 0.1155$$
$$\alpha_4 = 1.0242 \qquad D_{14} = 0.0479$$

$$\Delta_2 = 0.9997 \qquad \delta_2 = 0.02388$$
$$\Delta_4 = 0.9983 \qquad \delta_4 = 0.05776$$

$$1/2Q_2 = 0.02388 \qquad c_2 = 0.001377$$
$$1/2Q_4 = 0.05773 \qquad c_4 = 0.001385$$

$$1/D_{22} = 0.4993$$
$$1/D_{24} = 0.5007$$

$$\sin \beta_2 = \cos \beta_4 = \frac{1}{2}\sqrt{2 - \sqrt{2}}$$

$$\sin \beta_4 = \cos \beta_2 = \frac{1}{2}\sqrt{2 + \sqrt{2}}$$

(All the constants loaded into the computer were accurate to 16 significant figures, rounded off.)

The half-power frequencies of this filter lie at

$$\frac{\omega}{\omega_1} = \begin{cases} 1.0644512 \\ 0.9394512 \end{cases}$$

The frequencies at which the response is down 24.1 dB are

$$\frac{\omega}{\omega_1} = \begin{cases} 1.1327135 \\ 0.8828358 \end{cases}$$

These frequencies are only about 0.18 octave from ω_1 and not far outside the half-power band. Although this N = 4 design is rudimentary in comparison to most IF strips, the selectivity is substantial.

The difficulty of determining the impulse response of this N = 4 filter was one of the principal obstacles in this study. The impulse response of a conjugate pair of stages was discussed in Appendix B. That pair can be regarded as any one of the conjugate pairs in a Butterworth filter. The impulse response for N = 4 can be obtained by convolution of the impulse response of the j = 2 pair with the impulse response of the j = 4 pair. A review of the algebraic intricacy encountered in Appendix B will suggest the difficulty of collecting terms when this convolution is carried out. So far as we know the N = 4 impulse response has not previously been published.

The N = 4 bandpass impulse response is

$$h(t) = \frac{\omega_1(1+\sqrt{2})}{Q_1 D_{24}} \left\{ \alpha_4 e^{-\frac{\alpha_4 V}{2Q_4}} \sin(\alpha_4 \Delta_4 V + \frac{\pi}{8} - c_4 + \delta_4) \right.$$

$$\left. - \frac{e^{-\frac{V}{2\alpha_4 Q_4}}}{\alpha_4} \sin\left(\frac{\Delta_4}{\alpha_4} V - \frac{\pi}{8} + c_4 + \delta_4\right) \right\}$$

$$- \frac{\omega_1}{Q_1 D_{22}} \left\{ \alpha_2 e^{-\frac{\alpha_2 V}{2Q_2}} \sin\left(\alpha_2 \Delta_2 V + \frac{3\pi}{8} - c_2 + \delta_2\right) \right.$$

$$\left. - \frac{e^{-\frac{V}{2\alpha_2 Q_2}}}{\alpha_2} \sin\left(\frac{\Delta_2}{\alpha_2} V - \frac{3\pi}{8} + c_2 + \delta_2\right) \right\}$$

where, as usual, $V = \omega_1 t$. Phase angles β_2 and β_4 are evident; $1 + \sqrt{2}$ is tan (β_4).

This expression has been normalized to remove the insertion loss at ω_1. The derivation was checked by convolving h(t) with steady state $\sin(\omega t)$ to obtain the correct amplitude and phase of $\Gamma(\omega)$.

The lowpass equivalent of the bandpass filter was not used directly, although it was used to show the "envelopes" of the impulse- and stepped-carrier responses in Figs. III-13 and III-14. It appears that scale length S for higher order Butterworth bandpass filters could be estimated from the lowpass equivalent responses, and thus lowpass equivalents are discussed below.

All Butterworth bandpass filters are (geometrically) symmetric about the center frequency, and they do possess lowpass equivalents. The frequency response of the lowpass equivalent must be adjusted to

$$\Gamma\Gamma^* = \frac{1}{1 + \left(\frac{\omega}{\omega_c}\right)^{2N}}$$

where

$$\omega_c = \frac{\omega_1}{2Q_1}$$

If N is even, the lowpass filter contains N/2 mutually isolated lowpass LCR stages, all tuned to $\omega_R = \omega_c$, but with different Q's. If N is odd, the lowpass filter contains (N-1)/2 such lowpass LCR stages and also one lowpass RC stage whose cutoff frequency is ω_c. No lowpass equivalent of a Butterworth bandpass filter will contain more than one RC stage. The foregoing properties and the necessary values of the lowpass Q's are completely determined from the equations needed to eliminate the undesired terms in the denominator of $\Gamma\Gamma^*$. We tabulate the value of the lowpass Q's and the corresponding Δ's for several values of bandpass N:

N	j	Q	Δ
2	2	$\sin \beta_2$	$\cos \beta_2$
3	2	$2 \sin \beta_2$	$\cos \beta_2$
4	2	$\sqrt{2} \sin \beta_2$	$\cos \beta_4$
4	4	$\sqrt{2} \sin \beta_4$	$\cos \beta_2$
5	2	$2 \sin \beta_2$	$\cos \beta_4$
5	4	$2 \sin \beta_4$	$\cos \beta_2$
6	2	$2 \sin \beta_2$	$\cos \beta_6$
6	4	$\sin \beta_4$	$\cos \beta_4$
6	6	$2 \sin \beta_6$	$\cos \beta_2$
7	2	0.554958132[a]	$\cos \beta_6$
7	4	0.801927736	$\cos \beta_4$
7	6	2.246979605	$\cos \beta_2$
8	2	$\sqrt{4+2\sqrt{2}} \sin \beta_2$	$\cos \beta_8$
8	4	$\sqrt{4-2\sqrt{2}} \sin \beta_4$	$\cos \beta_6$
8	6	$\sqrt{4-2\sqrt{2}} \sin \beta_6$	$\cos \beta_4$
8	8	$\sqrt{4+2\sqrt{2}} \sin \beta_8$	$\cos \beta_2$

[a]For N = 7, the Q's are determined by a cubic equation.

To avoid confusion with the symbols, Q_2, Q_4, Δ_2, Δ_4 in the bandpass filter, we employ subscripts a and b corresponding to j = 2, 4. The step and impulse responses of the lowpass equivalent of the N = 4 bandpass filter are

$$H(t) = 1 - (1 + \sqrt{2})e^{-\frac{\omega_c t}{2Q_a}}\sin(\Delta_a \omega_c t + \tfrac{\pi}{4}) - e^{-\frac{\omega_c t}{2Q_b}}\sin(\Delta_b \omega_c t - \tfrac{\pi}{4})$$

$$h(t) = \omega_c \cdot \left[(1+\sqrt{2})e^{-\frac{\omega_c t}{2Q_a}} \sin(\Delta_a \omega_c t + \tfrac{\pi}{8}) - e^{-\frac{\omega_c t}{2Q_b}} \sin(\Delta_b \omega_c t + \tfrac{3\pi}{8}) \right]$$

An obscure phenomenon arises in connection with this lowpass impulse response. The "envelope" shown in Fig. III-13 is twice the lowpass h(t) given here. Although this is the lowpass equivalent filter, and h(t) is its correct lowpass impulse response, the amplitude of the lowpass impulse response is only half the amplitude of the bandpass impulse response. The reason is obscure. An argument based on the frequency response of the lowpass filter below zero frequency (that is, the "other half" of the bandpass response) would lead to a factor of $\sqrt{2}$ rather than 2. The other $\sqrt{2}$ is found in the average value of sine squared. In any case, this amplitude coefficient is of no consequence because absolute signal level is rarely of concern when the lowpass equivalent is invoked.

It is of some interest to notice how only two lowpass terms mimic four bandpass terms in these transient responses. We see that

$$\frac{1}{Q_a} = \left(\alpha_4 + \frac{1}{\alpha_4}\right)\frac{Q_1}{Q_4}$$

$$\frac{1}{Q_b} = \left(\alpha_2 + \frac{1}{\alpha_2}\right)\frac{Q_1}{Q_2}$$

Consequently,

$$\exp\left[-\frac{\omega_c t}{2Q_a}\right] = \sqrt{\exp\left[-\frac{\alpha_4 \omega_1 t}{2Q_4}\right]\exp\left[-\frac{\omega_1 t}{2\alpha_4 Q_4}\right]}$$

$$\exp\left[-\frac{\omega_c t}{2Q_b}\right] = \sqrt{\exp\left[-\frac{\alpha_2 \omega_1 t}{2Q_2}\right]\exp\left[-\frac{\omega_1 t}{2\alpha_2 Q_2}\right]}$$

The lowpass exponents are twice the arithmetic means of the pairs of bandpass exponents, so the lowpass exponential is the geometric mean of the bandpass exponentials. Similar relationships exist among the Δ's.

Appendix F

ANTENNA THEORY

Maxwell's equations provide the differential equations that govern the classical behavior of electromagnetic waves. They furnish a vector wave equation that can be separated into two scalar wave equations for the electric and magnetic fields. (Acoustic problems lead to a scalar wave equation of the same form.) The scalar wave equation relates the spatial and temporal dependences of the field strength (or acoustic pressure). The temporal factor can be removed in steady-state conditions, and the resulting wave equation is called the Helmholtz equation.*

The general solution of the steady-state wave equation is well known: it consists of all possible waves running in all possible directions. The only problem in writing out the complete solution arises in choosing among the numerous possible mathematical representations of the waves. The general solution is useless as it stands. Every problem that arises in wave propagation consists of finding the subset of possible waves that are consistent with the conditions of that problem and eliminating those that are not. That is to say, one must discover the set of waves that conform to the electric or magnetic or acoustic boundary conditions on all portions of all physical surfaces, and that fulfill the necessary conditions of mathematical continuity, conservation of energy, and so on. These are boundary-value problems, and all problems of linear wave propagation are boundary-value problems.

If the physical shapes of the boundaries are just right, we know how to write out the solution of the boundary-value problem. "Just right" means that the surfaces must conform to one of the 13 orthogonal coordinate systems in which the wave equation is separable (e.g., Cartesian, cylindrical, spherical, ellipsoidal, oblate and prolate spheroidal, etc.). The limitation to 13 does not await somebody finding

---

*The Helmholtz equation, and the associated boundary conditions, are not applicable to transient problems.

another; it has been proved that, in 3-dimensional space, there cannot be another. Thus it is fair to say that all these possible solutions to 3-dimensional boundary value problems were obtained long ago and are given in various textbooks. Problems that do not conform to the 13 coordinate systems require that the solution be represented by combinations of basic solutions, usually as infinite series whose convergence may be questionable, or by making approximations.

Advances in this area, even for seemingly simple problems, are remarkably scarce. Two notable solutions were obtained by Nobel laureates: in 1896 Arnold Sommerfeld obtained the diffraction pattern of an infinitely long straightedge with incident plane waves; in 1948 Julian Schwinger and H. Levine obtained the solution for acoustic waves emerging from the end of an infinitely long cylindrical pipe. To this day no wholly satisfactory solution has been found for the dipole antenna.

The reader should be disabused by now of any possible belief that there are prospects for solving the boundary-value problems posed by most real antennas--or even highly idealized versions of them. Moreover, the physical shapes and material constants of real antennas are so inconvenient that numerical solutions are beyond the capacity of present-day computers. The problem might be compared to undertaking the complete solution for world-wide weather prediction. Finally, it should be noted that even such famous achievements as Sommerfeld's and Schwinger's dealt with steady-state single-frequency conditions. In this study, it is the transient behavior of the antenna that is important, which means that we need the time-dependent field throughout space or else we need properly to sum the steady-state solutions over all frequencies.

Approximation--often very drastic--is necessary when studying real antennas. For directional antennas such as those used for radar or sonar the usual approach is to turn to optical diffraction theory.

The physical arrangements of primary interest in optics lent themselves to comparatively convenient geometry. The problems concerning the diffraction pattern of light passing through a hole in a screen, or of a multislit diffraction grating, invited idealization wherein one considered an opaque screen, geometrically plane, of infinite extent. The hole (or holes) in the screen, however, present mixed geometry. A round hole lends itself best to cylindrical coordinates whereas the incident light lends itself to Cartesian. Much mathematical entertainment was found in patching these together. Clearly, the boundary conditions at the edge of the hole present difficulty. If the screen is infinitesimally thin (a zero-thickness screen cannot be opaque, but one might overlook that) the radius of curvature of the edge is zero and the electric field becomes infinite. To avoid this situation, one can give the edge some finite radius of curvature, but that introduces still another coordinate frame (say, torroidal), and the problem is much more difficult.

A number of papers have described, in various orders of approximation, the diffraction pattern of a round hole in an infinite screen. Much of the foundation of diffraction theory was laid beneath this problem. Interested readers will find the formal theory and the usual approximations developed in some of the standard textbooks.* Various powerful general theorems are used to transform the boundary-value problem from one form that cannot be solved exactly to another form that cannot be solved exactly. The formulation discussed here is associated with the names of Huygens, Fresnel, Helmholtz (acoustics), Kirchoff, and Green, together with many others. The theoretical foundations are usually framed around steady-state conditions and monochromatic light (infinite coherence length). The principles whereby the analysis could be extended to transients are well known but often are

---

*See *Principles of Optics*, Born and Wolf, Macmillan, Chapters VIII and XI. See also *Introduction to Physical Optics*, Goodman, McGraw-Hill, Chapters 3 and 4. Some of the formal difficulties in diffraction theory are suggested by Goodman's criticism (page 38) of an argument advanced by Born and Wolf (section 8.3.2) to justify setting a certain integral equal to zero.

not discussed; the steady state is hard enough. (Born and Wolf's book is a notable exception--they do discuss partial coherence in considerable detail.)

The Kirchoff formulation of diffraction theory permits the field at any point in front of the screen to be obtained by evaluating integrals of the field over the entire area of a closed mathematical surface surrounding the point. The closed surface that is adopted consists of the entire infinite area of the front face of the screen (including the hole) plus a hemisphere of infinite radius. It is argued that the strength of the field on the hemispherical surface contributes nothing (a delicate point; see Goodman, pp. 38, 39) and the original problem becomes one of determining the field everywhere in the plane of the screen.

However, the field in the plane of the screen cannot be determined exactly. Numerous papers have presented approximate solutions, all of which are tedious. The usual approach, and the one with which people are generally familiar, is to adopt the Kirchoff approximate boundary conditions to the Huygens-Fresnel-Kirchoff formulation. This is justified, on pragmatic grounds, by the argument that the diameter of the hole is very large compared with a wavelength, for which reason the approximation works pretty well in the region in which one is usually interested--at angles not too far from the principal axis and only through the first few diffraction fringes (side lobes).

There are two Kirchoff approximate boundary conditions, neither of which is physically possible because they violate continuity conditions:

1. The field strength is taken to be zero everywhere on the surface of the screen (i.e., the forward side of the screen is dark).
2. The field strength everywhere in the hole is taken to be that which would have been present if the screen were absent.

Assumption (1) sets equal to zero the integral over the infinite surface of the screen. Thus it is supposed that the field strength

at the field point can be determined from an integral carried only over the hole--the aperture. It is a sure hallmark that the Kirchoff approximation is being used (often without explicit mention) when the radiation field of an antenna purports to be obtained from a bounded integral over "the aperture."

Assumption (2) is a tacit assertion that we know what the field distribution is in the aperture, namely, that it is what we prescribe it to be. In R-1819-PR and in this report it is taken to be constant over the aperture, by prescription of the author, despite the fact that uniformity is physically impossible.

The Kirchoff approximate boundary conditions lead to the well-known standard directivity patterns such as sinZ/Z for the uniform line and $2J_1(Z)/Z$ for the uniform circle. Other, nonuniform, illumination conditions, still based on the Kirchoff approximation, lead to lobe-suppressed patterns, steered beams, and so on. These work well enough for ordinary purposes because detailed interest rarely extends beyond the first few side lobes. Indeed, it is not as widely known as it should be that this analytic basis is only approximate and cannot be pushed too far. However, it is generally known that real directivity patterns, even for simple source geometry, do not match the theory very well beyond the first few lobes. In TOA systems, where detailed interest must extend to angles far from the principal axis, this difficulty is unusually important.

However, for the purposes of this study, the most important difficulty does not lie in the possible shortcomings of the Kirchoff approximate boundary conditions. It is more important that the model of a radar antenna as a hole in an infinite opaque screen is wrong, irrespective of how accurately one determines the field strength on the screen and in the hole. This is most obvious if one considers the radiated field of the antenna at angles larger than 90 degrees from the axis: the optical model imposes a screen that prohibits turning through such large angles. If one accepts, say, sinZ/Z at face value, that function indicates that the antenna sends the same pattern into the rear hemisphere as into the forward hemisphere.

Turning to a time-domain consideration of the antenna, the Kirchoff approximate boundary conditions attribute to the aperture--and thus to the entire antenna structure--an impulse response that is exactly an impulse on the axis (for plane uniform illumination), and a square-sided flat-topped "box car" shape off axis (for a line source). These will be discussed and illustrated in further detail in Appendix G; this impulse response assumption assumes that no currents are induced in any portion of the structure. If such currents were induced (as they must be) they would radiate, and the signal received on the axis would have an oscillatory tail.

There is, at present, nothing to be done theoretically about these difficulties. There is no hope of improving very much upon the infinite screen model (or, perhaps, some other equally inappropriate smooth surface). If the screen is adopted, rather less harm is caused by also adopting the approximate boundary conditions. The only recourse would be through appropriate experimentation that could shed light on whether real antenna behavior may exacerbate the theoretical effects uncovered in this study. Such experiment might also provide some guidance to theorists on how to devise improved descriptions of antennas. For example, the structure of the antenna might possibly be modeled by a few properly arranged parasitic elements. At present we have no idea how those parasitics should be arranged for real antennas.

Appendix G

IMPULSE RESPONSE OF THE ANTENNA

The impulse response of the antenna, mentioned briefly in Appendix F, will be discussed more fully here for the particular case of the uniform plane rectangular source. Extension to other directional antenna configurations is straightforward.

Two different impulse response concepts should be distinguished: (1) the impulse response of the aperture itself as posited in the Kirchoff formulation and the Kirchoff approximate boundary conditions and (2) the impulse response of the antenna as a whole, including the frequency-selective feeds, but devoid of the assumption of an aperture. The first of these is a theoretical abstraction constructed to facilitate approximate solution of the optical boundary-value problem. The second is also an abstraction, but is the limit that could be approached closely by experimental observation of antennas.

Discussion of the aperture impulse response should commence with the observation that the Huygens-Fresnel-Kirchoff formulation addressed an optical problem in which there is a very large screen and a hole in the screen. The mathematical idealization to an infinite screen, a geometrically flat screen, and a perfectly conducting screen are merely limits that can be approached reasonably well in reality.

The usual formulation is expressed in terms of steady-state monochromatic light. Multiple frequencies, such as must be considered in transient analysis, are presented in terms of integrals over the frequency spectrum. The latter can be understood to offer access to a Fourier transformation whereby one could go over to the time domain. The single frequency formulation offers a means to calculate the diffraction pattern of the aperture--in electronic parlance, the directivity pattern at a selected single frequency.

If one makes the transformation to the time domain and considers what the Kirchoff formulation amounts to, he will see that it is a prescription to carry out a convolution integration of the incident illumination--usually taken to be a steady sine wave--with a function

that can be thought of as the impulse response of the aperture.* It is the calculation of that impulse response function that necessitates integration over the infinite surface of the screen. If that function were known, then we could calculate the diffraction pattern for any illumination whatever; that is what is tacitly contained in the formulation inasmuch as the integration over a frequency spectrum is only a transform of a temporal convolution. It should be emphasized that this aperture impulse response function is a mathematical prescription of how to perform a calculation, not a physical thing. It is a property associated with the screen and aperture, not the incident light.

An optical impulse is physically impossible because the integral over all time of an impulse is finite and a propagating electromagnetic field cannot deliver that finite value. That is a restriction on the incident light, not on the screen and aperture. The aperture impulse response--the prescription for how to obtain the diffraction pattern by convolution--amounts to the diffraction pattern of an incident impulse. (The aperture impulse response would also yield the steady-state electric field distribution resulting from a uniform electric field behind the screen--a zero-frequency light wave whose wavelength is infinite.)

The Kirchoff approximate boundary conditions provide an approximation to the aperture impulse response. They posit (1) that the signal reaching any point in the diffraction field resulting from a uniform normally incident impulse is made up exclusively of contributions arriving from the aperture, and (2) that every incremental element of the aperture contributes an incremental impulse of equal strength, diverging uniformly from that point (Huygens-Fresnel postulate). The

---

*In this report, the term impulse response is used, as is customary in circuit analysis, to describe the temporal response of a system to an input temporal impulse $\delta(t)$. Two other usages are: (1) In contemporary treatments of optical systems, the term impulse response is used for the *spatial* response of a system to a *spatial* impulse $\delta(x,y)$--that is, the response to a point source of monochromatic light at location (x,y). (2) Acoustical treatments often discuss the acoustic *pressure* response that results when a radiating surface undergoes a temporal *velocity* impulse. Such a pressure response is occasionally termed an impulse response; that response is seen to be the time derivative of the impulse response as the term is used here.

signal at any point in space is taken to be the resultant of those incremental impulses, taking account of the differences of transit time. Thus, the approximate boundary conditions lead to the following aperture impulse response:

- At large (infinite) distance on the normal to the center of the aperture--i.e. on the axis--there are no relative time delays and the resultant is an impulse.
- Off axis there are relative time delays; the arrivals occur over a bounded time interval and do not accumulate to an impulse, but instead to a pulse whose duration and height are finite.

If the aperture is a uniform line or rectangle, the off-axis impulse response is a square-sided flat-top pulse of unit area. That is, the off-axis impulse response is a pair of steps with opposite sign separated by the transit time difference. As the off-axis angle increases, the length of the pulse increases and the height decreases in such a way that the area remains constant. Convolution of that impulse response with a steady-state sinusoid yields the well-known sinZ/Z directivity pattern. If the aperture is a uniform circle, the off-axis impulse response is a half-ellipse. Convolution with a sinusoid then yields the $2J_1(Z)/Z$ pattern usually attributed to the disc.

The restriction that light cannot have a zero-frequency component prohibits direct observation of the impulse response of real optical apertures. There are, however, pulsed lasers that deliver stable repeated subpicosecond pulses. With these one might conduct experiments that provide insight into the impulse responses of optical apertures. Because the mathematical underpinnings of diffraction theory are fundamentally sound and it is believed that the theory models the physical world adequately, it is doubtful that such experiments would yield unexpected results. Nevertheless, since the experiments would not be costly, they might be worthwhile.

Consider now the application of the optical aperture model to the problem of calculating the impulse response of the whole antenna. It will be recalled that we supposed that the internal structure of the source acts as a bandpass filter, lest the antenna be credited with the capability to deliver a net electric charge at large distance (see Section III-3). To simplify the analysis, it was supposed that the bandwidth-limiting circuit is a single-pole filter tuned to frequency ω_0/Δ_0 with Q equal to Q_0.* That assumption will suffice here, although it leads to somewhat unsatisfactory results; the frequency responses of antennas are undoubtedly more complicated in detail. An optical impulse is physically impossible, but that is a restriction imposed on the illumination reaching the aperture, not on the aperture itself. Here the bandpass filter is interposed between the terminals of the antenna and the aperture, and serves to restrict the illumination that reaches the aperture.

To obtain the impulse response of the whole antenna, we apply an impulse to the antenna terminals. The waveform passed onward to illuminate the aperture *is the impulse response of the bandpass filter:*

$$\frac{\omega_0}{\Delta_0^2 Q_0} e^{-\frac{\omega_0 t}{2\Delta_0 Q_0}} \cos(\omega_0 t + \delta_0) \; ; \; t > 0$$

This waveform is shown in Fig. G-1. It begins with a step of height $\omega_0/\Delta_0 Q_0$; note, however, that if we had assumed the presence of two or more cascaded bandpass stages, the impulse response would commence at zero.

The optical model assumes that the antenna can be regarded as a rectangular hole in an opaque screen, with this waveform arriving in

---

*In this appendix, the radiation of a modulated carrier whose frequency is ω_0 is not considered. Nevertheless, frequency ω_0 is defined by the tuning of this filter. Consequently we can continue to use $Z = (\omega_0 L/2c)\sin\theta$ as a measure of angular position with respect to the aperture.

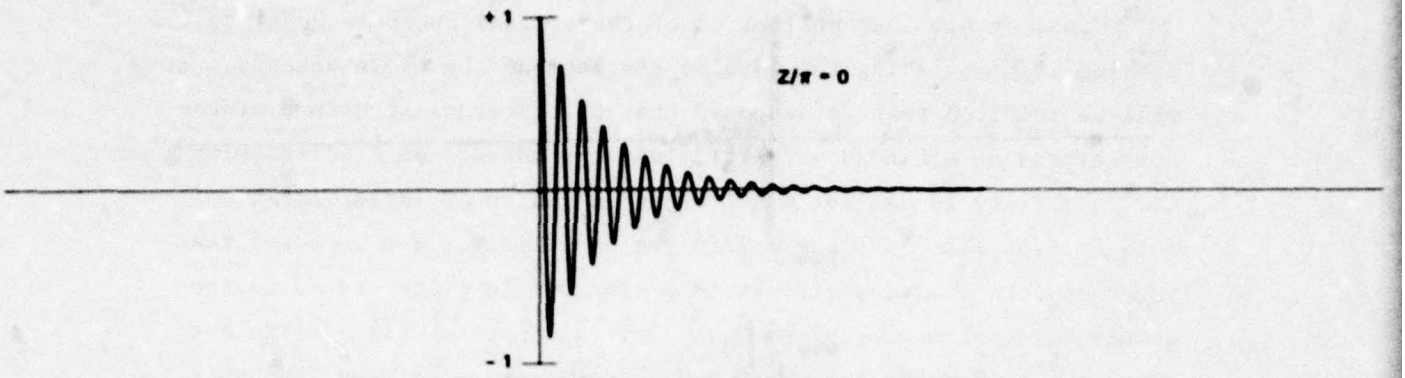


Fig. G-1 — Bandwidth-limited aperture impulse response implicit in the approximate Kirchoff boundary conditions; response on source axis is the impulse response of the bandwidth-limiting circuit

a normally incident plane wave. In keeping with the approximate boundary conditions, it is assumed that all portions of the aperture are illuminated simultaneously and equally, and that all of the signal reaching the far field arrives directly and exclusively from the aperture. The impulse response of the whole antenna in any direction is to be obtained by convolution of this waveform with the approximate aperture impulse response in that direction.

On the source axis, the approximate aperture impulse response is an impulse. (The strength of the impulse varies as 1/R to reflect the inverse square law, but we overlook that aspect of the behavior.) Convolution of a waveform with an impulse yields the waveform itself; consequently the impulse response of the antenna on axis is just the illumination waveform given above and shown in Fig. G-1. The signal has an oscillatory tail because of the bandpass filter, not because the aperture impulse response is thought to have such a tail.

Off axis, at angular position Z, the approximate aperture impulse response is a step of height $1/2Z$ at time $-Z/\omega_0$, followed by a step of height $-1/2Z$ at time $+Z/\omega_0$. Convolution of the illuminating waveform with this aperture impulse response yields the waveforms shown in Figs. G-2 and G-3. In these figures the vertical scale has been adjusted to offset the amplitude factor $1/2Z$.

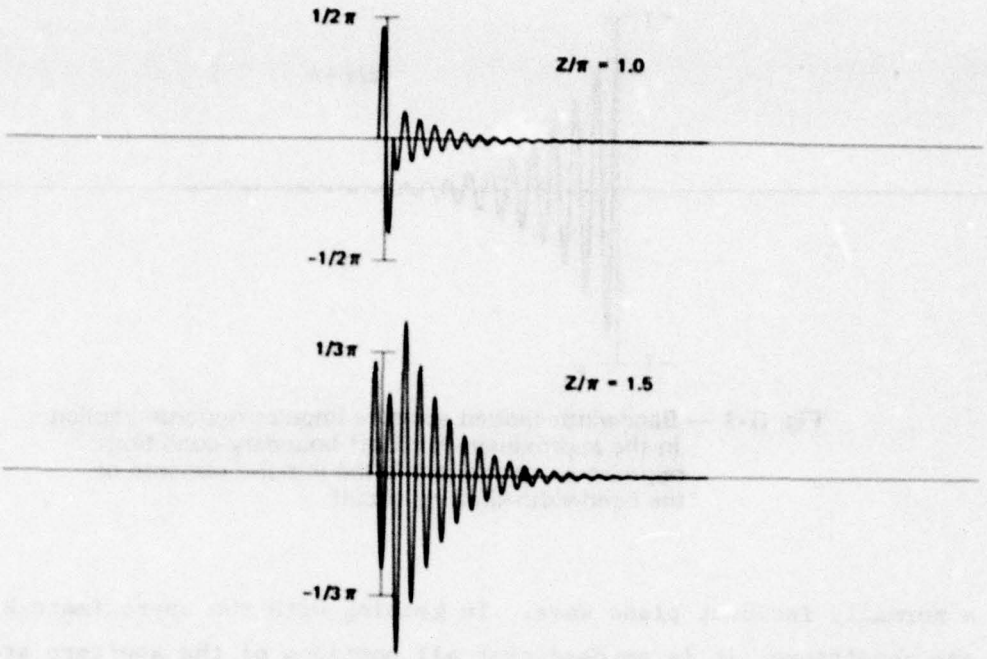


Fig. G-2 — Kirchoff-approximate bandwidth-limited aperture impulse response in first pattern null and first side lobe. (Z defined in terms of center frequency of filter)

The off-axis impulse responses of the antenna shown in Figs. G-2 and G-3 consist of two oppositely directed *step* responses of the band-pass circuit, separated in time by $2Z/\omega_0$. It will be noted that the off-axis impulse responses commence at zero, whereas the axial response commences with a step. This abrupt change with Z is not realistic, and bespeaks a shortcoming of the assumption that a single-pole band-pass circuit will suffice to characterize the frequency response of the source. However, the simple model will suffice for present purposes.*

Figures G-1, G-2, and G-3 show the impulse response of the antenna in five directions, as implied by the optical aperture model and the approximate boundary conditions. If any particular signal were applied

*See footnote at end of Section III-3.

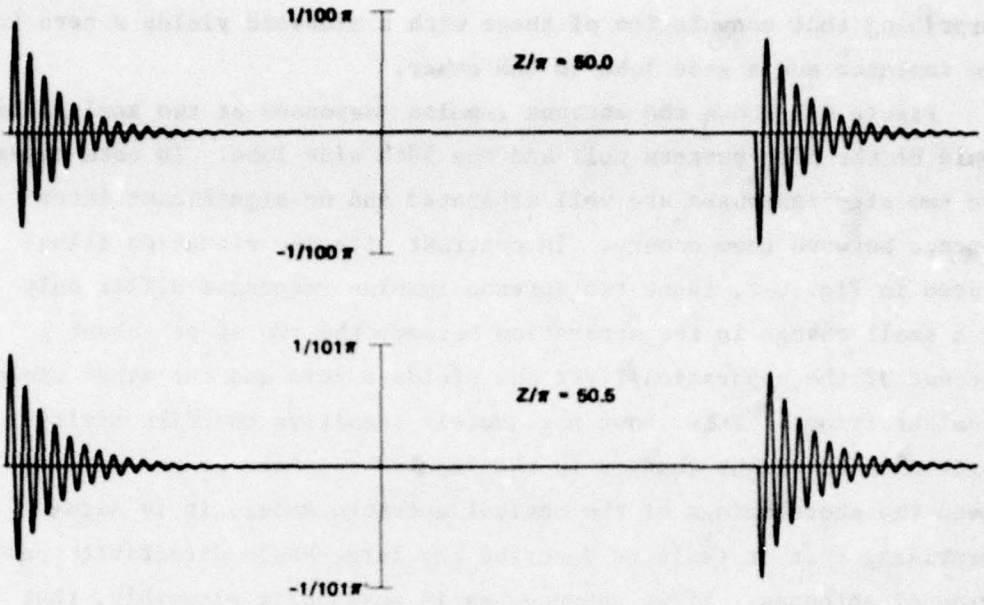


Fig. G-3 — Counterpart of Fig. G-2 at large angles; 50th null and 50th side lobe. Aperture impulse response is a pair of oppositely directed step responses of the bandwidth-limiting filter.

to the antenna terminals, then the resulting far-field signal in these directions could be obtained by convolving the input signal with these impulse responses. Indeed, waveform $F_{1,2}$ obtained in Section III-4 is obtained by convolving the input signal

$$u(0)\,\frac{\sqrt{1 + 3\Delta_0^2}}{2\Delta_0}\,\sin(\omega_0 t + \psi)$$

(See Section III-3 for an explanation of this choice of amplitude.)

Figure G-2 shows the antenna impulse responses at two angles that would be the first null and first side lobe (optically, the first bright fringe) if the antenna were to radiate a steady sinusoid at frequency ω_0. Here the two oppositely directed step responses that comprise the antenna impulse response interfere strongly, and the two antenna impulse responses are quite different. It does not seem

surprising that convolution of these with a sinusoid yields a zero in one instance and a side lobe in the other.

Figure G-3 shows the antenna impulse responses at two angles that would be the 50th pattern null and the 50th side lobe. In both cases the two step responses are well separated and no significant interference between them occurs. In contrast with the situation illustrated in Fig. G-2, these two antenna impulse responses differ only by a small change in the separation between the two steps (about 1 percent of the separation), yet one yields a zero and the other yields a bright fringe. This shows how acutely sensitive the directivity pattern is to slight changes in the impulse response of the antenna. Given the shortcomings of the optical aperture model, it is hardly surprising that it fails to describe the large-angle directivity patterns of antennas. If we suppose, as is physically plausible, that the correct aperture impulse response exhibits an oscillatory top and falls off in an oscillatory tail, then we can expect a very different wide-angle directivity pattern. Such a change in the antenna impulse response would also change, perhaps significantly, the strength and the waveform of the "extra" signal that is of concern in this study.

An experimental investigation could examine the impulse response of an antenna directly, without reference to an aperture and a related aperture impulse response. It would be unnecessary to posit the existence of some particular close-in waveform that is regarded as "the illumination." A careful investigation on a good antenna range might be worthwhile, not only to clarify the questions concerning TOA raised in this report, but also to advance our general understanding of antenna behavior.

Appendix H

THE APERTURE AS A FILTER

A directional aperture behaves as a spatially varying frequency selective filter, as noted earlier in this report and in R-1819-PR. Although the analysis presented here is carried out entirely in the time domain, the frequency domain behavior of the aperture is of some interest.

The Kirchoff approximate boundary conditions that were discussed in Appendix F are usually adopted to calculate the directivity pattern of the antenna at any given frequency, and that same approximation is used here. For the simple case of the uniform line source, the aperture impulse response in the far field is merely a square-sided flat-top pulse of height $c/(L\sin\theta)$. The pulse starts at time $-L\sin\theta/2c$ and ends at time $+L\sin\theta/2c$. (This has been normalized so a steady-state sinusoid has unit amplitude on axis.)

The transfer function of the antenna aperture is the Fourier transform of the aperture impulse response:

$$\Gamma(\omega) = \frac{c}{L\sin\theta} \int_{-\frac{L\sin\theta}{c}}^{+\frac{L\sin\theta}{c}} e^{i\omega t} dt$$

$$= \frac{\sin\left(\omega \frac{L\sin\theta}{2c}\right)}{\omega \frac{L\sin\theta}{2c}}$$

The transfer function exhibits in the frequency domain the same functional form as the directivity pattern exhibits in space: $\sin X/X$. The aperture behaves as a lowpass filter at angles off the axis, and the transfer function has zeros at every frequency for which

$$\frac{\omega L\sin\theta}{2c}$$

is an integer multiple of $\pi(\neq 0)$. At any particular angle θ, the transfer function is zero at every frequency that makes Z/π equal to an integer $(\neq 0)$.

The alternating algebraic sign of this transfer function presents the same difficulty as the directivity pattern insofar as the phase characteristic of the filter is concerned. That phase question was discussed in detail in R-1819-PR, where it was shown that this sinX/X form is an approximation that is approached as the propagation distance approaches infinity. The phase ambiguity is resolved by consideration of the higher order terms in the series, and it is found that each alteration of sign reflects a discontinuous phase jump of $+\pi$ as the function passes through the origin. Thus, the phase characteristic of the antenna aperture at any chosen angle θ is an endless "staircase" in which each step has height π.

We remark upon a point concerning the energy spectra of signals F_0 and $F_{1,2}$. The energy spectrum of F_0 is finite at all frequencies except ω_0, where it is infinite. Alternatively, the power spectrum of F_0 is zero at all frequencies except ω_0, where it is a delta function.

The energy spectrum of $F_{1,2}$ is equal to the energy spectrum of F_0 multiplied by the conjugate square of the transfer function of the aperture. The spectrum of $F_{1,2}$ has a more complicated shape than that of F_0 because of the various lobes of $[\text{sinX/X}]^2$, and has many zeros not present in the spectrum of F_0. The locations of the zeros change with θ; at large angles there will usually be several zeros at frequencies lower than ω_0.

At frequency ω_0, the transfer function of the antenna is zero at angles for which Z/π is an integer $\neq 0$, that is, in the pattern nulls at ω_0. At those locations the energy spectrum of $F_{1,2}$ is non-zero and finite. The combination of infinite energy content in F_0 and zero transfer function yields finite energy content in $F_{1,2}$ in the pattern nulls.

Appendix I

DESIGN OF THE ENVELOPE DETECTOR

The design of an envelope detector is ordinarily routine. It is customary to adopt a cutoff frequency high enough to pass the desired bandwidth, and to incorporate enough filter stages to reduce the RF ripple to an acceptable level. Detailed analysis of the tradeoff between the cutoff frequency and the number of stages is not usually needed. However, in a TOA system that seeks to discern arrival by the passage of the leading edge through a threshold, the steepness of the leading edge at the threshold is unusually significant. The transient response of the circuits ahead of the envelope detector--principally the IF strip--establishes a maximum possible steepness of the leading edge, and it is desirable that the envelope detector preserve that steepness as well as is reasonably possible.

The inherent steepness at which the output of the envelope detector itself can rise can be judged from the step response of the detector. The step response of (1 + S) cascaded mutually isolated identical RC stages is (see Appendix C)

$$H(t) = 1 - \left[1 + \omega_c t + \frac{(\omega_c t)^S}{2!} + \ldots + \frac{(\omega_c t)^S}{S!}\right] e^{-\omega_c t}$$

where $\omega_c = 1/RC$. The slope of H(t) at time t is given by the impulse response of the detector:

$$h(t) = \omega_c \frac{(\omega_c t)^2}{S!} e^{-\omega_c t}$$

It is convenient here to adopt, as a suitable measure, the value of h(t) at the time when $H(t) = 1/2$, that is, the steepness of the step response at the time when the step response of the detector is at half amplitude. Tabulating the dependence on the number of stages:

Number of stages	*h(t) when H(t) = 1/2*
1	$0.5\ \omega_c$
2	$0.3133\ \omega_c$
3	$0.2466\ \omega_c$
4	$0.2098\ \omega_c$

Adding more stages while holding ω_c constant slows the transient response of the detector and consequently reduces the receiver output steepness below that inherent in the IF output signal. The reduction caused by adding stages can be offset by raising ω_c. However, raising ω_c increases the ripple, and one's choice of a tolerable level of ripple provides a prescription of how to trade between S and ω_c.

To assess the dependence of ripple upon ω_c and S we will adopt, as a measure of ripple amplitude, the peak-to-peak ripple voltage that results when the input to the filter is a steady-state signal

$$\frac{\pi}{2} \left|\sin(\omega_1 t)\right|$$

This is a full-wave rectified sine wave such that the output dc from the detector is 1 volt.

For brevity let $r = \omega_c/\omega_1$, and define a dimensionless measure of time, ϵ. By definition, $0 \leq \epsilon \leq 1$; as ϵ varies over this span, the expressions below trace one complete period of the output waveform. One unit of ϵ corresponds to a half period of the sine wave $\sin(\omega_1 t)$.

For a one-stage filter (S = 0) the output waveform from the filter is

$$\frac{\pi r}{1 + r^2}\left\{\frac{e^{-\epsilon\pi r}}{1 - e^{-\pi r}} + \frac{r \sin\epsilon\pi - \cos\epsilon\pi}{2}\right\}$$

For a two-stage filter (S = 1) the output waveform is

$$\frac{\pi r^2 e^{-\epsilon\pi r}}{\left(1 + r^2\right)\left(1 - e^{-\pi r}\right)}\left[\epsilon\pi + \frac{2r}{1 + r^2} + \frac{\pi e^{-\pi r}}{1 - e^{-\pi r}}\right] - \frac{\pi}{2}\frac{r^2}{1 + r^2}\sin(\epsilon\pi + \theta)$$

where $\tan\theta = \dfrac{2r}{1-r^2}$

For a three-stage filter (S = 2) the output waveform is

$$\frac{\pi r^3 e^{-\epsilon\pi r}}{\left(1+r^2\right)\left(1-e^{-\pi r}\right)}\left[\frac{(\epsilon\pi)^2}{2}+\frac{2\epsilon\pi r}{1+r^2}-\frac{\left(1-3r^2\right)}{\left(1+r^2\right)^2}\right.$$

$$+\frac{\epsilon\pi^2\, e^{-\pi r}}{1-e^{-\pi r}}+\frac{\pi^2}{2}\cdot\frac{e^{-\pi r}\left(1+e^{-\pi r}\right)}{\left(1-e^{-\pi r}\right)^2}$$

$$\left.+\frac{2\pi r\, e^{-\pi r}}{\left(1+r^2\right)\left(1-e^{-\pi r}\right)}\right]+\frac{\pi}{2}\left(\frac{r}{\sqrt{1+r^2}}\right)^3\cos(\epsilon\pi+\theta)$$

where $\tan\theta = \dfrac{r\left(3-r^2\right)}{1-3r^2}$

The integrals of these waveforms over $0 < \epsilon < 1$ are unity; that is, the average of the ripple is zero. There is one maximum and one minimum in the interval $0 < \epsilon < 1$, but it is impractical to solve for the locations and values of those extrema.

For any given choice of r, the ripple amplitude falls rapidly as more stages are added. However, if r is adjusted to hold the peak-to-peak ripple amplitude constant as more stages are added, then the increase in r--and associated improvement of rise time--becomes quite slow as the number of stages is increased. To illustrate, we tabulate, for several choices of peak-to-peak ripple amplitude, the dependence of r on the number of stages.

millivolts of ripple	$r = \omega_c/\omega_1$		
	1 stage S = 0	2 stages S = 1	3 stages S = 2
0.5	0.000756	0.038521	0.144674
1	0.001512	0.054489	0.182546
1.5	0.002268	0.066749	0.209223
2	0.003024	0.077092	0.230536
3	0.004536	0.094458	0.264422
4	0.006048	0.109118	0.291552

The choice r = 0.003024 yields 2 millivolts of ripple when one stage is used, whereas that same r yields only 4.6 nanovolts of ripple when three stages are used. But if the ripple is held constant at 2 millivolts, then going from one stage to two allows r to increase by a factor of 25.5 (from 0.003024 to 0.07709), whereas going from two stages to three allows a further increase of only a factor of 3 (from 0.07709 to 0.2305).

The foregoing results can be combined to illustrate the improvement of slope with additional stages at any selected level of ripple amplitude:

Number of stages	h(t) when H(t) = 1/2	
	1 millivolt ripple	2 millivolts ripple
1	$0.000756\ \omega_1$	$0.001512\ \omega_1$
2	$0.017072\ \omega_1$	$0.024154\ \omega_1$
3	$0.045015\ \omega_1$	$0.056849\ \omega_1$

The corresponding slope of the IF strip (as judged from the steepness of the step response of the lowpass equivalent filter) is $0.023816\ \omega_1$. Because the slope of the envelope detector must be appreciably greater than this if it is not to slow down the receiver, at least three stages will be needed, and preferably more, if about 2 millivolts of ripple are accepted.

In the design studied here, four stages were adopted with $r = 0.375$. Thus, the RC cutoff was set at 30 MHz--considerably above the 10 MHz IF bandwidth. For this choice the slope is $0.078678\ \omega_1$--about 3.3 times the steepness of the IF strip. The rise steepness of the whole receiver is very nearly as good as that of the IF strip alone.

The dependence of peak-to-peak ripple amplitude on r is, for all these lowpass filters, somewhat complicated because the values of ϵ at which the ripple extrema occur shift with r. However, as $r \to 0$, the leading terms approach:

1 stage:	$661.348\ r$ millivolts
2 stages:	$337.096\ r^2$ millivolts
3 stages:	$166.364\ r^3$ millivolts

One can now estimate that, for four stages, the leading term is approximately $82\ r^4$ millivolts. (It was not thought to be worthwhile to obtain the four-stage waveform.) This estimate suggests that the peak-to-peak ripple amplitude in the four-stage filter with $r = 0.375$ is about 1.6 millivolts. The ripple seen in the computer output is about this much.

Appendix J

TOA SYSTEM ERROR

The layout of a simple 3-receiver (2-dimensional) TOA system is illustrated in Fig. J-1. The true location of the source is at the origin of the X,Y coordinate system. The X axis is in the downrange direction from the middle receiver, and the Y axis is in the corresponding (right-handed) crossrange direction. Errors made by the system in computing the location of the source will be expressed as small distances δX and δY.

The three receivers, numbered 0, 1, 2 in the figure, are at true distances ρ_0, ρ_1, ρ_2 from the source. It is assumed here that these

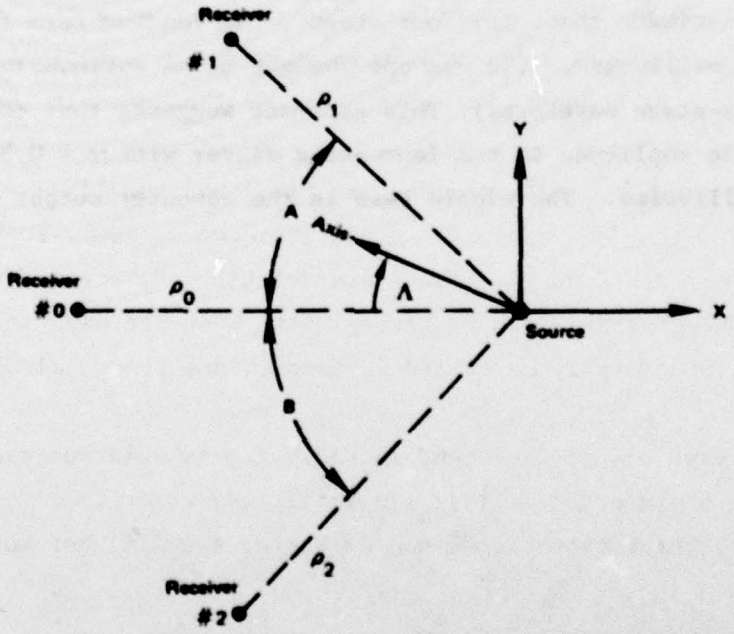


Fig. J-1 — Layout of an elementary TOA system

distances are large compared with δX and δY. The receiver array subtends angles A and B at the source.

The angular orientation of the directional source is described by angle Λ, measured from the -X axis. For any choice of Λ the three receivers lie at different locations within the directivity pattern of the source:

$$Z_0 = \frac{\pi L}{\lambda} \sin(\Lambda)$$

$$Z_1 = \frac{\pi L}{\lambda} \sin(A - \Lambda)$$

$$Z_2 = \frac{\pi L}{\lambda} \sin(B + \Lambda)$$

The nth receiver *should* receive a source pulse at a time determined by ρ_n. It is immaterial to this discussion what the numerical value of that time is, although the observed time must be used in the system to calculate the source location. If the nth receiver reports arrival to occur at a time that differs from the time that should have been observed, that time difference can be interpreted as equivalent to an error $\delta\rho_n$ in the distance ρ_n. (Late arrival time corresponds to positive $\delta\rho_n$.) The displacements of arrival time that are discussed in this report are generally expressed in meters, and these values constitute $\delta\rho_n$.

Arrival time disparities tend to cause the calculated location of the source to be in error. If $\delta\rho_n$ is small, the error that is made depends on all three disparities and on angles A and B, but not upon the distances ρ_n:

$$\delta X = \left[\frac{(\delta\rho_0 - \delta\rho_2) \sin A - (\delta\rho_1 - \delta\rho_0) \sin B}{\sin A + \sin B - \sin(A + B)} \right]$$

$$\delta Y = \left[\frac{(\delta\rho_0 - \delta\rho_2)(1 - \cos A) + (\delta\rho_1 - \delta\rho_0)(1 - \cos B)}{\sin A + \sin B - \sin(A + B)} \right]$$

For any given subtense (A + B), the system sensitivity to error is minimized when A = B. That arrangement is termed symmetric in this report. It should be noted that disposition of the three receivers at the corners of an isosceles triangle does not, of itself, assure that A = B; it is also necessary that the source be on the axis of symmetry. Conversely, angles A and B may be equal even if the receivers lie on a scalene triangle because the distances ρ_n are immaterial.

When A = B the expressions for the downrange and crossrange errors take a simple form:

$$\delta X = \left[\frac{\delta\rho_1 + \delta\rho_2 - 2\delta\rho_0}{2(1 - \cos A)} \right]$$

$$\delta Y = \left[\frac{\delta\rho_2 - \delta\rho_1}{2 \sin A} \right]$$

The computer experiments carried out during this study used A = B = 45 degrees. For that case

$$\delta X = (1 + \sqrt{1/2})[\delta\rho_1 + \delta\rho_2 - 2\delta\rho_0]$$

$$\delta Y = \sqrt{1/2}\ [\delta\rho_2 - \delta\rho_1]$$

If, under these conditions, receivers 1 and 2 both observe arrival early by scale length S, but receiver 0 observes the expected arrival time, then

$$\delta X = -(2 + \sqrt{2})S = \delta R$$

$$\delta Y = 0$$

If receiver 1 observes arrival early by scale length S, but the other two observe the expected times, then

$$\delta X = -(1 + \sqrt{1/2})S$$

$$\delta Y = \sqrt{1/2}\ S$$

and the radial error is

$$\delta R = \sqrt{2 + \sqrt{2}}\ 'S$$

These distances are indicated by tick marks on Figs. II-1 and II-2, and are seen to mark characteristic features of the error distributions.